BEYOND REPRODUCIBILITY: KNOCKING ON SUSTAINABILITY'S DOOR

KIM PHILIPP JABLONSKI

DISS. ETH NO. 28629

BEYOND REPRODUCIBILITY: KNOCKING ON SUSTAINABILITY'S DOOR

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

KIM PHILIPP JABLONSKI
M. Sc. Computational Life Science
Jacobs University
Bremen, Germany

born on 12 March 1995
citizen of Germany

accepted on the recommendation of

Prof. Dr. Niko Beerenwinkel
Prof. Dr. Peter Bühlmann
Prof. Dr. Caroline Uhler

2022

ABSTRACT

The following thesis presents three independent studies which were carried out as part of the author's doctoral studies in the Computational Biology Group at the Department of Biosystems Science and Engineering at ETH Zurich in Basel. These projects deal with the development of statistical methods for the detection of pathway dysregulations, and the processing and analysis of next-generation sequencing data with a particular focus on the importance of benchmarking the methods' performances in a sustainable way.

The first two studies are based on the fact that cancer is a heterogeneous disease where the same phenotype can arise from different mutational patterns and propose novel methods for the computation of pathway enrichments. The first study takes a causal approach and computes edge-specific pathway dysregulations while the second study computes global pathway dysregulation scores while accounting for term-term relations. Both studies include an extensive benchmark workflow which tests both the performance on synthetic and real data sets as well as runs exploratory analyses. The third study describes the development of a pipeline for the analysis of viral high-throughput sequencing data and an extensive benchmark of global haplotype reconstruction methods.

The dissertation is organized in the following way. The first chapter provides an overview of different workflow management systems which can be used to create reproducible benchmarking workflows, a comment on the distinction between reproducible and sustainable data science, and their relevance in the fields of cancer genomics as well as virology.

The second chapter presents *dce*, a computational method for the edge-specific detection of pathway dysregulations using a causal framework.

The third chapter presents *pareg*, a regression-based method which addresses the issue of large and redundant pathway databases by incorporating term-term relations into the enrichment computation. It accomplishes this goal by adding regularization terms to the loss function of a generalized linear model.

The fourth chapter presents a scalable, reproducible and transparent pipeline for the analysis of viral sequencing data as well as a benchmark of global haplotype reconstruction methods.

The fifth chapter concludes the thesis by summarizing its findings as well as suggesting potential future directions.

KURZFASSUNG

Die folgende Doktorarbeit präsentiert drei unabhängige Studien, die im Zuge des Doktorstudiums des Autors in der Computational Biology Gruppe im Department of Biosystems Science and Engineering der ETH Zürich in Basel verfasst wurden. Die Projekte handeln von der Entwicklung statistischer Methoden für das Erkennen von Dysregulationen biologischer Prozesse, und dem Prozessieren und der Analyse von Sequenzierdaten mit einem besonderen Fokus auf der Wichtigkeit der Bewertung der Performanz der Methoden in einer nachhaltigen Art und Weise.

Die ersten beiden Studien basieren auf der Grundlage das Krebs eine heterogene Krankheit ist, bei der der gleiche Phänotyp von unterschiedlichen Mutationsmustern abstammen kann, und schlagen neuartige Methoden für die Berechnung der Dysregulation biologischer Prozesse vor. Die erste Studie verfolgt einen kausalen Ansatz und berechnet verbindungsspezifische Prozessdysregulationen, während die zweite Studie globale Prozessdysregulationsstärken berechnet und dabei Zusammenhänge zwischen biologischen Prozessen miteinbezieht. Beide Studien beinhalten ausführliche Arbeitsabläufe für die Bewertung der Performanz mit sowohl synthetischen als auch echten Datensätzen und explorative Analysen. Die dritte Studie beschreibt die Entwicklung eines Arbeitsablaufes für die Analyse viraler Sequenzierdaten und eine ausführliche Performanzbewertung globaler Haplotyprekonstruktionsmethoden.

Diese Dissertation ist wie folgt aufgebaut. Das erste Kapitel gibt einen Überblick über verschiedene Arbeitsablaufmanagementsysteme die genutzt werden können um reproduzierbare Arbeitsabläufe zu erstellen, einen Kommentar zu dem Unterschied zwischen reproduzierbarer und nachhaltiger Datenwissenschaft und dessen Relevanz in den Gebieten der genomischen Krebsforschung und Virologie.

Das zweite Kapitel präsentiert *dce*, eine rechnerische Methode für die Bestimmung verbindungsspezifischer Dysregulationen biologischer Prozesse mit einem kausalen Rahmenkonzept.

Das dritte Kapitel präsentiert *pareg*, eine regressionsbasierte Methode, die das Problem großer und redundanter Datenbanken biologischer Prozesse adressiert, indem es Abhängigkeiten zwischen Prozessen in die Berechnung miteinbezieht. Es erreicht dieses Ziel durch das Hinzufügen von Regula-

risierungstermen zu der Verlustfunktion eines Verallgemeinerten linearen Models.

Das vierte Kapitel präsentiert einen skalierbaren, reproduzierbaren und transparenten Arbeitsablauf für die Analyse viraler Sequenzierdaten und das Testen globaler Haplotyprekonstruktionsmethoden.

Das fünfte Kapitel schließt diese Arbeit ab, indem es die Ergebnisse zusammenfasst und zudem mögliche, zukünftige Forschungsrichtungen aufzeigt.

CONTENTS

1

INTRODUCTION

> *We say that a theory is falsified only if we have accepted basic statements which contradict it [..]. This condition is necessary, but not sufficient; for we have seen that non-reproducible single occurrences are of no significance to science. [..] We shall take it as falsified only if we discover a reproducible effect which refutes the theory.*
>
> — Karl Popper [1]

The expansion of scientific knowledge, which in turn allows making insightful predictions and controlling complex systems, has been in large parts thanks to the application of the scientific method [2]. While the validity and scope of its application have been debated, it is generally seen as highly useful and influential [3]. One crucial component of the scientific process is the ability to reproduce previously established results. The goal of reproducibility thus encompasses both the corroboration of previous findings and the extension thereof. While it used to include, for example, checking the correctness of manual computations, the advent of computers has shifted the field towards the rigorous checking of experimental and computational workflows [4–6]. The importance of this has been widely demonstrated by the concerningly low reproducibility rates in a multitude of research areas [7–11], notably including cancer research [12] or bioinformatics [13] and other diverse fields such as psychology [14], climate science [15] and quantum computing [16].

To overcome these issues, novel findings need to be reproduced and new computational methods need to be benchmarked reliably. Most importantly, this requires the usage of representative synthetic and real data sets as well as the implementation of workflows which are flexible enough to accommodate new tools [17–19]. Both of these two factors lead to a standardized way of presenting research findings and their eventual acceptance by the scientific community. While the former factor is highly context-dependent and varies between disciplines, the latter can in general be accomplished by using an appropriate workflow management system (WMS).

The goals of this thesis are to (a) show the need to properly handle confounding factors when computing edge-specific pathway dysregulations,

(b) to demonstrate how modeling functional pathway overlaps improves the performance of pathway enrichment methods, (c) to underline the requirement of sustainable workflow development when conducting large-scale analyses and (d) to highlight the importance of purposefully designing benchmarking workflows when developing new bioinformatics methods. Chapter 2 and chapter 3 deal with the development of novel statistical methods to compute pathway dysregulations typically caused by diseases. In chapter 2 we show how a causal perspective can improve the intra-pathway detection of dysregulations by accounting for confounding factors, and in chapter 3 we motivate the usefulness of including term-term relations in the statistical enrichment computation. Chapter 4 summarizes how improvements in V-pipe, a workflow designed for the analysis of next generation sequencing (NGS) data from viral pathogens, enable the analysis of hundreds of thousands of SARS-CoV-2 samples and shows how it can be used to conduct a global haplotype reconstruction benchmark. Finally, chapter 5 finishes this thesis with concluding remarks and provides potential future directions.

The following sections give a brief introduction to the WMS used to conduct these studies, a distinction between reproducible and sustainable data science and their biological contexts in the fields of cancer genomics and virology.

1.1 CHOOSING THE RIGHT WORKFLOW MANAGEMENT SYSTEM

The choice of the most appropriate WMS is not a trivial one, as plenty of options exist [20]. One needs to balance the complexity which a flexible WMS induces with the rigidity a simple WMS provides. Figure 1.1 shows four workflow implementations using different WMSs.

At one end of the spectrum, WMSs such as Galaxy [21] provide a straightforward graphical interface in a web browser (fig. 1.1a). Due to their online presence, they make sharing workflows an integral part of their platform. While this approach allows end users to quickly get started, it limits the ability of power users to customize the workflows in detail and to run them on any architecture they desire.

WMSs such as Popper [22] remove the dedicated web interface for designing workflows and replace it with a declarative format where the workflow logic is implemented using a YAML [23] configuration file (fig. 1.1b). This approach increases the flexibility of developing the workflow which does not rely on visiting a website anymore, but also increases complexity by

requiring workflow developers to learn a new configuration language and use the command line for execution.

At the other end of the spectrum, WMSs such as SciPipe [24] are programming language-specific libraries which provide classes and functions for creating a workflow (fig. 1.1d). They require a full understanding of the respective programming language and are thus difficult to use. At the same time, they provide the highest degree of flexibility and virtually any logical requirement can be implemented.

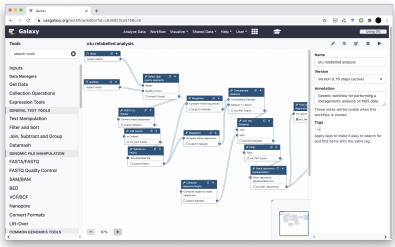
Somewhere in the middle of this spectrum, sharing both the flexibility of programming language-based WMSs and relative ease-of-use of configuration file-based ones, exist WMSs such as Nextflow [25] and Snakemake [26]. They make use of a domain specific language (DSL), which is based on a programming language but employs additional syntax features which make developing and understanding workflows easier (fig. 1.1c). In the case of Snakemake, this DSL is based on Python [27] which is a popular language in general scripting tasks as well as bioinformatics and deep learning specifically. Snakemake adds syntactical sugar which allows to clearly define individual components of a workflow and dependencies between them. In addition, it takes care of various tasks commonly related to workflow management, such as cluster execution, workflow modularization and parameter exploration.

Consequently, using a WMS of this kind is usually the best idea, unless special workflow requirements, such as, for example, the use of a specific implementation language, need to be fulfilled.

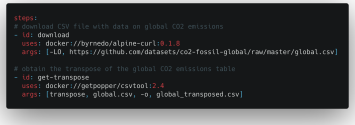
1.2 MOVING BEYOND REPRODUCIBILITY

Reproducibility is a necessary but not sufficient condition for the development of workflows which benefit not only the original creator but also the scientific community as a whole [28]. In fact, reproducibility is only one of three pillars needed to achieve truly sustainable data science (fig. 1.2).

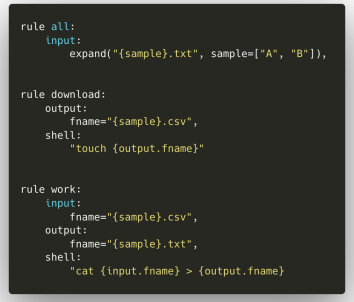
Reproducibility describes the execution of a workflow and encompasses automation, scalability and portability. Automation requires the workflow to be runnable without involved manual interventions by the researcher. In the best case, this should make it possible to run the workflow with the execution of a single command which then provides all required resources, such as, for example, databases with biological metadata or sequencing data to be analyzed, and then runs the workflow. Scalability refers to the ability of the workflow to run on different computer architectures, ranging



(a) Example workflow using the web-based WMS Galaxy from `https://galaxyproject.org/`.



(b) Example workflow using the configuration file-based WMS Popper from `https://popper.readthedocs.io/en/latest/sections/cn_workflows.html`.



(c) Example workflow using the domain specific language-based WMS Snakemake.

(d) Example workflow using the programming language-based WMS SciPipe from `https://scipipe.org/`.

FIGURE 1.1: Example workflows in four different WMSs highlighting the differences between their philosophies.

from Raspberry Pis over personal laptops to high-performance computing clusters, and by doing so being able to process varying amounts of data. This prevents locking in the workflow to a specific hardware environment. Portability ensures that all workflow dependencies are available at precisely defined versions during execution. This ensures that the workflow does not suddenly start producing different results because one of its dependencies was updated and has now slightly changed functionality.

The second pillar of sustainability is adaptability. It describes the ability to modify an existing workflow and adapt it to new research questions. This includes scalability and portability from the previous pillar and adds readability as a requirement. In this context, scalability makes it possible to process orders of magnitudes more samples and larger samples than initially envisioned, while portability ensures that new functionality can be easily implemented without worrying about compilation and software compatibility issues. To modify an existing workflow, it is necessary to understand each included component, how they depend on each other and what side effects occur. That is, the workflow needs to be readable.

Finally, the third pillar, transparency, is related to understanding a given workflow and also includes readability but adds traceability and documentation requirements. Traceability ensures that all parameters used to configure the workflow, all source code used to run the analysis and all workflow components traversed during execution are well defined and easily accessible. The existence of documentation makes sure that people not familiar with the code are able to get an overview of its functionality while also clearing up confusing parts for programmers who wish to understand the workflow and its results in more detail.

While following each pillar on its own is already a worthwhile pursuit which helps workflow development, only the joint application of all their teachings and being aware of their interconnections makes it possible to achieve truly sustainable data science.

Besides this taxonomy of sustainable data science, there exist other classification schemes which try to capture different aspects of how to improve computational workflow and method development. The principles of FAIR data, i.e., findable, accessible, interoperable and reusable data, can also be applied to workflows [29]. For example, projects such as the Interoperable Workflow Intermediate Representation [30] or the Common Workflow Language [31] aim at standardizing workflow descriptions and thus improving findability and interoperability. At the same time, different definitions of the term reproducibility exist and are often contrasted with replicability

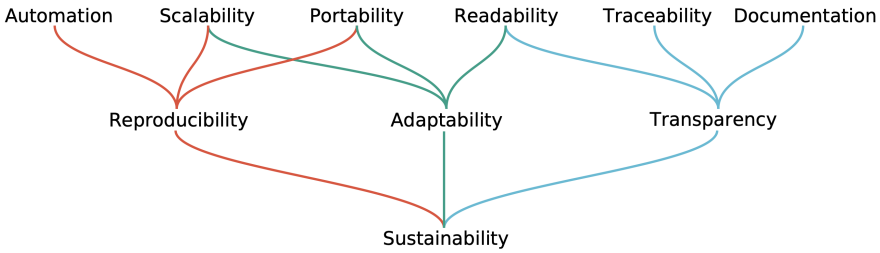


FIGURE 1.2: Hierarchical model of sustainable workflow development and data science from figure 1 in [28]. Red lines connect concepts related to workflow execution, green lines concepts related to workflow modification and blue lines concepts related to workflow understanding. All three are necessary to achieve sustainability in data science.

and repeatability [6]. While there exist some differences in their respective definitions between publications, they can typically be ordered by their degree of generality. While repeatability is the least general and usually refers to the same researchers repeating the same experiment in order to generate the same results, the other two terms are less consistently defined in literature. Both reproducibility and replicability usually refer to different researchers conducting the experiment, but replicability is often interpreted as the more general term. For example, [32] defines both replication and reproduction as experiments which are conducted by independent researchers while reaching the same final conclusions as the original experiment. In their view, the two terms differ in their requirements for implementation details, used hardware and produced raw data. According to their definition, when reproducing an experiment all these elements have to be kept the same. When replicating an experiment however, these elements should be different. In summary, the definition "Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators" proposed in [6], captures the same spirit as in the previously described hierarchy of sustainable data science.

1.3 SUSTAINABLE DATA SCIENCE AND ITS APPLICATIONS IN BIOINFORMATICS

Chapter 2 and chapter 3 present novel statistical methods for computing pathway dysregulations by integrating intra- and inter-pathway relations

into the model respectively. At the same time, they provide extensive benchmarking workflows to validate the results. In the following, we motivate why this is important.

Cancer is a complex, genomic disease which is caused by alterations of the genomes of cancer cells [33]. These alterations can be point mutations of single base positions of the genome, insertions and deletions of short stretches of DNA or copy number variations. The effects on the cancerous cell population can be summarized as a set of hallmarks which includes organizing principles such as resisting apoptosis and evading growth suppressors [34–36]. These cell populations are highly diverse and feature a high inter- as well as intra-patient heterogeneity: the same phenotype can have vastly different genetic causes. One way of reducing the high dimensionality of this genetic space is to consider the dysregulation of functional groups of genes, called pathways, instead of individual mutational events [37, 38]. This approach is called pathway enrichment analysis and promises improved performances in survival analysis, diagnostics and treatment design.

While this difficult biological setting in itself makes understanding cancer and thus improving diagnostics as well as treatment designs highly non-trivial, the situation is exacerbated by additional factors hindering reproducibility [12]. It starts at the lack of high quality model organisms [39] and suboptimal experimental design [40], but continues well into the field of computational oncology [41]. To counter this, specifically designed computational resources [42], analysis guidelines [43] and collaborative cloud environments [44] have been proposed. While they help to streamline the overall computational analysis, individually benchmarking the employed methods is not accomplished. In particular, the lack of reliable pathway enrichment benchmarks has been recognized [45] and various benchmark workflows were created [17, 46–48]. They typically combine a synthetic study with real data sets and have difficulties with simulating truly realistic and representative data in the former and knowing the appropriate ground truth in the latter case.

It thus becomes clear that the robustness and reliability of pathway enrichment methods used in contemporary computational oncology research is not being properly validated and the field is in need of sustainable benchmarking workflows.

Chapter 4 shows how V-pipe, a bioinformatics pipeline for the analysis of NGS data from short viral genomes, was extended to process large numbers of SARS-CoV-2 samples both from clinical and wastewater sources,

as well as a benchmark of global haplotype reconstruction methods used to estimate viral diversity. The rest of this section motivates the relevance of these developments.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a single-stranded RNA genome with 29,903 nucleotides and at least 13 open reading frames [49]. It is the cause of the respiratory disease COVID-19 and quickly developed into a global pandemic [50]. Due to this worldwide impact, it is crucial to track its geographical spread, infectious dynamics and mutational progression.

An important factor influencing such related properties including disease progression [51], transmission risk [52] and transmission heterogeneity [53] is the inter- as well as intra-host diversity of the virus. This diversity can be analyzed on the level of single mutations, local haplotypes which incorporate co-occurring mutations in small windows of sequencing read-length size and global haplotypes which cover the whole genome [54, 55]. These measures are of great relevance to public health. In principle, the high quality reconstruction of global haplotypes would be most beneficial. It however poses a set of challenges which still need to be overcome for it to become truly useful. They start with issues during sample preparation and sequencing but also include many computational challenges [56, 57]. For example, low-frequency mutations are difficult to distinguish from technical errors, there exists no unique solution to the reconstruction problem where genomic regions of variable genetic diversity exist next to each other, the true number of underlying haplotypes is generally unknown and needs to be estimated, and events more complicated than single point mutations, such as recombinations, are difficult to handle in current methods.

As a consequence, the analysis of viral sequencing data and estimation of its diversity levels requires great care during sample preparation and sequencing as well as during the subsequent computational investigation. This can be best achieved by organizing computational workflows in a sustainable way and properly benchmarking the included methods.

2

IDENTIFYING CANCER PATHWAY DYSREGULATIONS USING DIFFERENTIAL CAUSAL EFFECTS

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.

A novel computational approach, called Differential Causal Effects (*dce*), which compares normal to cancerous cells using the statistical framework of causality, is introduced. The method allows detecting 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. The approach is extended to handle unobserved dense confounding, where each latent variable, such as, for example, batch effects or cell cycle states, affects many covariates. *dce* outperforms competing methods on synthetic data sets and on CRISPR knockout screens. Its latent confounding adjustment properties are validated on a GTEx dataset. Finally, in an exploratory analysis of breast cancer data from TCGA, known and new genes involved in breast cancer progression are identified. 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.

The author's contributions to the following manuscript were the development of the statistical model, the implementation of the software package, as well as the synthetic performance evaluation and exploratory analysis. The paper is published as [58].

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 Beerenwinkel[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, 4058, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Basel, 4058, Switzerland
[3]Seminar for Statistics, ETH Zürich, Zürich, 8092, Switzerland
[*]To whom correspondence should be addressed
[$]Equal contribution

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 [1]. 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 [2–4].

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 [5, 6]. 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 two-fold, the expression of its child Y increases four-fold. Thus, X has a causal effect on Y of 2 [7]. Understanding how these causal networks are perturbed in tumors is necessary for prioritizing drug targets, understanding inter-patient heterogeneity, and detecting driver mutations [8].

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 [9]. 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 [10]. The underlying structure of gene interactions can thus be either estimated from the data used for the enrichment analysis [11, 12], or obtained from existing databases. Canonical pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [13] can then be incorporated as prior knowledge to guide the enrichment analysis using topological information of gene connectivity [14–17].

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 [16]. An exception is PARADIGM, which records an inferred activity for each entity in the pathway under consideration for a given patient sample [18]. 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 [19–21], 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 [13, 22–25]. 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. 2.1).

This allows us to identify pathway perturbations at the individual edge level while controlling for confounding factors using the statistical 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

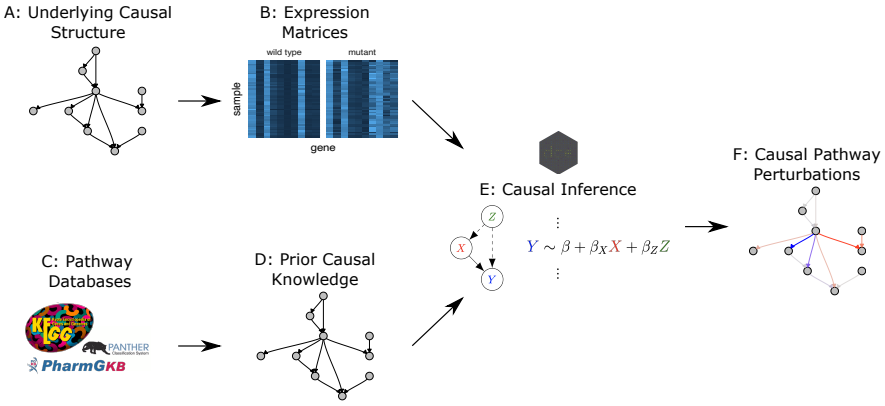


FIGURE 2.1: A causal network of genetic interactions in a biological pathway (A) is responsible for the observed wild type expression levels in a cell (B: wild type). A disease can lead to perturbations of these pathways and in turn generate altered expression levels (B: mutant). Pathway databases such as KEGG [13], PharmGKB [23] and Panther [24] curate genetic interaction data (C) and thus provide networks of putative causal interactions (D). Given the observed wild type and disease expression levels as well as the causal structure, *dce* fits a generalized linear model (GLM) for each edge to estimate differential causal effects (E). In the given example, the differential causal effect from X on Y (solid edge) is estimated using the valid adjustment set $\{Z\}$ (as determined from the dashed edges). These differential causal effects correspond to causal perturbations, i.e., differences in causal effects between two conditions. For example, an increase of causal effect strength from wild type to mutant is marked in blue, whereas the negative differential causal effects are marked in red (the transparency of an edge corresponds to the magnitude of the associated effect). These features are important for diagnosis and treatment design (F).

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) data set 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 [26]. 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.

