


# Demographic confounders distort inference of gene regulatory and gene co-expression networks in cancer

Anna Ketteler and David B. Blumenthal 

Corresponding author. David B. Blumenthal, Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens-Str. 61, 91052 Erlangen, Germany. Tel.: +49 9131 8570676; E-mail: david.b.blumenthal@fau.de

## Abstract

Gene regulatory networks (GRNs) and gene co-expression networks (GCNs) allow genome-wide exploration of molecular regulation patterns in health and disease. The standard approach for obtaining GRNs and GCNs is to infer them from gene expression data, using computational network inference methods. However, since network inference methods are usually applied on aggregate data, distortion of the networks by demographic confounders might remain undetected, especially because gene expression patterns are known to vary between different demographic groups. In this paper, we present a computational framework to systematically evaluate the influence of demographic confounders on network inference from gene expression data. Our framework compares similarities between networks inferred for different demographic groups with similarity distributions obtained for random splits of the expression data. Moreover, it allows to quantify to which extent demographic groups are represented by networks inferred from the aggregate data in a confounder-agnostic way. We apply our framework to test four widely used GRN and GCN inference methods as to their robustness w. r. t. confounding by age, ethnicity and sex in cancer. Our findings based on more than 44000 inferred networks indicate that age and sex confounders play an important role in network inference for certain cancer types, emphasizing the importance of incorporating an assessment of the effect of demographic confounders into network inference workflows. Our framework is available as a Python package on GitHub: <https://github.com/bionetslab/grn-confounders>.

**Keywords:** systems medicine; demographic confounders; gene regulatory networks; gene co-expression networks

## INTRODUCTION

Deciphering molecular mechanisms that regulate gene expression is a major task in the biomedical sciences. A broad range of computational methods has been proposed for this purpose, allowing efficient analyses of large-scale gene expression datasets. One class of computational methods commonly used in this context is methods for gene regulatory network (GRN) and gene co-expression network (GCN) inference [1–6]. In the inferred networks, genes are represented by nodes and regulatory or co-expression patterns between genes are represented by edges and are estimated using statistical inference, machine learning or correlation analysis. The distinction between GRNs and GCNs is mostly based on whether the network is directed (GRN) or undirected (GCN), i. e. based on whether an edge between two genes represents a (directed) gene regulation or an (undirected) gene co-expression.

A topic that has been paid considerably little attention in this context is the influence of demographic confounders on network inference. This is surprising, because various studies suggest differences in gene expression across different demographic groups, especially in the context of cancer. Several studies have revealed age-related differential gene expression patterns across multiple cancer types [7–11]. For instance, Wu *et al.* [7] identified differential

gene expression patterns between young and old patients in 16 cancer types and show that, in these 16 cancer types, age is associated with survival. In view of sex-specific differences in cancer incidence and prognosis, Dong *et al.* [12] strongly recommend the investigation of the impact of sex on a molecular level. Yang *et al.* [13] identified sex-specific molecular subtypes as major predictors for survival in glioblastoma and point to the need to consider sex-related molecular patterns in clinical treatments. Similarly, differences in cancer mortality and incidence across different ethnic groups [14, 15] suggest the consideration of ethnicity as a confounding factor in network inference. While some of the observed disparities can be explained by differences in access to high-quality treatment [16], there is also evidence for a molecular underpinning. For instance, Aguilar *et al.* [17] found ethnicity-specific expression and somatic mutation profiles in breast cancer patients.

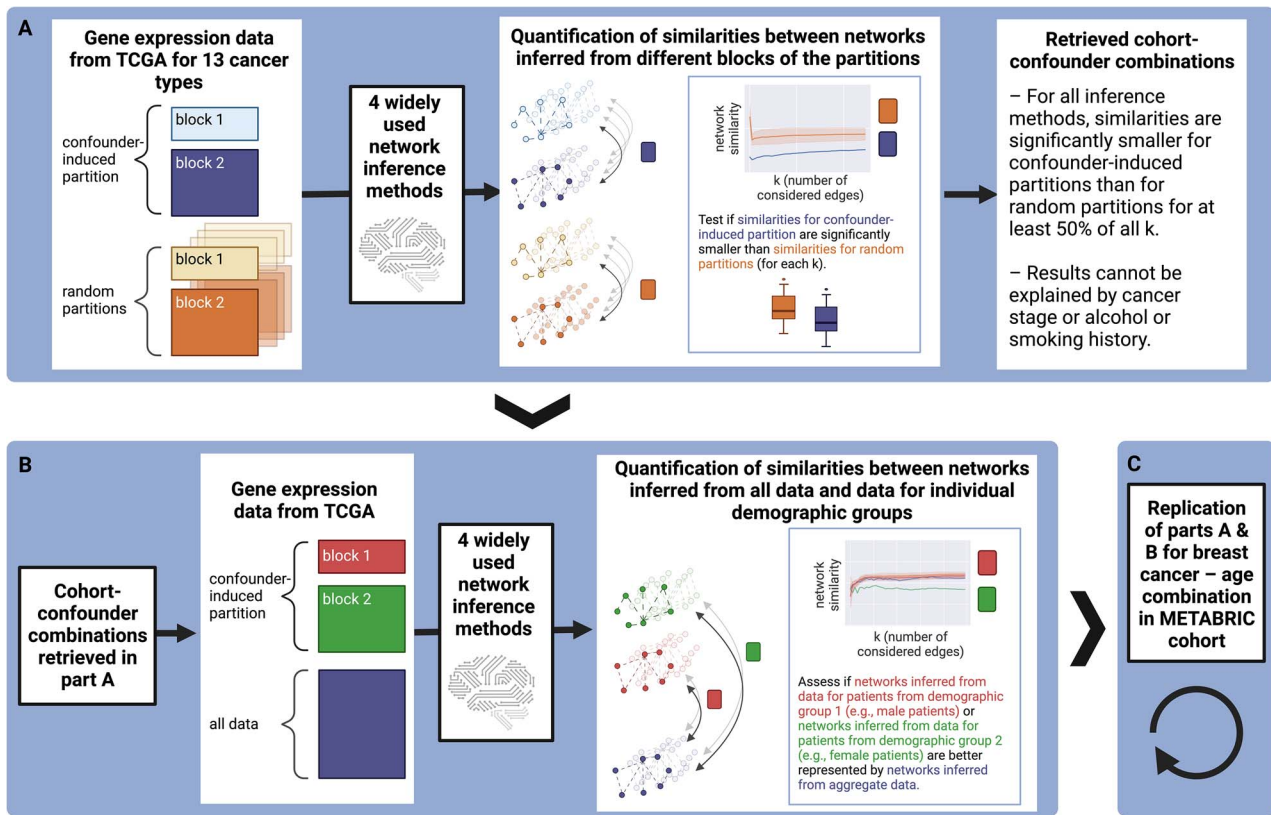
To the best of our knowledge, only two studies exist which investigate the effect of demographic confounders on network inference: Lopes-Ramos *et al.* [18] report sex differences in single-sample GRNs inferred with LIONESS [19] in the context of colon cancer. In [20], the same group of researchers extends their findings to 29 human healthy tissues. However, Lopes-Ramos *et al.* only consider sex as a potential demographic confounder and assess its effect on only one single-sample GRN inference method.

**Anna Ketteler** is an MSc student and student research assistant in the Biomedical Network Science Lab at the Department Artificial Intelligence in Biomedical Engineering of the Friedrich-Alexander-Universität Erlangen-Nürnberg.

**David B. Blumenthal** is a professor and head of the Biomedical Network Science Lab at the Department Artificial Intelligence in Biomedical Engineering of the Friedrich-Alexander-Universität Erlangen-Nürnberg. He obtained his PhD in Computer Science from the Free University of Bozen-Bolzano.

**Received:** June 16, 2023. **Revised:** September 19, 2023. **Accepted:** October 26, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)



**Figure 1.** Overview of our study, where we assessed the impact of demographic confounders sex, age and ethnicity on two GCN inference methods (WGCNA, CEMiTool) and two GRN inference methods (ARACNe-AP, GRNBoost2). (A) In the first step, we quantified similarities between networks inferred for different demographic groups to identify combinations of cancer types and confounders where the confounder significantly distorts the network inference, using gene expression data for 12 cancer types obtained from TCGA. (B) For the identified combinations, we then assessed to which extent the networks inferred for the individual demographic groups are (mis-)represented by networks inferred from all data in a confounder-agnostic way, as typically done in network inference studies. (C) Finally, for one cohort–confounder combination (BRCA–age), we replicated our results using data from an independent cohort.

Widely used cohort-based network inference methods are not tested. Moreover, Lopes-Ramos *et al.* do not provide a ready-to-use software package that would allow to run their analysis pipeline using other datasets or inference methods.

To close these gaps, we here systematically assess the effect of age, ethnicity and sex confounders on network inference from cancerous tissue, putting two GRN inference methods (ARACNe-AP [21], GRNBoost2 [22]) and two GCN inference methods (WGCNA [23], CEMiTool [24]) to the test (see Figure 1 for a schematic overview). At the core of our work, we propose a Python framework (available at <https://github.com/bionetslab/grn-confounders>) which can be used to assess the effect of confounding on GRN or GCN inference for any inference method and gene expression data that are provided with corresponding demographic annotations.

