Gene expression

Identifying cancer pathway dysregulations using differential causal effects

Kim Philipp Jablonski  1,2,†, Martin Pirkl  1,2,†, Domagoj Ćevid 3, Peter Bühlmann 3 and Niko Beerwenkel  1,2,*

1Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland, 2SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland and 3Seminar for Statistics, ETH Zürich, 8092 Zürich, Switzerland

†To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
*Associate Editor: Anthony Mathelier

Received on June 3, 2021; revised on November 5, 2021; editorial decision on November 30, 2021; accepted on December 14, 2021

Abstract

Motivation: Signaling pathways control cellular behavior. Dysregulated pathways, for example, due to mutations that cause genes and proteins to be expressed abnormally, can lead to diseases, such as cancer.

Results: We introduce a novel computational approach, called Differential Causal Effects (dce), which compares normal to cancerous cells using the statistical framework of causality. The method allows to detect individual edges in a signaling pathway that are dysregulated in cancer cells, while accounting for confounding. Hence, technical artifacts have less influence on the results and dce is more likely to detect the true biological signals. We extend the approach to handle unobserved dense confounding, where each latent variable, such as, for example, batch effects or cell cycle states, affects many covariates. We show that dce outperforms competing methods on synthetic datasets and on CRISPR knockout screens. We validate its latent confounding adjustment properties on a GTEx (Genotype–Tissue Expression) dataset. Finally, in an exploratory analysis on breast cancer data from TCGA (The Cancer Genome Atlas), we recover known and discover new genes involved in breast cancer progression.

Availability and implementation: The method dce is freely available as an R package on Bioconductor (https://bioconductor.org/packages/release/bioc/html/dce.html) as well as on https://github.com/cbg-ethz/dce. The GitHub repository also contains the Snakemake workflows needed to reproduce all results presented here.

Contact: niko.beerenwinkel@bsse.ethz.ch

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The complexity of cancer makes finding reliable diagnosis and treatment options a difficult task. Decades of research have improved our understanding of this intractable disease. However, many challenges remain due to its high variability and context specificity, e.g. regarding tissue and cell type (Nature Cancer, 2020). Patients with common cancer types in early stages show promising survival rates, even though rare subtypes still show low survival rates due to different traits like a more aggressive disease progression (Hawkes, 2019; Miller et al., 2019; Troester and Swift-Scanlan, 2009).

It has been hypothesized that cancer diversity can at least in part be explained by heterogeneous mutational patterns. These patterns influence the activity of biological pathways at the cellular level (Khakabimamaghani et al., 2019; Hanahan and Weinberg, 2011). For example, signaling pathways consist of several genes, which regulate certain cell programs, such as growth or apoptosis. The programs are driven by the causal interaction between the genes, e.g. the up-regulation of one causes the up-regulation of another gene. The causal effect (CE) determines the strength of this causal interaction, e.g. by increasing the expression of gene X twofold, the expression of its child Y increases fourfold. Thus, X has a causal effect on Y of 2 (Pearl, 2000). Understanding how these causal networks are perturbed in tumors is necessary for prioritizing drug targets, understanding inter-patient heterogeneity and detecting driver mutations (Vogelstein et al., 2013).

Traditionally, perturbed pathways are detected by assessing whether differentially expressed genes are members of the respective pathway more often than expected by chance. More sophisticated methods measure whether genes belonging to a pathway are localized at certain positions of a rank-ordered set of differentially expressed genes (Subramanian et al., 2005). In such cases, a pathway is interpreted as a simple set of genes and all topological information concerning the functional interconnectivity of genes is ignored. It has been recognized that interactions among genes can have a significant effect on the computation of pathway...
enrichments. Some tools consider, for example, gene expression correlations to account for confounding effects and control the type I error rate while retaining good statistical power (Wu and Smyth, 2012). The underlying structure of gene interactions can thus be either estimated from the data used for the enrichment analysis (Spirtes et al., 2000; Sedgewick et al., 2016), or obtained from existing databases. Canonical pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999) can then be incorporated as prior knowledge to guide the enrichment analysis using topological information of gene connectivity (Liu et al., 2019; Dutta et al., 2012; Tarca et al., 2009; Saez-Rodriguez et al., 2009).

While such enrichment methods go beyond treating pathways as plain gene sets and incorporate topological information of molecular interactions, they often only report a global pathway dysregulation score (Tarca et al., 2009). An exception is PARADIGM, which records an infarct activity for each entity in the pathway under consideration for a given patient sample (Vaske et al., 2010). It does, however, not model causal effects, but only quantifies whether there is some general association among the genes like correlation. Differential causal effects (DCEs) on biological pathways have already been investigated in a formal setting (Wang et al., 2018; He et al., 2019; Tian et al., 2016), where a DCE is modeled as the difference between CEs for the same edge under two conditions. These methods infer the gene network from observational data, which is a difficult task due to the combination of typically low sample size and noise of real data. An incorrect network can result in biased estimation of CEs and DCEs. Additionally, none of these methods make use of the estimated DCEs to compute a pathway enrichment score.