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 paper. 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 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 $X = (X_i)_{i=1}^{p}$ describing the expression of genes, a set of directed edges $E = (E_i)_{i=1}^{m}$ representing the causal structure and the structural equations $(f_i)_{i=1}^{n}$ describing how each variable X_i is generated from its parents $X_{pa(i)}$ in $\mathcal{G}$, $X_i \leftarrow f_i(X_{pa(i)}, \epsilon_i)$, where $(\epsilon_i)_{i=1}^{p}$ 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_j affects the expression of the child node X_i, 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 $\mathbb{E}[X_i \mid do(X_j = x)]$, which describe how the expected value of the variable X_i changes when we intervene and set the variable X_j to

some fixed value x. We define the causal effect (CE) of a variable X_j on its descendant X_i as

$$CE[X_i \mid do(X_j = x)] = \frac{d}{dx}\mathbb{E}[X_i \mid do(X_j = x)]. \tag{2.1}$$

This derivative equals β_x if, by changing the value of X_j from x to $x + \Delta x$, for some small value Δx, the value of X_i changes on average by $\beta_x \cdot \Delta x$. 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_j 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 cancerous 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:

$$X_i \leftarrow \gamma_0^{(i)} + \sum_{j \in pa(i)} \gamma_j^{(i)} X_j + \epsilon_i(X_{pa(i)}). \tag{2.2}$$

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 $\gamma_j^{(i)}$ correspond to the direct causal effects, whereas the total causal effects (2.1) 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 (2.2), the causal effect $CE[X_i \mid do(X_j = x)]$ does not depend on x. Furthermore, this causal effect corresponds to the coefficient β in the linear regression of X_i on X_j and an adjustment set $Z = (Z_k)_{k=1}^{|Z|}$,

$$X_i = \beta_0 + \beta X_j + \sum_{k=1}^{|Z|} \beta_k Z_k + \eta. \tag{2.3}$$

Here, β_0 denotes the intercept and η is random noise with mean zero [27, 28]. The adjustment set Z is a set of nodes in the pathway $\mathcal{G}$ which fulfills the Back-door criterion [7]. Hence, it holds that no element of Z is a descendant of X_j, and Z blocks every path between X_i and X_j 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_j on the gene expression X_i are respectively denoted as β^A and β^B under different conditions A and B, then the differential causal effect (DCE) δ is obtained as the difference

$$\delta = \beta^B - \beta^A. \tag{2.4}$$

Given a graph $\mathcal{G}$ describing a biological pathway and observations of the variables, we can compute all differential causal effects and identify interactions between any such two variables X_j and X_i that are different between the two conditions (fig. 2.1).

TESTING FOR SIGNIFICANCE. We can compute the DCE δ for the edge $X_j \to X_i$ 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 δ can be computed from all samples jointly by fitting the following linear model

$$\begin{aligned} X_i = &(\beta_0^A + (\beta_0^B - \beta_0^A)I) + (\beta^A + (\beta^B - \beta^A)I)X_j + \\ &\sum_{k=1}^{|Z|} \left(\beta_k^A + (\beta_k^B - \beta_k^A)I\right) Z_k + \eta \end{aligned} \tag{2.5}$$

with interaction terms $I \cdot X_j$ and $I \cdot Z_i$. The differential causal effect $\delta = \beta^B - \beta^A$ can be estimated by using the coefficient estimate corresponding to the interaction term IX_j in (2.5).

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 (2.2) 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 [29]. In this case, the linear model (2.5) 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 [30–32].

Besides assessing significance of DCEs for single edges, we can also calculate a global p-value measuring the overall dysregulation of a given pathway $\mathcal{G}$: we combine the p-values corresponding to different differential causal effects $\boldsymbol{\delta} = (\delta_i)_{i=1}^m$ by taking their harmonic mean [33].

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 [34, 35]. 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 [36], 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 [37].

Some information about latent factors can often be obtained from the principal components of the data [38]. This can be made rigorous under the linearity assumption (2.2) for our structural equation model $\mathcal{G}$, as follows. We assume that there are q latent variables $H_1, \ldots, H_q$ affecting our data. We extend the model (2.2) to include the latent confounding as follows:

$$X_i \leftarrow \gamma_0^{(i)} + \sum_{j \in pa(i)} \gamma_j^{(i)} X_j + \sum_{j=1}^{q} \delta_j^{(i)} H_j + \epsilon_i(X_{pa(i)}, H), \tag{2.6}$$

that is, the latent confounders $H_1, \ldots, H_q$ are additional source nodes in the DAG $\mathcal{G}$ and affect genes in the pathway linearly, analogously to (2.2). Not every gene needs to be affected ($\delta_j^{(i)}$ could be zero), but the methodology works better when many genes are affected, see discussion below. By writing the structural equations (2.6) in matrix form, where we define the matrices $\Gamma^0_{ji} = \gamma_0^{(i)}$, $\Gamma_{ji} = \gamma_j^{(i)}$, $\Delta_{ji} = \delta_j^{(i)}$ and $E(X,H)_{ji} = \epsilon_i(X_{pa(i)}, H)_j$, we obtain

$$X_{n\times p} \leftarrow \Gamma^0_{n\times p} + X_{n\times p}\Gamma_{p\times p} + H_{n\times q}\Delta_{q\times p} + E(X,H)_{n\times p}, \tag{2.7}$$

which gives

$$X = \underbrace{\Gamma^0}_{\text{intercepts}} + H \underbrace{\Delta(I-\Gamma)^{-1}}_{\text{loadings} \in \mathbb{R}^{q\times p}} + \underbrace{E(X,H)(I-\Gamma)^{-1}}_{\text{random noise with mean = 0}}, \tag{2.8}$$

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. 2.2). 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. 2.2, left). Therefore we obtain the confounding proxies $\hat{H}$ as the scores

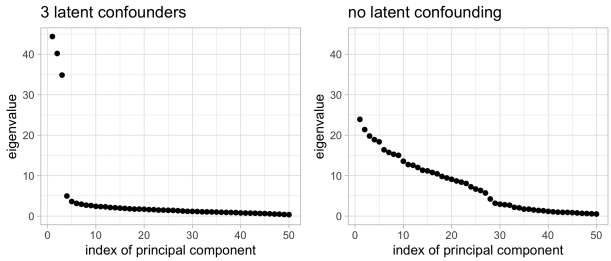


FIGURE 2.2: The scree plot (of synthetic data generated as described in the Methods section) shows that in presence of latent confounding as in (2.6), the first q principal components explain much more variability of the data, which we exploit for confounding adjustment.

of the first $\hat{q}$ principal components of the design matrix combining the data from both conditions.

The confounding proxies $\hat{H}$ are then simply added to the adjustment set Z, see equations (2.3) and (2.5). In this way, the Back-door adjustment not only adjusts for the confounding variables observed in the DAG $\mathcal{G}$ as before, but also helps reducing the bias induced by latent confounding.

The deconfounding methodology relies on the assumption that every confounding variable affects many variables in the dataset, i.e., the confounding is dense [39]. 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 covariates each latent factor H_i affects, the more information we have about it in the data and thus the confounding proxies $\hat{H}$ 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. 2.2). This helps estimating the number $\hat{q}$ of the confounding proxies used. As a default choice, we use a permutation method that can be shown to work well under certain assumptions [40] 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 $\mathcal{G}$. 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 $\mathcal{G}$ is a DAG. In practice, we sample edges from a binomial distribution with probability $\hat{p}$ for the upper triangle of $\mathcal{G}$. We further sample the coefficients $\gamma_j^{(i)}$ for every edge as in (2.2) from a uniform distribution $\mathcal{U}(-\gamma_{\max}, \gamma_{\max})$. We generate the data for network $\mathcal{G}$ in the following way. For a node X_i, we set the mean expression count

$$\mu_i = v - \vec{1} \cdot \left(\min_i v_i - \iota \right), \tag{2.9}$$

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,

$$v = \sum_{j \in pa(i)} \gamma_j^{(i)} X_j \tag{2.10}$$

where $\gamma_j^{(i)}$ represents the direct effect of X_j on X_i, $\iota > 0$ is a small shift, and $\vec{1}$ is a vector of ones. Subtracting the minimum ensures positive values of the mean for each data point. Then, a realization of X_i is drawn from the Poisson distribution $\text{Pois}(\mu_i)$. We introduce negative binomial noise by drawing a realization of each source node in $\mathcal{G}$ from the negative binomial distribution $\text{NB}(\mu, \theta)$ with a general mean μ and dispersion θ. 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 $\mathcal{G}$ 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 β^A we sample the new edge weight uniformly such that

$$\beta^B - \beta^A \sim \mathcal{U}\left([-\delta_{\max}, -\delta_{\min}] \cup [\delta_{\min}, \delta_{\max}]\right). \tag{2.11}$$

This ensures that the absolute difference between the two edge weights lies in $[\delta_{\min}, \delta_{\max}]$.

We also simulate latent variables. They are neither included in the data nor the network $\mathcal{G}$, but have (unknown) outgoing edges to all genes in the data set 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 [20, 41], a differential gene expression approach (*dge*) and random guessing. *cor* is provided by the R package `stats` [42]. For *pcor* we use the general matrix inversion from the R package `MASS` [43] 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 [20] to run *fggm*. For *fggm* we transform each gene expression count g to $\log(g+1)$. We use the differential expression result from `edgeR` [29] 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*.

We run all methods on simulated data for various modeling parameters. The default parameters are a network $\mathcal{G}$ 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 $\mathcal{G}$ (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 as 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 data set of $10,000$ genes including the genes in the network $\mathcal{G}$ 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 can 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 [44, 45]. The data set we analyze is a CRISPR knockout screen with global gene expression profiles as the read-out. We can use the known knockout information of these experiments as ground truth information for a performance evaluation of our method. In [45], this approach was used to systematically analyze the response of an integrated endoplasmic reticulum (ER) stress response pathway to the combinatorial knockout of the three transmembrane sensor proteins ATF6, EIF2AK3 and ERN1. Each considered combinatorial knockout (ATF6, ATF6+EIF2AK3, ATF6+EIF2AK3+ERN1, ATF6+ERN1, EIF2AK3, EIF2AK3+ERN1, ERN1) was treated either with a DMSO control, tunicamycin, or thapsigargin.

We download the raw gene expression count data from NCBI GEO (accession: GSE90546). The repository provides us with a mapping of guide and cell barcodes, and gene expression counts for all cells. We use this information to identify gene knock-outs for each cell and to create a gene expression count matrix of the individual cells labeled by their corresponding knockouts.

We download all pathway networks from KEGG and retain those which contain at least one of the three transmembrane sensor proteins. This results in the pathways *hsa04137*, *hsa04140*, *hsa04141*, *hsa04210*, *hsa04932*, *hsa05010*, *hsa05016*, *hsa05017*, *hsa05160*, *hsa05162*, *hsa05168*.

For each combination of the three treatments, seven (combinatorial) knockouts and 11 pathways, we compute DCEs if the respective knocked-out gene is contained in the respective pathway. In total, this yields 128 conditions for each of which we run our method.

We compare the performance of *dce* to both *cor* (correlation) and *pcor* (partial correlation). For the two correlation methods, we estimate the significance of whether a difference in correlation is different from zero using a permutation test. The performance of each method is evaluated using the area-under-curve (AUC) metric for the receiver-operating-characteristic

(ROC) curve. The false and true positive rates for the ROC curve are computed from the p-value per edge as in the synthetic benchmark.

DECONFOUNDING VALIDATION ON GTEX DATA. From the Genotype-Tissue Expression (GTEx) project [26], we obtain gene expression data for the samples belonging to many different human tissue types. For any pathway, one can use *dce* for comparing the expression data between two different tissue types. This approach will detect the edges for which the causal effects differ between the tissues. While this biological scenario is much different to comparing perturbed and unperturbed, or normal and tumor samples, the concept of DCEs remains the same.

In line with the rest of the paper, we choose the breast cancer pathway (*hsa05224*) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [13] and compare mammary gland tissue with each of 29 other tissue types that contain at least 200 samples.

An interesting feature of the available data set is that one is given 23 confounding proxies including genotyping principal components, gender of donors and PEER (probabilistic estimation of expression residuals) factors [46]. For the original breast cancer pathway (*hsa05224*), we run *dce* twice: once with and once without the confounding adjustment, yielding two sets of DCEs. Afterwards, we extend the pathway by adding the confounding proxies as the source nodes that have no incoming edges and have outgoing edges to all other nodes in the pathway. *dce* with and without confounding adjustment is then run on the extended pathway. This again yields two sets of DCEs. Finally, for both variants of *dce* (with and without confounding adjustment), we compute Pearson correlation between the obtained p-values for the original and the extended pathway in order to measure how well our confounding adjustment (which does not use any information about the confounding) is able to capture the effect of the known confounders.

EXPLORATORY ANALYSIS WITH TCGA DATA. We retrieve gene expression matrices from The Cancer Genome Atlas (TCGA) [47]. The rows of these matrices are indexed by genes and the columns by samples. The entries are from the data category Transcriptome Profiling, data type Gene Expression Quantification, experimental strategy RNA-Seq and workflow type HTSeq-Counts. Pathway structures in the form of adjacency matrices are obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [13].

Unlike the Perturb-seq dataset, data obtained from TCGA is observational instead of interventional. We do thus not have any ground truth information and perform an exploratory analysis. For a given cancer type, the associated samples are first grouped into normal and tumor samples. The tumor samples are subsequently stratified according to their stage. The clinical data needed to stratify the samples is readily available on TCGA as metadata for each gene expression matrix. In particular, we download all normal and tumor gene expression samples from TCGA for breast cancer (*TCGA-BRCA*) and selected all stages with a sufficient number of samples (stage I: 202 samples, stage II: 697 samples, stage III: 276 samples; normal: 113 samples). We use the breast cancer pathway (*hsa05224*) from KEGG which contains 147 nodes and 509 edges. We then compute DCEs between the normal condition and each of the three stages of the tumor condition, respectively.

RESULTS

In this section, we first show the performance of *dce* and its competitors on simulated data and a CRISPR data set. Next, we evaluate the deconfounding performance using the GTEx data set. 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.

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. 2.3a). 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 computed precision and recall for a p-value cutoff of 0.05 (figs. 2.12 and 2.13). 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

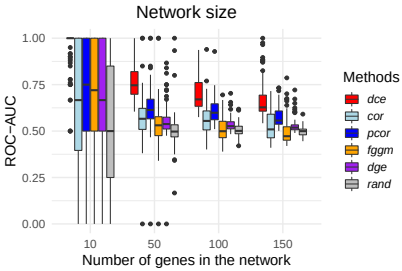
between the network of the wild type samples and the disease samples differ by at most 1. *dce* has difficulty estimating large differences as well as very small differences. However, it still significantly outperforms all other methods, which again show similar performance to random guessing for large effects (fig. 2.3b).

In additional simulations, *dce* shows increasing ROC-AUC for decreasing dispersion and increasing number of samples (figs. 2.6 and 2.7) as is expected due to decreasing noise. We found constant ROC-AUC of *dce* over varying ranges of library size (fig. 2.8). Different prevalence of positive edges has little effect on the ROC-AUC of *dce* (fig. 2.9). *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 data set (fig. 2.10). This is because without latent confounding adjustment one has a large number of false positives due to the confounding bias (fig. 2.11). 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 (fig. 2.14). 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.

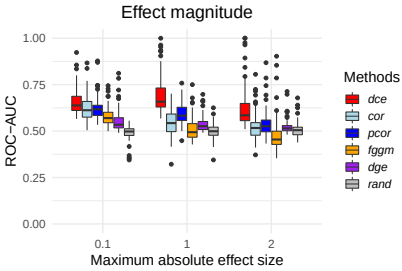
dce relies heavily on the given network $\mathcal{G}$. Hence, we investigate how well *dce* performs if $\mathcal{G}$ contains false edges or is missing true edges. We find that *dce* is robust to additional false edges in the network, but starts breaking down if true edges are missing in larger fractions (fig. 2.15).

Validation experiments using CRISPR knockout data

To benchmark our method using real-life data generated by Perturb-seq [45], 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 (fig. 2.15). 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.



(a) Network size. The ROC-AUC decreases as the number of pathway genes increases. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.



(b) Effect magnitude. Large causal effects reduce the ROC-AUC of DCE.

FIGURE 2.3: Performance benchmark. *dce* is compared to several competitors for varying network size (a) and effect magnitude (b) over 100 synthetic data sets each. *dce* achieves the highest ROC-AUC, which decreases for large networks $\mathcal{G}$ and very large or small differential effects. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

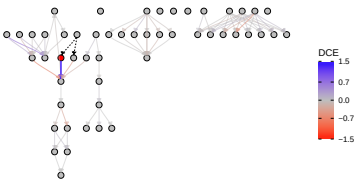
Figure 2.4a shows an example of this procedure for one of the conditions described above. The CRISPR knockout gene is highlighted in red and a positive DCE of $\sim$1.3 can be observed on the edge connecting ATF6 and DDIT3. This can be seen in more detail in fig. 2.4b. As this edge is adjacent to the knocked out gene ATF6, it is classified as a true positive for an effect size threshold of $|0.5|$. Following an analogous argument, the edge from EIF2AK3 to EIF2S1 is classified as a false positive.

We find that *dce* is significantly better (Wilcoxon signed-rank test [48] p-value $\leq 10^{-5}$) at recovering the knockout effects with a median ROC-AUC of 0.63 compared to 0.51 for *cor* and 0.53 for *pcor* (fig. 2.4c). To better understand the variability of the performance measure, we also investigate how performance varies when stratified by treatment and knockout gene (fig. 2.17). For example, for the knockout gene ATF6 the ROC-AUC of *dce* decreases from 0.89 for treatment 1 to 0.67 for treatment 2. This can be explained by the higher variability of the gene expression counts under treatment 2 (standard deviation of gene expression counts for treatment 1 is 0.88, and 0.99 for treatment 2), as the p-value estimation becomes less stable. This pattern can also be observed for other performance shifts between treatments. We note that *cor* outperforms *dce* for the knockout of ATF6 in treatment 2, as the permutation test is able to better account for the variance of the expression data in this case. This is due to the fact that the permutation test relies on fewer assumptions than the significance test in our joint model. In all other cases, *dce* is either better or roughly as good as the competing methods. We conclude that overall *dce* is able to better recover the dysregulations of single as well as combinatorial knockouts when compared to methods based on correlations.

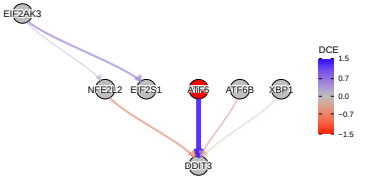
Deconfounding validation on GTEx data

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 Methods section, is used, we observe that the estimated DCEs between two different tissue types differ much less between the original and extended pathways (fig. 2.16).

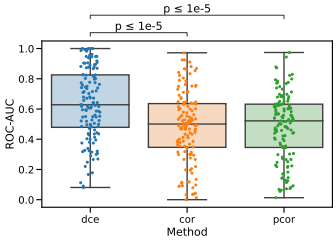
Similarly, the resulting p-values are also much more stable, as measured by the Pearson correlation between the negative logarithmic p-values computed for the original and extended pathway (fig. 2.4d). The correlation is



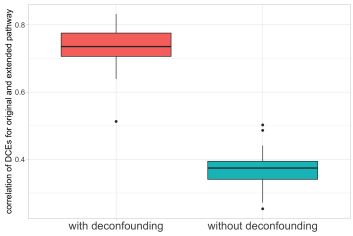
(a) Protein processing in the endoplasmic reticulum pathway for Homo Sapiens from KEGG (ID *hsa04141*). Each node corresponds to a gene and each edge to an interaction between two genes. Each edge is colored according to the effects size of DCEs computed for the experimental data for knocking out ATF6 and using DMSO control. The width of an edge corresponds to its absolute DCE (the wider an edge, the larger the absolute DCE). Black dashed edges are drawn when one of the two connected nodes has zero coverage (and thus no DCE can be estimated). The gene knocked out in the CRISPR experiment is highlighted in red.



(b) Zoomed-in version of fig. 2.4a with focus on the genes ATF6, ATF6B, NFE2L2, XBP1, DDIT3, EIF2AK3, EIF2S1. These genes constitute the neighborhood of the knocked-out gene ATF6 and illustrate the edge classification scheme used in the performance evaluation. Assume an effect size threshold of $|0.5|$. The edge ATF6→DDIT3 has a DCE of $\sim$1.3 and is adjacent to the knocked-out gene. Consequently, it is classified as a true positive. Both the edge EIF2AK3→EIF2S1 and NFE2L2→DDIT3 have a DCE whose absolute value is larger than 0.5 and are not adjacent to the knocked-out gene. They are thus classified as false positives. All remaining edges are classified as true negatives.



(c) Summary of the performance of the *dce*, *cor* and *pcor* methods in the form of ROC-AUCs for the recovery of the knocked-out genes in all considered pathways. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median. Additionally, each data point is indicated with a dot whose x position has been randomly shifted to improve visibility. The method *dce* shows the best performance with a ROC-AUC of 0.63 (standard deviation (std): 0.23) compared to 0.51 (std: 0.23) for *cor* and 0.53 (std: 0.22) for *pcor*. The significance of the difference between the boxplots has been estimated using the Wilcoxon signed-rank test [48].



(d) Correlation of the estimated $-\log_{10}(\text{p-value})$ for the original and the extended pathway, obtained by adding the provided confounding variables as the source nodes. The p-values are computed by the *dce* method with and without confounding adjustment and correspond to the significance of the estimated DCEs between the mammary gland tissue and 29 other tissue types. The variant of the *dce* with confounding adjustment shows the best performance with average correlation 0.73 (standard deviation (std): 0.07) compared to 0.37 (std: 0.06) for the version without confounding adjustment. The p-value for the difference between the boxplots equals 6.7×10^{-11} when using the Wilcoxon signed-rank test [48].

FIGURE 2.4: Overview of the CRISPR (figs. 2.4a to 2.4c) and GTEx (fig. 2.4d) benchmark results.

consistently larger when using the confounding adjustment, which is important since the latent confounding in general causes many false positives in the analysis.

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 (figs. 2.5a to 2.5c). The raw DCE values were transformed to a symmetric logarithm for greater visibility with the following formula

$$\text{symlog}(x) = \begin{cases} \log_{10}(x) + 1 & \text{if } x > 1 \\ -\log_{10}(-x) - 1 & \text{if } x < -1 \\ x & \text{otherwise} \end{cases} \tag{2.12}$$

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. 2.5d).

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 $< 10^{-20}$), as well as of WNT11→FZD7 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 [49–51] and in breast cancer in particular [52, 53].

Interactions between the FGF (Fibroblast Growth Factor) and FGFR (Fibroblast Growth Factor Receptor) protein complexes show strong negative effect sizes in all three stages (DCE < -100 for most members of these complexes). In particular, the FGF6→FGFR1 link features negative DCEs of $-1279, -665, -1961$, while the FGF8→FGFR1 link features negative DCEs

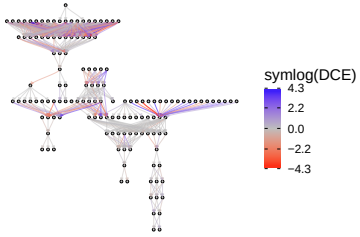
of $-402, -336, -285$, in the stages I, II, III respectively. This pair has already been recognized as a promising therapeutic target for breast cancer treatment [54].