## RESULTS

### Overview of the study

Figure 1 provides an overview of our study (see METHODS for details). We used bulk gene expression data with demographic information for primary tumor samples of 12 cancer types, which we obtained from The Cancer Genome Atlas (TCGA) [25, 26]: Breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP),

lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pheochromocytoma and paraganglioma (PCPG), rectum adenocarcinoma (READ) and stomach adenocarcinoma (STAD). Confounder-induced partitions of the datasets were constructed based on the patients' ages, sexes and ethnicities. See Table 1 for an overview of the resulting partitions for the individual cohorts. Similarly, we constructed partitions induced by the third variables cancer stage at diagnosis, smoking history and alcohol history for cohorts where these data are available (Supplementary Table 2).

In the first step (Figure 1A), for each confounder-induced partition  $\mathcal{C}$ , we separately inferred GCNs and GRNs for all blocks and quantified the similarities between the inferred networks by computing the mean Jaccard index (JI) of the  $k$  top-ranked edges, for varying  $k \in \{10, 110, \dots, 4910\}$ . For each partition, this procedure was repeated 10 times to account for non-deterministic network inference methods. We hence obtained a distribution  $JI_k(\mathcal{C})$  of 10 mean JIs for each  $k$ . Moreover, for each confounder-induced partition  $\mathcal{C}$ , we applied the same protocol but without repetition to 100 size-matched random partitions  $\mathcal{R}$ , leading to a distribution  $JI_k(\mathcal{R})$  of 100 mean JIs for each  $k$ .

The demographic confounder under consideration distorts the tested network inference method on the given data whenever the networks inferred from the different demographic groups are significantly less similar than expected by chance. To test if this is the case, we applied the one-sided Mann–Whitney  $U$  (MWU) test with alternative hypothesis  $JI_k(\mathcal{C}) < JI_k(\mathcal{R})$ . To assess if significant results might be explained by correlation of the

**Table 1:** Number of samples contained in the different blocks of the confounder- and third-variable-induced partitions and number of samples removed due to missing values in the column corresponding to the confounder. The block identifiers for the age and sex confounders are taken from the identifiers used in the TCGA phenotype file, with BAA, NHPI and AIAN abbreviating “Black or African American”, “Native Hawaiian or Other Pacific Islander” and “American Indian or Alaska Native”, respectively. For the age-based stratification of the TCGA cohorts, we retained samples from patients from the first (young) and fourth (old) age quartiles and discarded all other samples (age ranges covered by the first and fourth quartiles are detailed in [Supplementary Table 1](#)). The METABRIC breast cancer cohort used for replication was stratified into young and old patients using the age ranges of the first and fourth age quartiles of the TCGA BRCA cohort. For all confounders and cohorts, we discarded blocks containing less than 20 samples when inferring the networks. For the METABRIC cohort, no ethnicity information is available

| Cohort    | Age   |     |     | Ethnicity |       |     |      |       |     | Sex    |      |     |
|-----------|-------|-----|-----|-----------|-------|-----|------|-------|-----|--------|------|-----|
|           | Young | Old | N/A | AIAN      | Asian | BAA | NHPI | White | N/A | Female | Male | N/A |
| TCGA BRCA | 297   | 273 | 1   | 1         | 61    | 183 | 0    | 757   | 95  | 1084   | 12   | 1   |
| TCGA CESC | 80    | 76  | 0   | 7         | 20    | 30  | 2    | 209   | 36  | 304    | 0    | 0   |
| TCGA COAD | 118   | 115 | 2   | 1         | 11    | 61  | 0    | 223   | 173 | 221    | 246  | 2   |
| TCGA GBM  | 39    | 39  | 1   | 0         | 5     | 10  | 0    | 138   | 2   | 54     | 100  | 1   |
| TCGA HNSC | 130   | 110 | 1   | 2         | 10    | 47  | 0    | 426   | 15  | 133    | 367  | 0   |
| TCGA KIRC | 145   | 127 | 0   | 0         | 8     | 56  | 0    | 463   | 7   | 186    | 348  | 0   |
| TCGA KIRP | 80    | 62  | 3   | 2         | 6     | 60  | 0    | 205   | 15  | 76     | 212  | 0   |
| TCGA LUAD | 140   | 124 | 0   | 1         | 7     | 53  | 0    | 397   | 66  | 282    | 242  | 0   |
| TCGA LUSC | 134   | 118 | 9   | 0         | 9     | 30  | 0    | 349   | 113 | 130    | 371  | 0   |
| TCGA PCPG | 45    | 45  | 0   | 1         | 6     | 20  | 0    | 147   | 4   | 101    | 77   | 0   |
| TCGA READ | 44    | 41  | 1   | 0         | 1     | 6   | 0    | 80    | 79  | 75     | 90   | 1   |
| TCGA STAD | 100   | 92  | 4   | 0         | 74    | 11  | 1    | 238   | 51  | 134    | 241  | 0   |
| METABRIC  | 379   | 712 | 0   | –         | –     | –   | –    | –     | –   | 1965   | 0    | 0   |

demographic confounders with third variables, we carried out the same protocol for partitions of the data induced by cancer stage, alcohol history or smoking history, and controlled for dependence of the demographic confounder with these variables by performing separate  $\chi^2$ -tests (see [Supplementary Figure 1](#) for details).

For the second part of the protocol ([Figure 1B](#)), we only considered such cohorts and confounders for which, for all methods, the MWU test yielded a significant  $P$ -value (at significance level 5%) for at least 50% of all  $k$  and for which we did not detect a dependency between the confounder and cancer stage, alcohol history or smoking history that might explain these results. We then assessed whether the networks inferred from the aggregate data are more representative for one of the demographic groups induced by the confounder. For this, we compared the networks inferred from all data with the networks inferred for the different demographic groups by computing JIs for the top  $k$  edges, for varying  $k \in \{10, 110, \dots, 4910\}$ . To control for the effect of different sample numbers for the different demographic groups, we performed the same tests for the blocks of a size-matched random partition. We repeated this procedure 10 times. In each iteration, we re-sampled the size-matched random partition and re-inferred the networks from the aggregate data, from the blocks of the confounder-induced partition, and from the blocks of the random partition. For each  $k$  and each demographic group  $l$  induced by the confounder, we hence obtained distributions  $Jl_k^l(C)$  and  $Jl_k^l(R)$  of 10 JIs quantifying the similarities between the confounder-agnostic network inferred from the aggregate data and, respectively, the networks for demographic group  $l$  and for size-matched random samples.

Finally ([Figure 1C](#)), we focused on the cancer type-confounder combination BRCA-age, where the first part of the test protocol yielded highly significant results for all methods and the second part of the protocol revealed that older patients are worse represented by networks inferred from all samples than younger patients. For this combination, we replicated our

analyses using gene expression and age data from the METABRIC cohort [27].

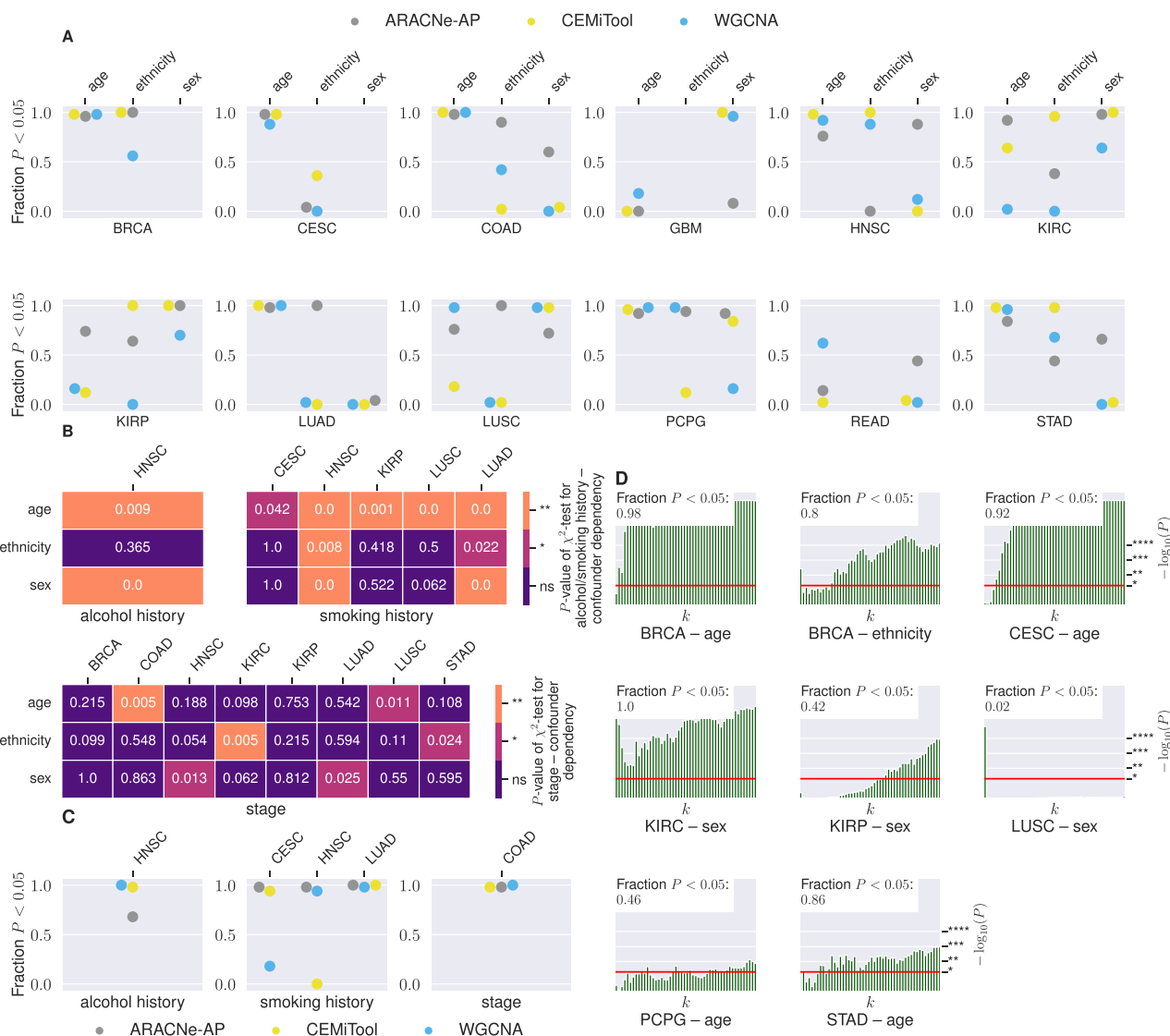
### Similarities between networks inferred for different demographic groups

