TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

EXTRACTING INFORMATION FROM HIGH-THROUGHPUT GENE EXPRESSION DATA WITH PATHWAY ANALYSIS AND DECONVOLUTION

Maria Jaakkola

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 668 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2022

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

EXTRACTING INFORMATION FROM HIGH-THROUGHPUT GENE EXPRESSION DATA WITH PATHWAY ANALYSIS AND DECONVOLUTION

Maria Jaakkola

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. AI OSA – TOM. 668 | ASTRONOMICA – CHEMICA – PHYSICA – MATHEMATICA | TURKU 2022

University of Turku

Faculty of Science
Department of Mathematics and Statistics
Applied Mathematics
Doctoral Programme in Exact Sciences

Supervised by

Professor, Research Director, Laura Elo
Turku Bioscience Centre, University of Turku and Åbo Akademi University
Institute of Biomedicine, University of Turku

Reviewed by

Assistant professor, Joanna Żyła
Departament of Data Science and Engineering, Silesian University of Technology

Professor, Dario Greco
Faculty of Medicine and Health Technology, Tampere University

Opponent

Professor, Sorin Drăghici
Department of Compter Science and Department of Obstetrics and Gynecology, Wayne State University

ISBN 978-951-29-8845-7 (PRINT)
ISBN 978-951-29-8846-4 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)
Painosalama, Turku, Finland, 2022

To my grandmothers,
scientists at heart

UNIVERSITY OF TURKU
Faculty of Science
Department of Mathematics and Statistics
Applied Mathematics
JAAKKOLA, MARIA: Extracting information from high-throughput gene expression data with pathway analysis and deconvolution
Doctoral dissertation, 135 pp.
Doctoral Programme in Exact Sciences
March 2022

ABSTRACT

Modern technologies allow for the collection of large biological datasets that can be utilised for diverse health-related applications. However, to extract useful information from such data, computational methods are needed. The field that develops and explores methods to analyse biological data is called *bioinformatics*. In this thesis I evaluate different bioinformatic methods and introduce novel ones related to processing gene expression data. Gene expression data reflects how active different genes are in a set of measured biological samples. These samples can be for example blood from human individuals, tissue samples from tumours and the corresponding healthy tissue, or brain samples from mice with different neural diseases. This thesis covers two topics, pathway analysis and deconvolution, related to downstream analysis of gene expression data. Notably, this summary does not repeat in detail the same points made in the original publications, but aims to provide a comprehensive overview of the current knowledge of the two wider topics. The original publications focus on comparing and evaluating the available methods as well as presenting new ones that cover some previously untouched features.

While the terms 'pathway analysis' and 'deconvolution' have been used with alternative definitions in other fields, in the context of this thesis, pathway analysis refers to estimating the activity of pathways, i.e. interaction networks body uses to react to different signals, based on given gene expression data and structural information of the relevant pathways. I focus on different types of analysis methods and their varying goals, requirements, and underlying statistical approaches. In addition, the strengths and weaknesses of the concept of pathway analysis are briefly discussed. The first two original publications I and II empirically compare different types of pathway methods and introduce a novel one. In the paper I, the tested methods are evaluated from different perspectives, and in the paper II, a novel method is introduced and its performance demonstrated against alternative tools.

Many biological samples contain a variety of cell types and here, deconvolution means computationally extracting cell type composition or cell type specific expression from bulk samples. The deconvolution sections of this thesis also focus on a general overview of the topic and the available computational methodology. As deconvolution is challenging, I discuss the factors affecting its accuracy as well as alternative wet lab approaches to obtain cell type specific information. The first original publication about deconvolution (publication III) introduces a novel method and

evaluates it against the other available tools. The second (publication IV) focuses on identifying cell type specific differences between sample groups, which is a particularly difficult task.

KEYWORDS: bioinformatics, transcriptomics, pathway analysis, deconvolution

TURUN YLIOPISTO
Matemaattis-luonnontieteellinen tiedekunta
Matematiikan ja tilastotieteen laitos
Sovellettu matematiikka
JAAKKOLA, MARIA: Extracting information from high-throughput gene expression data with pathway analysis and deconvolution
Väitöskirja, 135 s.
Eksaktien tieteiden tohtoriohjelma
Maaliskuu 2022

TIIVISTELMÄ

Moderni teknologia mahdollistaa laajojen biologisten data-aineistojen keräämisen, joita voidaan hyödyntää lukuisin tavoin terveyden ja hyvinvoinnin edistämiseksi. Suurten datojen hyödyntäminen edellyttää kuitenkin laskennallisia työkaluja. *Bioinformatiikka* on tieteenala, joka testaa ja kehittää erilaisia laskennallisia menetelmiä käsitellä biologista dataa. Tässä väitöskirjassa tutkin, testaan ja esittelen uusia bioinformatiikan menetelmiä, joilla voidaan analysoida dataa, joka kuvaa geenien ilmentymistä, eli ekspressiota. Biologiset näytteet, joiden geeniexpressiota tutkitaan voivat olla esimerkiksi verinäytteitä ihmisyksilöistä, kudosnäytteitä syöpäkasvaimesta ja vastaavasta terveestä kudoksesta tai aivonäytteitä hiiristä, joilla on erilaisia neurologisia sairauksia. Tämän väitöskirjan kaksi pääaihetta ovat reittianalyysi ja dekonvoluutio, jotka ovat laskennallisia tapoja jatkoanalysoida mitattua geenien ilmentymistä. Alkuperäisjulkaisut ovat liitteenä johdannon jälkeen, eikä niiden sisältöä käydä johdannossa läpi yksityiskohtaisesti, vaan siinä esitellään reittianalyysia ja dekonvoluutiota laajemmin ja pyritän antamaan niistä kattava yleiskatsaus tämanhetkisen tiedon valossa. Alkuperäisjulkaisuissa menetelmiä testataan ja vertaillaan eri tilanteissa, ja esitellään uusia menetelmiä jotka täydentävät havaitsemiani puutteita menetelmätarjonnassa.