We also find the interaction between EGFR (Epidermal Growth Factor Receptor) and PIK3CA (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha) to be significantly (p-values $< 10^{-14}$) dysregulated with a small negative DCE of approximately -0.2 in stages I and II but not III. EGFR$\rightarrow$PIK3CB shows similar behavior for stage II with a DCE of -0.12 and a p-value $< 10^{-15}$. While the small effect size suggests that there is only a small dysregulation of these interactions, the dysregulation of EGFR together with PIK3CA mutations have been recognized as independent prognostic factors in triple negative breast cancers [55].

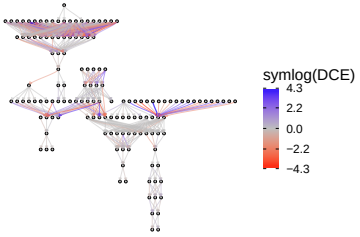
The interaction between DLL3 (Delta Like Canonical Notch Ligand 3) and NOTCH4 (Notch Receptor 4) features a significant DCE of $\sim$140 with p-values $< 10^{-6}$ in all three stages. The Notch signaling pathway has been shown to play an important role in Pancreatic ductal adenocarcinoma tumor cells, but has not been implicated in breast cancer [56]. 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/β-catenin signaling pathway and is important for cell development and growth regulation, has already been discussed in the context of breast cancer [57]. 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.

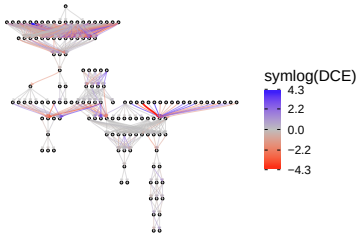
Overall, we are able to recover both interactions which are known to be dysregulated in breast cancer as well as novel ones. The former indicates that the prioritization of interactions given by *dce* is in accordance with current literature. The latter suggests that *dce* is also able to find dysregulated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



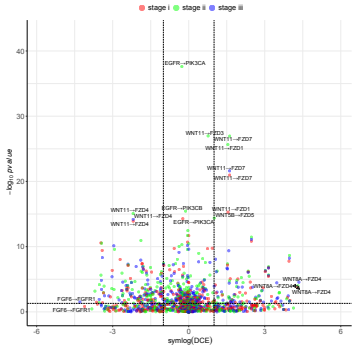
(a) DCEs for normal versus stage I samples.



(b) DCEs for normal versus stage II samples.



(c) DCEs for normal versus stage III samples.



(d) Volcano plot of effect size on the x-axis against the $-\log_{10}(\text{p-value})$ on the y-axis for all interactions over all three stages.

FIGURE 2.5: DCEs for *TCGA-BRCA normal* samples versus stage I, stage II, and stage III computed with the *hsa05224* pathway. In (a)-(c), edge thickness and opacity scale with absolute DCE size. More negative DCEs appear red, more positive DCEs appear blue. The color follows a symmetric logarithmic scale for values $|x| \geq 1$ and is linear otherwise. (d) shows a volcano plot for the symmetric logarithm of DCE against its associated $-\log_{10}(\text{p-value})$. DCE thresholds of 1 and -1 as well as a p-value threshold of 0.05 are denoted with grey dashed lines.

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 [58, 59]. 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 data sets 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 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 data sets 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.

DATA AVAILABILITY

The code used to construct the synthetic data sets is available as part of the R software package *dce*. The experimental data used in the Perturb-seq validation is available under the accession GSE90546 from NCBI GEO. GTEx data is publicly available through the GTEx portal. The experimental data used in the exploratory breast cancer analysis is 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.

CODE AVAILABILITY

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 [60] workflows needed to reproduce all results presented here.

ACKNOWLEDGEMENTS

Part of this work has been funded by SystemsX.ch, the Swiss Initiative in Systems Biology, under Grant No. RTD 2013/152 (TargetInfectX - Multi-Pronged Perturbation of Pathogen Infection in Human Cells), evaluated by the Swiss National Science Foundation, and by ERC Synergy Grant 609883 (to NB). The research of DC and PB has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461). The logo of *dce* was created with `https://github.com/dirmeier/ggpixel`.

AUTHOR INFORMATION

CONTRIBUTIONS KPJ and MP conceived the project. KPJ and MP developed the statistical model of *dce* and implemented the software package. DC contributed to the statistical methodology as well as software implementation. NB and PB supervised the study. KPJ and MP wrote the initial manuscript draft. All authors edited the manuscript.

CORRESPONDING AUTHORS Correspondence to Niko Beerenwinkel (niko.beerenwinkel@bsse.ethz.ch).

ETHICS DECLARATIONS

COMPETING INTERESTS The authors declare no competing interests.

BIBLIOGRAPHY

1. Nature Cancer. The global challenge of cancer. *Nature Cancer* **1**, 1 (2020).
2. Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *BMJ* **364** (2019).
3. Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., Jemal, A., Kramer, J. L. & Siegel, R. L. Cancer treatment and survivorship statistics, 2019. *CA: a cancer journal for clinicians* **69**, 363 (2019).
4. Troester, M. A. & Swift-Scanlan, T. Challenges in studying the etiology of breast cancer subtypes. *Breast Cancer Research* **11**, 104 (2009).
5. Khakabimamaghani, S., Ding, D., Snow, O. & Ester, M. Uncovering the subtype-specific temporal order of cancer pathway dysregulation. *PLoS computational biology* **15** (2019).
6. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646 (2011).
7. Pearl, J. *Causality: Models, Reasoning and Inference.* (Cambridge University Press, Cambridge, UK., 2000).
8. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. & Kinzler, K. W. Cancer genome landscapes. *science* **339**, 1546 (2013).
9. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).
10. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research* **40**, e133 (2012).
11. Spirtes, P., Glymour, C. N., Scheines, R. & Heckerman, D. *Causation, prediction, and search* (MIT press, 2000).
12. Sedgewick, A. J., Shi, I., Donovan, R. M. & Benos, P. V. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC bioinformatics* **17**, 307 (2016).
13. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* **27**, 29 (1999).

14. Liu, A., Trairatphisan, P., Gjerga, E., Didangelos, A., Barratt, J. & Saez-Rodriguez, J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *NPJ systems biology and applications* **5**, 1 (2019).

15. Dutta, B., Wallqvist, A. & Reifman, J. PathNet: a tool for pathway analysis using topological information. *Source code for biology and medicine* **7**, 10 (2012).

16. Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P. & Romero, R. A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75 (2009).

17. Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S. & Sorger, P. K. Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular Systems Biology* **5**, 331 (2009).

18. Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. & Stuart, J. M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237 (2010).

19. Wang, Y., Squires, C., Belyaeva, A. & Uhler, C. *Direct estimation of differences in causal graphs* in *Advances in Neural Information Processing Systems* (2018), 3770.

20. He, H., Cao, S., Zhang, J.-g., Shen, H., Wang, Y.-P. & Deng, H.-w. A Statistical Test for Differential Network Analysis Based on Inference of Gaussian Graphical Model. *Scientific Reports* **9**, 10863 (2019).

21. Tian, D., Gu, Q. & Ma, J. Identifying gene regulatory network rewiring using latent differential graphical models. *Nucleic acids research* **44**, e140 (2016).

22. Nishimura, D. BioCarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117 (2001).

23. Whirl-Carrillo, M., McDonogh, E., Herbet, J., Gong, L., Sangkuhl, K., Thotn, C., Altman, R. & Klein, E. *Pharmacogenomics Knowledge for Personlized Medicine. Clinical Pharmacology and Therpeutics 92, 4 (2012), 414–417* 2012.

24. Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T. & Thomas, P. D. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research* **49**, D394 (2021).

25. Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. & Buetow, K. H. PID: the pathway interaction database. *Nucleic acids research* **37**, D674 (2009).

26. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* The genotype-tissue expression (GTEx) project. *Nature genetics* **45**, 580 (2013).

27. Goldszmidt, M. & Pearl, J. *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* (1992), 661.

28. Pearl, J. Causal Diagrams for Empirical Research. *Biometrika* **82**, 669 (1995).

29. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881 (2007).

30. Eicker, F. *Limit theorems for regressions with unequal and dependent errors* in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1** (1967), 59.

31. Huber, P. J. *et al. The behavior of maximum likelihood estimates under nonstandard conditions* in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1** (1967), 221.

32. White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817 (1980).

33. Good, I. J. Significance tests in parallel and in series. *Journal of the American Statistical Association* **53**, 799 (1958).

34. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882 (2012).

35. Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, 1 (2013).

36. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR genomics and bioinformatics* **2**, lqaa078 (2020).

37. Ćevid, D., Bühlmann, P. & Meinshausen, N. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research* **21**, 232 (2020).

38. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40**, 646 (2008).

39. Guo, Z., Ćevid, D. & Bühlmann, P. Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. *arXiv preprint arXiv:2004.03758* (2020).

40. Dobriban, E. Permutation methods for factor analysis and PCA. *arXiv preprint arXiv:1710.00479* (2017).

41. Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., Wang, J. & Chen, W. FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLOS Computational Biology* **12**, 1 (2016).

42. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020).

43. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* Fourth. ISBN 0-387-95457-0 (Springer, New York, 2002).

44. Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P. & Lim, W. A. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173 (2013).

45. Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867 (2016).

46. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology* **6**, e1000770 (2010).

47. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113 (2013).

48. Wilcoxon, F. in *Breakthroughs in statistics* 196 (Springer, 1992).

49. Dijksterhuis, J. P., Baljinnyam, B., Stanger, K., Sercan, H. O., Ji, Y., Andres, O., Rubin, J. S., Hannoush, R. N. & Schulte, G. Systematic mapping of WNT-FZD protein interactions reveals functional selectivity by distinct WNT-FZD pairs. *Journal of Biological Chemistry* **290**, 6789 (2015).

50. Chien, A. J., Conrad, W. H. & Moon, R. T. A Wnt survival guide: from flies to human disease. *Journal of Investigative Dermatology* **129**, 1614 (2009).

51. Schulte, G. International union of basic and clinical pharmacology. LXXX. The class Frizzled receptors. *Pharmacological reviews* **62**, 632 (2010).

52. Yin, P., Wang, W., Gao, J., Bai, Y., Wang, Z., Na, L., Sun, Y. & Zhao, C. Fzd2 contributes to breast cancer cell mesenchymal-like stemness and drug resistance. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **28**, 273 (2020).

53. Koval, A. & Katanaev, V. L. Dramatic dysbalancing of the Wnt pathway in breast cancers. *Scientific reports* **8**, 1 (2018).

54. Santolla, M. F. & Maggiolini, M. The FGF/FGFR System in Breast Cancer: Oncogenic Features and Therapeutic Perspectives. *Cancers* **12**, 3029 (2020).

55. Jacot, W., Mollevi, C., Fina, F., Lopez-Crapez, E., Martin, P.-M., Colombo, P.-E., Bibeau, F., Romieu, G. & Lamy, P.-J. High EGFR protein expression and exon 9 PIK3CA mutations are independent prognostic factors in triple negative breast cancers. *BMC cancer* **15**, 1 (2015).

56. Song, H. & Zhang, Y. Regulation of pancreatic stellate cell activation by Notch3. *BMC cancer* **18**, 1 (2018).

57. Connor, A. E., Baumgartner, R. N., Baumgartner, K. B., Kerber, R. A., Pinkston, C., John, E. M., Torres-Mejia, G., Hines, L., Giuliano, A., Wolff, R. K., *et al.* Associations between TCF7L2 polymorphisms and risk of breast cancer among Hispanic and non-Hispanic white women: the Breast Cancer Health Disparities Study. *Breast cancer research and treatment* **136**, 593 (2012).

58. Song, D., Cui, M., Zhao, G., Fan, Z., Nolan, K., Yang, Y., Lee, P., Ye, F. & Zhang, D. Y. Pathway-based analysis of breast cancer. eng. *Am J Transl Res* **6**, 302 (2014).

59. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W., Liu, B., Lei, Y., Du, S., Vuppalapati, A., Luu, H. H., Haydon, R. C., He, T.-C. & Ren, G. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases* **5**, 77 (2018).

60. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).

SUPPLEMENTS

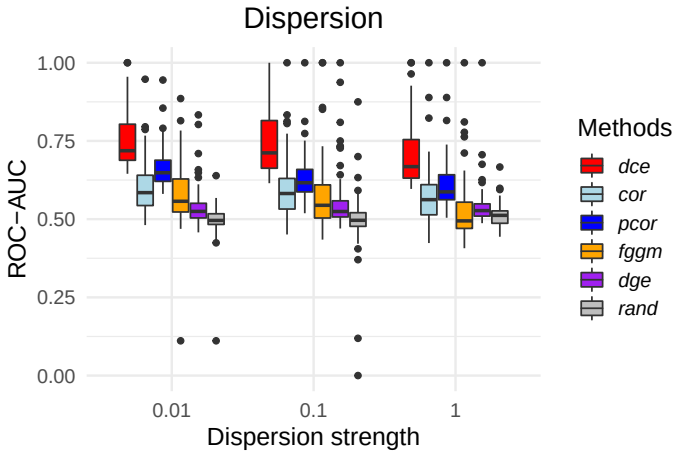


FIGURE 2.6: Dispersion. *dce* is compared to its competitors over 100 synthetic data sets with varying dispersion values. Performance decreases for higher dispersion values. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

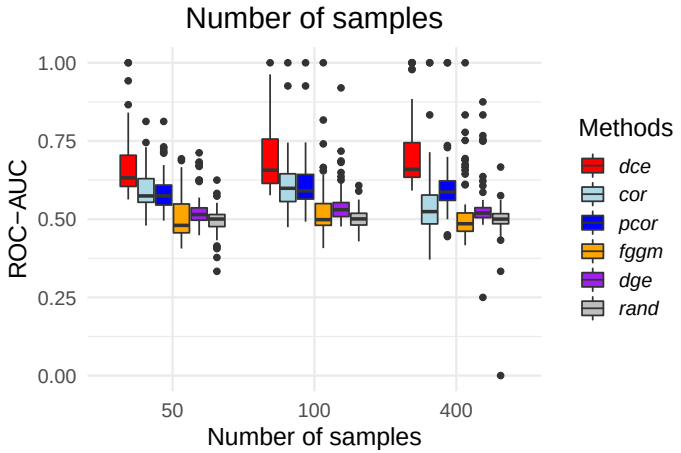


FIGURE 2.7: Sample size. *dce* is compared to its competitors over 100 synthetic data sets with varying sample sizes for one condition. The other conditions has a fixed sample size of 200. Performance decreases for lower sample sizes. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

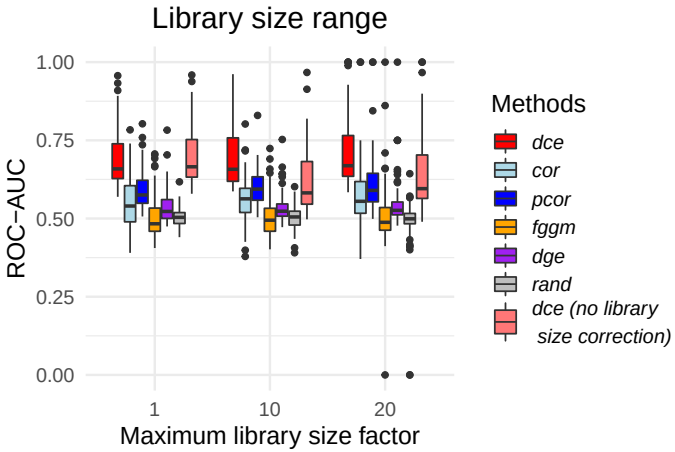


FIGURE 2.8: Library size. *dce* is compared to its competitors over 100 synthetic data sets with varying library size factors. Library size has little effect on the ROC-AUC of all methods. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

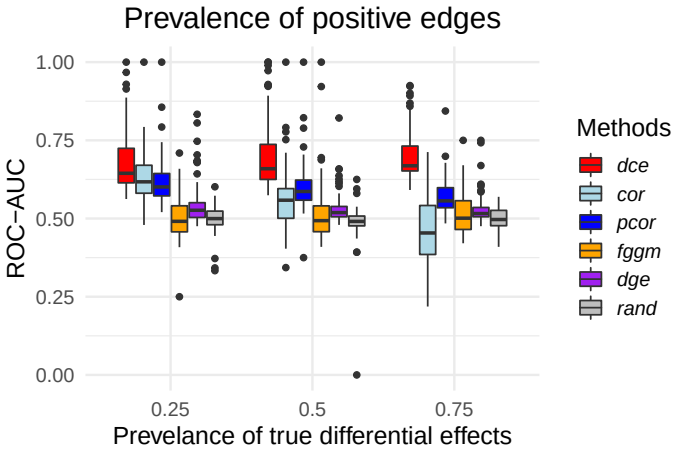


FIGURE 2.9: Prevalence. *dce* is compared to its competitors over 100 synthetic data sets with varying prevalence for DCE $\neq 0$. ROC-AUC decreases for all methods and higher prevalence except for *dce*. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

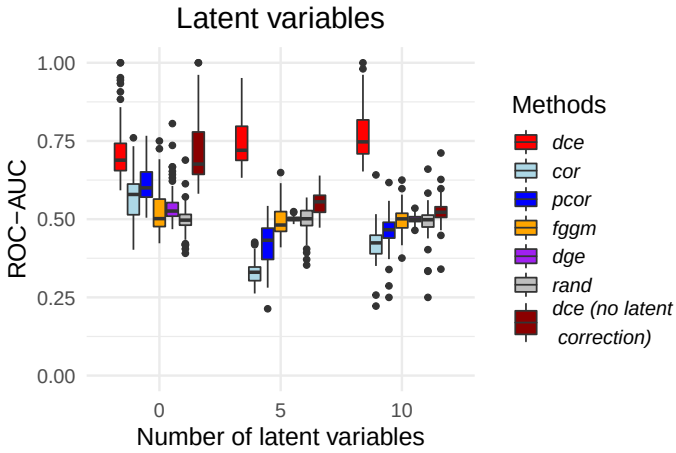


FIGURE 2.10: Latent variables. *dce* is compared to its competitors over 100 synthetic data sets with varying numbers of latent variables. *dce*'s ROC-AUC stays robust, if we account for latent variables, but drastically decreases, if we do not. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

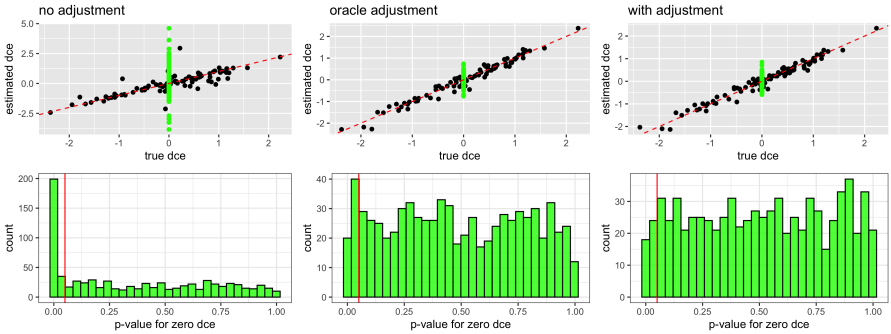


FIGURE 2.11: The performance of the *dce* without latent confounding adjustment (left), *dce* using true values of condounders (not known in practice) and *dce* with the latent confounding adjustment. Null DCEs are denoted in green, whereas the non-zero DCEs are denoted in black. This figure uses synthetic data with 300 genes, 300 observations and 3 latent confounders. Red line in the bottom row indicates the 0.05 threshold. The performance with the deconfounding step is close to the performance if we actually observed the latent confounders. Furthermore, it avoids increased number of falsely significant findings due to confounding bias (bottom row).

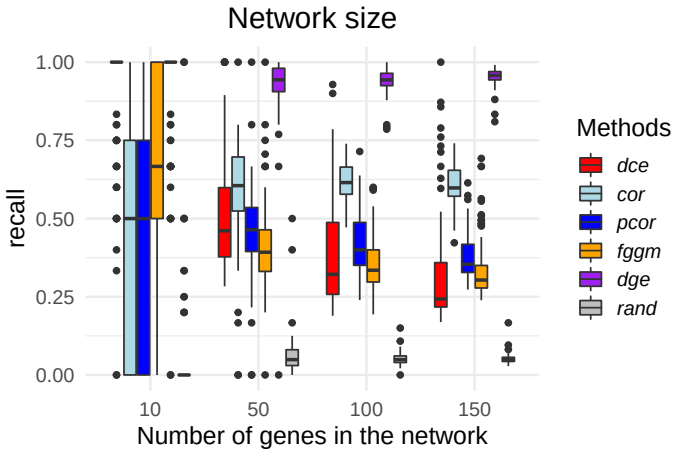


FIGURE 2.12: *dce* is compared to its competitors over 100 synthetic data sets over different network sizes. The accuracy measure is recall with a p-value cutoff of 0.05.

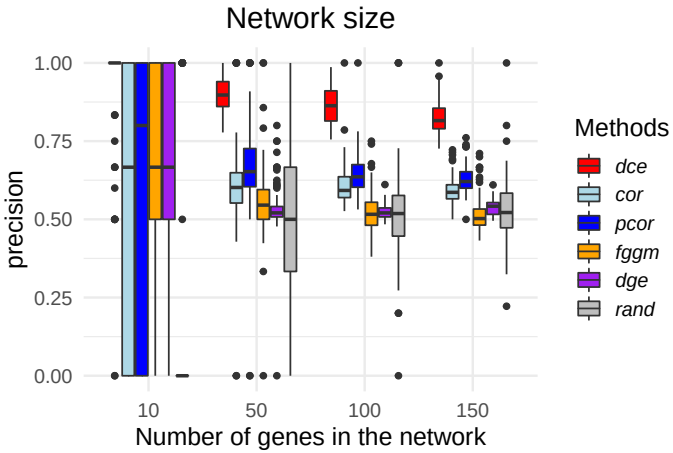


FIGURE 2.13: *dce* is compared to its competitors over 100 synthetic data sets over different network sizes. The accuracy measure is precision with a p-value cutoff of 0.05.

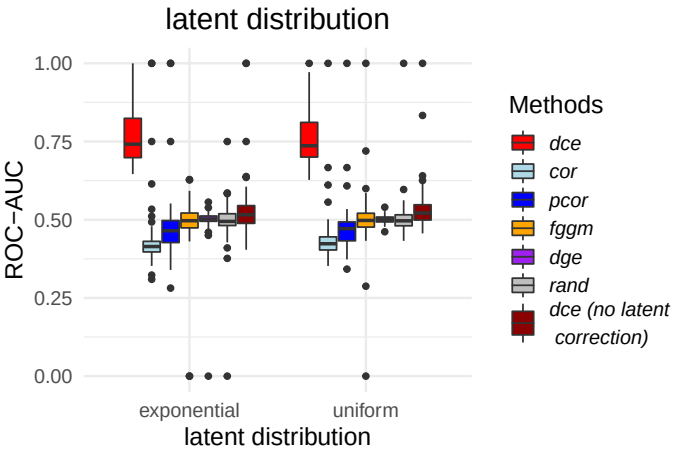


FIGURE 2.14: *dce* is compared to its competitors over 100 synthetic data sets with different distributions for effects of latent confounders. Performance does not change if we sample the effects from an exponential instead of a uniform distribution. Hence, we can account for latent confounders whether they affect all genes in the network uniformly or just some genes very strongly (exponential).