Here, we separate the problem of estimating the causal network and the CEs by replacing the former with the addition of prior knowledge in the form of biological pathways readily available in public databases (Ogata et al., 1999; Nishimura, 2001; Whirl-Carrillo et al., 2012; Mi et al., 2021; Schaefer et al., 2009). We make use of the general concept of causal effects in order to define differential CEs. Specifically, we estimate the CE of gene $X$ on gene $Y$ in normal samples and cancer samples and define the DCE as their difference. In particular, we compare the causal effects between two conditions, such as a malignant tissue from a tumor and a healthy tissue, to detect differences in the gene interactions. We propose Differential Causal Effects (dce), a new method which computes the DCE for every edge (i.e. molecular interaction) of a pathway for two given conditions based on gene expression data (Fig. 1).

This allows us to identify pathway perturbations at the individual edge level while controlling for confounding factors using the framework of causality. By including the additional covariates constructed from the principal components of the design matrix, we also provide a methodological extension of our method to handle potential unobserved confounding that is dense, i.e. where the confounding variable affects many (though not necessarily all) covariates. For example, batch effects from different experimental laboratories or cell cycle stages are not necessarily known, but are accounted for automatically. Our approach allows for computing pathway enrichments in order to rank all networks in large pathway databases to identify cancer specific dysregulated pathways. In this manner, we can detect pathways which play a prominent role in tumorigenesis and pinpoint specific interactions in the pathway that make a large contribution to its dysregulation and the disease phenotype.

We show that dce can recover significant DCEs and outperforms competitors in simulations. In a validation on real data, we apply dce to a public CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) dataset to recover differential effects in the network. We validate the methodological extension for latent confounding adjustment on simulated data and also on real data from the Genotype–Tissue Expression (GTEx) project (Lonsdale et al., 2013). In an exploratory study, we apply dce to breast cancer samples and compare the DCEs among different cancer stages. We identify dysregulated edges common across stages as well as stage-specific edges.

2 Materials and methods

In this section, we describe the Differential Causal Effects (dce) method. We briefly review the causality framework and then introduce the model and computation of DCEs, including under potential latent ‘dense’ confounding. We provide implementation details for obtaining both the estimates and their significance levels. Then, we describe the generating mechanism for synthetic data used throughout the article. We explain the setup of our Perturb-seq validation, as well as the validation of the latent confounding adjustment on the GTEx dataset. Finally, we describe the results of the exploratory The Cancer Genome Atlas (TCGA) analysis.

Causality of biological pathways. First, we give a quick review of causality in the context of biological pathways. A gene pathway can be represented as a structural equation model (SEM) consisting of a directed acyclic graph (DAG) $\mathcal{G}$ with nodes $X = (X_i)_{i=1}^m$ describing the expression of genes, a set of directed edges $E = (E_{ij})_{i,j=1}^m$ representing the causal structure and the structural equations $(f_{ij})_{j=1}^m$ describing how each variable $X_i$ is generated from its parents $X_{pa(i)}$ in $\mathcal{G}$, $X_i = f_i(X_{pa(i)}; \epsilon_i)$, where $(\epsilon_{ij})_{j=1}^m$ are jointly independent noise variables. The causal interpretation of an edge between any two nodes is as follows: changing the expression of a parent $X_i$ affects the expression of the child node $X_j$, which is propagated further to all descendants. The parental sets are given by the edge set $E$. Of particular interest are the interventional distributions for the SEM, in particular their expectations $E[X_i \mid do(X_i = x)]$, which describe how the expected value of the variable $X_i$ changes when we intervene and set the variable $X_i$ to some fixed value $x$. We define the causal effect (CE) of a variable $X_i$ on its descendant $X_j$ as

$$CE[X_i \mid do(X_i = x)] = \frac{d}{dx} E[X_j \mid do(X_i = x)].$$

This derivative equals $\beta_i$, if, by changing the value of $X_i$ from $x$ to $x + \Delta x$, for some small value $\Delta x$, the value of $X_j$ changes on average by $\beta_i \cdot \Delta x$. In the literature, the CE is often also referred to as the total causal effect, because it quantifies the overall effect of an intervention at variable $X_i$ on all of its descendants. We are interested in differential causal effects (DCE) defined as the differences between the causal effects of two conditions of interest, such as, e.g. two different cancer stages or healthy and cancereous samples.