[Figure 2A](#) shows the fractions of significant MWU  $P$ -values, grouped by cancer type and confounders, for three of the four tested network inference tools (ARACNe-AP, CEMiTool and WGCNA). Due to long runtimes of GRNBoost2, tests for GRNBoost2 were carried out only on a selection of such confounders and cohorts (see details below and [Supplementary Figure 1](#)). Consistently high fractions across all methods suggest a strong effect of the respective confounder on the gene regulatory or gene co-expression patterns in the respective cancer type. In the first step, we hence focused on all combinations of cohorts and confounders for which, for all three fast network inference methods, at least 50% of the MWU tests yielded significant  $P$ -values:

- Cohorts with strong effect of age confounder: BRCA, CESC, COAD, HNSC, LUAD, PCPG and STAD.
- Cohort with strong effect of ethnicity confounder: BRCA.
- Cohorts with strong effect of sex confounder: KIRC, KIRP and LUSC.

For the ethnicity confounder, there is only one cohort (BRCA), where all methods yield a fraction of significant MWU  $P$ -values larger than 0.5. Unsurprisingly, age and sex seem to have a larger effect on gene regulatory and gene co-expression patterns in cancer than ethnicity.

For demographic confounders where the sample stratification is correlated to a third variable, the effect observed for the confounder might result from implicitly approximating stratification via that variable. We hence tested correlation of the confounders with the three variables alcohol history, cancer stage and smoking history in cohorts where a strong effect of at least one confounder could be observed and where data on the third variables are



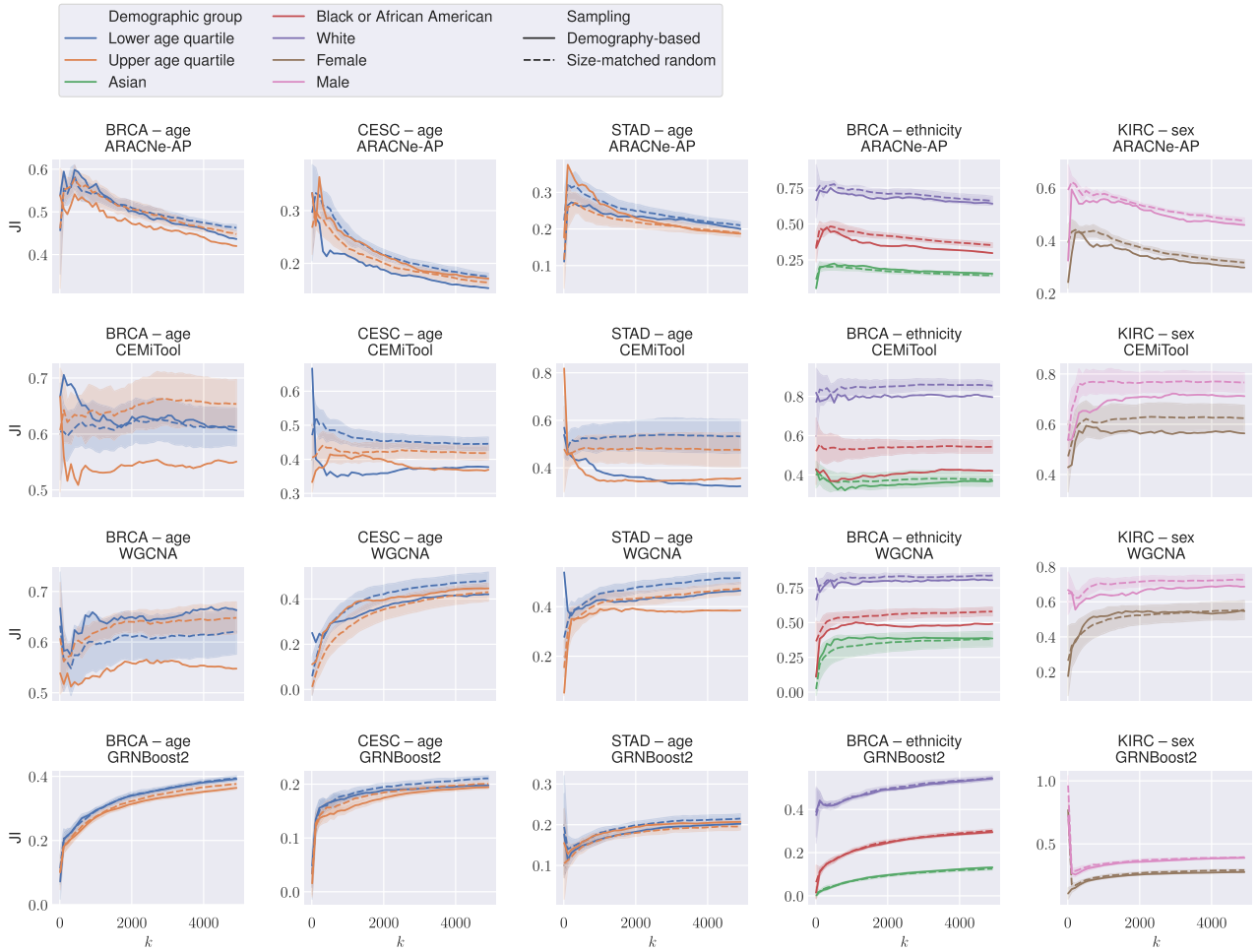
**Figure 2.** Results of the first part of our test protocol (Figure 1A). (A) Fractions of significant P-values at 5% significance level across  $k \in \{10, 110, \dots, 4910\}$  for each of the three fast network inference tools ARACNe-AP, CEMiTool and WGCNA, grouped by cohort for the age, ethnicity and sex confounders. P-values were obtained via the MWU test for the alternative hypothesis  $JI_k(C) < JI_k(R)$ . Since in the GBM and READ cohorts all blocks of the ethnicity except for one contain less than 20 samples, no results are reported for the ethnicity confounder. Similarly, the BRCA and CESC cohorts only contain 12 and 0 male samples, respectively; so no results are reported for the sex confounder. (B) P-values of the  $\chi^2$ -test between the demographic confounders and, respectively, alcohol history, cancer stage and smoking history. Missing P-values for some cohorts indicate that no data on alcohol history, cancer stage or smoking history are available. (C) Fractions of significant P-values at 5% significance level as shown in (A), grouped by cohort for the alcohol history, stage, and smoking history variables. Tests were conducted only for such cohorts for which more than 50% of the P-values in (A) are significant for at least one confounder and for which the P-values of the  $\chi^2$ -test shown in (B) are significant on a 5% level. (D) Negative log-transformed MWU P-values for varying  $k$  obtained for GRNBoost2 on a selection of the cohort–confounder combinations. The red lines show the significance threshold 0.05. Tests were conducted only for such cohort–confounder combinations for which more than 50% of the P-values in (A) are significant for at least one confounder and for which either the  $\chi^2$ -tests between the respective confounder and the three variables shown in (B) are not significant or the fraction of significant P-values shown in (C) for the particular variable is smaller than 50% for at least one method. The line plots visualizing the distributions  $JI_k(C)$  and  $JI_k(R)$  underlying the P-values shown in (A), (C) and (D) are provided in Supplementary Figures 2–13.

available. Figure 2B shows a heatmap for each variable with the P-values of  $\chi^2$ -tests used to assess whether the three demographic confounders are independent of the three variables. Indeed, in the HNSC cohort, age and alcohol history variables are correlated. In the CESC, HNSC and LUAD cohorts, the age confounder depends on the smoking history variable. In the COAD cohort, the age confounder significantly depends on the stage variable. Consequently, we ran the first part of the test protocol (Figure 1A) for the cohort–variable combinations HNSC–alcohol history, CESC–smoking history, HNSC–smoking history, LUAD–smoking history and COAD–stage to rule out that the observed effect of the age

confounder on the three fast network inference methods might be an effect of the respective dependent variable.