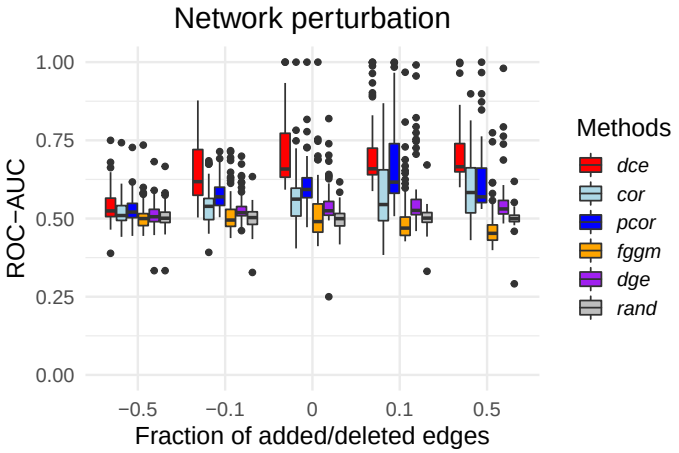


FIGURE 2.15: *dce* is compared to its competitors over 100 synthetic data sets with incorrect network information. Performance decreases for networks with missing edges, but stays robust, if additional edges are included. The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

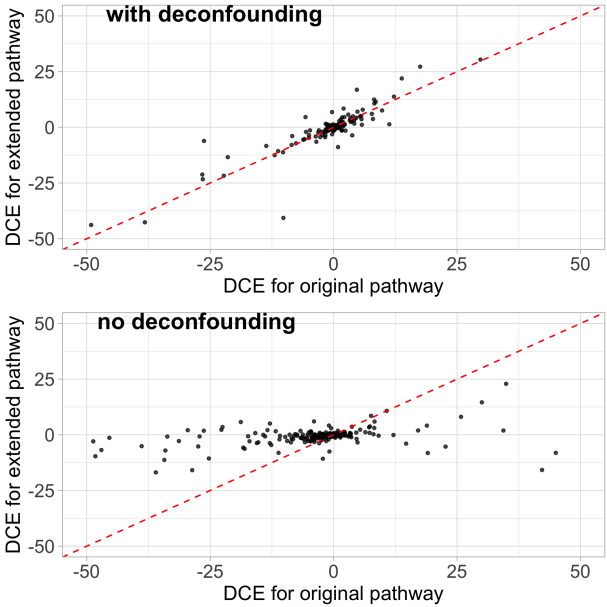


FIGURE 2.16: Comparison of estimated DCEs between the lung and mammary gland tissues for the original and the extended pathway, computed by the *dce* method with and without confounding adjustment. Correlation of the DCE estimates with confounding adjustment equals 0.833, significantly better thant 0.407 without any latent confounding adjustment.

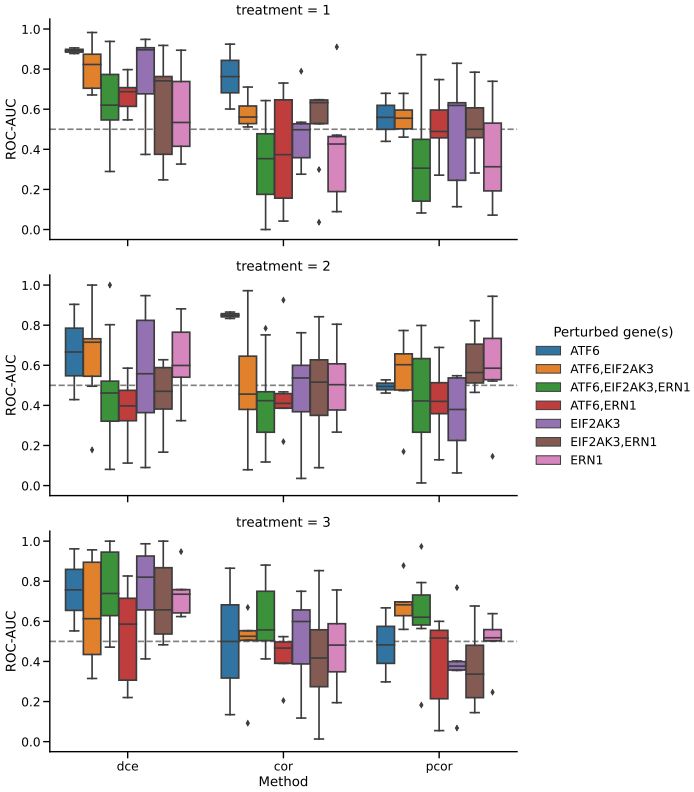


FIGURE 2.17: Summary of the performance of the *dce*, *cor* and *pcor* methods in the form of ROC-AUCs for the recovery of the knocked-out genes in all considered pathways. The performance measure has been stratified by perturbed gene and treatment (1: DMSO control, 2: tunicamycin, 3: thapsigargin). The whiskers of the boxplot correspond to the minimum and maximum of the data, the box denotes the first and third quartiles and the horizontal line within the box describes the median.

3

COHERENT PATHWAY ENRICHMENT ESTIMATION BY MODELING INTER-PATHWAY DEPENDENCIES USING REGULARIZED REGRESSION

Gene set enrichment methods are a common tool to improve the interpretability of gene lists as obtained, for example, from differential gene expression analyses. They are based on computing whether dysregulated genes are located in certain biological pathways more often than expected by chance. Gene set enrichment tools rely on pre-existing pathway databases such as KEGG, Reactome, or the Gene Ontology. These databases are increasing in size and in the number of redundancies between pathways, which complicates the statistical enrichment computation.

The following paper addresses this problem and develops a novel gene set enrichment method, called *pareg*, which is based on a regularized generalized linear model and directly incorporates dependencies between gene sets related to certain biological functions, for example, due to shared genes, in the enrichment computation. *pareg* is more robust to noise than competing methods. Additionally, the ability to recover known pathways as well as to suggest novel treatment targets in an exploratory analysis using breast cancer samples from TCGA is demonstrated. *pareg* is freely available as an R package on Bioconductor (`https://bioconductor.org/packages/release/bioc/html/pareg.html`) as well as on `https://github.com/cbg-ethz/pareg`. The GitHub repository also contains the Snakemake workflows needed to reproduce all results presented here.

The author's contributions to the following manuscript were the development of the statistical model, the implementation of the software package, as well as the synthetic performance evaluation and exploratory analysis. The paper is published as [59].

Coherent pathway enrichment estimation by modeling inter-pathway dependencies using regularized regression

Kim Philipp Jablonski[1,2] and Niko Beerenwinkel[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, 4058, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Basel, 4058, Switzerland
[*]To whom correspondence should be addressed

INTRODUCTION

The behavior of cells is governed by a complex interplay of molecules. Their functional dynamics are organized according to biological pathways [1]. Perturbations of pathways have been linked to certain diseases, such as, for example, cancer [2, 3]. Biological pathways can be obtained from pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Gene Ontology (GO), or Reactome [4–6]. It is important to note that pathways typically impose a structure of interactions in the form of a network on its contained molecules. While the nodes of this network typically correspond to genes, the edges correspond to interactions, such as signal transductions [7]. Another way of grouping genes in a meaningful way is to forgo the structure requirement and simply consider, for example, functionally related genes to be part of the same gene set.

Experiments investigating, for instance, differentially expressed genes between several conditions (e.g., wild-type versus mutant cell cultures) often produce a long list of genes of interest which is difficult to interpret [8, 9]. A common method for aggregating these lists of potentially interesting genes is to assess whether the genes preferentially appear in biologically relevant pathways. This reduces the amount of information which needs to be interpreted from individual genes to groups of genes, i.e., pathways, following a similar function.

There are several approaches to computing whether certain genes preferentially appear in certain gene sets. They can be roughly divided into the three groups: (a) singular enrichment analysis, (b) gene set enrichment analysis, and (c) modular enrichment analysis [10]. In a singular enrichment analysis, a list of genes resulting from a differential expression analysis is first partitioned into differentially expressed and not differentially expressed genes based on a threshold typically applied to effect size or p-value. These two groups of genes are then used to compute a pathway enrichment score individually. The gene set enrichment analysis lifts the requirement of a pre-selection of genes and considers all input genes without partitioning them into groups based on a threshold. Finally, the modular enrichment analysis computes the enrichment of each gene set not in isolation but rather by incorporating term-term relations into the statistical model. A term is a set of genes which are all involved in the same biological process and are thus functionally related. These term-term relations represent dependencies between gene sets, which can arise, for example, due to shared genes. This approach has the advantage of not requiring arbitrary thresholds to prepare

the input genes and is able to incorporate additional biological knowledge into the enrichment computation by imposing a structure on the gene set database. This additional biological knowledge can help maintain high statistical power in large, redundant gene set databases or structure the final visual presentation of enrichment scores [10].

One of the most basic approaches to compute singular enrichments is to use Fisher's exact test which is based on the hypergeometric distribution and requires a stratification of the input gene set [11]. There have been many extensions to this initial approach, including threshold-free methods such as the popular tool GSEA [12] which does not require an a priori stratification of the input and LRPath which formulates the enrichment computation as a regression [13].

Various methods have been proposed which follow the modular enrichment approach. topGO [14] is tailored to the tree structure of the gene sets provided by the Gene Ontology resource and removes local dependencies between GO terms which leads to better performance. By relying on the topology of a tree, it is not applicable to many other gene set sources. Another approach is to reduce the number of pathways which are included in the enrichment computation by removing redundant terms based on the notion of semantic similarity [15, 16]. RedundancyMiner [17] transforms the GO database prior to the enrichment computation by de-replicating redundant GO categories and thus tries to reduce the amount of noise introduced by overlapping pathways appearing in the enrichment analysis. These approaches rely on the directed acyclic graph structure of GO terms and cannot be generalized to other pathway databases. GENECODIS [18] incorporates relations between pathways into the enrichment computation by testing for the enrichment of co-occurring pathways. It can in principle be applied to any pathway database but is only available as a web-based tool and can thus not be easily used in automated workflows. The same limitation applies to ProfCom [19] which computes the enrichment of unions, intersections, and differences of pathways. In addition, it uses a greedy heuristic which does not guarantee to find an optimal solution for each case. MGSA [20] embeds all pathways in a Bayesian network and identifies enriched pathways using probabilistic inference. It does however not allow to explicitly model pathway relations. Finally, tools such as EnrichmentMap [21], ClueGO [22], REVIGO [23] and GOrilla [24] compute a singular enrichment score per pathway and subsequently visualize the result as a network of gene set clusters based on gene overlaps. This approach can be

applied to any gene set database but loses statistical power by executing the enrichment analysis and term-term relation inclusion in separate steps.

While many methods exist which try to overcome the issue of large redundant pathway databases, none of them, to the best of our knowledge, has accomplished this goal in a simultaneously database-agnostic, flexible and robust way. By not relying on the hierarchical structure of the Gene Ontology it is possible to create a method which is less restricted and can be used with other pathway databases that are more specialized to the experiment at hand. As there are various approaches to comparing pathways with each other, it is desirable for the enrichment algorithm to not be hard-coded to use a single specific pathway similarity measure but allow different ones based on the needs of the respective research question. The noise inherent to biological experiments leads to measurements of differential gene expression which can deviate from the underlying true differences. Robustness to the level of noise of the input data is thus a crucial property of pathway enrichment methods.

Here, we introduce a novel method called *pareg* for computing pathway enrichments which is based on regularized regression. It follows the ideas of GSEA as it requires no stratification of the input gene list, of MGSA as it incorporates term-term relations in a database-agnostic way, and of LRPath as it makes use of the flexibility of the regression approach. By regressing the differential expression p-values of genes on their membership to multiple gene sets while using LASSO and gene set similarity-based regularization terms, we require no prior thresholding and incorporate term-term relations into the enrichment computation. We show in a synthetic benchmark that this model is more robust to noise than competing methods, and demonstrate in an application to real data from The Cancer Genome Atlas (TCGA) [25] that it is able to recover known pathway associations as well as suggest novel ones.

METHODS

OVERVIEW. The input to *pareg* consists of (a) a list of genes, where each gene is associated to a single p-value obtained from a differential expression experiment and (b) a gene set database where a gene can be part of multiple gene sets simultaneously. *pareg*'s approach is general enough to support any kind of experimental value associated to the input genes. Pathway enrichments are then computed by regressing the differential expression p-value vector of input genes on a binary matrix indicating gene membership

to each gene set in the input database. The estimated coefficient vector captures the degree of association which gene sets have with p-values of differentially expressed genes; they can thus be regarded as an enrichment score. To induce sparsity in the coefficient vector and thus in the selected set of enriched pathways, we use the least absolute shrinkage and selection operator (LASSO) regularization term [26]. Term-term relations are included in the model using a network fusion penalty [27, 28].

REGRESSION APPROACH. We use a regularized multiple linear regression model to estimate gene set enrichment scores. Suppose we want to compute the enrichment of K pathways using N genes. Each gene g_i is associated with a p-value p_i from a differential expression analysis for $i = 1, \ldots, N$. We then define the response vector $\mathbf{Y}$ to be

$$\mathbf{Y} = (p_1, \ldots, p_N)^T \tag{3.1}$$

The binary regressor matrix $\mathbf{X}$ captures the membership information of each gene g_i, $i = 1, \ldots, N$, in pathway t_j, $j = 1, \ldots, K$,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} \tag{3.2}$$

with

$$x_{ij} = \begin{cases} 1 & \text{if gene } i \text{ is in pathway } j \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

In the resulting linear model $\mathbf{Y} = \mathbf{X}\beta$, the vector of coefficients $\beta = (\beta_1, \ldots, \beta_K)^T$ is estimated using stochastic gradient descent to minimize the objective function

$$\begin{aligned} \widehat{\beta} = \underset{\beta,\phi}{\arg\min} \big(& -\log(\mathcal{L}(\beta, \phi | \mathbf{Y}, \mathbf{X})) + \lambda \|\beta\|_1 \\ & + \psi \sum_{i=1}^{K} \sum_{j=1}^{K} \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\right\|_2^2 g_{ij} \big) \end{aligned} \tag{3.4}$$

where $\mathcal{L}(\beta, \phi | \mathbf{Y}, \mathbf{X})$ is the likelihood and $\mathbf{G} = (g_{ij})_{ij} \in (0,1)^{K \times K}$ a pathway similarity matrix, where g_{ij} describes the similarity between pathway i and j.

To model the p-values in the response vector, the likelihood is defined using the beta distribution [29]

$$\mathcal{L}(\beta, \phi | \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{N} \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \mathbf{Y}_i^{p-1} (1 - \mathbf{Y}_i)^{q-1} \right] \tag{3.5}$$

where $p = \mu\phi$ and $q = (1-\mu)\phi$ with mean $0 < \mu < 1$, precision parameter $\phi > 0$ and Gamma function $\Gamma(\cdot)$. The mean is then modeled as $g(\mu) = \mathbf{X}\beta$ where $g(\cdot)$ is a link function [30].

The optimal values for the regularization parameters λ (LASSO) and ψ (network fusion) are determined using cross-validation [28], which balances the effects of the LASSO and network fusion terms. The former term induces a sparse coefficient vector, i.e., it reduces the number of enriched pathways needed to explain the observed data. The latter term promotes assigning a similar enrichment score to (functionally) similar pathways.

PATHWAY SIMILARITY MEASURES. The goal of adding pathway similarities to the model is to group pathways in the enrichment computation. By doing so, redundant sets of functionally related pathways jointly drive the enrichment signal and reduce the influence of noisy measurements. Due to the flexibility of our model, this can be any similarity measure which can be stored as a real matrix.

As pathways are typically defined as lists of genes, the Jaccard similarity and overlap coefficients are common choices [21]. They group pathways which share many genes together and are thus a good measure of functional relation [31]. The overlap coefficient is particularly suited for pathway collections which feature a hierarchical structure.

In addition, when using the popular Gene Ontology [5] as a pathway database, semantic similarity measures exist. These measures incorporate the topological structure of the Gene Ontology and are better at inferring functional relations between pathways [32–34].

PRESENTATION OF ENRICHMENT RESULTS. The estimated coefficient vector β can be ordered descendingly by absolute value such that the most dysregulated and thus interesting pathways appear at the top of the list. A regression coefficient β_j of large absolute value corresponds to a strong dysregulation of pathway j.

In addition, we implement a network-based visualization of the enrichment result. Each node in this network corresponds to a pathway, and edges correspond to high pathway similarities. The nodes are colored by

the respective enrichment score of each pathway. This allows for the quick identification of functional modules as network clusters.

Finally, the result of *pareg* can be transformed to a format readily understood by the functional enrichment visualization R package enrichplot [35]. This enables the usage of many plotting functions, such as dot plots, tree plots and UpSet plots, as well as immediate access to newly implemented ones.

GENERATION OF SYNTHETIC DATA. The goal of the synthetic benchmark is to create a known set of dysregulated pathways which induces a set of differentially expressed genes, apply several enrichment methods (listed below) to this data set and evaluate how well each method is able to recover the initially dysregulated pathways. Thus, each synthetic data set consists of a list of genes with associated p-values obtained from a simulated differential expression experiment, as well as a respective ground truth set of pathways.

Given an existing term database $D = \{T_1, \ldots, T_K\}$ consisting of K terms $T_j = \{g_1, \ldots, g_{L_j}\}$, each made up of L_j genes g_i, we randomly sample a ground truth set of activated terms $D_A \subset D$. In order to model the joint activation of functionally related pathways, we apply a similarity sampling approach. Given a similarity matrix S with $0 \leq s_{ij} \leq 1$ and similarity factor $0 \leq \rho \leq 1$ we first uniformly sample a single term j. The next term is then drawn according to the probability vector $(1-\rho)U + \rho S_j$ where S_j is column j of S and denotes the similarity of term j to all other terms, and U is a vector of length $|S_j|$ with values $\frac{1}{|S_j|}$. This procedure is continued by setting j to the previously sampled term and repeated until the required number of terms has been sampled. For ρ close to 1 this results in similar pathways being sampled while ρ close to 0 leads to a uniformly random sample.

Next, we model synthetic differential expression p-values for the N genes $(g_1, \ldots, g_N)$ by sampling from a Beta distribution whose parameters are determined from a linear combination of a noisy gene-term membership matrix and a term activation vector. This mimics the real life setting where the dysregulation of a pathway is jointly driven by the dysregulated genes it contains.

In particular, we create the activation vector $\beta_A = (b_1, \ldots, b_K)^T$ with

$$b_k \sim \begin{cases} -1 & \text{if } T_k \in D_A \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

That is, we assign a non-zero coefficient to activated pathways. The gene-term membership matrix $\mathbf{X}_A$ is defined analogously to eqs. (3.2) and (3.3). To model the effect of noisy measurements, we remove the association between genes and activated terms in $\mathbf{X}_A$ by setting a fraction of η entries to 0. Next, we compute $\mu = g^{-1}(\mathbf{X}_A \beta_A)$ where g^{-1} is the logistic function and set $\phi = 1$ to parametrize the Beta distribution. To create the final synthetic data set $E = (D_A, \{(g_i, p_i), \ldots, (g_N, p_N)\})$, we sample the differential expression p-value p_i for gene i from $\mathcal{B}(\mu_i, \phi)$.

We run 20 replicates with 20 activated terms each and use all pathways with sizes between 50 and 500 in the biological process subtree of the Gene Ontology.

PERFORMANCE EVALUATION IN SYNTHETIC BENCHMARK. Due to the strong class imbalance in the experimental setup of pathway enrichments featuring few positives, i.e., dysregulated pathways, compared to the number of negatives, i.e., unaffected pathways, we use precision-recall (PR) curves to evaluate the performance of each pathway enrichment method [36, 37].

A term T_j is classified as a true positive (TP) if it is in D_A and is enriched according to a method and respective threshold. It is classified as a false positive (FP) if it is not a member of D_A but is estimated to be enriched. Analogously, a true negative (TN) is a term which is not in D_A and is not enriched, while a false negative (FN) is a term which is in D_A but is not detected by a method. Precision is then defined as $TP/(TP + FP)$ and recall as $TP/(TP + FN)$. By varying the threshold used to create the classifications we can then readily create PR curves. To obtain a numeric summary of a method's performance, we compute the area under the precision-recall curve (AUC).

REAL DATA APPLICATION. We conduct an exploratory analysis using cancer and normal samples from processed TCGA data available in the Gene Expression Omnibus entry GSE62944 [38]. We retrieve 113 tumor and matched normal samples for TCGA-BRCA (Breast Invasive Carcinoma). We then use limma [39] to run a differential gene expression analysis to compare tumor and normal samples. The obtained p-values and pathways from the biological process subtree of the Gene Ontology are then used as input to *pareg*. We use the Jaccard similarity to create a similarity matrix for all considered pathways. As in the synthetic benchmark, we use all

pathways with sizes between 50 and 500 in the biological process subtree of the Gene Ontology.

RESULTS

First, we compare the performance of *pareg* to competing methods using a synthetic benchmark study. Second, we conduct an exploratory analysis using a breast cancer data set from TCGA.

Synthetic benchmark

We compare the performance of *pareg* to other enrichment tools using a synthetic data set where the ground truth is known. To do so, we select a set of activated terms and generate differential gene expression p-values using a linear model. We vary the level of noise η used when generating synthetic data in order to simulate different real life situations where noise can arise from measurement errors (fig. 3.1).

In addition to *pareg*, we benchmark four other methods. MGSA is a Bayesian approach which embeds pathways in a Bayesian network and explicitly models the activation of sets of pathways [20]. It constitutes a modular enrichment method of competitive performance to *pareg* which does not depend on a particular pathway database. Fisher's exact test (FET) is a classical single-term enrichment method which is still commonly used and serves as a simple alternative in the comparison [11]. topGO's elim algorithm incorporates the GO tree structure into the enrichment computation and is a modular enrichment method which relies on using the Gene Ontology [14]. The null model serves as the baseline indicating how random guessing would perform. It assigns a random enrichment p-value between 0 and 1 to each pathway.

We observe that *pareg* consistently outperforms all competing methods over a wide range of parameter values (fig. 3.1). For varying levels of noise $\eta = 0, 0.25, 0.5$ and similarity factor $\rho = 0, 0.5, 1$, *pareg* achieves the highest mean areas under the precision-recall curve (PR-AUC) in all cases (figs. 3.1a and 3.1c). *pareg* clearly outperforms the singular enrichment method FET, which emphasizes that the proposed method of including term-term relations in the enrichment computation yields an advantage when working with large and redundant pathway databases. Out of all other benchmarked methods, MGSA performs closest to *pareg* indicating that its Bayesian model-based approach which explicitly handles term-term

relations in a database-agnostic way is to some extent able to deal with the clustered pathway database. topGO performs slightly worse than FET. It explicitly uses the GO tree structure and performs successive enrichment tests which are individually similar to FET. This approach is not able to appropriately process the clustering structure assumed in the synthetic benchmark which is not based on a tree.

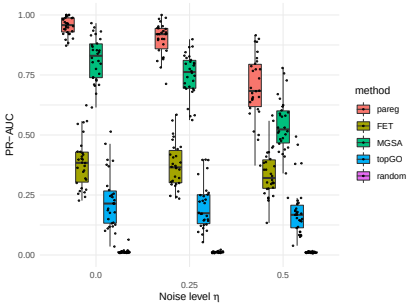
When increasing the noise level η, we observe that FET and topGO show a smaller decrease in performance than *pareg* and MGSA (fig. 3.1a). This is in line with the observation that the precision of FET and topGO remains nearly constant when fixing the recall (fig. 3.1b). For example, at a recall of 80% *pareg* has a median precision of 94% for $\eta = 0.25$ while MGSA has a median precision of 37%. FET and topGO have median precision values of 12% and 5% respectively. For *pareg* and MGSA, most PR-AUC is lost for large values of recall where FET and topGO show poor performance even for small η.