Linearity of the conditional mean. We model the relationship between the mean of any gene expression $X_i$ and its parents $X_{pa(i)}$ by a linear function:

$$X_i = \gamma_i^{(0)} + \sum_{j \in pa(i)} \gamma_{ij} x_j + \epsilon_i(X_{pa(i)}).$$

Conditionally on $X_{pa(i)}$, the error term $\epsilon_i(X_{pa(i)})$ has mean zero and variance depending on $X_{pa(i)}$. A prime example is any generalized linear model (GLM) with identity link function. The coefficients $\gamma_{ij}$ correspond to the direct causal effects, whereas the total causal effects (1) measure the aggregate effect over all directed paths from a certain variable $X_j$ to $X_i$ in $\mathcal{G}$.

Let us consider two arbitrary genes $X_i$ and $X_j$ in the pathway. Under the linearity assumption (2), the causal effect $CE[X_i \mid do(X_j = x)]$ does not depend on $x$. Furthermore, this causal effect corresponds to the coefficient $\beta_i$ in the linear regression of $X_i$ on $X_j$ and an adjustment set $Z = \{Z_k\}_{k=1}^{K}$,

$$X_i = \beta_i + \beta X_j + \sum_{k \in Z} \beta_k Z_k + \eta.$$

Here, $\beta_i$ denotes the intercept and $\eta$ is random noise with mean zero (Goldszmidt and Pearl, 1992; Pearl, 1995). The adjustment set $Z$ is a set of nodes in the pathway $\mathcal{G}$ which fulfills the Back-door criterion (Pearl, 2000). Hence, it holds that no element of $Z$ is a descendant of $X_j$, and $Z$ blocks every path between $X_j$ and $X_i$ that contains an edge with $X_j$ as the child. For example, the parent set $X_{pa(j)}$ always fulfills the Back-door criterion and we always use it as the adjustment set.
If the causal effects of the gene expression $X_i$ on the gene expression $X_k$ are, respectively, denoted as $\beta^A_i$ and $\beta^B_i$ under different conditions $A$ and $B$, then the differential causal effect (DCE) $\delta$ is obtained as the difference

$$\delta = \beta^B_i - \beta^A_i.$$  \hspace{1cm} (4)

Given a graph $G$ describing a biological pathway and observations, we can compute all differential causal effects and identify interactions between any such two variables $X_A$ and $X_B$ that are different between the two conditions (Fig. 1).

**Testing for significance.** We can compute the DCE $\delta$ for the edge $X_A \rightarrow X_B$ by fitting a joint model for both conditions, which also allows us to easily compute the significance of the estimates. Let $I$ be an indicator random variable, which is equal to 1, if the observation comes from condition $A$, and 0, if it comes from condition $B$. The DCE $\delta$ can be computed from all samples jointly by fitting the following linear model

$$X_i = (\beta^A_i + (\beta^B_i - \beta^A_i)I) + (\beta^B_i + (\beta^B_i - \beta^A_i)I)X_j + \sum_{k=1}^{m} (\beta^B_k + (\beta^B_k - \beta^A_k)I)Z_k + \eta$$ \hspace{1cm} (5)

with interaction terms $I \cdot X_j$ and $I \cdot Z_k$. The differential causal effect $\delta = \beta^B_i - \beta^A_i$ can be estimated by using the coefficient estimate corresponding to the interaction term $IX_i$ in (5).

Testing the significance of the estimated DCEs now corresponds to the well-known task of testing the significance of coefficient estimates in a linear model. However, some care is needed if the variances of the error terms $\epsilon_i(X_{pa}(i))$ in our structural Equations (2) indeed depend on the values of the predictors $X_{pa}(i)$, i.e. if there is a certain mean-variance relationship for the gene expression levels, as has been described for RNA-seq data (Robinson and Smyth, 2007). In this case, the linear model (5) is heteroscedastic and the usual formulae for standard errors of the coefficient estimates, that result in $t$-tests for the significance, do not apply. We, therefore, use heteroscedasticity-consistent standard errors that yield asymptotically valid confidence intervals and $P$-values regardless of the dependence of the noise level on predictor values (Eicker, 1967; Huber et al., 1967; White, 1980).

Besides assessing significance of DCEs for single edges, we can also calculate a global $P$-value measuring the overall dysregulation of a given pathway $G$: we combine the $P$-values corresponding to different differential causal effects $\delta = (\delta_i)_{i=1}^n$ by taking their harmonic mean (Good, 1958).

**Adjusting for latent confounding.** A fundamental assumption for most of causal inference methods is that there is no unobserved confounding, i.e. that there are no unmeasured factors affecting both the cause and the effect (Leek et al., 2012; Gagnon-Bartsch et al., 2013). Such unobserved confounders could be, for example, batch effects, cell cycle stages, varying laboratory conditions, different patient demographics, etc. Although some methods exist for accounting for measured confounding (Zhang et al., 2020), unobserved confounding is much more challenging. Presence of latent confounding can result in spurious correlations and false causal conclusions. Therefore, adjusting for potential latent confounding is crucial for making the method robust in applications to biological data (Cevird et al., 2020).