Figure 2C shows the fractions of significant MWU P-values obtained for these cohort–variable combinations, for the fast tools ARACNe-AP, CEMiTool and WGCNA. Alcohol history has a strong effect in the HNSC cohort, smoking history has a strong effect in the LUAD cohort, and cancer stage has a strong effect in the COAD cohort. Therefore, the cohort–confounder combinations HNSC–age and COAD–age were excluded from any further analyses, despite the strong effect of age on network inference in these cohorts.





**Figure 3.** Results of the second part of our test protocol (Figure 1B). The plots show Jaccard indices between the top- $k$  edges of networks inferred from gene expression data for, respectively, all patients and patients from specific demographic groups (solid line plots) or size-matched random subsamples. For the size-matched randomly sampled patient groups, mean JIs (dashed line plots) and standard deviations (shaded areas) are shown for each  $k$ .

For the remaining eight cohort-confounder combinations with strong confounder effects (BRCA-age, CESC-age, PCPG-age, STAD-age, BRCA-ethnicity, KIRC-sex, KIRC-sex and LUSC-sex), we then ran the test protocol sketched in Figure 1A also with the runtime-intensive GRN inference tool GRNBoost2. The results are shown in Figure 2D. For five out of eight combinations (BRCA-age, BRCA-ethnicity, CESC-age, KIRC-sex and STAD-age), the MWU tests were significant for more than 50% of the edge cutoffs  $k$  also for GRNBoost2. For these combinations, we hence consistently observe strong confounder effects across all tested methods which cannot be explained by an implicit approximation of the stratification by one of the three variables alcohol history, cancer stage or smoking history.

### Comparison of networks inferred from data for specific demographic groups to networks inferred from entire datasets

In the second part of our test protocol (Figure 1B), we assessed how similar networks inferred from gene expression data for specific demographic groups are to confounder-agnostic networks inferred from the entire datasets. For this, we focused on the five cohort-confounder combinations where we observed strong confounder effects for all tested network inference methods that cannot be explained by a dependency on cancer stage, smoking or alcohol history (see last subsection). The results are shown in Figure 3.

The first question which arises in this context is how well the confounder-agnostic networks represent the networks inferred for the different demographic groups. This question can be answered by examining the JI distributions  $JI_k^l(\mathcal{C})$  obtained for the different demographic groups (solid line plots in Figure 3). The first striking observation is that all demographic groups are represented relatively poorly with JIs between around 0.2 and 0.75. Moreover, for all three confounders, the networks for the demographic groups with the largest number of samples (Table 1) are best represented by the confounder-agnostic networks (lower quartile for the age confounder, White patients for the ethnicity confounder, male patients for the sex confounder).

These results raise the question to which extent the observed differences can be explained by the sizes of the demographic groups alone. To answer this question, we computed JI distributions  $JI_k^l(\mathcal{R})$  between the confounder-agnostic networks and networks inferred for randomly sampled patient groups whose sizes were matched to the sizes of the different demographic groups (dashed line plots with shaded error intervals in Figure 3). Whenever we have  $JI_k^l(\mathcal{C}) < JI_k^l(\mathcal{R})$  for a demographic group  $l$  across many  $k$  (solid line plots below dashed line plots with error intervals of the same color), this is evidence that (dis-)similarities between the confounder-agnostic networks and the networks inferred for the demographic group  $l$  cannot be explained by the sizes of the demographic group alone.

For the sex confounder and the KIRC cohort, we have  $JI_k^{\text{male}}(\mathcal{R}) > JI_k^{\text{female}}(\mathcal{R})$  across all  $k$  for all methods. This suggests that the observation that networks inferred for male patients are better represented by confounder-agnostic networks can largely be explained by the fact that the KIRC dataset contains more samples from male patients. However, not all the observed differences can be explained by the sizes of the demographic groups alone: For instance, for all methods except GRNBoost2, the networks for male KIRC patients are clearly less similar to confounder-agnostic networks than those for patients from size-matched random subgroups ( $JI_k^{\text{male}}(\mathcal{C}) < JI_k^{\text{male}}(\mathcal{R})$ , see pink line plots in column 6 of Figure 3).

For the ethnicity confounder in the BRCA cohort, we have  $JI_k^{\text{White}}(\mathcal{R}) > JI_k^{\text{BAA}}(\mathcal{R}) > JI_k^{\text{Asian}}(\mathcal{R})$ , which again reflects the numbers of samples for White, Black or African American, and Asian patients in the TCGA BRCA cohort. Interestingly, the sizes of the demographic groups largely explain the results obtained for the White and Asian patient groups but not those for Black or African American patients ( $JI_k^{\text{BAA}}(\mathcal{C}) < JI_k^{\text{BAA}}(\mathcal{R})$  for all  $k$  and all methods except GRNBoost2, see red line plots in column 5 of Figure 3).

For the age confounder, the  $JI$  distribution for the size-matched randomly sampled patient groups lie closer together, as the sizes of the two demographic groups (lower and upper quartiles) are nearly balanced. As for the other two confounders, not all variation can be explained by the sizes of the demographic groups alone. For instance, in the BRCA cohort, we consistently have  $JI_k^{\text{upper}}(\mathcal{C}) < JI_k^{\text{upper}}(\mathcal{R})$ , suggesting that older BRCA patients are more poorly represented by confounder-agnostic networks than expected by chance (see orange line plots in column 1 of Figure 3). For the CESC cohort, we observe a particularly poor representation of younger patients ( $JI_k^{\text{lower}}(\mathcal{C}) < JI_k^{\text{lower}}(\mathcal{R})$ , see blue line plots in column 2 of Figure 3).

## Replication using data from the METABRIC breast cancer cohort

To assess if our test protocol can reveal confounder effects that generalize to cohorts not used for discovery, we replicated the analyses for the age confounder and the TCGA BRCA cohort using gene expression data from METABRIC. Recall that our analyses on the TCGA BRCA cohort indicate that age strongly distorts GCN and GRN inference in breast cancer (Figure 2) and that this distortion cannot solely be explained by differing sizes of the individual demographic groups (Figure 3).

The results of our replication study are shown in Figure 4. For all four network inference methods and almost all edge ranks  $k$ , similarities between networks inferred for young and old patients are significantly smaller than similarities between networks inferred for size-matched random patient groups. Again, the subgroup-specific networks are poorly represented by confounder-agnostic networks, and again this effect cannot be explained by subgroup sizes alone (Supplementary Figure 15). Our replication study hence indicates that the observed effect of age on GRN and GCN inference in breast cancer generalizes across several cohorts.

## Method-centric view on the results

Another interesting question is whether some network inference methods are more sensitive to demographic confounders than others. By and large, this answer can be answered negatively: As can be seen in Figure 5A, the distributions of fractions of significant MWU P-values are similar for all three fast network inference methods for which we ran the first part of our test protocol for all cohort-confounder combinations. Moreover, the

fractions of significant MWU P-values obtained for the three tools are positively correlated (Figure 5B). Together, these results indicate that demographic confounders distort GCN and GRN inference independently of the employed method.

We do, however, observe a clear difference between the tested network inference methods when looking at the distributions of mean  $JIs$  between different blocks of random partitions (Figure 5C). Here, the two GRN inference methods ARACNe-AP and GRNBoost2 yield smaller mean  $JIs$  than the GCN inference methods WGCNA and CEMiTool, which indicates that ARACNe-AP and GRNBoost2 are more sensitive to random bias. Yet, the mean  $JIs$  are disturbingly small also for CEMiTool and WGCNA (recall that the blocks of the random partitions underlying the results shown in Figure 5C are i. i. d. samples from the same datasets). As previously observed for disease mechanism mining in protein-protein interaction networks [28, 29], random effects hence have a major effect on all four tested network inference methods.