Termeille 'reittianalyysi' ja 'dekonvoluutio' on useita tulkintoja alasta riippuen, mutta tämän väitöskirjan yhteydessä reittianalyysi tarkoittaa kehon sisäisten vuorovaikutusreittien aktiivisuuden arviointia hyödyntäen geenien mitattua ekspressiota ja reittien rakenteita. Väitöskirjan johdannon reittianalyysia käsittelevän osan painopiste on tilastollisissa menetelmissä sekä niiden eroissa, toimintaperiaatteissa ja -edellytyksissä. Lisäksi reittianalyysin konseptin heikkouksia ja vahvuuksia esitellään lyhyesti. Väitöskirjan kaksi ensimmäistä alkuperäisjulkaisua I ja II liittyvät reittianalyysiin, ensimmäisessä testataan erilaisia reittimenetelmiä empiirisesti ja arvioidaan millaisissa tilanteissa erityyppiset menetelmät toimivat hyvin. Toisessa taas esitellään uusi reittianalyysimenetelmä, demonstroidaan sen toimivuutta ja verrataan sitä vaihtoehtoisiin menetelmiin.

Monet biologiset näytteet sisätävät erilaisia soluja ja tämän väitöskirjan kontekstissa dekonvoluutio tarkoittaa eri solutyyppien määrien tai solutyyppikohtaisen geenien ilmentymisen arviointia laskennallisin keinoin. Myös tämän väitöskirjan dekonvoluutiota käsittelevissä luvuissa keskitytään yleiskatsaukseen aiheesta ja siihen kehitetyistä laskennallisista menetelmistä. Dekonvoluutio on vaativa analyysi,

joten sen onnistumiseen vaikuttavia tekijöitä ja vaihtoehtoisia tapoja saada solutyyppikohtaista tietoa näytteistä esitellään myös. Dekonvoluutioon liittyvissä alkuperäisjulkaisuissa 1) esitellään uusi menetelmä ja verrataan sitä muihin olemassaoleviin työkaluihin eri tilanteissa (julkaisu III), ja 2) tarkastellaan miten hyvin eri menetelmät tunnistavat solutyyppikohtaisia eroja näyteryhmien välillä (julkaisu IV).

ASIASANAT: bioinformatiikka, transkriptomiikka, reittianalyysi, dekonvoluutio

Acknowledgements

Firstly, I would like to thank my supervisor professor Laura Elo for the great balance between helping and supporting me, and letting me do things in my own (sometimes suboptimal) way and grow as a researcher. Your example has left a lasting impact on how I approach research. Besides great supervising, I have also had a very supporting working environment. It has been a pleasure to be associated with two units, Turku Bioscience Centre and the Department of Mathematics and Statistics; I have always felt welcome in both. This has been mostly thanks to the friendly staff of both units; researchers have treated me as a member of the community and the administrative staff have been most helpful never considering my concerns and questions as 'the other unit's problem'.

Big thanks also to the Doctoral Programme in Mathematics and Computer Sciences (MATTI), the Otto A. Malm foundation, the Finnish Cultural Foundation, and the Alfred Kordelin Foundation for funding me. Being able to conduct research as full time work rather than an evening hobby has been crucial for the progress of this thesis. I'm grateful to my pre-examiners assistant professor Joanna Żyła and professor Dario Greco, and opponent professor Sorin Drăghici for taking the time and effort to help me finalise this dissertation.

It has been an honour and a pleasure to work in the Computational Biomedicine research group. Seriously, it has been interesting and fun. I would like to thank all the current and former group members for your kind and capable help, inspirational example, and entertaining coffee talk. Many of you caused my classification algorithm to struggle between classes 'friends' and 'colleagues', and all of you would have deserved to be named and acknowledged personally, but unfortunately the space won't allow it. Special thanks to Mehrad Mahmoudian for his patient tech support and endless nature trivia, Riku Klén and Mikko Venäläinen for being my team leaders (i.e. first targets for favour requests), and my co-authors An Le, Tomi Suomi, Fatemeh Seyednasrollah, Arfa Mehmood, Aidan McGlinchey, Tommi Välikangas, and Thomas Faux. I have had several opportunities to collaborate with great wet lab researchers as well. I would like to thank Riitta Lahesmaa and her group, as well as Michael Courtney and his group for showing me glimpses of the wet side of research. It has been a privilege to work with you.