Some information about latent factors can often be obtained from the principal components of the data (Novembre and Stephens, 2008). This can be made rigorous under the linearity assumption (2) for our structural equation model $G$, as follows. We assume that there are $q$ latent variables $H_1, \ldots, H_q$ affecting our data. We extend the model (2) to include the latent confounding as follows:

$$X_i = \gamma^{(0)}_i \sum_{j \in \text{pa}(i)} \gamma^{(0)}_{ij} X_j + \sum_{j=1}^{q} \delta_j H_j + \epsilon_i(X_{pa}(i), H),$$ \hspace{1cm} (6)

that is, the latent confounders $H_1, \ldots, H_q$ are additional source nodes in the DAG $G$ and affect genes in the pathway linearly, analogously to (2). Not every gene needs to be affected ($\delta_j$ could be zero), but the methodology works better when many genes are affected, see discussion below. By writing the structural Equations (6) in matrix form, where we define the matrices $\Gamma^{(0)}_i = \gamma^{(0)}_{ij}, \Gamma_{ij} = \gamma^{(0)}_{ij}, \Delta_0 = \delta_j$, and $E(X,H) = \epsilon_i(X_{pa}(i), H),$ we obtain

$$X_{\text{lp}} = \Gamma^{(0)}_i X_{\text{lp}} + X_{\text{lp}} \Gamma_{x, \text{lp}} + H_{\text{lp}} \Delta_{x, \text{lp}} + E(X,H)_{\text{lp}},$$ \hspace{1cm} (7)

which gives

$$X = \frac{\Gamma^{(0)}_i}{\text{intercepts}} + H \Delta (I - \Gamma)^{-1} + E(X,H)(I - \Gamma)^{-1},$$ \hspace{1cm} (8)

which is the standard linear factor model with heteroscedastic
errors. From this representation, one can see that $H$ can be determined from the principal components of $X$ (Fig. 2). The scree plot for a toy example visualizes the effect of latent variables having a global effect on the data. The first principal components are clearly separated from the rest, if latent factors are present (Fig. 2, left). Therefore, we obtain the confounding proxies $H$ as the scores of the first $q$ principal components of the design matrix combining the data from both conditions.

The confounding methodology relies on the assumption that every confounding variable affects many variables in the dataset, i.e. the confounding is dense (Guo et al., 2020). This condition is to some extent necessary, because in the case when the latent confounders affect only a few covariates, it is not identifiable whether the resulting association between them could be causal or is due to confounding. We emphasize that not every covariate needs to be affected by each confounder. However, the more confounding factors each latent factor $H_i$ affects, the more information we have about it in the data and thus the confounding proxies $H$ capture the effect of the confounders $H$ better. Furthermore, the dense confounding assumption ensures that the scree plot, showing the singular values of the design matrix, has a spiked structure, as several latent factors can explain a relatively large proportion of the variance (Fig. 2). This helps estimating the number $q$ of the confounding proxies used. As a default choice, we use a permutation method that can be shown to work well under certain assumptions (Dobriban, 2017) and which compares the observed value of the variance explained by the principal components with its expected value over many random permutations of the values in each column of gene expression matrix $X$.

Algorithm and implementation in R. The presented methods are implemented in the R package dce which is freely available on Bioconductor. The function dce::dce takes as input the structure of a biological pathway, i.e. the adjacency matrix of a DAG, and two $n \times p$ matrices, with $n$ samples and $p$ genes, storing gene expression data for each of the two conditions, respectively. As output, the function returns the estimated DCEs, as well as standard errors and two-sided $P$-values for the DCE at each edge in the pathway together with the $P$-value measuring the overall pathway enrichment. The results can be easily transformed into a dataframe and plotted for further downstream analyses.

Generating synthetic data and benchmarking methods. We assess the behavior of dce and its competitors in a controlled setting by generating synthetic data with known DCEs (ground truth). We start by generating a random DAG $G$. Without loss of generality, we assume the nodes of the DAG to be topologically ordered, i.e. node $X_i$ can only be parent of node $X_j$, if $i < j$. This ensures that the network $G$ is a DAG. In practice, we sample edges from a binomial distribution with probability $p$ for the upper triangle of $G$. We further sample the coefficients $c_{ij}$ for every edge as in (2) from a uniform distribution $U(−\gamma_{\text{max}}, \gamma_{\text{max}})$. We generate the data for network $G$ in the following way. For a node $X_i$, we set the mean expression count

$$
\mu_i = v - \bar{x} \cdot (\min_j x_j - i),
$$

and then generate $X_i \sim \text{Pois}(\mu_i)$ as a vector of counts, corresponding to gene expression values from experiments like RNA-seq. The mean depends on its parents in a linear fashion,

$$
\nu = \sum_{j \in \text{pa}(i)} c_{ij} X_j
$$

where $c_{ij}$ represents the direct effect of $X_j$ on $X_i$, $i > 0$ is a small shift, and $\bar{x}$ is a vector of ones. Subtracting the minimum ensures positive values of the mean for each data point. Then, a realization of $X$ is drawn from the Poisson distribution $\text{Pois}(\mu)$. We introduce negative binomial noise by drawing a realization of each source node in $G$ from the negative binomial distribution $\text{NB}(\mu, \theta)$ with a general mean $\mu$ and dispersion $\theta$. We use this setup to control the variance across all nodes, which can blow up for descendants with larger means.