When increasing the similarity factor ρ, we see that *pareg* remains at roughly the same PR-AUC (fig. 3.1c) and only slightly decreases in precision at a fixed recall level (fig. 3.1d), while MGSA shows a stronger decrease in performance. For example, fixing recall to 80% at $\rho = 0.5$ yields a median precision of 94% for *pareg*. MGSA, FET and topGO have median precision values of 29%, 12% and 5% respectively. This indicates that *pareg* is better able to deal with varying levels of clustering in the set of dysregulated pathways. topGO exhibits a slight decline in performance as its tree-based approach is not able to handle the clustering structure induced by the Jaccard similarity measure. As FET does not incorporate term-term relations into the enrichment computation, we observe no dependence on ρ.

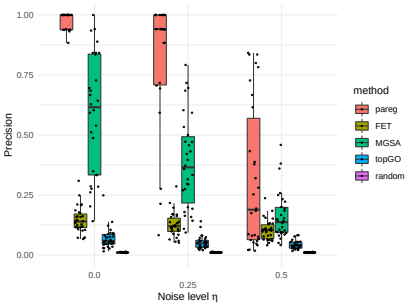
Exploratory analysis of breast cancer samples

To investigate the behavior of *pareg* on real data, we use it to run a pathway enrichment analysis on breast cancer (BRCA) samples from TCGA with terms from the Gene Ontology biological process subtree. We order the terms by their absolute enrichment level and list the top 25 results in fig. 3.2a as well as visualize the top 50 non-isolated results in a network (fig. 3.2b).

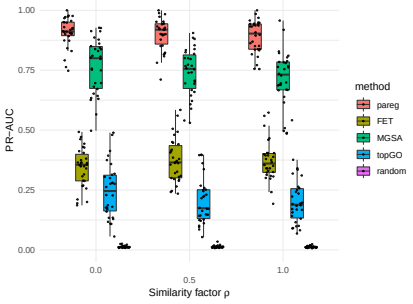
The largest cluster of the network visualization is made up of 8 nodes and features terms related to cell migration such as ameboidal-type cell migration and actin filament organization. It has been recognized that cancer cells can use amoeboid migration as their preferred migratory strategy [40]. In particular, it has been shown that treatment via endocrine therapy



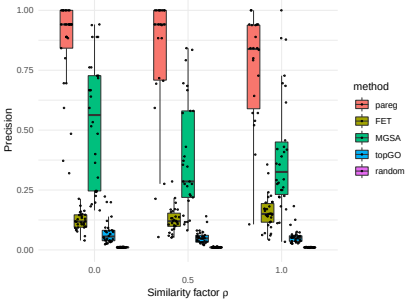
(a) Boxplots of precision-recall areas-under-the-curve (PR-AUC) for varying noise level η. Individual PR curves are given in figs. 3.3 to 3.5.



(b) Boxplots of precision values obtained when setting recall to 0.8 in figs. 3.3 to 3.5 for varying noise level η.

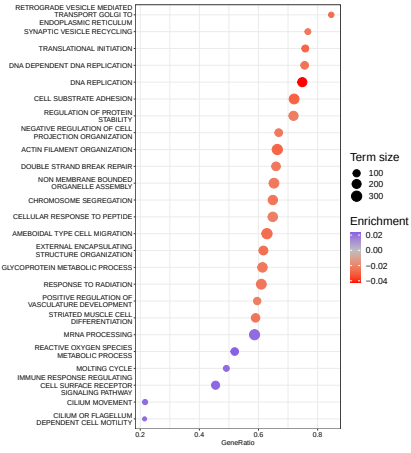


(c) Boxplots of precision-recall areas-under-the-curve (PR-AUC) for varying similarity factor ρ. Individual PR curves are given in figs. 3.6 to 3.8.

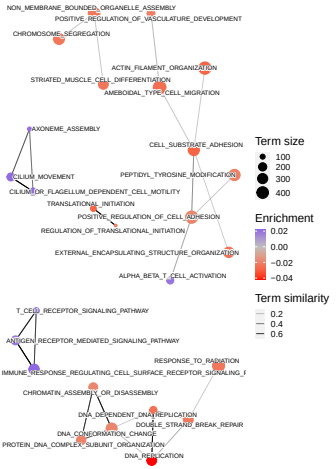


(d) Boxplots of precision values obtained when setting recall to 0.8 in figs. 3.6 to 3.8 for varying similarity factor ρ.

FIGURE 3.1: Summary of performance measures calculated for synthetic benchmark. Each point correspond to a single replicate.



(a) Top 25 terms ordered by absolute enrichment. The y-axis lists the terms while the x-axis denotes the fraction of significantly differentially expressed genes (p-value < 0.05) over the respective term size. The term size is the number of genes making up the term and also represented as the size of each circle. The color of each circle indicates the enrichment of the respective term where blue corresponds to positive and red to negative enrichment.



(b) Non-isolated terms of the 50 terms with largest absolute enrichment. Nodes correspond to terms and edges to Jaccard similarities greater than 0.1. The node color and size has the same meaning as in fig. 3.2a. The higher the opacity of an edge the larger the corresponding term similarity.

FIGURE 3.2: Summary of term enrichment results obtained for TCGA breast cancer samples (normal versus tumor) and the biological process subtree of the Gene Ontology.

inhibits this kind of migration in breast cancer. Furthermore, it has been shown that the organization of actin stree fibers promotes proliferation of pre-invasive breast cancer cells [41]. The dysregulation of cell adhesion dynamics has also been investigated in the literature [42] and is captured by the enrichment of the cell-substrate adhesion and positive regulation of cell adhesion terms. In addition, the peptidyl-tyrosine modification term is enriched. Tyrosine acts as a key player in the initiation of proteins to focal adhesion sites. Apart from this, the influence of tyrosine phosphatases for many different cancer types [43] and of tyrosine kinases specifically for breast cancer [44] has been recognized.

The second largest cluster made up of 7 nodes is thematically related to DNA replication and conformation changes. These processes are of high relevance to cancers in general [45] as well as breast cancer specifically [46]. Furthermore, the importance of double-strand break repair has been captured by the enrichment of the corresponding term [47].

A few smaller clusters remain. One cluster of three nodes contains the terms chromosome segregation, non-membrane-bounded organelle assembly and striated muscle cell differentiation. The importance of chromosomal stability and the impact of proteins which modulate it have been highlighted for breast cancer [48]. Furthermore, it has been observed that breast cancer cells exhibit non-random chromosome segregation [49]. In addition, striated muscle cell differentiation has been linked to the metastatic potential of breast cancer cells [50]. Another cluster of three nodes contains the terms cilium movement, cilium or flagellum-dependent cell motility and axoneme assembly. It has been shown that the expression of cilia is downregulated in various types of cancer, including breast cancer [51]. It furthermore has impact on the regulation of cancer development [52]. The related enrichment of the axoneme assembly terms suggests the importance of the assembly and organization of an axoneme. This constitutes a novel finding and suggests further experimental investigations. The last cluster with three nodes contains the terms T-cell receptor signaling pathway, antigen receptor-mediated signaling pathway and immune response-regulating cell surface receptor signaling pathway. Both the relevance of the T-cell receptor signaling [53] and immune response-regulating cell surface receptor signaling term [54] has been recognized. The possibility of investigating the antigen receptor-mediated signaling pathway for a Chimeric antigen receptor T cell therapy has very recently been considered [55]. Finally, the two node cluster contains the terms translational initiation and regulation of translational initiation. The regulation of translation via changed expression

of the eukaryotic translation initiation factor 3 has been observed to play a positive role in breast cancer progression [56].

In addition to the network clusters, we also detect individually enriched pathways (fig. 3.2a). We find the retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum term to be enriched. The potential implications of this apparatus have already been discussed [57], but have, to the best of our knowledge, not been linked to breast cancer specifically. The synaptic vesicle recycling term is also enriched. Its potential as a therapeutic target has been recognized [58], however not in the context of breast cancer. In both cases, our results suggest the novel finding that these pathways may be especially relevant to breast cancer and that further experimental validations in that direction would be interesting.

We also demonstrate the effectiveness of network regularization by comparing the enrichments to results obtained from running *pareg* without the network regularization term (fig. 3.9) and from FET (fig. 3.10). In both cases, much fewer clusters are observed, making the biological interpretation more difficult. This indicates that employing the network regularization term is useful for better understanding of the enrichment results.

DISCUSSION

We have developed a novel pathway enrichment method called *pareg* which is based on a regularized generalized linear model. It makes use of LASSO and network fusion penalty terms to produce a sparse and coherent list of enriched pathways. The network fusion term incorporates a pathway similarity network which models functional relations between pathways and clusters pathways as part of the enrichment computation in order to handle large and redundant pathway databases.

In a synthetic benchmark, we show that *pareg* is able to outperform single-term enrichment methods such as Fisher's exact test, a popular tool explicitly including the GO tree in its calculations as well as a model-based approach which embeds pathways in a Bayesian network.

In an exploratory analysis with breast cancer samples, we are able to recover many relevant pathways already known in literature, as well as suggest novel ones which pose interesting future targets for experimental validation.

We note that *pareg* assumes that a linear combination of gene-pathway memberships is driving the overall pathway dysregulation, an assumption which may reduce the algorithm's applicability in certain biological environ-

ments, such as, for example, the interactions between genes in myocardial infarction as measured by mRNA expression profiles [59].

Due to the flexibility of the regression approach, potential future work could go in many directions. Instead of modeling the response variable using a Beta distribution, one may use a beta-uniform mixture which has been suggested for p-values [60]. As the network fusion penalty depends on a general similarity matrix, different measures could be explored. For example, there exist a wide range of different semantic similarity measures which have been used to relate GO terms [34, 61–65]. Alternatively, similarity measures which embed sets of genes in protein-protein interaction networks and compare their localization have been shown to be useful for predicting disease status; they could be another viable choice [31, 66].

Furthermore, the potential effects of other regularization terms are interesting. Using an Elastic-Net term instead of LASSO or stability selection [67] could improve the sparsity of the coefficient vector. Instead of the network fusion term, regularizations such as hierarchical feature regression [68], regularized k-means clustering [69] or group LASSO [70] can be used to incorporate term-term relations and may exhibit more desirable statistical properties, such as stronger robustness to noise, smaller sample size requirements and faster convergence of the optimizer. Due to these regularization terms, it is not immediately possible to compute confidence intervals for each entry of the estimated coefficient vector. The de-biased LASSO approach [71] can be explored to get a better understanding of the uncertainty involved in the enrichment computation.

Finally, while there have been programming language specific efforts to standardize gene set enrichment benchmarking workflows [72], no widely accepted consensus has been found. The benchmarking workflow we implement is written in the workflow management system Snakemake [73] and thus allows easy integration of additional tools as well as reproducible execution on different back ends. We thus hope that other enrichment tools can use a similar approach to enable comparative benchmarks of new methodologies.

DATA AVAILABILITY

The code used to construct the synthetic data sets is available as part of the R/Bioconductor software package *pareg*. The experimental data used in the exploratory analysis is available as GSE62944 on the Gene Expression

Omnibus. The pathway database has been obtained from the Gene Ontology resource.

CODE AVAILABILITY

The method *pareg* is freely available as an R package on Bioconductor (`https://bioconductor.org/packages/release/bioc/html/pareg.html`) as well as on `https://github.com/cbg-ethz/pareg`. The GitHub repository also contains the Snakemake [73] workflows needed to reproduce all results presented here.

AUTHOR INFORMATION

CORRESPONDING AUTHORS Correspondence to Niko Beerenwinkel (niko.beerenwinkel@bsse.ethz.ch).

ETHICS DECLARATIONS

COMPETING INTERESTS. The authors declare no competing interests.

BIBLIOGRAPHY

1. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annual review of cell and developmental biology* **26**, 721 (2010).
2. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100**, 57 (2000).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646 (2011).
4. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* **27**, 29 (1999).
5. Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258 (2004).
6. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic acids research* **33**, D428 (2005).

7. Steffen, M., Petti, A., Aach, J., D'haeseleer, P. & Church, G. Automated modelling of signal transduction networks. *BMC bioinformatics* **3**, 1 (2002).

8. Simillion, C., Liechti, R., Lischer, H. E., Ioannidis, V. & Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC bioinformatics* **18**, 1 (2017).

9. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. Gene set analysis: challenges, opportunities, and future research. *Frontiers in genetics*, 654 (2020).

10. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1 (2009).

11. Fisher, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* **85**, 87 (1922).

12. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545 (2005).

13. Sartor, M. A., Leikauf, G. D. & Medvedovic, M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* **25**, 211 (2009).

14. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600 (2006).

15. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. & Wang, S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976 (2010).

16. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

17. Zeeberg, B. R., Liu, H., Kahn, A. B., Ehler, M., Rajapakse, V. N., Bonner, R. F., Brown, J. D., Brooks, B. P., Larionov, V. L., Reinhold, W., *et al.* RedundancyMiner: De-replication of redundant GO categories in microarray and proteomics analysis. *BMC bioinformatics* **12**, 1 (2011).

18. Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M. & Pascual-Montano, A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology* **8**, 1 (2007).

19. Antonov, A. V., Schmidt, T., Wang, Y. & Mewes, H. W. ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic acids research* **36**, W347 (2008).

20. Bauer, S., Gagneur, J. & Robinson, P. N. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research* **38**, 3523 (2010).

21. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one* **5**, e13984 (2010).

22. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z. & Galon, J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091 (2009).

23. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* **6**, e21800 (2011).

24. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 1 (2009).

25. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).

26. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267 (1996).

27. Cheng, W., Zhang, X., Guo, Z., Shi, Y. & Wang, W. Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics* **30**, i139 (2014).

28. Dirmeier, S., Fuchs, C., Mueller, N. S. & Theis, F. J. netReg: network-regularized linear models for biological association studies. *Bioinformatics* **34**, 896 (2018).

29. Ferrari, S. & Cribari-Neto, F. Beta regression for modelling rates and proportions. *Journal of applied statistics* **31**, 799 (2004).

30. Cribari-Neto, F. & Zeileis, A. Beta regression in R. *Journal of statistical software* **34**, 1 (2010).

31. Bass, J. I. F., Diallo, A., Nelson, J., Soto, J. M., Myers, C. L. & Walhout, A. J. Using networks to measure similarity between genes: association index selection. *Nature methods* **10**, 1169 (2013).

32. Guo, X., Liu, R., Shriver, C. D., Hu, H. & Liebman, M. N. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics* **22**, 967 (2006).

33. Ehsani, R. & Drabløs, F. TopoICSim: a new semantic similarity measure based on gene ontology. *BMC bioinformatics* **17**, 1 (2016).

34. Zhao, C. & Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific reports* **8**, 1 (2018).

35. Yu, G. enrichplot: Visualization of Functional Enrichment Result. *R package version* **1** (2022).

36. Davis, J. & Goadrich, M. *The relationship between Precision-Recall and ROC curves* in *Proceedings of the 23rd international conference on Machine learning* (2006), 233.

37. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**, e0118432 (2015).

38. Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H. & Piccolo, S. R. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31**, 3666 (2015).

39. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).

40. Graziani, V., Rodriguez-Hernandez, I., Maiques, O. & Sanz-Moreno, V. The amoeboid state as part of the epithelial-to-mesenchymal transition programme. *Trends in cell biology* (2021).

41. Tavares, S., Vieira, A. F., Taubenberger, A. V., Araújo, M., Martins, N. P., Brás-Pereira, C., Polónia, A., Herbig, M., Barreto, C., Otto, O., *et al.* Actin stress fiber organization promotes cell stiffening and proliferation of pre-invasive breast cancer cells. *Nature communications* **8**, 1 (2017).

42. Maziveyi, M. & Alahari, S. K. Cell matrix adhesions in cancer: the proteins that form the glue. *Oncotarget* **8**, 48471 (2017).

43. Motiwala, T. & Jacob, S. T. Role of protein tyrosine phosphatases in cancer. *Progress in nucleic acid research and molecular biology* **81**, 297 (2006).

44. Biscardi, J. S., Ishizawar, R. C., Silva, C. M. & Parsons, S. J. Tyrosine kinase signalling in breast cancer: epidermal growth factor receptor and c-Src interactions in breast cancer. *Breast cancer research* **2**, 1 (2000).

45. Jia, R., Chai, P., Zhang, H. & Fan, X. Novel insights into chromosomal conformations in cancer. *Molecular cancer* **16**, 1 (2017).

46. Ghimire, H., Garlapati, C., Janssen, E. A., Krishnamurti, U., Qin, G., Aneja, R. & Perera, A. Protein conformational changes in breast cancer sera using infrared spectroscopic analysis. *Cancers* **12**, 1708 (2020).

47. Bau, D.-T., Mau, Y.-C., Ding, S.-l., Wu, P.-E. & Shen, C.-Y. DNA double-strand break repair capacity and risk of breast cancer. *Carcinogenesis* **28**, 1726 (2007).

48. Garcia, J. & Lizcano, F. KDM4C activity modulates cell proliferation and chromosome segregation in triple-negative breast cancer. *Breast cancer: basic and clinical research* **10**, BCBCR (2016).

49. Liu, W., Jeganathan, G., Amiri, S., Morgan, K. M., Ryan, B. M. & Pine, S. R. Asymmetric segregation of template DNA strands in basal-like human breast cancer cell lines. *Molecular Cancer* **12**, 1 (2013).

50. Nikulin, S., Zakharova, G., Poloznikov, A., Raigorodskaya, M., Wicklein, D., Schumacher, U., Nersisyan, S., Bergquist, J., Bakalkin, G., Astakhova, L., *et al.* Effect of the Expression of ELOVL5 and IGFBP6 genes on the metastatic potential of breast cancer cells. *Frontiers in genetics* **12**, 769 (2021).

51. Higgins, M., Obaidi, I. & McMorrow, T. Primary cilia and their role in cancer. *Oncology letters* **17**, 3041 (2019).

52. Fabbri, L., Bost, F. & Mazure, N. M. Primary cilium in cancer hallmarks. *International journal of molecular sciences* **20**, 1336 (2019).

53. Shah, K., Al-Haidari, A., Sun, J. & Kazi, J. U. T cell receptor (TCR) signaling in health and disease. *Signal transduction and targeted therapy* **6**, 1 (2021).

54. Rezaei-Tavirani, M., Zamanian-Azodi, M., Bashash, D., Ahmadi, N. & Rostami-Nejad, M. Breast cancer interaction network concept from mostly related components. *Galen Medical Journal* **8**, e1298 (2019).

55. Yang, Y.-H., Liu, J.-W., Lu, C. & Wei, J.-F. CAR-T Cell Therapy for Breast Cancer: From Basic Research to Clinical Application. *International Journal of Biological Sciences* **18**, 2609 (2022).

56. Grzmil, M., Rzymski, T., Milani, M., Harris, A., Capper, R., Saunders, N., Salhan, A., Ragoussis, J. & Norbury, C. An oncogenic role of eIF3e/INT6 in human breast cancer. *Oncogene* **29**, 4080 (2010).

57. Spang, A. Retrograde traffic from the Golgi to the endoplasmic reticulum. *Cold Spring Harbor perspectives in biology* **5**, a013391 (2013).

58. Li, Y. C. & Kavalali, E. T. Synaptic vesicle-recycling machinery components as potential therapeutic targets. *Pharmacological reviews* **69**, 141 (2017).

59. Hartmann, K., Seweryn, M., Handleman, S. K., Rempała, G. A. & Sadee, W. Non-linear interactions between candidate genes of myocardial infarction revealed in mRNA expression profiles. *BMC genomics* **17**, 1 (2016).

60. Pounds, S. & Morris, S. W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**, 1236 (2003).

61. Jiang, J. J. & Conrath, D. W. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008* (1997).

62. Lin, D. *et al. An information-theoretic definition of similarity.* in *Icml* **98** (1998), 296.

63. Resnik, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research* **11**, 95 (1999).

64. Schlicker, A., Domingues, F. S., Rahnenführer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics* **7**, 1 (2006).

65. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274 (2007).

66. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J. & Barabási, A.-L. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

67. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417 (2010).

68. Pfitzinger, J. Cluster Regularization via a Hierarchical Feature Regression. *arXiv preprint arXiv:2107.04831* (2021).

69. Sun, W., Wang, J. & Fang, Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics* **6**, 148 (2012).

70. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49 (2006).

71. Xia, L., Nan, B. & Li, Y. A revisit to de-biased lasso for generalized linear models. *arXiv preprint arXiv:2006.12778* (2020).

72. Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics* **22**, 545 (2021).

73. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).

SUPPLEMENTS

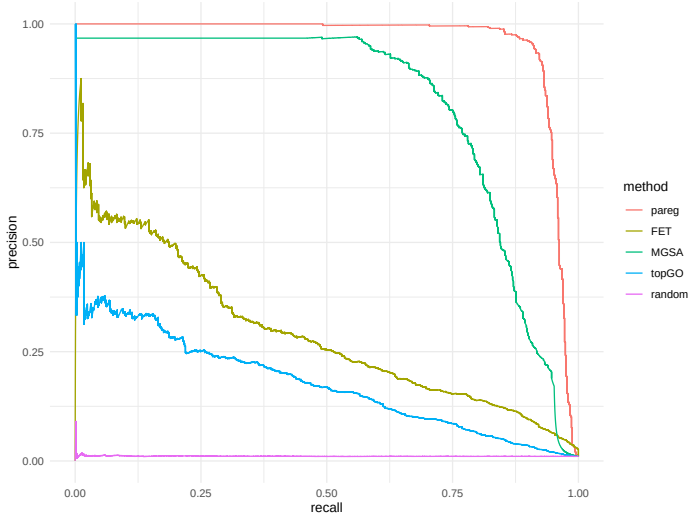


FIGURE 3.3: Precision-Recall (PR) curves aggregated over all replicates for noise level $\eta = 0$.

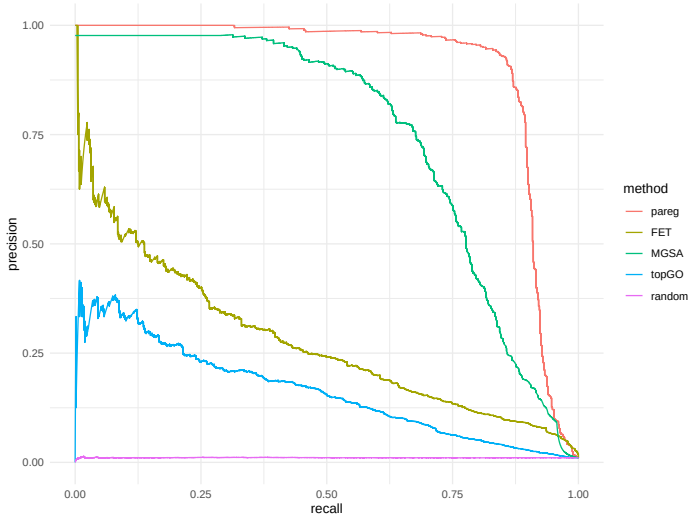


FIGURE 3.4: Precision-Recall (PR) curves aggregated over all replicates for noise level $\eta = 0.25$.