Besides research related support, I have also received plenty of other forms of encouragement and good company. Therefore, I would like to thank my friends and

extended family. Aino, Ben, Arla, Arli, Petra, Narges & Mohammad, and the Harju family, you all introduce different interesting and fun things to my life, it is quite hard to come up with a topic that would be boring with you. I would also like to thank the Peltomäki family; Tytti, Tapio, Janne, and Sauli & Carita, it is always a pleasure to spend time with you and I have always felt accepted and welcome in your family. Sauli, I miss you. It hurts that my words will no longer reach you, but I will try to mimic your wide-scale curiosity towards life. Then I would like to express my gratitude to my immediate family. My parents Tiina and Heikki, while you have been proud and encouraging about my decision to go into research, it is a great relief to know that you would have been equally proud and encouraging if I would have chosen to do something totally different. You have got my back. Aleksi, Mirja, and Oskari, it is good to have siblings as, besides obvious reasons, you break my social bubble and introduce very different mindsets to my world. Finally, Jarkko, ”thank you” sounds insufficient after all your encouragement, companionship, and support over the years, ”I love you” covers it better.

October 22, 2021
Maria Jaakkola

Table of Contents

Abbreviations

API	application programming interface
csDEG	cell type specific DEG
csDMP	cell type specific DNA methylation profile
csGEP	cell type specific gene expression profile
DE	differentially expressed
DEG	differentially expressed gene
DNA	deoxyribonucleic acid
FACS	fluorescence-activated cell sorting
FCS	functional class scoring
FDR	false discovery rate
GEO	gene expression omnibus
GO	gene ontology
IPA	ingenuity pathway analysis
KEGG	Kyoto encyclopedia of genes and genomes
lncRNA	long non-coding RNA
ORA	over representation analysis
PBMC	periferial blood mononuclear cells
PCA	principal component analysis
RNA	ribonucleic acid
RNAseq	RNA sequencing
sc-RNAseq	single cell RNAseq
T1D	type 1 diabetes

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Maria K. Jaakkola and Laura L. Elo. Empirical comparison of structure-based pathway methods. Briefings in bioinformatics, 2016; 2: 336-345.

II Maria K. Jaakkola, Aidan J. McGlinchey, Riku Klén, and Laura L. Elo. PASI: A novel pathway method to identify delicate group effects. PloS one, 2018; 7: pages.

III Maria K. Jaakkola and Laura L. Elo. Computational deconvolution to estimate cell type-specific gene expression from bulk data. NAR genomics and bioinformatics, 2021; 1: pages.

IV Maria K. Jaakkola and Laura L. Elo. Estimating cell type specific differential expression using deconvolution. Briefings in bioinformatics, 2021; bbab433.

1 Introduction

1.1 General introduction

Modern technology enables the collection of very broad biological data. While such datasets hold considerable potential for great new discoveries, without suitable computational tools to analyse them, they are just collections of data. Bioinformatics aims to answer that need by providing tools to extract knowledge from such datasets, or to make predictions from them with black box approaches. Typical goals of bioinformatic applications include, but are not limited to, predicting the onset of a disease, identifying subtypes of a disease, predicting the likely response to a particular treatment, and learning how different molecules function and interact. These goals can serve multiple purposes, and for example identifying different subtypes of a disease can be utilised for either selecting the correct treatment for an individual with the disease, learning more about the disease and its mechanisms, or preparing for phenomena (e.g. strong symptoms, secondary disease, or long recovery time) related to the particular subtype. Thus, the underlying motivations are humane: better quality of life and expansion of human knowledge.

This thesis is about computational methods to extract information from large gene expression datasets. The two main topics, whose methodology I study, are pathway analysis and computational deconvolution. Pathway analysis is among the standard approaches to analyse gene expression data and it aims to estimate activities (or differences in them) of known biological pathways. Deconvolution aims to extract information regarding individual cell types that have contributed RNA to the measured samples. Both analysis types can be applied to all the scenarios introduced in the previous paragraph. The focus of this thesis is on comparing the available methods, introducing new ones that provide answers to the observed needs, and evaluating which features affect the performance of the methods and how accurate the results can be expected to be.

1.2 Goals and publications

The overall goal of this thesis is to assist researchers applying bioinformatic tools by a) providing new accurate and robust methods with a simple user interface, and b) comparing and evaluating the available methods so that the researchers can choose a method best suited for their particular dataset, and estimate how accurate the results

can be expected to be. The two analysis types covered here are pathway analysis and deconvolution. The detailed research questions are listed below:

Goal 1 Evaluate how accurate the current state-of-the-art pathway methods are, compare the methods that utilise pathway structures to those that do not, and investigate when the methods perform well and when not.

Goal 2 Develop a robust pathway method that utilises pathway structures and provides deregulation scores for all samples separately.

Goal 3 Test different expression deconvolution methods to estimate cell type specific expression profiles from bulk data.

Goal 4 Develop an expression deconvolution method that is robust against outliers.

Goal 5 Compare different methods to estimate cell type specific differentially expressed genes from bulk data containing several cell types, and investigate when the methods perform well and when not.