After sampling the data $D_A$ for the nodes of network $G$ under condition $A$, we resample a certain fraction of edge weights in order to generate new data $D_B$ under condition $B$. For a fixed edge weight $\beta$ we sample the new edge weight uniformly such that

$$
\beta^B = \beta^A \sim U([-\delta_{\text{max}}, \delta_{\text{min}}] \cup [\delta_{\text{min}}, \delta_{\text{max}}]).
$$

This ensures that the absolute difference between the two edge weights lies in $[\delta_{\text{min}}, \delta_{\text{max}}]$. We also simulate latent variables. They are neither included in the data nor the network $G$, but have (unknown) outgoing edges to all genes in the dataset with non-zero effects. Hence, these latent variables have global effects on the data, e.g. emulating batch effects.

We compare dce to correlation (cor), partial correlation (pcor), the method Fast Gaussian Graphical Models (fggm) tailored to DCEs (Wang et al., 2016; He et al., 2019), a differential gene expression approach (dge) and random guessing. cor is provided by the R package stats (R Core Team, 2020). For pcor, we use the general matrix inversion from the R package MASS (Venables and Ripley, 2002) to compute the precision matrix. fggm is based on partial correlation, but additionally tries to learn the network structure to adjust for confounding effects. We use the R code provided by the authors (He et al., 2019) to run fggm. For fggm, we transform each gene expression count $g$ to $\log(g + 1)$. We use the differential expression result from edgeR (Robinson and Smyth, 2007) as input for dge. We compute the DCE for the edge between two genes $x$ and $y$ as the difference of the log foldchanges of both genes. We compute the corresponding $P$-value for the same edge as the minimum of the $P$-values for both genes $x$ and $y$. We provide pcor with the same adjustment set of confounding variables as dce.

Fig. 2. The scree plot (of synthetic data generated as described in the Materials and Methods section) shows that in presence of latent confounding as in (6), the first $q$ principal components explain much more variability of the data, which we exploit for confounding adjustment.
We run all methods on simulated data for various modeling parameters. The default parameters are a network \( G \) of 100 genes, 200 samples for both sample conditions, an absolute magnitude in effect differences between the two conditions of 1, mean of 100 negative binomial distributed counts with a dispersion of 1 for the source genes in the network \( G \) (no parents), a true positive rate of 50% (edges which have different effects between the two conditions), and library size factors for each sample in the interval \([1,10]\). The library size factor accounts for different sequencing depth among the samples, i.e. for one sample including more reads because more RNA was available even though the gene expression was the same in samples with less RNA. We account for different library sizes over all samples by computing Transcripts Per Kilobase Million (TPM).

Overall, we simulate a full dataset of 10, 000 genes including the genes in the network \( G \) to allow for the realistic estimation of the library size. As a performance measure, we use the area under the receiver operating characteristic (ROC-AUC). We count the number of true/false positive and false negative DCEs based on the edges in the ground truth network and the significant \( P \)-values for different significance levels. Based on these true/false positives, we compute the ROC curve and its AUC. For both correlation methods, we use a permutation test to compute empirical \( P \)-values.

Validation using Perturb-seq. Perturb-seq, a CRISPR-Cas9-based gene knockout method, can be used to inhibit the expression of multiple target genes on a single-cell level (Qi et al., 2013; Adamson et al., 2016). The dataset we analyze is a CRISPR-Cas9-based gene knockout study, which was conducted to inhibit the expression of multiple target genes in the context of single-cell RNA sequencing. We can compute the ROC curve and its AUC for both correlation methods, we use a permutation test to compute empirical \( P \)-values.

In this section, we first show the performance of \( dce \) and its competitors on simulated data and a CRISPR dataset. Next, we evaluate the deconfounding performance using the GTEx dataset. Finally, we use \( dce \) for an exploratory analysis of breast cancer data from TCGA and show the progression of pathway dysregulation over different cancer stages.

3 Results

In this section, we first show the performance of \( dce \) and its competitors on simulated data and a CRISPR dataset. Next, we evaluate the deconfounding performance using the GTEx dataset. Finally, we use \( dce \) for an exploratory analysis of breast cancer data from TCGA and show the progression of pathway dysregulation over different cancer stages.