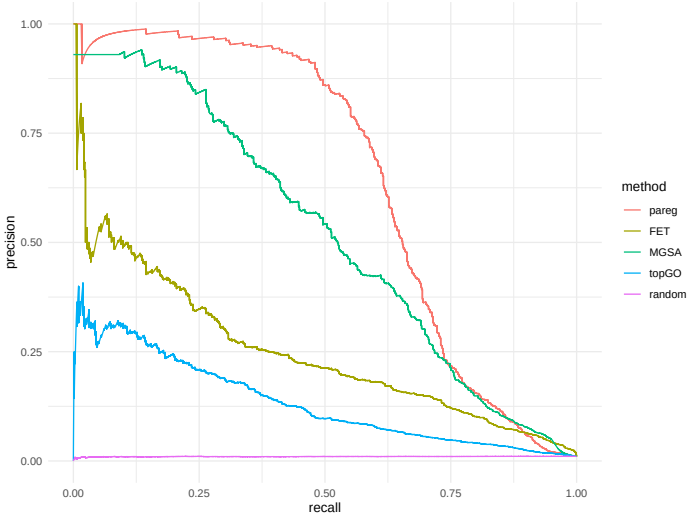


FIGURE 3.5: Precision-Recall (PR) curves aggregated over all replicates for noise level $\eta = 0.5$.

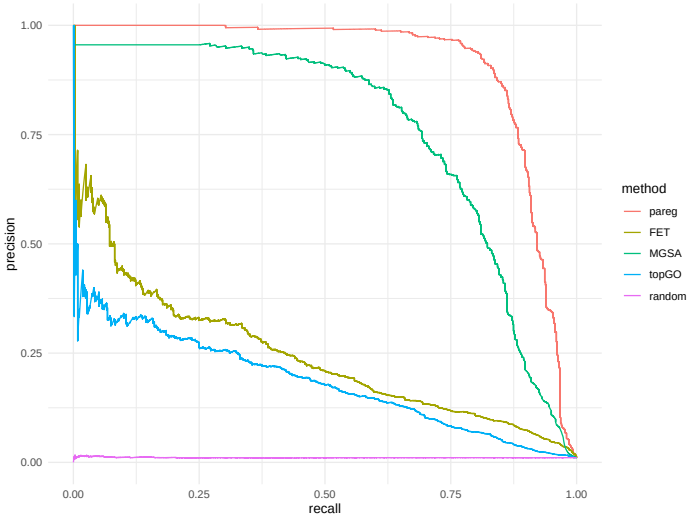


FIGURE 3.6: Precision-Recall (PR) curves aggregated over all replicates for similarity factor $\rho = 0$.

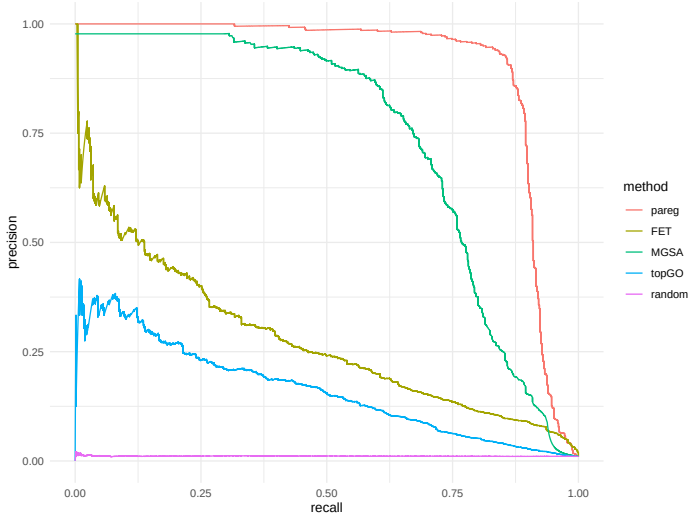


FIGURE 3.7: Precision-Recall (PR) curves aggregated over all replicates for similarity factor $\rho = 0.5$.

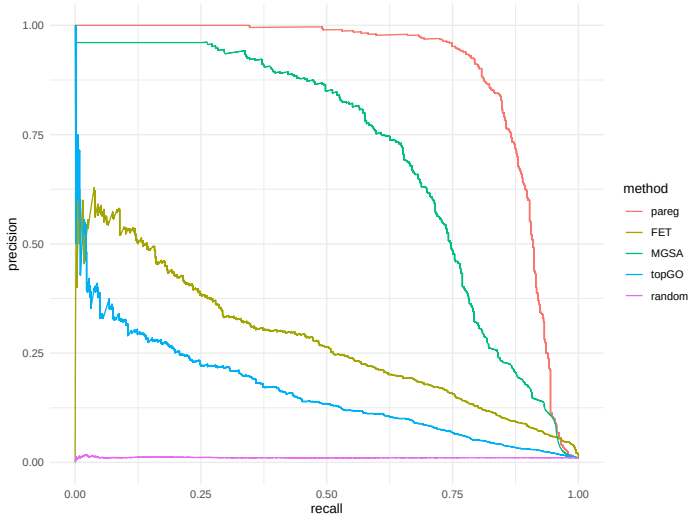


FIGURE 3.8: Precision-Recall (PR) curves aggregated over all replicates for similarity factor $\rho = 1$.

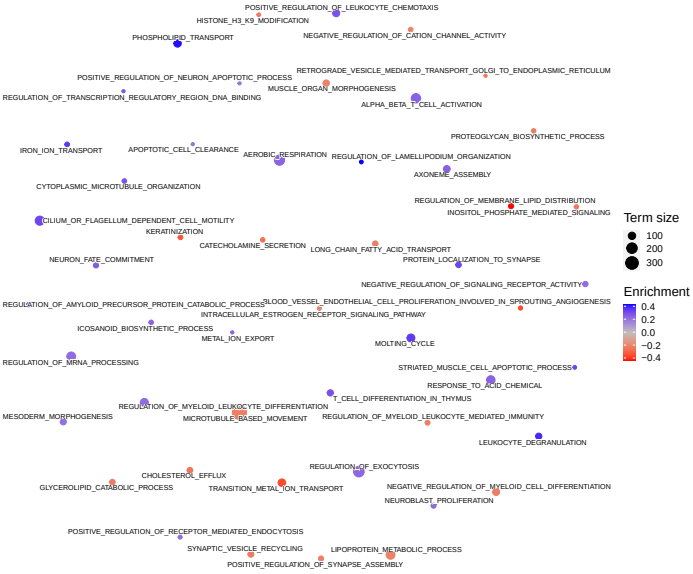


FIGURE 3.9: Term network for *pareg* without network regularization with same parameters as in fig. 3.2b except for also including isolated terms.

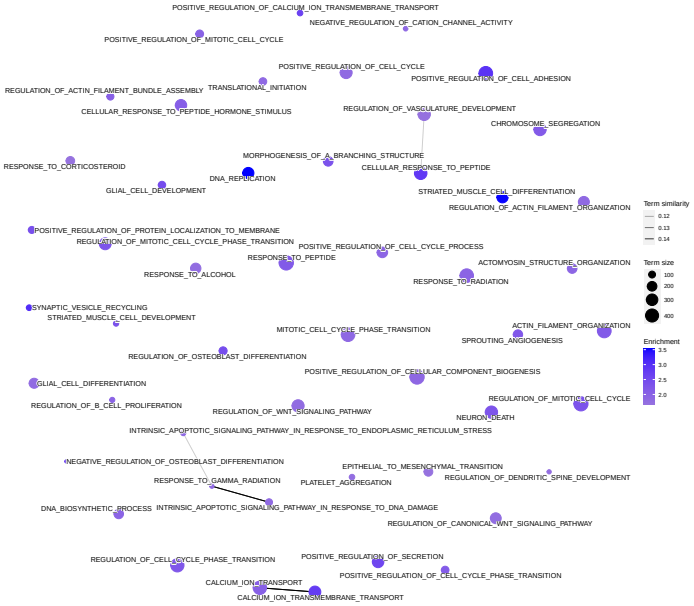


FIGURE 3.10: Term network for FET with same parameters as in fig. 3.2b except for also including isolated terms. The enrichment score is the negative decadic logarithm of the p-value.

4

THE NEXT GENERATION OF V-PIPE: TOWARDS SUSTAINABLE DATA PROCESSING WORKFLOWS

The large amount of diverse viral sequencing data sets generated by next-generation sequencing technologies poses a set of challenges for computational workflows. These challenges include the need for stringent quality control of sequencing reads, the ability to adapt the processing workflow to more samples with higher coverage and the possibility to adapt single steps of the workflow to application-specific needs. Such setups enable the analysis of viral diversity which is important in epidemiological and clinical settings.

Here, we show how V-pipe, a pipeline designed for analysing next-generation sequencing data of short viral genomes, has been extended to become more reproducible, adaptable and transparent. We demonstrate its ability to process hundreds of thousands of samples using high-performance computing environments, how automated testing and software deployment allows for rapid propagation of new releases to end-users and, finally, how these developments bring V-pipe closer to being a sustainable data processing workflow. One of V-pipe's core functionalities is the estimation of viral diversity which can be computed as global haplotypes, i.e., genome-length assemblies of all strains occurring in a sample. As this is still an unsolved problem with many existing approaches, we conduct a benchmark study of global haplotype reconstruction methods. We apply a set of global haplotype reconstruction methods to both synthetic and real data sets with varying parameters, to highlight how each method's performance depends on the biological setting it is applied to. The focus of this benchmark workflow is to make it easy to add additional methods and data sets in the future, in order to make a continuously ongoing benchmarking effort possible.

The author's contributions to the following manuscript were the development of new V-pipe features related to functionality and surrounding infrastructure, discussions on how to best advertise V-pipe to the scientific community, helping with public V-pipe tutorials and user support requests. In addition, they designed and implemented the global haplotype reconstruction benchmark.

The next generation of V-pipe: towards sustainable data processing workflows

Kim Philipp Jablonski[1,2], Ivan Topolsky[1,2],
Lara Fuhrmann[1,2], Benjamin Langer[1], Niko Beerenwinkel[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, 4058, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Basel, 4058, Switzerland
[*]To whom correspondence should be addressed

INTRODUCTION

Since the advent of next-generation sequencing (NGS) technologies, large amounts of RNA-sequencing data are being generated which can no longer be easily analyzed on personal computers [1]. As this availability of high-coverage data sets brings interesting research opportunities but also computational challenges, many new processing and analysis tools are being developed.

In particular, new possibilities of characterizing viral variants and analyzing the genetic diversity of viral sequencing samples have emerged [2, 3]. While inter-host variability describes how viral strains differ between separate hosts and can be used, for example, to build phylogenetic trees, intra-host variability measures the diversity of viral strains within a single host, provides adaptive advantages for the virus and is thus especially relevant to understanding infectious dynamics and treatment options [4, 5]. The viral strains observed in a single host are referred to as the set of haplotypes. They can be measured as local haplotypes, where mutations co-occurring at genomic distances up to a single read length are analyzed together to better differentiate between true mutations and technical errors. Due to improved sequencing error rates and higher coverage, it is also possible to recover low-frequency haplotypes at full genome length, a process called global haplotype reconstruction. This is more complicated as multiple reads need to be merged together to cover a whole genome, but also provides a better measure of viral diversity [6].

There are various issues which typically arise when applying computational methods for the reconstruction of global haplotypes to large data sets and sharing the results afterwards. As these tools are written in various programming languages (Python, R, Perl, C, C++, Java, Rust, ...), installing them on different machines requires setting up multiple, potentially conflicting software environments. Due to the complexity of the data, these tools are usually executed as part of a processing workflow which takes care of data retrieval, quality control, running the method and visualizing the results. In such cases it is difficult to organize the workflow in a reproducible way which can be easily understood by others and is quickly adapted to novel research settings. As the methodologies and data sources can be heterogeneous, understanding the performance of each tool and benchmarking them in a realistic way is difficult. As a consequence, it is important to provide data analysis procedures as publicly available workflows implemented in proper workflow management systems and use tools

which have been benchmarked appropriately. This approach also facilitates the continuous re-evaluation of the benchmarking workflow with new and updated parameter settings. This is needed as new methods are being developed which have to be compared to already existing ones, new test data sets become available either synthetically with new simulation setups or real with new experimental setups, and completely new application domains can appear which requires an adaption of the existing benchmarking workflow.

Various workflows have been proposed which try to accomplish these goals. ViralFlow is a workflow to analyze SARS-CoV-2 Illumina amplicon sequencing data [7]. While it is able to scale from personal computers to high-performance clusters, it is not easily possible to extend the workflow to, for example, analyses other types of viruses or include different tools. [8] presents a benchmarking workflow which compares several haplotype caller methods on several parameter settings, including mutation rates and samples sizes. However, as before, it is not easily possible to extend this workflow with new methods and data sets.

V-pipe is a pipeline designed for analyzing NGS data of short viral genomes [9]. It combines multiple tools related to quality control, sequence alignment, consensus sequence assembly, haplotype calling, result visualization and publication into an easy to execute workflow. In particular, it started to address the aforementioned issues by implementing the workflow in Snakemake [10]. However, due to a deprecated cluster integration system and HIV-specific organization of the workflow structure, it was difficult to scale up to diverse, large-scale data sets. The diversity comes from the need to apply such analysis workflows to different types of viruses. The onset of the SARS-CoV-2 pandemic has shown that such requirements can quickly appear and are important to public health. In addition, the sequencing samples can come from diverse sources, such as clinical or wastewater settings and can thus require application-specific processing steps which need to be supported in the same workflow. Another influence of the SARS-CoV-2 pandemic is that a substantial increase in sequencing capacities has led to unprecedentedly large numbers of samples becoming publicly available. Analysis workflows need to be able to handle such large amounts of data in order to be beneficial to public health and epidemiological advances.

Here, we propose a substantial rewrite of V-pipe which significantly improves the workflow's ability to perform sustainable data analysis as outlined in [10]. Reaching sustainability includes improved reproducibility, scalability, adaptability and readability of the workflow. We highlight which

improvements have been implemented to achieve these properties and describe how they have been applied to large-scale analysis projects. In particular, we show how rewriting the workflow in a more efficient way enables it to process hundreds of thousands of samples, how automated source code testing makes it possible to quickly make new functionalities and bug fixes available to end-users and how its modular design allows to quickly implement application-specific features. Finally, we show how one of V-pipe's core functionalities, the estimation of viral diversity using global haplotypes, is still an unsolved problem where many computational approaches exist. We conduct an extensive benchmark study where we apply a set of global haplotype reconstruction methods to both synthetic and real data sets in order to highlight in which settings current methods fail to estimate the set of true haplotypes. We implement this benchmark workflow such that it is easily possible to add new methods and test data sets. This enables an ongoing benchmarking effort which can quickly adapt to future developments and ensures that V-pipe employs state-of-the-art viral diversity estimation methods.

METHODS

Global haplotype reconstruction benchmark

GENERATION OF SYNTHETIC DATA SETS. The synthetic data sets are generated in two steps, first, viral haplotype populations are generated, in the second steps reads are simulated (fig. 4.1). If no master sequence is provided by the user, it is generated by drawing bases uniformly at random for each position based on the user-provided genome length. The benchmarking framework includes options to simulate haplotype populations based on a mutation rate or pairwise distances.

In the case of haplotype generation based on mutation rates, substitutions, deletions and insertions are introduced into the master sequence based on the user-specified rates μ. This follows the coalescent model assumptions. The frequency composition of those haplotypes in the population is derived from haplotype frequencies $f = (f_1, ..., f_K)$ inputted by the user. This method generates haplotype populations that have a star-like phylogeny. These simulation settings allow testing the reconstruction limits of the different methods.

In the case of haplotype generation by pairwise distances, we simulate hierarchical relationships among the haplotypes by generating two groups

of closely related haplotypes that share a common ancestor. First, using the user-specified between-group pairwise distance d_{12} two haplotypes are generated from the master sequence. Second, for each haplotype, child-haplotypes are generated by introducing mutations based on the respective within-group pairwise distance (d_1 and d_2 respectively) and group size (n_1 and n_2 respectively). The frequency distribution of the generated haplotypes is obtained from a geometric series with a given ratio (default: 0.75, this results in a few high-frequency and many low-frequency haplotypes being present) or drawn from a Dirichlet distribution with user-povided concentration parameters α_i. Given a user-specified per-position coverage and read length, paired-end reads are simulated in shotgun-mode using ART Illumina read simulator [11].

INCLUSION OF REAL DATA SETS. In addition to synthetic data sets where the ground truth is known but the degree of realism can be debated, real data sets are included in the benchmark. We test the global haplotype reconstruction methods on sequencing reads from the 5-virus-mix [6]. It provides Illumina MiSeq reads for a mixture of the five HIV-1 strains HXB2, 89.6, JR-CSF, NL4-3 and YU-2 and thus gives an estimate of ground truth which can be used in the performance evaluation. The benchmark is designed to make the addition of further real data sets easily possible. In the benchmark study we test the performance of the global haplotype reconstruction methods with the 5-virus-mix at the read subsampling levels $1, 0.75, 0.5, 0.25$.

PERFORMANCE EVALUATION. To evaluate the performance of each method in the global haplotype reconstruction benchmark, we compute precision and recall scores for the recovery of ground truth global haplotypes for each method in each condition. To do so, we consider the ground truth set of haplotype sequences and the set of sequences produces by a method. For each predicted sequence, we check if there exists a ground truth sequence with a relative edit distance below a predefined threshold γ. We define the relative edit distance ED_{rel} as

$$ED_{rel} = \frac{ED}{\max(L_{pred}, L_{true})} \tag{4.1}$$

where ED is the edit distance between a predicted and ground truth haplotype which have lengths L_{pred} and L_{true} respectively. If yes, this counts as a true positive, otherwise as a false positive. To compute the number of false negatives, we iterate over all ground truth sequences. If a ground

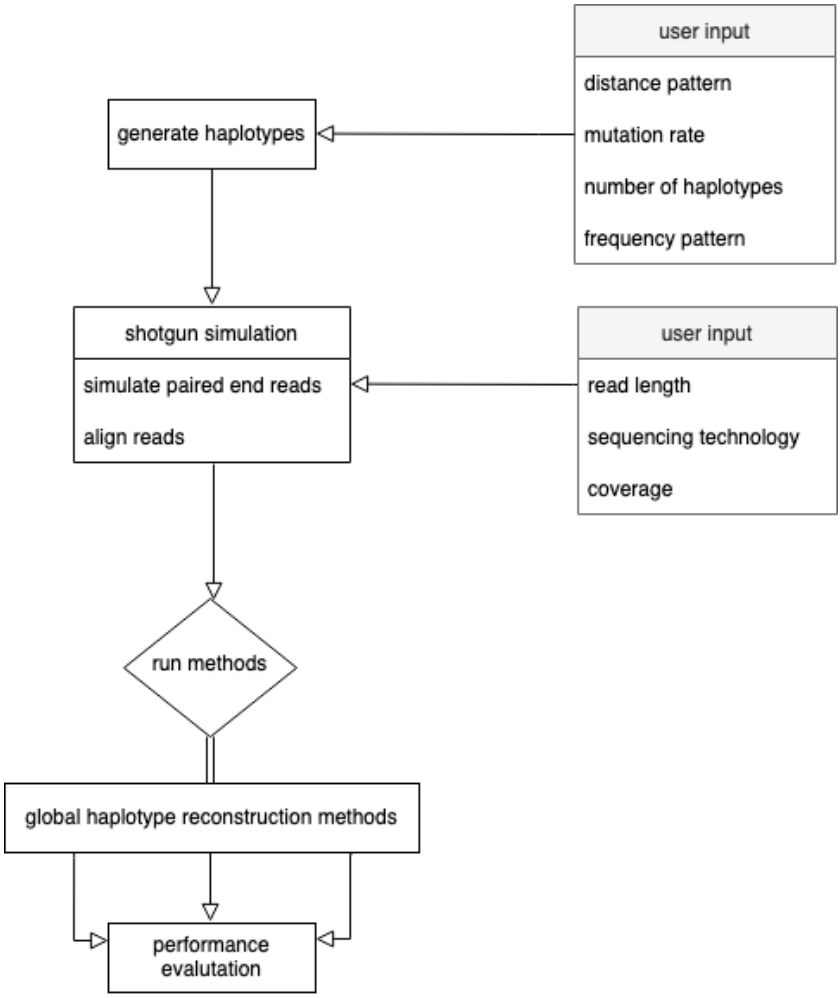


FIGURE 4.1: Workflow for the performance evaluation of global haplotype reconstruction methods: 1. Generation of haplotype population based on user input, 2. Simulation of paired-end Illumina sequencing reads, 3. Run global haplotype reconstruction methods, 4. Performance evaluation.

truth sequence has no matching, i.e., relative edit distance below a certain threshold, predicted sequence we count a false negative. From this, we compute precision as $TP/(TP+FP)$ and recall as $TP/(TP+FN)$. We use $\gamma = 0.01$ as the relative edit distance threshold in the benchmark study.

Two-dimensional embeddings of haplotype sequences are generated by applying multidimensional scaling with precomputed edit distances between all sequences.

We use MetaQUAST to compute measures of assembly quality for the reconstructed haplotypes [12]. In particular, we compute the N50 score which, in this context, equals the length of the shortest haplotype which together with all larger haplotypes, covers at least half the genome.

RESULTS

V-pipe is a bioinformatics workflow which combines various tools for analyzing viral NGS data. It focuses on estimating the viral diversity of samples at SNP, local haplotype, and global haplotype levels. In addition, it provides many helpful components for pre- and post-processing the generated data by, for example, conducting automated quality control tests or generating interactive result visualizations.

In the following, we will describe how we have extended V-pipe to become a more sustainable data processing workflow, applied it to large-scale projects and conducted a global haplotype reconstruction benchmark.

Towards sustainability

In order to enable sustainable data analysis using V-pipe, we follow the hierarchy of sustainability proposed in [10] and extend the workflow to make it more reproducible, adaptable and transparent. It has been widely recognized that this is crucial to scientific progress but often lacking in current literature [13, 14].

Reproducibility allows other researchers to execute an existing workflow and obtain the exact same results as the original workflow authors. To achieve this goal, we make the workflow more portable between different computing architectures by defining all software dependencies in Conda environments. That way, it can be executed without complicated, manual installation procedures. We reach better scalability by using efficient programming techniques to execute jobs on the cluster, specifying cluster resources dynamically such that they can be adapted to specific

data requirements which both help with deploying the workflow on new cluster environments, validating user configuration files using *jsonschema* in order to catch potential runtime errors during startup and provide sensible default values, splitting tasks which were previously performed in a single, centralized job over multiple compute nodes and computing various summary statistics in a per-sample distributed fashion in order to be able to scale up to the analysis of $> 100,000$ samples. Furthermore, to make large-scale analyses of public data sets easier, we have added an input data retrieval functionality which requires a set of SRA accessions [15] as input and automatically downloads all data files needed to run the whole workflow. In addition, new scripts are made available which facilitate the unattended mass-import of raw files as produced by Illumina's demultiplexing software into the structure that V-pipe expects as its input. To automate quality control steps needed to assert the sensibility of input data and produced results, we have added a component for frameshift insertion/deletion (indel) checks to the workflow. It identifies frameshift indels in the consensus sequence of the sample and reports a summary of those which includes the read support, the gene region of the indel, as well as if it is located in a homopolymeric region. This report helps to assess the plausibility of the frameshift indels and is used, for example, to submit SARS-CoV-2 sequences to GISAID [16]. GISAID requires a quality check of those frameshifts before publishing new submissions. This allows for the quick publication of newly generated sequences with less manual checks of the alignments by the submitters and is helpful for public health initiatives [17]. To standardize the consensus sequence generation and make V-pipe's results more comparable to other data processing workflows, we added a workflow component to compute consensus sequences using BCFtools [18] and trim primers from the alignment using iVar [19].