In publication I, we evaluated different pathway methods and particularly compared the approaches using pathway structures to those not using them (Goal 1). We observed that if the data contains large changes between the sample groups (cancer data was used as an example in the paper), methods using pathway structure at a rough level, namely SPIA [1] and Cepa [2], detected the same pathways from different datasets. In cases where the differences between the sample groups were subtle (type 1 diabetes (T1D) represented this case in the paper), none of the tested methods performed well. In publication II, we introduced and demonstrated the novel pathway method PASI, which provides sample specific pathway scores and performs reasonably well even with challenging data that contain only subtle changes (Goal 2). PASI was evaluated together with the two alternative tools and also the effect of some practical aspects, like sample size and uncertainty in pathway structure, were investigated. In publication III, we compared different expression deconvolution methods to estimate cell type specific gene expression profiles (csGEPs) from bulk data (Goal 3) and introduced the novel method Rodeo (Goal 4). In the comparison, Rodeo was particularly robust against outliers in the data. However, in cases of small sample size and heterogeneous sample donors, we observed that none of the tested methods performed well. Another key observation was that supervised methods outperformed the unsupervised ones. Finally, in publication IV, we tested different methods to identify cell type specific differentially expressed genes (csDEGs) from bulk data (Goal 5). We observed that methods designed for this task outperformed the general model and deconvolution methods designed to estimate csGEPs. Another important observation was that methods designed for methylation data also performed well with RNAseq data, assuming that their input requirements can be met. Besides comparing

the tested methods using simple gold standard data, we tested which aspects of the data affected the obtained accuracy. The most important observations were that 1) csDEGs from rare cell types cannot be reliably estimated from bulk data with any of the tested computational approaches, and 2) individual heterogeneity has a great impact on the accuracy of the estimated csDEGs. In the paper, we also explained how the end-user can evaluate the underlying individual heterogeneity of the bulk data.

1.3 Organisation of the thesis

This thesis consists of the summary part and the original publications attached at the end. The summary part is organised into five chapters: Introduction (this), Background, Pathway analysis, Deconvolution, and Discussion, followed by references. In the background section, the basics of study design, measuring gene expression, and typical downstream analyses are presented to give context for pathway analysis and deconvolution. The topic chapters for pathway analysis and computational deconvolution start with a topic-specific introduction covering key terminology and definitions, as well as some closely related themes. Different goals, input requirements, approach subtypes, and other relevant aspects of the topic are then discussed in the middle part. These general sections are followed by sections about available computational methods, and finally, the topics are briefly summarised. In the Discussion chapter, the general conclusions are presented together with challenges, limitations, impact, and further research related to this dissertation.

2 Background

2.1 Study settings and data types

Biological data has an enormous amount of health-related applications and the techniques to obtain it have improved drastically over the past few decades. As the quantity and size of available biological data grows with increasingly accurate and affordable technologies, the more important the computational approaches to analyse it become. The growth of sample size has led to the possibility of using sophisticated statistical approaches to extract information from the data. Different biological samples to be analysed can be taken from the donor individuals, and blood from human individuals is the most common, though not the only, sample type in this thesis. In most experiments the sample type to measure is selected based on the expected relevance to the condition of interest, but sometimes practical aspects drive the sample type selection. For example, it is very difficult to take tissue samples from certain organs (e.g. pancreas) without harming the sample donor, which often leads into utilising other tissues or species other than human. Another example is studies aiming to identify predictors or markers that can be used in large scale screening. As the intended application of the results demands large scale sample extraction, the selected method should be easy and cost efficient, which favours samples like urine or blood. As individuals are heterogeneous, samples from many donors are needed for statistical conclusions about the condition of interest. The sources of heterogeneity can include known factors that can be controlled, such as age, gender, and ethnicity, but also those that are hard to regulate without considerable effort as well as unknown ones. It is a common study setting to compare individuals with the condition of interest to those without it. In such a setting of two sample groups, the samples from individuals with the condition are called *case* samples and the samples from individuals without the condition are called *control* or *reference* samples. The issue of known sources of individual variation can be controlled with *paired* or *matched* samples, which means that for every case sample there is a control sample from an individual with the same age, gender, etc. Also more than two sample groups can be compared, which is common in studies involving e.g. multiple subtypes of a disease or different treatments to a condition.

Proteins are folded amino acid chains that perform a wide range of tasks in a body. They are constructed in protein synthesis (Figure 1), in which the sections of

DNA corresponding to the protein (genes) are coded into a messenger RNA strand, which is used as the instruction to build a chain of amino acids that will then fold into a 3D structure characteristic of that protein. RNA strands are also called *transcripts*. Protein synthesis and its regulation is a complex process and it contains far more steps and molecules than described here. A gene is said to be *expressed* if RNA is prepared based on it. Studies of the presence of different genes, RNA, and proteins are called genomics, transcriptomics, and proteomics, respectively. These, and other *omics*, are used to study molecular changes in a body under different circumstances. Other omics include, but are not limited to epigenomics, metabolomics, and lipidomics. The focus of this thesis is on transcriptomics, i.e. gene expression, and here the term RNA always refers to messenger RNA despite the existence of other types of RNA.

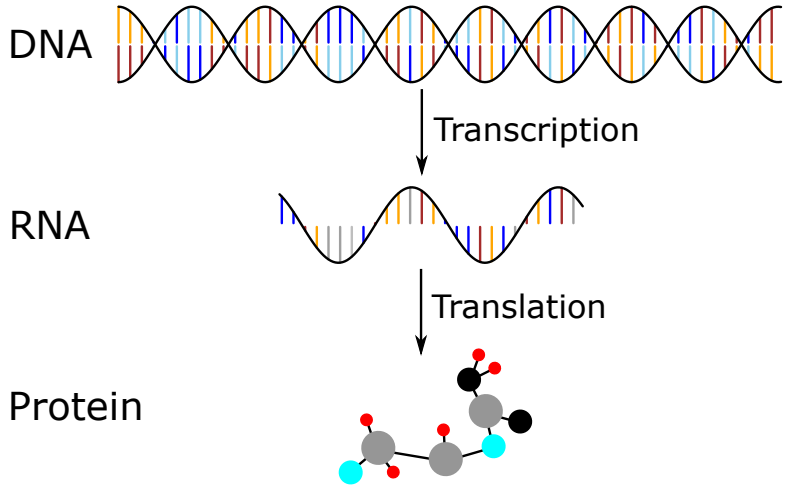

Figure 1. In protein synthesis genetic information (DNA) is first copied into an RNA strand in the nucleus. Then the RNA is transported into a ribosome, where it is used as instructions to construct an amino acid chain. Finally the produced amino acid chain folds into a 3D structure (protein).