3.1 Simulation study

Pathway databases contain networks of different sizes. We first investigate the influence of network size on the ability of each method to recover ground truth differential causal effects. \( dce \) achieves the highest ROC-AUC for all four network sizes considered (10, 50, 100 and 150 genes). Methods which do not account for known confounding variables perform similar to random guessing for large networks (Fig. 3a). However, \( dce \) also outperforms \( pcor \) with an AUC of 0.61 versus 0.55. Variability is very high for competitors and size ten. The methods either successfully recover all of the very few effects or none at all. As an alternative performance assessment, we also compute \( dce \) for a P-value threshold of 0.05 (Supplementary Figs S7 and S8). While the true positive rate decreases for large networks, precision is relatively robust and \( dce \) avoids a high rate of false positives.
Second, we assess how the magnitude of differential causal effects affects the identification of significant differences. We sample the magnitudes from the set \{0.1, 1, 2\}. For example, for a magnitude of 1 the edge weights between the network of the wild-type samples and the disease samples differ by at most 1. dce has difficulty estimating large differences as well as very small differences. However, it still significantly outperforms all other methods, which again show similar performance to random guessing for large effects (Fig. 3b).

In additional simulations, dce shows increasing ROC-AUC for decreasing dispersion and increasing number of samples (Supplementary Figs S1 and S2) as is expected due to decreasing noise. We found constant ROC-AUC of dce over varying ranges of library size (Supplementary Fig. S3). Different prevalence of positive edges has little effect on the ROC-AUC of dce (Supplementary Fig. S4). dce with latent variable adjustment performs similarly to dce without latent variable integration if we do not simulate any latent variables. But dce significantly outperforms dce without latent variable integration for five and ten latent variables influencing the dataset (Supplementary Fig. S5). This is because without latent confounding adjustment one has a large number of false positives due to the confounding bias (Supplementary Fig. S6). Sampling the effects of latent variables from an exponential distribution with default rate 1 instead of a uniform distribution does not result in much difference in ROC-AUC of dce (Supplementary Fig. S9). This shows that even if only some and not all genes in the graph are strongly affected by the latent confounders, we can still successfully account for it.

dce relies heavily on the given network \( G \). Hence, we investigate how well dce performs if \( G \) contains false edges or is missing true edges. We find that dce is robust to additional false edges in the network, but starts breaking down if true edges are missing in larger fractions (Supplementary Fig. S10).

### 3.2 Validation experiments using CRISPR knockout data

To benchmark our method using real-life data generated by Perturbseq (Adamson et al., 2016), we ask whether we can recover the CRISPR knockout from single-cell RNA-seq data using pathways from KEGG which contain the knocked-out genes. Hence, we assume that these pathways capture the causal gene interactions governing the response of the cell to the experimental intervention. As seen in the synthetic benchmark, slight deviations of the observed network from the true underlying network have no major impact on the performance of our method (Supplementary Fig. S10). By interpreting a CRISPR knockout as an intervention of the causal pathway, we define the positive class to consist of all edges adjacent to a knocked-out gene, and the negative class as all other genes. Consequently, a true positive occurs when an edge adjacent to a CRISPR knocked-out gene is (significantly) associated to a non-zero DCE.

### 3.3 Deconfounding validation on GTEx data

To validate the extension of our methodology for latent confounding adjustment, we investigate the robustness of our estimates when the confounding variables are latent, compared to when they are added to the pathway as the source nodes. When the confounding adjustment, as described in the Materials and Methods section, is used, we observe that the estimated DCEs between two different tissue types differ much less between the original and extended pathways (Supplementary Fig. S11).

Similarly, the resulting \( P \)-values are also much more stable, as measured by the Pearson correlation between the negative logarithmic \( P \)-values computed for the original and extended pathway (Fig. 4d). The correlation is consistently larger when using the confounding adjustment, which is important since the latent confounding in general causes many false positives in the analysis.
3.4 Exploratory analysis of TCGA data

To demonstrate the ability of our method to recover known cancer-related pathway dysregulations as well as to discover new genes of potential biological and clinical relevance, we compute DCEs using breast cancer gene expression data from TCGA on the breast cancer pathway obtained from KEGG. The results for each stage are then visualized on the pathway structure (Fig. 5a–c). The raw DCE values were transformed to a symmetric logarithm for greater visibility with the following formula:

\[
\text{symlog}(x) = \begin{cases} 
\log_{10}(x) + 1 & \text{if } x > 1 \\
-\log_{10}(-x) - 1 & \text{if } x < 1 \\
x & \text{otherwise}
\end{cases}
\]

(12)

Roughly 40% of all investigated interactions (614 out of 1527) show no difference in causal effects (|DCE| < 1 and P-value > 0.05) between normal and stage condition for all stages. In the following, we will discuss cases with large effect sizes and significant P-values (Fig. 5d).