Adaptability refers to making it easy for other researchers to build upon an existing workflow and extend it for their application- and domain-specific needs. We have refactored V-pipe's configuration setup in such a way that the base configuration is virus-agnostic while virus-specific settings (specific reference sequences, different alignment tools, etc.) can be easily plugged in. This makes it possible to quickly adapt V-pipe to novel viruses, such as, for instance, Influenza or Human Papilloma viruses, without requiring complicated workflow changes as well as share virus configurations with others to make collaborations easier. For example, we provide HIV- and SARS-CoV-2-specific configuration setups which select appropriate reference files, read alignment software and post-processing

steps. It is easily possible to write such configuration files for other viruses in the future. To make sure new features can be quickly added to the workflow without compromising its correctness, we track the development using git and run automated integration and unit tests using GitHub Actions workflows on every commit submitted to the repository. We use small, exemplary data sets from different viruses in our tests to make sure that V-pipe is able to successfully run from start-to-end, and also explicitly test specific parts of the workflow to ensure reliability of notoriously unstable components. To help with common post-processing steps, we have added scripts to assist the user with uploading generated consensus sequences and compressed raw reads to databases as well as created interactive visualizations which summarize cohort and sample-specific information. This includes an alignment visualization which helps understand surprising SNPs, a combined coverage and SNP plot which summarizes a sample's diversity, the display of a phylogenetic tree which puts the analyzed samples into the context of a larger population, and links to other resources, such as, for example, the ability to model and visualize the structure of genes and their individual SNPs using SWISS-MODEL [20] with a single click. To help with privacy concerns before publication of raw reads in databases like SRA, we have added a component which depletes samples from all host, e.g., human, reads before preparing the compressed upload files. We have also added the computation of various diversity measures. To assess the different aspects of within-sample diversity, we included the most commonly used indices [21], namely the position-wise Shannon Entropy, average Shannon Entropy, mean mutation frequency, standard error of the mean mutation frequency, sum of mutation frequency, population nucleotide diversity, number of all mutations and mutation spectrum (e.g., the distribution of mutation frequencies). Another way of measuring viral diversity is the computation of global haplotypes. To show how to easily add new software components achieving this goal, we added PredictHaplo [22] a well-performing global haplotype reconstruction method as an example.

Transparency describes the ability of quickly understanding a given workflow. We have rewritten V-pipe's documentation as dynamic scripts which allows testing them as part of the automated test suite mentioned before and makes sure they always represent the latest release version and do not contain outdated information. To allow users to quickly benefit from new features as well as get new users quickly up and running, we provide three deployment methods: 1) a Bash script which automatically creates the required Conda environments, installs all dependencies and initializes a

project structure, 2) the ability to use Snakemake's *snakedeploy* tool to install V-pipe, and 3) a Docker container [23] which is automatically generated for every new release and for the *master* branch of the git repository.

Applications to large-scale projects

The effectiveness of the previously described improvements can observed in the successful application of V-pipe to various large-scale projects.

Because of the automation that workflows like V-pipe provide, they have been valuable tools to sequence large numbers of samples as part of national pandemic surveillance programs. V-pipe has been successfully applied as part of the Swiss SARS-CoV-2 Sequencing Consortium [24] and has provided a substantial part of the data in the national surveillance efforts [25, 26].

One of the successful applications of V-pipe has been the surveillance of viruses through the sequencing of wastewater. This category of samples contains mixtures of multiple viral sub-variants and workflows targeting diversity analysis are prime candidates for handling them. V-pipe was used to analyze the sequencing data which was used in the subsequent statistical modeling. In particular, the spread of SARS-CoV-2 has been monitored this way [27], and the surveillance of wastewater has enabled early detection of new variants such as Omicron B.1.1.529 [28].

The unprecedented availability of sequencing data from clinical SARS-CoV-2 samples and associated meta-data makes global analyses of diversity patterns very interesting [29]. V-pipe was used to process sequencing data and compute diversity measures.

Global haplotype reconstruction benchmark

One core functionality of V-pipe is the estimation of viral diversity. While this diversity can be measures in many ways, the reconstruction of global haplotypes recovers viral strains of whole genome length and is thus especially interesting for analyzing viral samples and, for instance, characterizing the contained virus population. However, at the same time, this reconstruction is experimentally and computationally challenging [30].

In order to better understand which global haplotype reconstruction methods are best suited for inclusion into V-pipe and application to contemporary research questions, we create a Snakemake based workflow as part of V-pipe which automatically applies a set of global haplotype reconstruc-

tion tools to various synthetic and real data sets, computes their respective performances and summarizes the results. We note that adding new tools and new data sets to this benchmark is very easy and only requires the addition of a single file and no further modifications of the workflow. This also supports the goal of making V-pipe a sustainable workflow, as robustly benchmarking included software components makes the workflow more adaptable to new data sets and its results more reliable.

In this benchmark study, we generate two synthetic data sets and use one real data set. In the first synthetic data set, we consider a genome of length 10000 with reads of length 200. We then generate two groups of haplotypes such that group one has size $n_1 = 5$ and group two has size $n_2 = 5$, the sequence distance within group one is $d_1 = 50$, the sequence distance within group two is $d_2 = 20$, and the sequence distance between the two groups is $d_{12} = 200$. We vary the coverage at $500, 1000, 5000, 10000$ in order to investigate how well the methods are able to recover low-frequency haplotypes as the coverage decreases. In the second synthetic data set, we consider a genome of length 10000 with reads of length 200 at a constant coverage of 1000. We then use the six haplotype population parameter settings $n_1 = 5, n_2 = 5, d_{12} = 200, d_1 = 50, d_2 = 20$, $n_1 = 5, n_2 = 5, d_{12} = 400, d_1 = 100, d_2 = 100$, $n_1 = 10, n_2 = 20, d_{12} = 200, d_1 = 50, d_2 = 20$, $n_1 = 10, n_2 = 20, d_{12} = 400, d_1 = 100, d_2 = 100$, $n_1 = 5, n_2 = 50, d_{12} = 200, d_1 = 50, d_2 = 20$ and $n_1 = 5, n_2 = 50, d_{12} = 400, d_1 = 100, d_2 = 100$ in order to investigate how well the methods are able to recover different types of haplotype populations with different diversity levels. For the real data set, we use the 5-virus-mix which contains the HIV-1 strains HXB2, 89.6, JR-CSF, NL4-3 and YU-2.

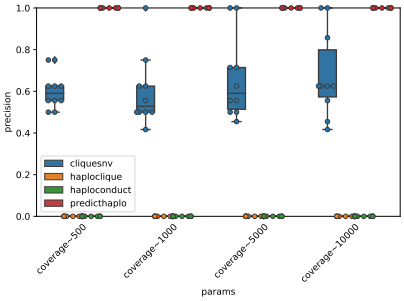
We consider all methods discussed in [30] for which a Conda package is available. They are aBayesQR [31], CliqueSNV [32], HaploClique [33], HaploConduct [34], PEHaplo [35], PredictHaplo [22], QuasiRecomb [36], and RegressHaplo [37]. From the benchmark study we exclude aBayesQR because the program fails to parse the input sequencing reads, PEHaplo because it fails execution during the result assembly, QuasiRecomb as it terminates during startup and Regresshaplo because not all dependencies of its conda package are available.

The remaining tools are HaploConduct, HaploClique, PredictHaplo and CliqueSNV which are all reference-based global haplotype reconstruction methods. This means that they rely on the existence of a viral reference sequence which is similar to the haplotypes expected to occur. The input reads are then typically mapped against this reference sequence which

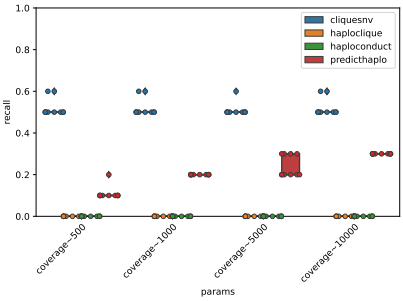
makes reconstructing global haplotypes easier, because read positions relative to the genome are available, but also introduces a bias, as haplotypes which are dissimilar to the reference might not be captured. For the real data set, we exclude HaploClique because it needed more than 50GB of memory to run.

For the case of constant haplotype population parameters and varying coverage, we observe that PredictHaplo achieves a precision of 1 in all cases, CliqueSNV stays around 0.6 with a slight increase with larger coverage, while HaploClique and HaploConduct show a precision of 0 (fig. 4.2a). This indicates that all hapotypes predicted by PredictHaplo are correct according to a relative edit distance threshold of $\gamma = 0.01$, while all sequences predicted by HaploClique and HaploConduct are incorrect. We observe that CliqueSNV features the highest recall of 0.64 which remains constant over all coverage values (fig. 4.2b). PredictHaplo's recall is 0.1 for the lowest coverage of 500 and increases up to 0.3 for the highest coverage of 10000. HaploClique and HaploConduct do not recover any true haplotypes and have a recall of 0 in all cases. Consequently, the recall performance of graph-based methods such as CliqueSNV is less dependent on the coverage level when compared to a probabilistic method such as PredictHaplo. As expected from these precision and recall metrics, CliqueSNV and PredictHaplo feature N50 scores of the full genome length of 10000 indicating that all their reconstructed haplotypes cover the whole genome for the whole range of coverage values (fig. 4.2c). In contrast, haplotypes reconstructed by HaploClique and HaploConduct do not even cover a quarter of the whole genome. We see that CliqueSNV consistently requires the least amount of time to run, while PredictHaplo needs over an hour for the highest coverage (fig. 4.2d). An exemplary overview of a single reconstruction result visualized using multi-dimensional scaling with edit distances between sequences can be seen in fig. 4.2e. The two ground truth haplotype clusters features 5 sequences each. PredictHaplo recovers a single haplotype for each of the clusters, while CliqueSNV finds multiple haplotypes for one of the clusters but none for the other. This recapitulates the higher recall of CliqueSNV compared to PredictHaplo.

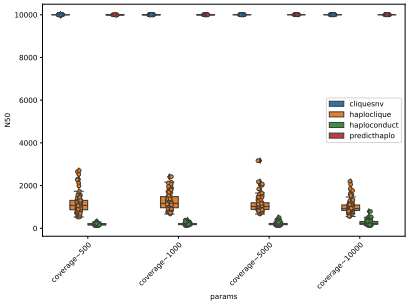
When keeping the coverage constant and varying the haplotype population parameters, we again observe perfect precision of 1 for PredictHaplo, precision values from below 0.6 up to 1 for CliqueSNV and precision of 0 for HaploClique and HaploConduct in all cases (fig. 4.3a). CliqueSNV's precision is lowest small group sizes ($n_1 = 5$) as it tends to overestimate the number of haplotypes in such cases. In all cases except for two small



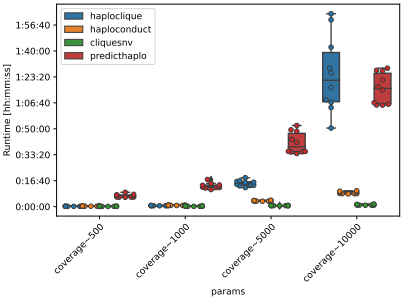
(a) Precision metric where each point corresponds to a replicate.



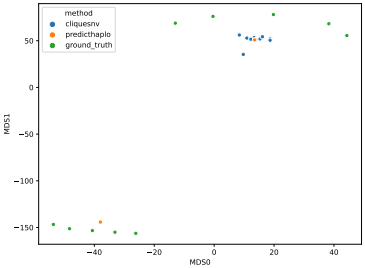
(b) Recall metric where each point corresponds to a replicate.



(c) N50 metric where each point corresponds to a replicate.



(d) Run time metric where each point corresponds to a replicate.



(e) MDS plot where each point corresponds to a single sequence for $n_1 = 5, n_2 = 5, d_{12} = 200, d_1 = 50, d_2 = 20$ with coverage 1000. HaploClique and HaploConduct results were excluded due to their poor performance.

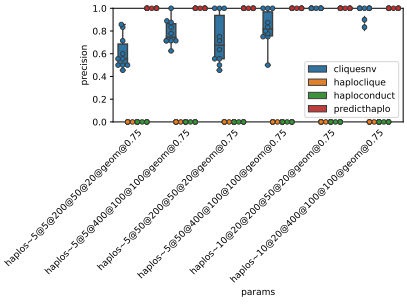
FIGURE 4.2: Performance overview for haplotype population generated with parameter setting $n_1 = 5, n_2 = 5, d_{12} = 200, d_1 = 50, d_2 = 20$ and varying coverage $500, 1000, 5000, 10000$.

haplotype groups $n_1 = n_2 = 5$ and large inter-group sequence distances of $d_{12} = 400$, CliqueSNV features a larger recall than PredictHaplo (fig. 4.3b). Both CliqueSNV and PredictHaplo get their highest recall for similar group sizes with $n_1 = n_2 = 5$ and their lowest recall dissimilar group sizes such as $n_1 = 5, n_2 = 50$. This indicates that both tools are not able to appropriately deal with haplotype populations which consist of multiple groups with imbalanced sizes. CliqueSNV's generally higher recall than PredictHaplo's, as before, is due to CliqueSNV estimating a larger amount of haplotypes than PredictHaplo. HaploClique and HaploConduct remain at a recall of 0. The N50 score looks similar to before, CliqueSNV and PredichtHaplo cover the whole genome while HaploClique and HaploConduct do not even reach a quarter of it (fig. 4.3c). The run time of all methods remains roughly the same over all parameter settings with PredictHaplo taking the longest in all cases with slightly more than 10 minutes per run (fig. 4.3d). In one particular example, we now observe that PredictHaplo predicts a single haplotype per cluster while CliqueSNV finds multiple ones per cluster (fig. 4.3e). In addition, CliqueSNV identifies a single haplotype not belonging to any of the two clusters.

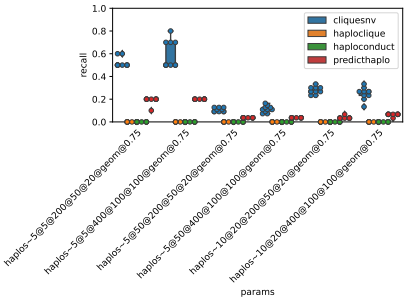
For the real data, we observe that precision and recall remain in the range of $0.2 - 0.4$ over all subsampling fractions for PredictHaplo (figs. 4.4a and 4.4b). CliqueSNV and HaploConduct mostly remain at 0 in both cases, except for CliqueSNV which performs better in one of the replicates for a subsampling fraction of 0.5.

As before, PredictHaplo's and CliqueSNV's reconstructions cover nearly the whole genome while HaploConduct only reaches less than a fifth (fig. 4.4c). In all cases the run time decreases due to the decreased coverage with more subsampling (fig. 4.4d). PredictHaplo takes the longest with over 3 hours in the unsampled data set while CliqueSNV takes the least amount of time in all cases. In one exemplary reconstruction case, both PredictHaplo and CliqueSNV find the correct number of haplotypes but PredictHaplo's predictions are closer to the ground truth sequences (fig. 4.4e). This recapitulates the precision and recall plots and highlights that when both tools find the same correct number of haplotypes, PredictHaplo tends to be closer to the ground truth.

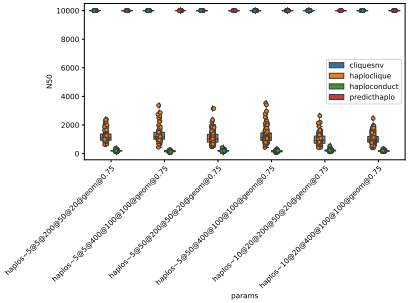
In summary, we have shown that CliqueSNV had the least run time in all three benchmark cases and the best recall performance in the two synthetic data sets, while PredictHaplo shows better precision in the two synthetic data sets. This can mostly be explained by CliqueSNV typically recovering a larger amount of haplotypes than PredictHaplo. PredictHaplo



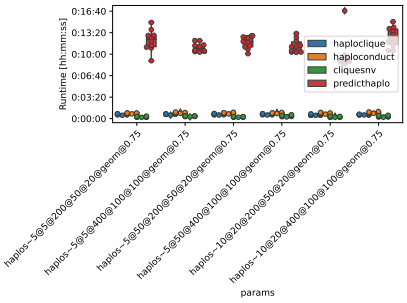
(a) Precision metric where each point corresponds to a replicate.



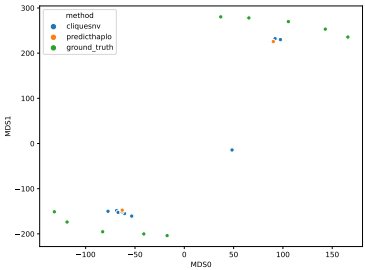
(b) Recall metric where each point corresponds to a replicate.



(c) N50 metric where each point corresponds to a replicate.

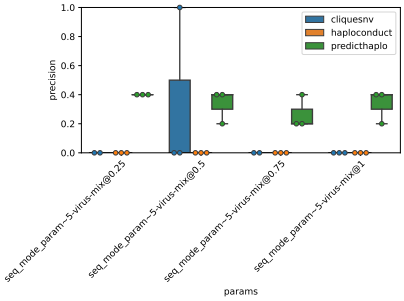


(d) Run time metric where each point corresponds to a replicate.

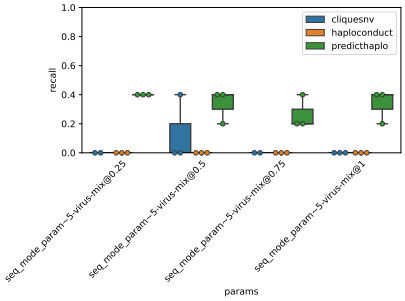


(e) MDS plot where each point corresponds to a single sequence for $n_1 = 5, n_2 = 5, d_{12} = 400, d_1 = 100, d_2 = 100$ with coverage 1000. HaploClique and HaploConduct results were excluded due to their poor performance.

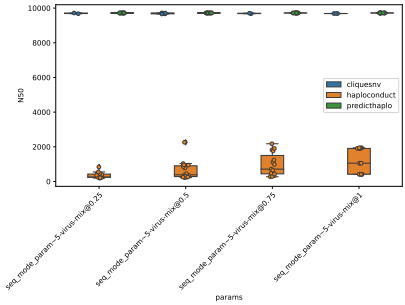
FIGURE 4.3: Performance overview for haplotype population generated with varying parameter settings $n_1 = 5, n_2 = 5, d_{12} = 200, d_1 = 50, d_2 = 20$, $n_1 = 5, n_2 = 5, d_{12} = 400, d_1 = 100, d_2 = 100$, $n_1 = 10, n_2 = 20, d_{12} = 200, d_1 = 50, d_2 = 20$, $n_1 = 10, n_2 = 20, d_{12} = 400, d_1 = 100, d_2 = 100$, $n_1 = 5, n_2 = 50, d_{12} = 200, d_1 = 50, d_2 = 20$ and $n_1 = 5, n_2 = 50, d_{12} = 400, d_1 = 100, d_2 = 100$, and coverage 1000.



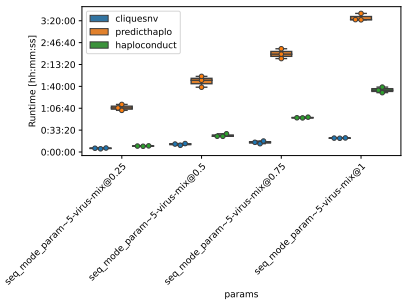
(a) Precision metric where each point corresponds to a replicate.



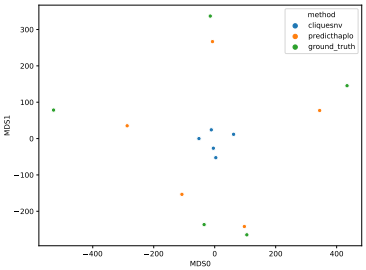
(b) Recall metric where each point corresponds to a replicate.



(c) N50 metric where each point corresponds to a replicate.



(d) Run time metric where each point corresponds to a replicate.



(e) MDS plot where each point corresponds to a single sequence for a subsampling fraction of 0.5. HaploConduct results were excluded due to their poor performance.

FIGURE 4.4: Performance overview for haplotype population obtained from the 5-virus-mix with subsampling fractions $0.25, 0.5, 0.75, 1$.

was better able to reconstruct global haplotypes with the real data set both in terms of precision and recall. This could be due to PredictHaplo's default parameters being tuned to HIV-1 strains as they were used in its original publication as well. In addition, we highlight the difficulty of reconstructing global haplotypes in the presence of haplotype clusters of different sizes. HaploClique and HaploConduct were generally not able to recover any haplotypes.

CONCLUSIONS

We have presented a substantial update of V-pipe, a pipeline designed for analyzing next-generation sequencing data of short viral genomes, with the aim of making it a more sustainable workflow. In particular, we describe how we make it more reproducible by following Snakemake's best-practice guidelines, more adaptable by adding virus-specific configuration files which can be quickly exchanged and more transparent by providing automatically tested usage examples which are available online.

We demonstrate the effectiveness of these improvements by highlighting two large-scale projects in which V-pipe was used to process many thousands of samples over multiple years in an automated fashion. For example, V-pipe is an integral part of the Swiss wastewater sequencing efforts.

One of V-pipe's core functionalities is the estimation of viral diversity. A very effective yet difficult to compute way of measuring viral diversity is the reconstruction of global haplotypes from NGS data as they describe the intra-host strain distribution at full length of the viral genome. We conduct a benchmark study of current global haplotype reconstruction methods in order to better understand the performance and limitations of current methods. As this field is still quickly advancing and no solution generally accepted as state-of-the-art has been found, we focus on making the addition of new tools and test data sets to the workflow as straight-forward as possible. Adding new methods is as easy as writing a single script which defines how to execute the tool and how to install it from Conda. New data sources can be either synthetic or derived from real experimental samples. In the synthetic case, different haplotype evolution modeling assumptions can be specified in a flexible way. Real data sources can be automatically downloaded as part of the workflow and pre-processed, for example by subsampling reads, specifically test interesting applications.

In summary, we have made it easier for V-pipe to be applied to more samples by other researchers while keeping its execution robust and its

workflow structure open to modifications. We have created a benchmark for one of V-pipe's core functionalities which can be continuously updated when new methods and data sets appear.

In the future, we will extend V-pipe to support even more different kinds of viruses, make it more robust to unpredictable failure points in cluster environments and further improve interoperability with data providers and consumers.

BIBLIOGRAPHY

1. Pereira, R., Oliveira, J. & Sousa, M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. *Journal of clinical medicine* **9**, 132 (2020).
2. Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S. & Palù, G. Next-generation sequencing technologies in diagnostic virology. *Journal of Clinical Virology* **58**, 346 (2013).
3. Capobianchi, M., Giombini, E. & Rozera, G. Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection* **19**, 15 (2013).
4. Ko, H.-Y., Li, Y.-T., Chao, D.-Y., Chang, Y.-C., Li, Z.-R. T., Wang, M., Kao, C.-L., Wen, T.-H., Shu, P.-Y., Chang, G.-J. J., *et al.* Inter-and intra-host sequence diversity reveal the emergence of viral variants during an overwintering epidemic caused by dengue virus serotype 2 in southern Taiwan. *PLoS neglected tropical diseases* **12**, e0006827 (2018).
5. Bonnaud, E. M., Troupin, C., Dacheux, L., Holmes, E. C., Monchatre-Leroy, E., Tanguy, M., Bouchier, C., Cliquet, F., Barrat, J. & Bourhy, H. Comparison of intra-and inter-host genetic diversity in rabies virus during experimental cross-species transmission. *PLoS pathogens* **15**, e1007799 (2019).
6. Giallonardo, F. D., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S., Campbell, N. K., Joos, B., Lecca, M. R., *et al.* Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic acids research* **42**, e115 (2014).
7. Dezordi, F. Z., Neto, A. M. d. S., Campos, T. d. L., Jeronimo, P. M. C., Aksenen, C. F., Almeida, S. P., Wallau, G. L. & Network, F. C.-1. G. S. ViralFlow: a versatile automated workflow for SARS-CoV-2 genome assembly, lineage assignment, mutations and intrahost variant detection. *Viruses* **14**, 217 (2022).