2.2 Two approaches to measure gene expression

High-throughput gene expression data is one answer to the need for large scale data to do statistical analysis on multiple genes and samples. *High-throughput* means that the measuring device measures/analyses multiple features (genes associated with the RNA) from multiple samples at once. Gene expression data consists of numeric values reflecting the amount of RNA originating from different genes, i.e. describing how actively the genes are read for protein synthesis. There are two main high-throughput techniques to obtain gene expression data, *microarrays* and *RNA sequencing* (RNAseq). The rough idea of microarrays is to build a chip, whose parts

bind to only certain RNA (or artificial single stranded DNA constructed from it) and use it to estimate to what extent different RNA fragments are present in a sample. As the name suggests, in RNAseq the RNA strands present in samples are sequenced and mapped to the corresponding genes. Microarrays are an older technique than RNAseq, and while most of the new data is from RNAseq, a considerable proportion of publicly available data is still from microarray studies. RNAseq has some advantages over microarray experiments including 1) lower background noise enabling more accurate detections including those from low-expression transcripts, and 2) the possibility to detect new transcripts as sequences are not dependent on reference genomes built into the technology [3; 4; 5]. The lower noise level of RNAseq leads to higher number of observed altered expression levels as detections of low abundance transcripts are possible, and more dynamic coverage, i.e. a wider range of measured expression values [4]. Additionally, the possibility to detect transcripts that are not part of the pre-defined reference genome enables the differentiation of biologically critical isoforms and the identification of genetic variants [3; 4; 5]. Both microarrays and RNAseq include several experimental and computational steps to form the final estimates of gene expression levels.

2.3 Downstream analysis of gene expression data

There are many ways to further analyse the obtained high-throughput gene expression data. The simplest approaches aim to identify genes whose expression level reflects some property of the samples. These can be genes that are differentially expressed (DE) between sample groups, or genes that correlate with a sample property like disease severity. Detecting differentially expressed genes (DEGs) is a popular approach whenever the study includes sample groups and many methods have been developed to do it [6; 7; 8; 9; 10; 11]. If the sample size is too small to provide statistical power for an actual DE test, a simple fold change between the sample groups is typically reported.

Besides evaluating each gene separately, their combinations can be utilised for more sophisticated analyses. Different model fitting approaches, machine learning methods, and analyses of gene groups are common examples of such analyses. The gene groups can be *pathways* (reaction and regulation networks that a body uses to transport information or to react to changing situations), gene neighbourhoods (closely located co-expressed genes), or merely sets of genes associated with the same biological functions and conditions in the literature. Many gene group analyses can be utilised to generate new features, which can contain valuable insight themselves and/or be used for further basic analyses like differential expression. These features can have diverse interpretations, such as a score predicting/estimating something (time to diagnosis, survival time, sensitivity to side effects...), activity of a pathway or other biological process, or some property of the measured tissue sample

itself. Two examples of such tissue properties are the purity of a tumour sample and the cell type composition (deconvolution) if the sample contains RNA from different types of cells. In most of the analyses of either measured genes or some generated features, statistical significance of the findings is reported as p-values or as false discovery rates (FDR) if the multiple test correction is required.

3 Pathway analysis

3.1 Introduction to pathway analysis

3.1.1 Basic terminology and definitions in pathway analysis

In the context of pathway analysis, the terms *pathway*, *network*, and *gene set* are related and here I describe their differences. As the name suggests, gene sets are lists of genes associated with the same function. Gene sets are very general and they don't include any topological information about how the genes are related to each other. However, in practice the identification of gene sets that are enriched with DEGs is usually referred to as a form of pathway analysis [12], despite the slight inaccuracy. The difference between a network and a pathway is more vague as they both contain interactions between biological units, and there is no exact definition available. However, pathways are usually rather well known and validated in literature and often contain detailed information about their interaction mechanisms, whereas the support for network structure can be much lighter, like large-scale screening [13]. Many pathway methods ignore plenty of the sophisticated information about pathways and they could also be called network analysis methods. Here I systematically use the term 'pathway' instead of 'network' or 'gene set' despite the level and confidence of structural information, as is often the case in the literature when different approaches are discussed from a wider perspective [12; 14; 15].

Pathways consist of *nodes* (also called units or sometimes vertices) and *interactions* (arcs, links, or edges) between them. The network structure of these pathway components is called *pathway topology* or *pathway structure*. The nodes can be for example genes, transcripts, proteins, protein complexes, chemical compounds, or even other pathways. If they are some kind of gene products, they are usually annotated by the genes encoding them. Interactions are typically either reactions or inhibiting or activating relations. The level of available details about nodes and interactions vary based on the source of the pathway information (see Section 3.1.3) and different pathway methods can also either dismiss or require certain information. Especially for interactions, the level of available details varies a lot between sources. In the simplest form they are just undirected links between nodes, but they can also contain detailed information such as regulatory mechanisms.