Throughout all stages, interactions between the WNT (Wingless/Int1) and FZD (Frizzled) protein complexes exhibit significant, non-zero DCEs indicating a strong dysregulation of the breast cancer pathway. Most notably, we observe a highly significant dysregulation of WNT11 → FZD1, WNT11 → FZD3 and WNT11 → FZD7 in stage II (P-value < 10^{-14}) and in stages I and II. Additionally, the interaction between WNT8A and FZD4 features a strongly positive DCE of ~2000 in all three stages. These observations are expected, because the interactions between the WNT and FZD protein complexes have been implicated in disease formation in general (Dijksterhuis et al., 2015; Chien et al., 2009; Schulte, 2010) and in breast cancer in particular (Yin et al., 2020; Koval and Katanaev, 2018).

The interaction between DLL3 (Delta Like Canonical Notch Ligand 3) and NOTCH4 (Notch Receptor 4) features a significant DCE of ~140 with P-values < 10^{-6} in all three stages. The Notch signaling pathway has been shown to play an important role in Pancreatic ductal adenocarcinoma tumor cells, but has not been
implicated in breast cancer (Song and Zhang, 2018). Our finding suggests that stromal cells located in the breast may play an important role for disease progression throughout all stages.

For the interaction between TCF7L2 (Transcription Factor 7 Like 2) and CCND1 (Cyclin D1), we observe a significant negative DCE of −11.9 with a $P$-value of $< 10^{-6}$ in stage III. The role of TCF7L2, which participates in the Wnt/$\beta$-catenin signaling pathway and is important for cell development and growth regulation, has already been discussed in the context of breast cancer (Connor et al., 2012). However, its interaction with CCND1 has, to the best of our knowledge, not been investigated in the literature. Due to the down-regulation in the diseased condition for stage III, we suggest that an improved understanding of the underlying biological reasons might provide insights into the late-stage behavior of breast cancer.

Overall, we are able to recover both interactions which are known to be dysregulated in breast cancer as well as novel ones. The former indicates that the prioritization of interactions given by dce is in accordance with current literature. The latter suggests that dce is also able to find dysregulated interactions which up to now have only been recognized for other diseases but may play an important role for breast cancer.

4 Discussion

We have presented a new method, dce, to compute differential causal effects between two conditions using a regression approach. dce enables the edge-specific identification of signaling pathway dysregulations. This piece of information can help to further our understanding of subtle differences on the molecular level in seemingly similar cancer types.

dce assumes a linear relationship among pathway genes. The linear model is solved using network information to account for additional genes confounding the linear relationship between gene pairs. The network information is included via prior knowledge from literature. dce also accounts for latent confounders in the model, which are unknown and not included in the gene network. They are assumed to linearly affect a large number of measured covariates. We have successfully applied dce to normalized gene expression counts (TPM) in all analyses. However, dce is a general framework, which makes no strong assumption on the data and can be applied to other data types.

We have shown in our simulations that dce is able to detect changes in causal effects even in the presence of noise and for certain ranges of effect sizes. For a wide array of parameter choices, dce outperforms methods using (partial) correlation, fggm and an approach based on differential expression. Especially in the case of latent confounders, we showed that dce with the integration of latent variables outperforms dce without, except if no latent confounders were used to simulate the data. In this case, both methods are equally accurate. Hence, we recommend the integration of latent variables in the model as the default configuration.

In addition to the synthetic benchmark, we have also validated our method on real data derived from Perturb-seq experiments. We have shown that dce is able to recover the experimental knockouts with better performance than correlations and partial correlations.

For breast cancer, we have shown that not all parts of the signaling pathway are perturbed and characteristic hotspots exist. Some causal effects between two genes are invariant to stage information, while other causal effects can vary in either magnitude or even sign of their effect size. This indicates that certain areas of such pathways are more relevant than others. This phenomenon has also been observed in other studies (Song et al., 2014; Feng et al., 2018). Some parts of a pathway seem to be either more conserved or just not relevant to tumorigenesis. This provides interesting opportunities to identify drugs which target certain parts of a pathway and might...
explain their efficacy. However, we want to stress that not all dysregulated edges will be relevant for causing cancer, just like not all mutations are cancer-causing mutations. Additionally, the robustness of our method depends on the availability of enough samples. In many cases, few are available and make our approach infeasible. While dce performs still better than random for even 10 samples, it is significantly worse than for higher sample sizes.

In summary, we have proposed a novel application of the concept of differential causal effects which describe the differences in causal effects between two conditions and developed a regression approach to compute those differences. We demonstrate their robustness in a simulation study, and point out interesting results in application to real data, e.g. we show that some dysregulated edges are consistent among breast cancer tumor stages I-III, but that other dysregulations are unique to each stage.