8. Eliseev, A., Gibson, K. M., Avdeyev, P., Novik, D., Bendall, M. L., Pérez-Losada, M., Alexeev, N. & Crandall, K. A. Evaluation of haplotype callers for next-generation sequencing of viruses. *Infection, Genetics and Evolution* **82**, 104277 (2020).

9. Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J. & Beerenwinkel, N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* **37**, 1673 (2021).

10. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10** (2021).

11. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593 (2012).

12. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088 (2016).

13. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533** (2016).

14. Sayre, F. & Riegelman, A. The reproducibility crisis and academic libraries. *College & Research Libraries* **79**, 2 (2018).

15. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., *et al.* The European nucleotide archive. *Nucleic acids research* **39**, D28 (2010).

16. Bogner, P., Capua, I., Lipman, D. J. & Cox, N. J. A global initiative on sharing avian flu data. *Nature* **442**, 981 (2006).

17. Nadeau, S. A., Vaughan, T. G., Beckmann, C., Topolsky, I., Chen, C., Hodcroft, E., Schär, T., Nissen, I., Santacroce, N., Burcklen, E., *et al.* Swiss public health measures associated with reduced SARS-CoV-2 transmission using genome data. *medRxiv* (2021).

18. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

19. Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S., *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome biology* **20**, 1 (2019).

20. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research* **31**, 3381 (2003).

21. Fuhrmann, L., Jablonski, K. P. & Beerenwinkel, N. Quantitative measures of within-host viral genetic diversity. *Current opinion in virology* **49**, 157 (2021).

22. Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N. & Roth, V. HIV haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM transactions on computational biology and bioinformatics* **11**, 182 (2013).

23. Merkel, D. *et al.* Docker: lightweight linux containers for consistent development and deployment. *Linux j* **239**, 2 (2014).

24. *Swiss SARS-CoV-2 Sequencing Consortium* `https://bsse.ethz.ch/cevo/research/sars-cov-2/swiss-sars-cov-2-sequencing-consortium.html`. Accessed: 2022-07-22.

25. Chen, C., Nadeau, S. A., Topolsky, I., Manceau, M., Huisman, J. S., Jablonski, K. P., Fuhrmann, L., Dreifuss, D., Jahn, K., Beckmann, C., *et al.* Quantification of the spread of SARS-CoV-2 variant B. 1.1. 7 in Switzerland. *Epidemics* **37**, 100480 (2021).

26. Chen, C., Nadeau, S., Topolsky, I., Beerenwinkel, N. & Stadler, T. Advancing genomic epidemiology by addressing the bioinformatics bottleneck: Challenges, design principles, and a Swiss example. *Epidemics* **39**, 100576 (2022).

27. Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Bänziger, C., Devaux, A. J., Stachler, E., Caduff, L., *et al.* Detection and surveillance of SARS-CoV-2 genomic variants in wastewater. *MedRxiv* (2021).

28. Bagutti, C., Hug, M. A., Heim, P., Pekerman, L. M., Hampe, E. I., Hübner, P., Fuchs, S., Savic, M., Stadler, T., Topolsky, I., *et al.* Wastewater monitoring of SARS-CoV-2 shows high correlation with COVID-19 case numbers and allowed early detection of the first confirmed B. 1.1. 529 infection in Switzerland: results of an observational surveillance study. *Swiss Medical Weekly* (2022).

29. Kuipers, J., Batavia, A. A., Jablonski, K. P., Bayer, F., Borgsmüller, N., Dondi, A., Dragan, M.-A., Ferreira, P., Jahn, K., Lamberti, L., *et al.* Within-patient genetic diversity of SARS-CoV-2. *BioRxiv* (2020).

30. Jablonski, K. P. & Beerenwinkel, N. in *Virus Bioinformatics* 51 (Chapman and Hall/CRC, 2021).

31. Ahn, S. & Vikalo, H. *aBayesQR: a Bayesian method for reconstruction of viral populations characterized by low diversity* in *International Conference on Research in Computational Molecular Biology* (2017), 353.

32. Knyazev, S., Tsyvina, V., Shankar, A., Melnyk, A., Artyomenko, A., Malygina, T., Porozov, Y. B., Campbell, E. M., Switzer, W. M., Skums, P., *et al.* CliqueSNV: an efficient noise reduction technique for accurate assembly of viral variants from NGS data. *bioRxiv*, 264242 (2020).

33. Töpfer, A., Marschall, T., Bull, R. A., Luciani, F., Schönhuth, A. & Beerenwinkel, N. Viral quasispecies assembly via maximal clique enumeration. *PLoS computational biology* **10**, e1003515 (2014).

34. Baaijens, J. A. & Schönhuth, A. Overlap graph-based generation of haplotigs for diploids and polyploids. *Bioinformatics* **35**, 4281 (2019).

35. Chen, J., Zhao, Y. & Sun, Y. De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding. *Bioinformatics* **34**, 2927 (2018).

36. Töpfer, A., Zagordi, O., Prabhakaran, S., Roth, V., Halperin, E. & Beerenwinkel, N. Probabilistic inference of viral quasispecies subject to recombination. *Journal of Computational Biology* **20**, 113 (2013).

37. Leviyang, S., Griva, I., Ita, S. & Johnson, W. E. A penalized regression approach to haplotype reconstruction of viral populations arising in early HIV/SIV infection. *Bioinformatics* **33**, 2455 (2017).

5

CONCLUSIONS & OUTLOOK

In this work, we have presented two studies which develop novel methods for detecting pathway dysregulations in diseases and a third study which shows how proper workflow design can enable large-scale and sustainable data analysis pipelines as well as be used to benchmark global haplotype reconstruction methods.

The first method, *dce*, recognizes the difficulty of performing both causal discovery and causal inference at the same time and replaces the discovery part with prior information obtained from pathway network databases. As input, it takes a pathway network structure and gene expression vectors from two conditions to be compared. It produces a list of edge-specific pathway perturbations caused by the disease condition compared to the wild-type condition. We show that using a causal approach for modeling these dysregulations outperforms more naive approaches and competing tools in a synthetic benchmark. In an exploratory analysis of publicly available Breast cancer samples from The Cancer Genome Atlas (TCGA) we are able to recover pathways which are already recognized in existing literature and thus validate our approach, as well as identify previously unknown pathway associations with different stages of Breast cancer progression which pose putative treatment targets and suggest further experimental validations. The presented method is general enough to not only apply to biological data, but to other data types as well. The method is freely available as an R/Bioconductor package, well documented and can be easily used by others. An interesting future direction would be to combine the prior pathway network data with causal graphs estimated from observational data. This could further improve the method's performance, in particular in cases where the assumed causal network has many incorrect or missing edges. It would furthermore be interesting to investigate how to handle biological networks with cycles, which violate the DAG assumption, in the causal framework, for example, by interpreting the gene expression data in a time series format. The regression approach could also be reconsidered in order to make the optimization procedure more robust, especially for low sample counts, or by introducing regularization terms.

The second method, *pareg*, produces a list of pathways associated with a dysregulation score which allows to infer which pathway was most af-

fected by a condition compared to another baseline condition. The method takes a pathway database containing lists of genes and per-gene p-values coming from a differential gene expression experiment as input. It formulates the enrichment computation as a regularized regression where the regularization terms are LASSO and a network fusion penalty. The LASSO term ensures a sparse and interpretable set of enriched pathways, while the network fusion penalty incorporates term-term relations into the model and provides more robust enrichment measures in large and redundant pathway databases. A synthetic benchmark is used to show that *pareg* outperforms competing methods over a range of representative simulation parameters and thus demonstrates the viability of the regularized regression to include term-term relations. In an exploratory analysis, *pareg* is applied to public Breast cancer samples from TCGA and is able to recover dysregulated pathway clusters which both recapitulate literature knowledge and suggest novel pathways. As before, the tool is freely available as an R/Bioconductor package, well documented and can be easily used by others. Interesting next steps could include an augmentation of the regression model by adding other regularization terms, for example, the Elastic-Net term, using other pathway similarity measures based on semantic similarities or protein-protein-interaction network embeddings, and modeling the response variable with a beta-uniform-mixture which has been shown to model p-value distributions more closely.

These two first studies investigated the detection of pathway dysregulations from two different perspectives. *dce* considered intra-pathway dependencies while *pareg* considered inter-pathway dependencies. In the first setting, the network structure induced by relations between genes in, for example, signal transduction pathways, was modeled with a causal framework and allowed to reduce the impact of confounding effects which obscures the true biological signal. In the second setting, the network structure induced by relations between pathways, for example, due to shared genes, was modeled with a regularized regression and allowed to reduce the detrimental impact of large redundant pathway databases which makes the estimation of enrichment scores more robust. In both cases, the run time of the methods on data sets of realistic size, i.e., hundreds of edges and up to thousands of samples for *dce*, and over a thousand of pathways and tens of thousands of genes for *pareg* when running without cross-validation, was less than 10 hours and thus feasible to run on personal computers. When running *pareg* with cross-validation, the total run time depends on the resolution of the parameter grid. As testing different regularization parameter

combinations is easily parallelizable, this scales well on high-performance computing clusters but can become prohibitive on personal machines. In such cases, small parameter grids should be used.

In the third study, it is shown how the workflow management system Snakemake can be used to create the workflow *V-pipe* which can process hundreds of thousands of samples in a reproducible, understandable and adaptable way. By following the latest workflow creation guidelines and programming best practices, the execution of the workflow on different underlying architectures, such as high-performance computing clusters, is enabled, while keeping dependency versions constant. Extensive documentation and automated testing of the software stack ensure that other scientists are always able to rely on them and use *V-pipe* to conduct their own research. By clearly separating the components responsible for different tasks in the data processing pipeline (e.g., quality control, alignment, visualizations), it is made possible to exchange parts without having to understand the whole workflow structure or to extend the workflow by adding research-specific post-processing steps. Furthermore, it is shown how this framework can be used to benchmark global haplotype reconstruction methods, a notoriously difficult task, in a rigorous way. The processing of around 1,000 SARS-CoV-2 samples takes approximately 5 hours on a high-performance computing cluster which is able to parallelize appropriately. Due to this high workload, such large-scale analyses are not possible on personal machines.

The three presented studies have made significant advances in their respective fields of research. The first two by developing novel models for detecting pathway dysregulations, and the third one by enabling large-scale analyses of viral sequencing data. In all three cases, the sustainability of the developed methods has been demonstrated by generating the results from sustainable Snakemake workflows. As described in chapter 1, this builds the basis for sustainable data science and is crucial to ensure the correctness, longevity and generalizability of these results. In particular, all workflows are well automated and can be executed with a single command (`snakemake -prj1 --use-conda --profile lsf`). The scalability is ensured by providing runtime and memory resources for each workflow component and using Snakemake cluster profiles which allow executing the same workflow on different cluster environments by simply switching a commandline parameter. The workflow is portable as each rule is associated with a Conda environment which defines all software dependencies with exact versions. Snakemake's domain specific language, which is based on

Python, makes the workflow readable, while having the statistical models published as Bioconductor packages maintains high coding standards. The benchmarking workflows employ parameter spaces to span many possible configurations in a traceable way. Specifically, this means that model parameters are specified in an external file which is not hidden in the workflow code, are prominently displayed during the execution of each workflow component and are always accessible during inspection of the final results. At the same time, documentation is provided in the form of web resources for the workflows and dynamic R vignettes for the statistical packages.

As a consequence, the studies covered in this thesis are indeed reproducible but also adaptable as well as transparent, and aim to be a work of fully sustainable data science. They thus serve as an example that exciting research, proper software engineering and workflow development are not in conflict but actually benefit each other. This thesis hopefully convinces the reader that conducting sustainable data science is very advantageous to currently ongoing research as well as the whole scientific community in the future, and well worth the small additional effort needed to achieve it.

BIBLIOGRAPHY

1. Popper, K. *The logic of scientific discovery* (Routledge, 2005).
2. Betz, F. in *Managing Science* 21 (Springer, 2011).
3. Andersen, H. & Hepburn, B. Scientific method (2015).
4. Claerbout, J. F. & Karrenbach, M. in *SEG technical program expanded abstracts 1992* 601 (Society of Exploration Geophysicists, 1992).
5. Plesser, H. E. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics* **11**, 76 (2018).
6. Gundersen, O. E. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A* **379**, 20200210 (2021).
7. Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).
8. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533** (2016).
9. Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. A manifesto for reproducible science. *Nature human behaviour* **1**, 1 (2017).
10. Sayre, F. & Riegelman, A. The reproducibility crisis and academic libraries. *College & Research Libraries* **79**, 2 (2018).
11. Kapoor, S. & Narayanan, A. *Leakage and the Reproducibility Crisis in ML-based Science* 2022.
12. Begley, C. G. & Ellis, L. M. Raise standards for preclinical cancer research. *Nature* **483**, 531 (2012).
13. Baggerly, K. A. & Coombes, K. R. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 1309 (2009).
14. Collaboration, O. S. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
15. Benestad, R. E., Nuccitelli, D., Lewandowsky, S., Hayhoe, K., Hygen, H. O., Van Dorland, R. & Cook, J. Learning from mistakes in climate research. *Theoretical and applied climatology* **126**, 699 (2016).

16. Frolov, S. *Quantum computing's reproducibility crisis: Majorana fermions* 2021.

17. Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics* **22**, 545 (2021).

18. Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., *et al.* Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).

19. Nado, Z., Band, N., Collier, M., Djolonga, J., Dusenberry, M., Farquhar, S., Filos, A., Havasi, M., Jenatton, R., Jerfel, G., Liu, J., Mariet, Z., Nixon, J., Padhy, S., Ren, J., Rudner, T., Wen, Y., Wenzel, F., Murphy, K., Sculley, D., Lakshminarayanan, B., Snoek, J., Gal, Y. & Tran, D. Uncertainty Baselines: Benchmarks for Uncertainty & Robustness in Deep Learning. *arXiv preprint arXiv:2106.04015* (2021).

20. Di Tommaso, P. *Awesome Pipeline* `https://github.com/pditommaso/awesome-pipeline` (2022).

21. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* **46**, W537 (2018).

22. Jimenez, I., Sevilla, M., Watkins, N., Maltzahn, C., Lofstead, J., Mohror, K., Arpaci-Dusseau, A. & Arpaci-Dusseau, R. *The popper convention: Making reproducible systems evaluation practical* in *2017 ieee international parallel and distributed processing symposium workshops (ipdpsw)* (2017), 1561.

23. Ben-Kiki, O., Evans, C. & Ingerson, B. Yaml ain't markup language (yaml™) version 1.1. *Working Draft 2008* **5**, 11 (2009).

24. Lampa, S., Dahlö, M., Alvarsson, J. & Spjuth, O. SciPipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines. *GigaScience* **8**, giz044 (2019).

25. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. & Notredame, C. Nextflow enables reproducible computational workflows. *Nature biotechnology* **35**, 316 (2017).

26. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520 (2012).

27. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

28. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10** (2021).

29. Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M. R., Peters, K. & Schober, D. FAIR computational workflows. *Data Intelligence* **2**, 108 (2020).

30. Plankensteiner, K., Montagnat, J. & Prodan, R. *IWIR: a language enabling portability across grid workflow systems* in *Proceedings of the 6th workshop on Workflows in support of large-scale science* (2011), 97.

31. Crusoe, M., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanic, N., Ménager, H., Soiland-Reyes, S. & Goble, C. Methods included: Standardizing computational reuse and portability with the Common Workflow Language. CoRR abs/2105.07028. *arXiv preprint arXiv:2105.07028* (2021).

32. Peng, R. D., Dominici, F. & Zeger, S. L. Reproducible epidemiologic research. *American journal of epidemiology* **163**, 783 (2006).

33. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719 (2009).

34. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100**, 57 (2000).

35. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646 (2011).

36. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer discovery* **12**, 31 (2022).

37. Iorio, F., Garcia-Alonso, L., Brammeld, J. S., Martincorena, I., Wille, D. R., McDermott, U. & Saez-Rodriguez, J. Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. *Scientific reports* **8**, 1 (2018).

38. Khakabimamaghani, S., Ding, D., Snow, O. & Ester, M. Uncovering the subtype-specific temporal order of cancer pathway dysregulation. *PLoS computational biology* **15** (2019).

39. Francia, G. & Kerbel, R. S. Raising the bar for cancer therapy models. *Nature biotechnology* **28**, 561 (2010).

40. Rubin, E. H. & Gilliland, D. G. Drug development and clinical trials—the path to an approved cancer drug. *Nature reviews Clinical oncology* **9**, 215 (2012).

41. De Anda-Jáuregui, G. & Hernández-Lemus, E. Computational oncology in the multi-omics era: state of the art. *Frontiers in oncology* **10**, 423 (2020).

42. Ruschhaupt, M., Huber, W., Poustka, A. & Mansmann, U. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical applications in genetics and molecular biology* **3** (2004).

43. Emmert-Streib, F., Dehmer, M. & Yli-Harja, O. Ensuring Quality Standards and Reproducible Research for Data Analysis Services in Oncology: A Cooperative Service Model. *Frontiers in cell and developmental biology* **7**, 349 (2019).

44. Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., *et al.* The Cancer Genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer research* **77**, e3 (2017).

45. Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C. & Draghici, S. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology* **4**, 278 (2013).

46. Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T. M., Mukherjee, S. & Ancona, N. Comparative study of gene set enrichment methods. *BMC bioinformatics* **10**, 1 (2009).

47. Alavi-Majd, H., Khodakarim, S., Zayeri, F., Rezaei-Tavirani, M., Tabatabaei, S. M. & Heydarpour-Meymeh, M. Assessment of gene set analysis methods based on microarray data. *Gene* **534**, 383 (2014).

48. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics* **15**, 504 (2014).

49. Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265 (2020).

50. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20**, 533 (2020).

51. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344 (2006).

52. Poon, L. L., Song, T., Rosenfeld, R., Lin, X., Rogers, M. B., Zhou, B., Sebra, R., Halpin, R. A., Guan, Y., Twaddle, A., *et al.* Quantifying influenza virus diversity and transmission in humans. *Nature genetics* **48**, 195 (2016).

53. Zhang, Y., Leitner, T., Albert, J. & Britton, T. Inferring transmission heterogeneity using virus genealogies: Estimation and targeted prevention. *PLoS computational biology* **16**, e1008122 (2020).

54. Jablonski, K. P. & Beerenwinkel, N. in *Virus Bioinformatics* 51 (Chapman and Hall/CRC, 2021).

55. Fuhrmann, L., Jablonski, K. P. & Beerenwinkel, N. Quantitative measures of within-host viral genetic diversity. *Current Opinion in Virology* **49**, 157 (2021).

56. Beerenwinkel, N., Günthard, H. F., Roth, V. & Metzner, K. J. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in microbiology* **3**, 329 (2012).

57. Posada-Cespedes, S., Seifert, D. & Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus research* **239**, 17 (2017).

58. Jablonski, K. P., Pirkl, M., Ćevid, D., Bühlmann, P. & Beerenwinkel, N. Identifying cancer pathway dysregulations using differential causal effects. *Bioinformatics* **38**, 1550 (2022).

59. Jablonski, K. P. & Beerenwinkel, N. Coherent pathway enrichment estimation by modeling inter-pathway dependencies using regularized regression. *bioRxiv* (2022).

ACKNOWLEDGEMENTS

I would like to thank my supervisor Niko Beerenwinkel for giving me the opportunity to conduct my doctoral studies in his group, his advice and support. I would also like to thank my committee members Caroline Uhler and Peter Bühlmann for reviewing this thesis and providing guidance throughout my PhD. Special thanks to all my collaborators without whom my research would not have been possible. Many thanks to current and past members of the Computational Biology and NEXUS groups for being so helpful during discussions and being extremely fun to hang out with both during and outside work. Thanks to the world for housing me, thanks to family and friends for joining the ride.

CURRICULUM VITAE

PERSONAL DATA

Name	Kim Philipp Jablonski
Email	kim.philipp.jablonski@gmail.com
Date of Birth	March 12, 1995
Place of Birth	Bremen, Germany
Citizen of	Germany

EDUCATION

2017 – 2022	**ETH Zürich, Switzerland** *Computational Biology* Doctoral Studies
2015 – 2017	**Jacobs University, Germany** *Computational Life Sciences* Master of Science
2012 – 2015	**Jacobs University, Germany** *Applied and Computational Mathematics* Bachelor of Science

PUBLICATIONS

1. Jablonski, K. P., Pirkl, M., Ćevid, D., Bühlmann, P. & Beerenwinkel, N. Identifying cancer pathway dysregulations using differential causal effects. *Bioinformatics* **38**, 1550 (2022).
2. Jablonski, K. P. & Beerenwinkel, N. Coherent pathway enrichment estimation by modeling inter-pathway dependencies using regularized regression. *bioRxiv* (2022).
3. Jablonski, K. P., Topolsky, I., Fuhrmann, L., Langer, B. & Beerenwinkel, N. The next generation of V-pipe: towards sustainable data processing workflows. *tba* (202?).
4. Jablonski, K. P. & Beerenwinkel, N. in *Virus Bioinformatics* 51 (Chapman and Hall/CRC, 2021).
5. Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., *et al.* Sustainable data analysis with Snakemake. *F1oooResearch* **10** (2021).
6. Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Bänziger, C., Devaux, A. J., Stachler, E., Caduff, L., *et al.* Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. *Nature Microbiology*, 1 (2022).
7. Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J. & Beerenwinkel, N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics* (2021).
8. Kuipers, J., Batavia, A. A., Jablonski, K. P., Bayer, F., Borgsmüller, N., Dondi, A., Dragan, M.-A., Ferreira, P., Jahn, K., Lamberti, L., *et al.* Within-patient genetic diversity of SARS-CoV-2. *BioRxiv* (2020).
9. Fuhrmann, L., Jablonski, K. P. & Beerenwinkel, N. Quantitative measures of within-host viral genetic diversity. *Current Opinion in Virology* **49**, 157 (2021).
10. Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., Melidou, A., Neher, R. A., O'Toole, Á., Pereyaslov, D., *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance* **25**, 2001410 (2020).

11. Nadeau, S., Beckmann, C., Topolsky, I., Vaughan, T., Hodcroft, E., Schaer, T., Nissen, I., Santacroce, N., Burcklen, E., Ferreira, P., *et al.* Quantifying SARS-CoV-2 spread in Switzerland based on genomic sequencing data. *medRxiv* (2020).

12. Chen, C., Nadeau, S. A., Topolsky, I., Manceau, M., Huisman, J. S., Jablonski, K. P., Fuhrmann, L., Dreifuss, D., Jahn, K., Beckmann, C., *et al.* Quantification of the spread of SARS-CoV-2 variant B. 1.1. 7 in Switzerland. *Epidemics* **37**, 100480 (2021).

13. Nadeau, S., Vaughan, T. G., Beckmann, C., Topolsky, I., Chen, C., Hodcroft, E., Schär, T., Nissen, I., Santacroce, N., Burcklen, E., *et al.* Swiss public health measures associated with reduced SARS-CoV-2 transmission using genome data. *medRxiv* (2021).