In this thesis I focus on computationally estimating behaviour of whole pathways based on high-throughput transcriptomic data, but also briefly introduce several re-

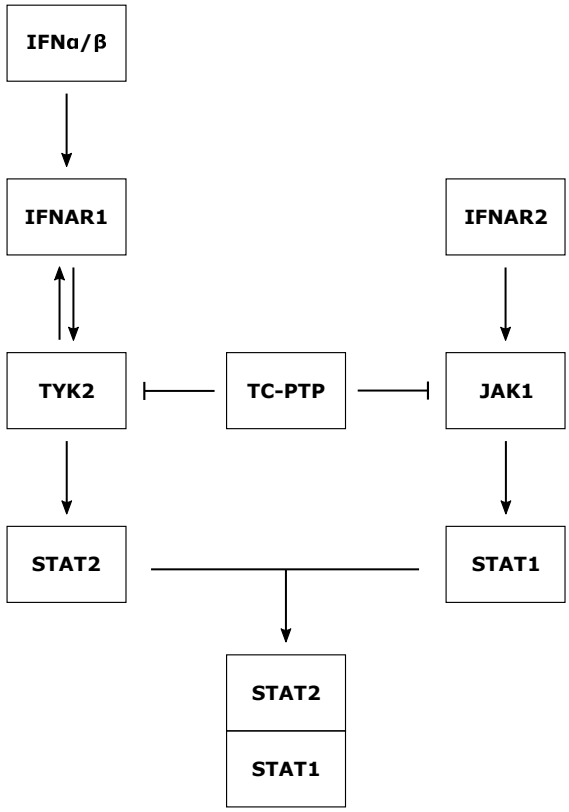

Figure 2. Part of Interferon signaling pathway as an example of a pathway. The structure is simplified from the representation in IPA database. Activating interactions are usually marked with arrows and inhibiting ones with bar-headed lines.

lated topics, such as the analysis of subpathways [16; 17] and multi-omic pathway analysis. An important subtype of pathway analysis not introduced in this thesis is genome wide association studies which often focus on single nucleotide polymorphisms [18; 19; 20; 21]. Rather than gene expression levels, they focus on investigating pathway-level differences in the genome itself, which are of interest especially in cancer studies [13].

3.1.2 Branches of pathway analysis

Pathway analysis is a wide topic with multiple different branches that include, but are not limited to

1. coarse estimation of pathway activity utilizing high-throughput data,

2. detailed mechanistic modelling of a single pathway,

3. study of pathway structures.

Of these branches, the estimation of pathway activities based on high-throughput data is the one covered in this thesis, and the other two are only briefly introduced here. However, a strict classification of different types of pathway analysis is not meaningful as different branches often overlap. For example, a study of pathway structures can include mechanistic modelling [22], and an unknown molecule altering a pathway can be considered as a change in either pathway activity or structure [23].

In coarse pathway analysis, pathway activities or differences in them are estimated for a wide range of pathways based on high-throughput input data using different computational methods. While this approach is less detailed and accurate than mechanistic modelling, it provides insight to the bigger picture as multiple different pathways related to different biological processes are analysed at once. Coarse statistical analysis is also more straightforward for the end user than mechanistic modelling as the analysis is well automated and there is no need for estimating different parameters or concentrations. Strengths and weaknesses of this branch are introduced in Sections 3.2.1 and 3.1.4, respectively.

While approaches based on high-throughput data typically cover plenty of pathways on a very general level, mechanistic modelling focuses on one pathway in detail. In mechanistic modelling each interaction of the pathway, such as transportation over a membrane, binding with other molecules, or activation by dephosphorylation, is modelled using known or estimated concentrations of present molecules and reaction rate parameters. This leads to a set of differential equations that may or may not be analytically solvable, depending on the reactions involved in the model. Typical objectives of mechanistic modelling include investigating why a pathway behaves unexpectedly under a certain condition, studying quick phenomena difficult to capture with high-throughput means, and supporting hypotheses of regulatory mechanisms. If the pathway under consideration is large and/or contains complex regulation structures, mechanistic models can become computationally very heavy as the size of the set of differential equations increases. However, there are implementations, like cupSODA [24], to tackle this issue. Practical benefits and challenges of mechanistic modelling are introduced in summary paper [25]. Mechanistic models are sometimes called kinetic models in the literature.

The study of pathway structures covers topics such as the identification of previously unknown pathway structures, the completion of the already known ones through the discovery of new interactions or nodes, and the gathering of additional information of for example regulatory mechanisms of interactions [26]. This branch includes a wide range of studies from detailed wet lab research into molecular interactions (e.g. [27]) to automated computational methods to infer the most likely interactions from high-throughput omic data (e.g. [28]). As pathway structures may vary

according to biological condition [29], studying pathway structures under different circumstances is common. Cancer signalling is among the most studied examples of this [30; 31].

3.1.3 Availability of known pathway structures

Pathway analysis requires information about the pathways which can be found from databases collecting it. There are plenty of heterogeneous databases for pathway structures and network interactions and Table 1 lists some of the most commonly used. Meta database pathguide.org [32] introduces hundreds of available pathway databases and there are also multiple studies introducing, summarising, and comparing them [32; 33; 34; 35; 36]. Pathway databases are sometimes called knowledge bases in the literature. Besides databases for structural information, there is also a wide range of other types of assets available, such as visualisation tools, but they are out of the scope of this thesis.