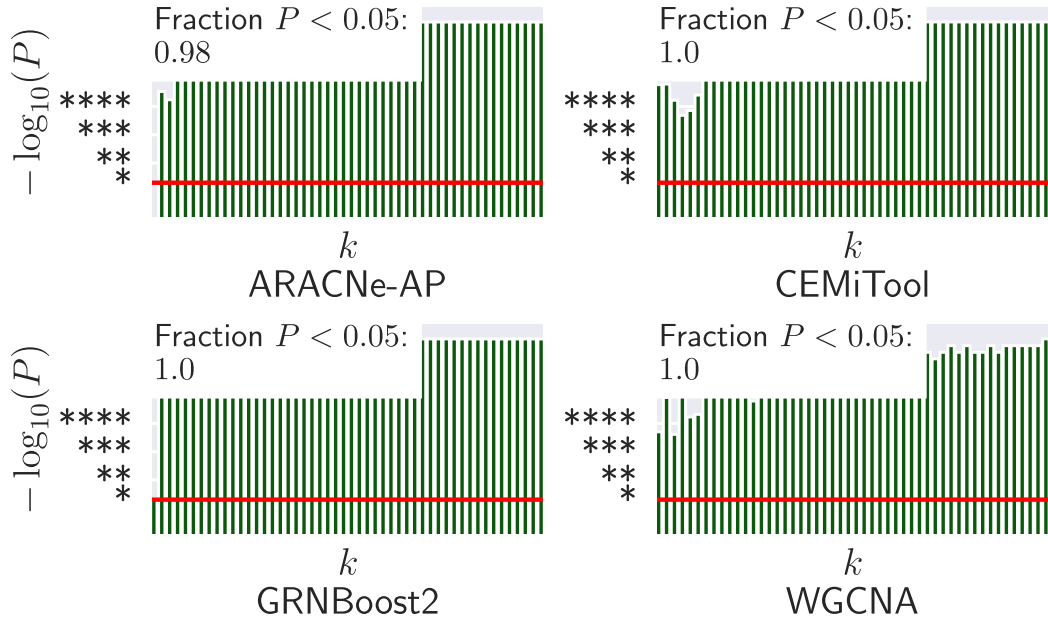
## DISCUSSION

Studies using network inference methods such as WGCNA, CEMiTool, ARACNe-AP and GRNBoost2 are usually conducted without stratifying for demographic variables such as age, sex or ethnicity. However, our results show that the GRNs and GCNs inferred using four different computational methods differ substantially between different demographic groups, in some cohorts with high agreement between the tested methods. The effects of sex and age confounders stand out, leading to under-representation of certain groups that cannot be explained by block sizes alone. At the same time, the confounder-agnostic networks inferred from the aggregate data—i. e. the networks commonly used in network inference studies—often do not properly represent any of the individual demographic groups. In this section, we discuss the key implications of our results and sketch possible approaches and obstacles toward controlling for confounding in GCN and GRN inference studies.

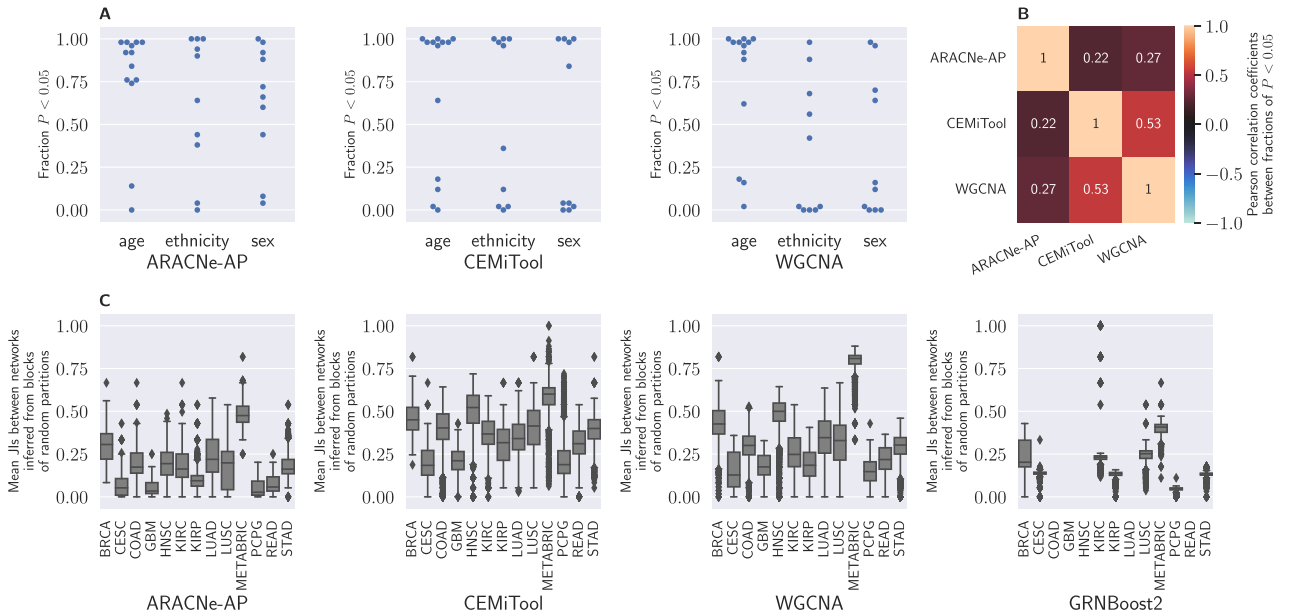
## Uncovered confounder effects align with previous findings

Several studies align with our findings that age is a strong confounder for GCN and GRN network inference in CESC, TCGA BRCA and METABRIC cohorts. For instance, Azim *et al.* [30] and Gómez-Flores-Ramos *et al.* [31] have identified several differential molecular characteristics when comparing older and younger breast cancer patients, some of which are linked to survival. Lu *et al.* [32] associate DNA methylation age with chronological age in cervical squamous cell carcinoma (a subtype of CESC) and show that DNA methylation age is a predictor for survival. Further studies [33, 34] support the role of age as a prognostic factor in cervical squamous cell carcinoma.

Peired *et al.* [35] reviewed sex differences in kidney cancer. Despite pointing out that most kidney cancers do not present sex-related differences in genetics, they portray the correlation of a higher incidence of certain renal cell carcinoma subtypes in females with a mutation in the TSC1 or TSC2 gene, uterine tumors and DNA damage induced by gynecological surgery. Overall, the different types of kidney cancer occur significantly more often in men, suggesting that the possibility of sex-specific molecular differences should be considered. In line with these findings, our results indicate that sex is a strong confounder for GRN and GCN inference in the KIRC cohort.



**Figure 4.** Results of replication study on the METABRIC breast cancer cohort (negative log-transformed MWU  $P$ -values for varying  $k$ ). For all four network inference methods, we obtained significant MWU  $P$ -values for almost all  $k$ , showing that the strong effect of the age confounder on network inference discovered in the TCGA BRCA cohort generalizes to the METABRIC cohort. Line plots visualizing the distributions  $J_{I_k}(C)$  and  $J_{I_k}(R)$  underlying the  $P$ -values are provided in [Supplementary Figure 14](#).



**Figure 5.** (A) Fractions of significant  $P$ -values at 5% significance level across  $k \in \{10, 110, \dots, 4910\}$  for each of the three fast network inference tools ARACNe-AP, CEMiTool and WGCNA, aggregated across all cohorts. For each method-cohort combination, one dot corresponds to one cohort.  $P$ -values were obtained via the MWU test for the alternative hypothesis  $J_{I_k}(C) < J_{I_k}(R)$ . Overall, the distributions of the fractions are similar for the three methods, showing that no method is clearly more or less sensitive to demographic confounders than the others. (B) Pearson correlation coefficients between the fractions of significant  $P$ -values obtained for the three methods. The obtained positive correlations again indicate that the effects of demographic confounders on GRN and GCN inference are not method-specific. (C) Distributions of mean JIs  $J_{I_k}(R)$  between networks inferred from different blocks of random partitions generated for the first part of the test protocol ([Figure 1A](#)), aggregated across all runs and all  $k$  for each method-cohort combination.

Roelands et al. [14] investigate molecular differences among breast cancer patients of African, Arab and European ancestry. Their findings suggest that differences in cancer-related pathways between the ethnic groups lead to disparities in clinical outcomes, which aligns with the effect of the ethnicity confounder observed in our results for the TCGA BRCA cohort. Similarly, a recent epidemiological study [15] suggest that Black or African American

triple-negative breast cancer patients have a higher risk of mortality than White patients, even after correcting for treatment-regimen and socio-economic background. This is again in line with our results, where networks inferred from gene expression data for Black or African American breast cancer patients are particularly poorly represented by confounder-agnostic networks.

## Limitations

It is important to note that our study is subject to at least three limitations. Firstly, we controlled for a possible effect of third variables by considering cancer stage, as well as smoking and alcohol history. However, data about these variables are available only for a subset of the considered cohorts. Moreover, there may be other third variables distorting our results (e. g. diet, medication, comorbidities) for which no data are available in TCGA and METABRIC. We hence cannot exclude that some of the effects we attributed to demographic confounders may in fact be caused by correlated third variables.

Secondly, it has previously been reported that some samples in TCGA are misannotated [36]. To visually inspect samples with possibly incorrect sex annotations, we used MODMatcher [36] to jointly plot the expression of the Y-linked genes RPS4Y1 and DDX3Y (Supplementary Figure 16), which are higher expressed in males than in females [37]. The plots show that male and female samples are mostly well separated but also that, for some cohorts, individual samples might have a wrong sex annotation. Similar misannotations might of course be also present for the other confounding variables considered in this study. Therefore, we cannot exclude that some of our results are slightly distorted by sample misannotations.