Our simulations show the need for sufficiently large datasets when dealing with large pathways. Additionally, dce relies on correct network information. While very robust to incorrect edges in the network, dce’s performance breaks down significantly when edges are missing from the network. We have also simulated data from DAGs only and this assumption is made throughout all analyses. In reality, biological pathways include cycles, which could affect dce’s performance breaks down significantly when edges are missing from the network. We have also simulated data from DAGs only and this assumption is made throughout all analyses. In reality, biological pathways include cycles, which could affect the result of dce. Similarly, we rely on the assumption that all causal effects are propagated linearly. Other types of causal effects could affect dce as well. That is, the expression of a gene could depend on the expression of its parents in a non-linear fashion. The linearity of our model might also hinder dce from reaching better performance in case of very large or very small effect sizes.

Future research should focus on modifying the regression to adapt it to small datasets and make it more robust, for example, by enforcing sparsity through the introduction of $L_1$ or $L_2$ norms on the coefficients to avoid outliers produced by artifacts in the data.

References

Adamson, B. et al. (2016) A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. Cell, 167, 1867–1882.

Cevred,D. et al. (2020) Spectral deconfounding via perturbed sparse linear models. J. Mach. Learn. Res., 21, 232.

Chen, W.J. et al. (2009) A Wnt survival guide: from flies to human disease. J. Investig. Dermatol., 129, 1614–1627.

Connor, A.E. et al. (2012) Associations between tcf7l2 polymorphisms and risk of breast cancer among Hispanic and non-Hispanic white women: the breast cancer health disparities study. Breast Cancer Res. Treat., 136, 593–602.

Dijkstraen,J.P. et al. (2015) Systematic mapping of wnt-frz protein interactions reveals functional selectivity by distinct wnt-frz pairs. J. Biol. Chem., 290, 6789–6798.

Dobriban, E. (2020). Permutation methods for factor analysis and PCA. The Annals of Statistics, 48, 2824–2847.

Dutta, B. et al. (2012) Pathnet: a tool for pathway analysis using topological information. Source Code Biol. Med., 7, 10.

Eicker,F. (1967) Limit theorems for regressions with unequal and dependent errors. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, Berkeley, CA, pp. 59–82.

Feng, Y. et al. (2018) Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes Dis., 5, 77–106.

Gagnon-Bartsch,J.A. et al. (2013) Removing Unwanted Variation from High Dimensional Data with Negative Controls. Tech Reports from Dep Stat Unv California, Berkeley, pp. 1–112.

Goldezi,M., and Pearl,J. (1992) Rank-based systems: a simple approach to belief revision, belief update, and reasoning about evidence and actions. In: Proceeding of the 3rd Conference on Knowledge Representation, pp. 661–672.

Good,I.J. (1958) Significance tests in parallel and in series. J. Am. Stat. Assoc., 53, 799–813.

Guo,Z. et al. (2020) Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. arXiv preprint arXiv: 2004.03758.

Hanahan,D., and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. Cell, 144, 646–674.

Hawkes,N. (2019) Cancer survival data emphasize importance of early diagnosis. BMJ, 364, 1408.

Ht,H. et al. (2019) A statistical test for differential network analysis based on the inferences of Gaussian graphical model. Sci. Rep., 9, 10863.

Huber,P.J. et al. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. University of California Press, pp. 221–233.

Jacob,W. et al. (2015) High EGFR protein expression and exon 9 pki,kca mutations are independent prognostic factors in triple negative breast cancers. BMC Cancer, 15, 1–10.

Khakabimamaghani,S. et al. (2019) Uncovering the subtype-specific temporal order of cancer pathway dysregulation. PLoS Comput. Biol., 15, e1007451.

Koval,A., and Katanaev,V.L. (2018) Dramatic dysbalancing of the wnt pathway in breast cancers. Sci. Rep., 8, 1–10.

Leek,J.T. et al. (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28, 882–883.

Liu,A. et al. (2019) From expression footprints to causal pathways: contextualizing large signaling networks with carnivial. JIP Syst. Biol. Appl., 5, 40.

Lonsdale,J. et al. (2013) The genotype-tissue expression (GTEx) project. Nature Genet., 45, 580–585.

Mi,H. et al. (2021) Panther version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive api. Nucleic Acids Res., 49, D394–D403.

Miller,K.D. et al. (2019) Cancer treatment and survivorship statistics, 2019. CA Cancer J. Clin., 69, 363–385.

Author contributions

K.P.J. and M.P. conceived the project. K.P.J. and M.P. developed the statistical model of dce and implemented the software package. D.C. contributed to the statistical methodology as well as software implementation. N.B. and P.B. supervised the study. K.P.J. and M.P. wrote the initial manuscript draft. All authors edited the manuscript.

Conflict of Interest: none declared.

Data availability

The code used to construct the synthetic datasets is available as part of the R software package dce. The experimental data used in the Perturb-seq validation are available under the accession GSE90546 from NCBI GEO. GTEx data are publicly available through the GTEx portal. The experimental data used in the exploratory breast cancer analysis are available under the accession TCGA-BRCA from The Cancer Genome Atlas. The pathway structures have been obtained from the Kyoto Encyclopedia of Genes and Genome.