Table 1. Several available pathway/network structure databases

Name	Comment	Reference
Biocarta		[37]
BioGRID		[38]
DAVID	Collects data from other databases	[39; 40]
Ingenuity	Commercial	
IntAct	includes also MINT [41] interactions	[42]
KEGG		[43]
NDEx	Includes also the content of a discontinued NCI-PID [44]	[45]
NetPath	Focus on immune and cancer signaling pathways	[35]
MetaCyc		[46]
Omnipath	Several resources merged into one enormous network	[17]
PANTRHER		[47]
Pathway Commons	Collects data from other databases	[48]
Pathway Studio	Commercial	
Reactome		[49; 50]
WikiPathways		[51]

Each of the databases have their own features and scopes and, therefore, the choice of pathway database is important [52]. Common differences between pathway databases are related to criteria for inclusion and how the data is collected, stored, maintained, and accessed. Criteria for inclusion can concern either whole pathways or their parts. For example, a database might be specialised to certain species or pathway types, like signalling or metabolic pathways. Parts of pathways to include can be selected based on e.g. interaction type and accept only for example protein-protein interaction. Further details of more technical differences include

- **collection:** Pathways can be manually curated from literature, extracted from

other available databases using some selection criteria, or computationally inferred from data.

- **storage:** The available databases differ from each other based on what they store (e.g. directions and mechanisms of regulation interactions available) and how they store it. There are a few commonly used standard languages, such as SBGN [53] and BioPAX [54], to encode the pathway information, as well as several comparisons and summaries about them [55; 56].

- **maintenance:** Some pathway databases are updated frequently, even weekly in the case of KEGG, whereas others are unmaintained making them obsolete while still accessible.

- **access:** The accessing of pathway information either by manually browsing for it online, or computationally via API can be implemented differently. There are also different criteria for access (none, registration, or pay-wall).

Due to the numerous differences between databases, the same pathway might be annotated slightly differently. In this thesis, KEGG is selected as the primary source of pathway structures as it provides pathways in directed format (i.e. the direction of a regulatory relationship is indicated), it is frequently maintained and updated, and it allows a user to easily access its content via API. The KEGG pathway database started as a database for metabolism pathways and, while nowadays including a wide range of other pathways, still includes an especially large subset of them.

3.1.4 Challenges in pathway analysis

Pathway analysis encounters difficulties from multiple sources. These can be related to the pathway methods and their usage, to the limitations of the available input data, to the pathways and databases annotating them, or to the very concept of pathway analysis. Here I introduce several pathway, database, and input data related challenges. Difficulties related to pathway methods and their validation are discussed in Section 3.3 together with other methodology related topics. In the literature, challenges in pathway analysis and especially the weaknesses of different types of methods are often discussed in review studies [36; 12; 34; 57], and Kelder et al. present wider context for pathway analysis and its usage [58].

One of the main issues with pathway analysis is that many of the known pathways are tissue specific so it is questionable how reliably changes in their activity can be detected from other measured tissues. For example, the pathway 'Salivary secretion' (KEGG id hsa04970) is located in the salivary glands. Can its activity be detected from blood samples? While blood transports many molecular products, the RNA to construct those molecules (i.e. the measured part) is not necessarily visible in it. It is straightforward to leave out pathways whose parts are not expressed in the input

data, but unfortunately the situation is usually not that simple as many of the genes in an unrelated pathway are expressed in the measured tissue, they simply play a different role in it. In small pathways these alternative roles of genes in the measured tissue may result in false positive findings. Larger pathways are more resistant to this issue as it is unlikely that all/most of their genes would a) have another role in the measured tissue, and b) those roles would be systematically up or down regulated. Instead, if the behaviour of a big pathway is altered, but the change is not visible in the measured tissue, it will go unnoticed. However, this is a minor issue as the conclusion that the pathway is not differentially expressed in the measured tissue is still true. The researcher interpreting the results should bear in mind that this does not indicate that the pathway could not be altered in the tissue it is located. To our knowledge, none of the existing pathway methods process any filtering of pathways that cannot be detected reliably from the measured tissue and there is not much public discussion about the topic either. However, some methods generally exclude very small pathways from the analysis due to their overlap with bigger pathways and statistical instability. Reimand et al. summarise different issues related to very large or very small pathways in their study [57].

Another challenge is that results of protein-protein interactions are not visible at the transcriptomic level. For example, if protein A destroys protein B, RNA expression level of protein B remains unaffected. However, in the pathway analysis process this may falsely look like interaction $A \dashv B$ is inactive. Analysing proteomic data or combining it with transcriptomic data avoids this issue. While some pathway databases, such as KEGG, provide detailed information about the interaction level (protein-protein in this example), many pathway methods do not utilise it. Selection of interaction types to use is not obvious as it means balancing between using only interactions (and nodes) whose activity can be reliably estimated from the data, and on the other hand, not wasting a vast amount of structural information. Another issue related to transcriptomic data is its static nature; as transcriptomic data reflects the amount of RNA present at the moment of extracting it, it is unlikely to capture quick and dynamic processes. This issue is not limited to pathway analysis, but is present in all analyses using transcriptomic data.