Thirdly, we would like to stress that all results reported in this study are cancer-type-specific. To assess if demographic confounders also affect GRN and GCN inference in cancer types not considered here, researchers can use the Python test framework we developed for this study. Thanks to its modular design, our framework can also easily be extended to test the effect of demographic confounders on network inference methods other than WGCNA, CEMiTool, ARACNe-AP and GRNBoost2.

## Implications

Our study has several important practical implications: Firstly and most importantly, we showed that confounder-agnostic networks often poorly represent networks inferred for different demographic groups, with JIs ranging from around 0.2 to 0.75. This implies that individual demographic groups are often not fully captured in confounder-agnostic networks, potentially leading to an incomplete understanding of gene regulatory interactions in specific populations. In line with existing studies which highlight the importance of close-up analyses in network medicine [38], our findings hence emphasize the need to establish data dis-aggregation as a standard procedure in network inference from gene expression data. Concretely, we suggest a two-step protocol:

- Step 1: Quantify to which extent network inference is distorted by demographic confounders in your specific use case, e. g. using the Python package we developed for this study.
- Step 2: If step 1 reveals a strong effect of demographic confounders, infer and interpret networks separately for the individual demographic groups.

A second related implication is that methods such as the one proposed by Parsana *et al.* [39] which correct for confounding effects in network inference should be used very carefully. Instead of correcting for demographic confounders, our results suggest that it might be more promising to infer separate confounder-specific molecular networks: While a high inter-method variability of the results would have suggested that susceptibility to demographic confounding is largely due to specifics of the network inference methods (and hence should be controlled for), we obtained remarkably consistent results, especially for the age and the sex confounders. This indicates that the gene expression data indeed often contain biologically rooted age- and sex-specific

signals and that picking up on those signals is hence a feature (and not a bug) of GRN and GCN inference methods.

Thirdly, our study provides an additional argument for the importance of studies that generate molecular data for currently under-represented demographic groups [40, 41]. For instance, in the TCGA datasets used for this study, certain ethnic groups are under-represented to the extent that their corresponding blocks had to be omitted from the tests to meet the minimum sample size requirements of the tested inference methods. In these cases, data dis-aggregation cannot be applied to control for confounding, as the groups are too small to be utilized individually for network inference.

Fourthly, our study implies that, whenever possible, demographic information should be made available along with molecular data. For instance, the main reason we choose TCGA datasets for this study is that, in TCGA, demographic information is provided for all samples (although some uncertainty is cast on the precision of the data, especially with regard to the ethnicity confounder, since the demographic annotations are based on self-reports of the patients). Such information is lacking for many other publicly available omics datasets [41], impeding data dis-aggregation and confounder-aware network inference at the outset.

Finally, we observed as a side result that the tested GRN and GCN inference methods are very sensitive to random bias: Even when networks are inferred from equally sized i. i. d. samples from the same cohort, the obtained mean JIs are mostly smaller than 0.5 (Figure 5). This suggests that measures to increase robustness (e. g. ensemble inference or bootstrapping) should be included in GCN and GRN inference workflows.

## Outlook

While we used bulk gene expression data for this study, huge amounts of single-cell RNA sequencing (scRNA-seq) data have become available during the past years. Consequently, various methods have been proposed that infer GRNs or GCNs from scRNA-seq data (sometimes in combination with other single-cell omics data modalities such as scATAC-seq) [42]. A natural question for future work is hence if demographic confounders similarly affect GCN and GRN inference from scRNA-seq data. Given sufficiently large multi-sample scRNA-seq datasets with demographic information, this question can in principle be answered using the framework proposed in this article by propagating demographic annotations from samples to cells. When doing so, it will be crucial to account for possible effects of sample-specific cell type compositions, which can be achieved using techniques similar to our approach to correct for the effects of cancer stage, smoking and alcohol history.

## METHODS

### Compared network inference methods

The presented tests were carried out using two different GRN inference methods (ARACNe-AP [21], GRNBoost2 [22]) and two different GCN inference methods (WGCNA [23], CEMiTool [24]). As criteria for method selection, we required that the methods only use gene expression data as input, that well-documented reference implementations are available and that they are sufficiently fast to allow inferring the very large number of networks required for our test protocol within a reasonable time frame. The four tools were selected because they respect these criteria and are widely used in the community: As of August 2023, WGCNA has been cited more than 15000 times, GRNBoost2 and ARACNe-AP have been cited more than 2500 times (plus more than 1500 and



2900 citations for their predecessors GENIE3 [4] and ARACNe [43]) and CEMiTool has been cited more than 200 times (citation counts from Google Scholar). In total, 44880 networks were inferred for this study.

ARACNe-AP is an improved and more runtime-efficient version of ARACNe. It models the genes in the expression dataset as random variables and estimates pairwise mutual information (MI) of the genes using a Gaussian kernel estimator. Indirect regulations are removed via scanning all triangles  $(u, v)$ ,  $(v, w)$ ,  $(u, w)$  and removing the edge  $(u, w)$  if its MI is smaller than the MIs of  $(u, v)$  and  $(v, w)$ . Edges are scored using P-values obtained via bootstrapping tests.

GRNBoost2 is a tree-based ensemble method which is very similar to the widely used tool GENIE3. Like GENIE3, it computes a random forest for each gene in the dataset, where the expression values of possible regulators (transcription factors) are used as design variables. The obtained feature importance scores are interpreted as the scores of edges from the predicting transcription factor to the target gene. A final ranked edge list is obtained by merging the edge lists of all forests and then sorting in decreasing order w. r. t. the edge scores. Unlike GENIE3, GRNBoost2 uses gradient boosting when learning the random forest by iteratively adding shallow decision trees to the ensemble until an early-stopping criterion is met. This significantly decreases the runtime, which is the reason why we decided to use GRNBoost2 instead of GENIE3.

Both WGCNA and CEMiTool use the absolute correlation matrix of all gene expression profiles as adjacency matrix of the inferred GCN. To facilitate preprocessing, CEMiTool applies a variance-stabilizing transformation to the gene expression data and implements gene filtering based on the inverse gamma distribution. Note that WGCNA and CEMiTool are deterministic tools, while ARACNe-AP and GRNBoost2 involve randomization.

## Gene expression and demographic data from TCGA

We used gene expression and demographic data for twelve cancer types (see Table 1) from TCGA, which we downloaded from the UCSC Xena Hub [26]. We only kept protein-coding genes (obtained from the HGNC Database [44]). For the tested GRN inference methods, we used all known human transcription factors [45] as potential regulators. The data were further filtered by only keeping samples from primary tumor tissue (i. e. samples with entry “Primary Tumor” in the column “sample\_type.sample” of the phenotype files) that appear in both the gene expression and the phenotype data. Moreover, we removed genes with a standard deviation of zero across all samples. Since the gene expression data provided by the UCSC Xena Hub are  $\log(x + 1)$ -transformed and the authors of CEMiTool and ARACNe-AP recommend not to log-transform the data, we reversed the transformation for these methods.

The following columns of the phenotype files were used to obtain information about demographic confounders and third variables:

- Age: “age\_at\_initial\_pathologic\_diagnosis”.
- Sex: “gender.demographic”.
- Ethnicity: “race.demographic”.
- Cancer stage: “tumor\_stage.diagnoses”.
- Alcohol history: “alcohol\_history.exposures”.
- Smoking history: “tobacco\_smoking\_history”.

Before partitioning, we removed samples from the dataset for which the respective column is empty or contains “not reported”.

For the cancer stage variable, we further removed stage-X samples. The sex and ethnicity variables contain categorical information, which we used to partition the samples into blocks of demographic groups. To allow stable network inference, blocks containing less than 20 samples (the minimum sample size according to the documentations of WGCNA and CEMiTool) were removed. For the stage variable, we proceeded similarly, but instead of removing blocks with fewer than 20 samples, we merged the blocks for stage III and stage IV (for stage I and II, all datasets where stage information is available contain at least 20 samples). In order to form a partition by age, we split the datasets into age quartiles and then removed the second and third quartiles. That is, for each dataset, we obtained two age-induced blocks: one for the first (lower) and one for the fourth (upper) quartile. Alcohol history of a patient is encoded as a binary variable in the TCGA datasets, with values “Yes” and “No”. Thus, the induced partition contains two blocks. Smoking history is encoded with numeric values ranging from 1 to 5 (1: lifelong non-smoker; 2: current smoker; 3: current reformed smoker for more than 15 years; 4: current reformed smoker for less than or equal to 15 years; 5: current reformed smoker, duration not specified).

## Gene expression and demographic data from METABRIC

We used the METABRIC breast cancer dataset containing microarray gene expression data of breast cancer patients for replication. The dataset was obtained from the cBio portal [46, 47], along with a phenotype file containing age and sex information of the samples. As in the TCGA datasets, we only kept the protein-coding genes from HGNC and used all known human transcription factors [45] as potential regulators for the tested GRN inference methods. Again, only samples from primary tumor tissue (i. e. samples with entry “Primary” in the column “Sample Type” of the phenotype file) that appear in both the gene expression and the phenotype data were used. Genes with a standard deviation of zero across all samples were removed. As in the TCGA datasets, the METABRIC BRCA gene expression data are  $\log(x + 1)$ -transformed, hence the transformation was reversed for CEMiTool and ARACNe-AP.

## Method-specific pre-processing

Method-specific pre-processing was carried out as described by the authors and considered part of the methods. Therefore, we applied method-specific pre-processing separately to each block of the partitions. For WGCNA, we followed the instructions for FPKM data given in the software FAQ provided by the authors. Genes with an expression signal of zero were removed, and of the remaining genes, only the 50% most variable ones were used for network construction. For CEMiTool, we followed the instructions given for high-throughput RNA sequencing data in the supplementary material. As recommended, we applied CEMiTool’s in-built variance-stabilizing transformation. For ARACNe-AP, zero-expressed genes were removed from the data. Note that the authors of ARACNe-AP use RPKM data to benchmark their method. Since UCSC Xena Hub does not host RPKM data, we instead used FPKM data. For GRNBoost2, genes with a standard deviation of zero were removed from the gene expression data. The remaining genes were normalized to unit variance to ensure comparability of edges with different target genes. We used the default parameters to run GRNBoost2, as recommended by the authors.

## Partitioning the data and inferring the networks

Let  $\mathcal{C} = \{C^l \mid l \in \{1, \dots, L\}\}$  be the sample partition induced by one of the demographic confounders or third variable described above on one of the employed datasets (note that some samples might have been removed to ensure block sizes  $|C^l| \geq 20$ ).  $L$  is the number of induced subgroups, i. e.  $L = 2$  for age, sex and alcohol history, and  $L = 3$  for cancer stage. For the ethnicity confounder, we have  $L = 2$  or  $L = 3$ , depending on for how many ethnic groups the datasets contain at least 20 samples (see Table 1). For smoking history, we have  $L = 4$ , since the group 5 (“current reformed smokers, duration not specified”) contains less than 20 samples for all cohorts. Given  $\mathcal{C}$ , we generated 100 size-matched random partitions  $\mathcal{R}_i = \{R_i^l \mid l \in \{1, \dots, L\}\}$  with  $\bigcup \mathcal{C} = \bigcup \mathcal{R}_i$  and  $|R_i^l| = |C^l|$ , for all  $(l, i) \in \{1, \dots, L\} \times \{1, \dots, 100\}$ .

For each of the four tested network inference methods and each block  $C^l \in \mathcal{C}$  of the partition induced by the phenotype variable, we then inferred 10 networks  $G_i^l(C) = (V, E_i^l(C))$ ,  $i = \{1, \dots, 10\}$ , by running the inference method 10 times on the block  $C^l$ . Here,  $V$  is the set of protein-coding genes and  $E_i^l(C)$  is the set of edges inferred for block  $l$  in the  $i^{\text{th}}$  run. We repeated the network inference to account for the fact that some of the inference methods (ARACNe-AP and GRNBoost2) are non-deterministic. Similarly, for each block  $R_i^l$  of each of the 100 size-matched random partitions, we inferred one network  $G_i^l(\mathcal{R}) = (V, E_i^l(\mathcal{R}))$ . For each dataset, network inference method and phenotype variable, we hence inferred 110 networks for each subgroup induced by the phenotype variables: 10 for the subgroup itself and 100 for size-matched random sample sets. Moreover, for each dataset and network inference method, we inferred 10 confounder-agnostic networks  $G_i = (V, E_i)$ ,  $i \in \{1, \dots, 10\}$ , by running the inference method 10 times on the aggregate dataset.

## Quantifying network similarities

For each  $k \in \{10, 110, \dots, 4910\}$ , we computed a distribution  $J_k(C) = (J_{k,i}(C))_{i=1}^{10}$  of 10 mean JIs which quantify the similarity of the top- $k$  edges of the networks  $G_i^l(C)$  inferred from different blocks of the phenotype-induced partition  $\mathcal{C}$ . The mean JIs were computed as

$$J_{k,i}(C) = \binom{L}{2}^{-1} \sum_{l=1}^{L-1} \sum_{l'=l+1}^L \frac{|E_{i,k}^l(C) \cap E_{i,k}^{l'}(C)|}{|E_{i,k}^l(C) \cup E_{i,k}^{l'}(C)|}, \quad (1)$$

where  $E_{i,k}^l(C)$  and  $E_{i,k}^{l'}(C)$  are the sets of the top- $k$  highest scored edges contained in  $E_i^l(C)$  and  $E_i^{l'}(C)$ , respectively (if  $|E_i^l(C)| < k$  or  $|E_i^{l'}(C)| < k$ , we set  $E_{i,k}^l(C) = E_i^l(C)$  or  $E_{i,k}^{l'}(C) = E_i^{l'}(C)$ , respectively). Distributions  $J_k(\mathcal{R}) = (J_{k,i}(\mathcal{R}))_{i=1}^{100}$  of mean JIs quantifying the similarities of networks inferred from different blocks of the size-matched random partitions were computed analogously.

We also quantified similarities between the networks inferred from the different blocks of the partitions and the networks inferred from the entire datasets. For the  $i^{\text{th}}$  block of the confounder-induced partition  $\mathcal{C}$ , we computed JI distributions  $J_k^l(C) = (J_{k,i}^l(C))_{i=1}^{10}$  as

$$J_{k,i}^l(C) = \frac{|E_{i,k}^l(C) \cap E_{i,k}|}{|E_{i,k}^l(C) \cup E_{i,k}|}, \quad (2)$$

where  $E_{i,k}^l(C)$  is defined as above and  $E_{i,k}$  is the set of top- $k$  highest scored edges from the edge set  $E_i$  of the network  $G_i$  inferred in the  $i^{\text{th}}$  run on the entire dataset. For the  $l^{\text{th}}$  blocks of the random partitions, JI distributions  $J_k^l(\mathcal{R}) = (J_{k,i}^l(\mathcal{R}))_{i=1}^{10}$  were computed

analogously. Note that only the first 10 random partitions were used here, since we only inferred 10 networks  $G_i$  from the aggregate datasets.

## Statistical tests

For each cancer type, network inference method, phenotype variable and edge cutoff  $k \in \{10, 110, \dots, 4910\}$ , we compared the distributions  $J_k(C)$  and  $J_k(\mathcal{R})$  using a one-sided MWU test with alternative hypothesis  $J_k(C) < J_k(\mathcal{R})$ . We hence obtained a  $P$ -value for each  $k$ , which tells us whether the top- $k$  edges of networks inferred from different blocks of confounder-induced partitions are significantly less similar to each other than expected by chance. Then, we computed the fractions of significant  $P$ -values at a 5% significance level across all  $k$  (Figure 2A). Moreover, for such cancer types and demographic confounders for which at least 50% of the  $P$ -values are significant for all three methods, we carried out a  $\chi^2$ -test to check for dependencies between the demographic confounders and three possible third variables, alcohol history, smoking history and cancer stage. If the  $\chi^2$ -test with at least one variable is significant for a specific confounder, disaggregating the data by the confounder causes implicit disaggregation by the variable. In this case, a significant MWU  $P$ -value for the confounder-induced partition might be explained by the third variable if also the MWU  $P$ -value for the third variable-induced partition is significant.

## Implementation

We developed a Python package implementing the functionalities of the test protocol which can be installed locally from a `pyproject.toml` file and the source code using `pip`. The package consists of a `TestRunner` class to be instantiated with all parameters of the tests to be performed, most importantly, the names of the algorithms, cancer types and confounders. An abstract `NetworkInferenceWrapper` interface provides the requirements of the concrete wrapper classes, each wrapping one of the tested network inference methods. The wrappers execute the tested network inference methods—either directly or via system calls to the original executable or to a caller-script implemented in the programming language of the respective inference method. Statistical tests are implemented with `SciPy` [48]. `Matplotlib` [49] and `seaborn` [50] are used for visualization.

The package contains a runner script `run_tests.py` that can be used as-is to reproduce the presented results, as well as on other TCGA datasets or on custom data. Developers can set up individual tests by inheriting from the `NetworkInferenceWrapper`. User parameters can be specified with a command line tool or by providing configuration files. Users can also use the package to test new methods using the `CustomGCNWrapper` to wrap methods that infer networks with undirected edges, or the `CustomGRNWrapper` for inference of networks with directed edges. The custom wrappers can be called from `run_tests.py` using the algorithm identifiers “CUSTOMGCN” or “CUSTOMGRN”, respectively. The package, runner scripts and a detailed README are available at <https://github.com/bionetslab/grn-confounders/>.

### Key Points

- Sex, age and ethnicity strongly confound gene regulatory and gene co-expression network inference in several cancer types.

- The age confounder stands out, since it has a strong impact on network inference at particularly high consensus of the network inference methods.
- Under-representation of some demographic subgroups in networks inferred from aggregate data cannot be explained by sample size alone.

## SUPPLEMENTARY DATA

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

## ACKNOWLEDGEMENTS

Figure 1 and Supplementary Figure 1 were generated with BioRender.com.

## FUNDING

AK and DBB were supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the CompLS funding concept [031L0309A (NetMap)].

## DATA AVAILABILITY

The gene expression data and corresponding phenotype data of the tested cohorts were obtained from the GDC Hub hosted by UCSC Xena: <https://xenabrowser.net/datapages/> (version 18.0 from the GDC data portal, last updated on 28 August 2019). Gene expression and phenotype data from the METABRIC cohort were obtained from the cBio portal: [https://www.cbioportal.org/study/summary?id=brca\\_metabric](https://www.cbioportal.org/study/summary?id=brca_metabric). We used the list of protein-coding genes published at <https://www.genenames.org/download/statistics-and-files/> (protein-coding gene from the "Statistics" category). The list of known human transcription factors used by the GRN methods was obtained from <http://humantfs.cbr.utoronto.ca/download.php> (human TFs, full database). The package source code and scripts to reproduce the reported results are available at <https://github.com/bionetslab/grn-confounders/>. To facilitate reproducibility, the repository contains the script `download_test_data.py`, which allows to automatically download all test datasets. An AIME report [51] to further enhance reproducibility is available at <https://aime.report/Wj2jFr>.

## AUTHOR CONTRIBUTIONS STATEMENT

AK and DBB conceived and designed this study and wrote the manuscript. AK implemented the Python package and carried out the analyses. DBB supervised this work.

## REFERENCES

1. Elkon R, Linhart C, Sharan R, et al. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res*2003;**13**(5):773–80.
2. Basso K, Margolin AA, Stolovitzky G, et al. Reverse engineering of regulatory networks in human B cells. *Nat Genet*2005;**37**(4):382–90.
3. Faith JJ, Hayete B, Thaden JT, et al. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*2007;**5**(1):e8.
4. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*2010;**5**(9):e12776.
5. Sanz-Pamplona R, Berenguer A, Cordero D, et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol Cancer*2014;**13**(1).
6. Hasankhani A, Bahrami A, Sheybani N, et al. Differential co-expression network analysis reveals key hub-high traffic genes as potential therapeutic targets for COVID-19 pandemic. *Front Immunol*2021;**12**.
7. Yingcheng W, Wei J, Chen X, et al. Comprehensive transcriptome profiling in elderly cancer patients reveals aging-altered immune cells and immune checkpoints. *Int J Cancer*2018;**144**(7):1657–63.
8. Shah Y, Verma A, Marderstein AR, et al. Pan-cancer analysis reveals molecular patterns associated with age. *Cell Rep*2021;**37**(10):110100.
9. Chatsirisupachai K, Lesluyes T, Paraoan L, et al. An integrative analysis of the age-associated multi-omic landscape across cancers. *Nat Commun*2021;**12**(1).
10. Li CH, Haider S, Boutros PC. Age influences on the molecular presentation of tumours. *Nat Commun*2022;**13**(1).
11. Lee W, Wang Z, Saffern M, et al. Genomic and molecular features distinguish young adult cancer from later-onset cancer. *Cell Rep*2021;**37**(7):110005.
12. Dong M, Cioffi G, Wang J, et al. Sex differences in cancer incidence and survival: a pan-cancer analysis. *Cancer Epidemiol Biomarkers Prev*2020;**29**(7):1389–97.
13. Yang W, Warrington NM, Taylor SJ, et al. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. *Sci Transl Med*2019;**11**(473).
14. Roelands J, Mall R, Almeer H, et al. Ancestry-associated transcriptomic profiles of breast cancer in patients of african, Arab, and european ancestry. *npj Breast Cancer*2021;**7**(1).
15. Cho B, Han Y, Lian M, Colditz G, Weber JD, Ma C, and Liu Y. Evaluation of racial/ethnic differences in treatment and mortality among women with triple-negative breast cancer. *JAMA Oncol*, **7**(7):1016–1023, July 2021.
16. Esnaola NF, Ford ME. Racial differences and disparities in cancer care and outcomes. *Surg Oncol Clin N Am*2012;**21**(3):417–37.
17. Aguilar B, Abdilleh K, Acquah-Mensah GK. Multi-omics inference of differential breast cancer-related transcriptional regulatory network gene hubs between young black and white patients. *Cancer Genet*2023;**270–271**: 1–11.
18. Lopes-Ramos CM, Kuijjer ML, Ogino S, et al. Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer Res*2018;**78**(19):5538–47.
19. Kuijjer ML, Tung MG, Yuan G, et al. Estimating sample-specific regulatory networks. *iScience*2019;**14**:226–40.
20. Lopes-Ramos CM, Chen C-Y, Kuijjer ML, et al. Sex differences in gene expression and regulatory networks across 29 human tissues. *Cell Rep*2020;**31**(12):107795.
21. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*2016;**32**(14):2233–5.
22. Moerman T, Santos SA, González-Blas CB, et al. GRNBoost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*2018;**35**(12):2159–61.

23. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9**(1).
24. Russo PST, Ferreira GR, Cardozo LE, et al. CEMiTool: a bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 2018; **19**(1):56.
25. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013; **45**(10):1113–20.
26. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020; **38**:675–8.
27. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; **486**(7403):346–52.
28. Bernett J, Krupke D, Sadegh S, et al. Robust disease module mining via enumeration of diverse prize-collecting steiner trees. *Bioinformatics* 2022; **38**(6):1600–6.
29. Sarkar S, Lucchetta M, Maier A, et al. Online bias-aware disease module mining with ROBUST-web. *Bioinformatics* 2023; **35**(6):btad345.
30. Azim HA, Nguyen B, Brohée S, et al. Genomic aberrations in young and elderly breast cancer patients. *BMC Med* 2015; **13**(1):266.
31. Gómez-Flores-Ramos L, Castro-Sanchez A, Peña-Curiel O, Mohar A. Molecular biology in young women with breast cancer: from tumor gene expression to dna mutations. *Revista de investigacion Clinica* 2017; **69**(4).
32. Xiaofan L, Zhou Y, Meng J, et al. Epigenetic age acceleration of cervical squamous cell carcinoma converged to human papillomavirus 16/18 expression, immunoactivation, and favourable prognosis. *Clin Epigenetics* 2020; **12**(1).
33. Rivard C, Stockwell E, Yuan J, et al. Age as a prognostic factor in cervical cancer: a 10-year review of patients treated at a single institution. *Gynecol Oncol* 2016; **141**:102.
34. Meanwell CA, Kelly KA, Wilson S, et al. Young age as a prognostic factor in cervical cancer: analysis of population based data from 10 022 cases. *BMJ* 1988; **296**(6619):386–91.
35. Peired AJ, Campi R, Angelotti ML, et al. Sex and gender differences in kidney cancer: clinical and experimental evidence. *Cancer* 2021; **13**(18):4588.
36. Yoo S, Huang T, Campbell JD, et al. MODMatcher: multi-omics data matcher for integrative genomic analysis. *PLoS Comput Biol* 2014; **10**(8):e1003790.
37. Vakilian H, Mirzaei M, Tabar MS, et al. DDX3Y, a male-specific region of Y chromosome gene, may modulate neuronal differentiation. *J Proteome Res* 2015; **14**(9):3474–83.
38. Sadegh S, Skelton J, Anastasi E, et al. Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond. *Nat Commun* 2023; **14**(1):1662.
39. Parsana P, Ruberman C, Jaffe AE, et al. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol* 2019; **20**(1):94.
40. Knight HE, Deeny SR, Dreyer K, et al. Challenging racism in the use of health data. *Lancet Digit Health* 2021; **3**(3):e144–6.
41. Bond KM, McCarthy MM, Rubin JB, and Swanson KR. Molecular omics resources should require sex annotation: a call for action. *Nat Methods*, **18**(6):585–588, June 2021.
42. Badia-I-Mompel P, Wessels L, Müller-Dott S, et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023; **24**:739–54.
43. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006; **1**(7 Suppl):S7.
44. HUGO Gene Nomenclature Committee (HGNC). European molecular biology laboratory, European bioinformatics institute (EMBL-EBI), and Wellcome genome campus. *Hgnc database* 2023; Accessed on January 2023.
45. Lambert SA, Jolma A, Campitelli LF, et al. The human transcription factors. *Cell* 2018; **172**(4):650–65.
46. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012; **2**(5):401–4.
47. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013; **6**(269):pl1.
48. Virtanen P, Gommers R, Oliphant TE, et al. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 contributors. *SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods* 2020; **17**:261–72.
49. Hunter JD. Matplotlib: a 2d graphics environment. *Comput Sci Eng* 2007; **9**(3):90–5.
50. Waskom ML. Seaborn: statistical data visualization. *J Open Source Softw* 2021; **6**(60):3021.
51. Matschinske J, Alcaraz N, Benis A, et al. The AIME registry for artificial intelligence in biomedical research. *Nat Methods* 2021; **18**:1128–31.