Pathways and databases providing them can also include some caveats besides the obvious issues related to unknown, altered, or erroneous pathway structures. As mentioned before, the structure of a pathway can vary from database to another. One explanation for this is related to difficulties in defining limits of a pathway [52]. As even the most downstream nodes tend to further regulate some other molecules, it is non-trivial to define where the pathway ends. If ambiguity in pathway limits is a problem, one approach to tackle this is to define huge meta pathways, such as Omnipath [17], and then focus on subpathway analysis. Another reason for differences in pathway structures between databases is that the pathways need to be collected from diverse original studies discovering them, which can be tedious work and, due to

major differences between the original studies, involve subjective decision making. Besides pathway structures depending on the database, the annotation coverage could also be improved. Despite the wide selection of known pathways, only a minority of commonly measured genes (or expression associated with them) are annotated in any given pathway [48].

3.2 Objectives and approaches in pathway analysis

3.2.1 Motivation and goals

Pathway analysis has several advantages compared to investigating individual genes. Most of these are related to three main benefits:

1. **Robustness**, as pathways consist of multiple nodes, they are more robust against random variation than single genes [59]. As pathways summarise bigger functional units, they can also be detected even if different parts of them would be altered in different samples, making the gene-level observations too unsystematic to be identified as interesting findings.

2. **Reduction of data**, if thousands of genes are identified as DE or otherwise interesting, it is hard to interpret such a high number of findings. Pathway analysis reduces the amount of findings to a more understandable scale. This is particularly important in cancer studies as they often involve drastic changes in gene expression [13].

3. **Insight into the underlying phenomena**, as pathways reflect biological processes, the detected pathways can hint what is going on in the samples in the bigger picture, better than individual genes.

In addition, pathway analysis allows for combining different omics [60], which is discussed separately in subsection 3.2.3. Challenges and weaknesses of pathway analysis were discussed separately in Section 3.1.4

Typical goals of pathway analysis are related to detecting differences between samples or sample groups, or identifying the underlying biological processes. The simplest example is the identification of pathways behaving differently between case and control samples and then further investigating the potential causes and sources of those differences, based on which pathways are altered in case samples. If samples are analysed separately rather than in case and control groups, their pathway scores can be used to classify or cluster samples. In the case of classification, identifying pathways that tell the sample groups apart can provide interesting insight into the classes (such as different subtypes of breast cancer [61]). On the other hand, such subtypes of a condition could be identified from clustered samples. Another typical application is detecting pathways whose altered activity could predict something

(e.g. cancer relapse, severe disease symptoms, drug resistance, or development of a disease). In practice this means comparing samples from donors who will get the condition of interest in the future to those from donors who won't.

3.2.2 Different interpretations of the output values

A pathway method typically provides either a list of DE pathways or, more commonly, pathway scores possibly with their significance levels. These scores can be provided for each sample separately (*sample-level* analysis) or for sample groups (*group-level* analysis). In group-level pathway analysis the aim is to investigate which pathways have differing activity between tested sample groups, whereas in sample-level analysis the estimated pathway values are obtained for each tested sample. Typically, group-level pathway analysis tools require a gene expression matrix and group labels as input and they return an output value for each pathway. The input for sample-level analysis is similar to group-level except group information is not always required. However, the typical output is a matrix of pathways and samples. Group-level methods are far more common, but nowadays multiple sample-level pathway methods are also available [61; 62; 63; 64; 65; 66]. Sample-level analysis can be preferred in several cases, if for example the study does not include sample groups, the condition is known to have high individual heterogeneity, or there are sample subgroups that would go unnoticed if the samples are somehow pooled. On the other hand, group-level analysis results are more straightforward to interpret as they do not necessarily require any further downstream analysis. When constructing the group-level scores the underlying gene expression needs to be summarised either before the pathway analysis, in which case the input can be for example a list of DEGs, or during it. Ackermann et al. provide a summary of different early approaches to construct group-level scores [67].

The output values can represent two different properties, *deregulation* or *activity*. Deregulation scores (e.g. [61; 1]) reflect the normality of the pathway behaviour and they usually require control samples to define the normal level. Besides the method-specific deregulation scores, some measure of statistical significance, such as p-value or FDR, is usually returned so that a user can evaluate which pathways' behaviour is significantly altered in case samples. As the name suggests, an activity score describes how active the pathway is. Usually, activity scores are calculated by summarising scaled expression values of pathway nodes [68; 59]. The methods differ from each other by the summary and scaling techniques used. The problem with these types of approaches is that they assume all genes in the pathway to be highly expressed when the pathway is active, which is not the case with different inhibiting structures present in many pathways. In our own pathway method PASI (publication II), this pitfall of activity score is avoided by estimating if each node should be highly or lowly expressed when the pathway is active, and the measured

values are multiplied by +1 or -1 accordingly. Also some other methods utilising pathway structure, for example SPIA [1], overcome the issue. Both score types can be used in sample-level and group-level pathway analysis. In general, deregulation scores are more common than activity scores.

Besides analysis of whole pathways, methods for subpathway analysis have also been developed [69; 70; 71; 72; 73]. There are several biological questions to motivate subpathway analysis. In cases where a disease or other biological condition alters some part of a pathway, the downstream pathway will also be altered. With subpathway analysis it can be identified which part of a pathway is associated with the initiation of the disease and which are just progressions of that [74]. Another example is analysis of large pathways, where the sizes of the pathways make it harder to do biologically accurate interpretations from whole-pathway results. This issue is common with metabolic pathways [75].

3.2.3 Special input data types: multi-omics and single cell

Despite transcriptomic data being on the focus of this thesis, each omic has its own strengths. For example, as proteins are the actual functional units in a body, their amount and activities can reveal more to-the-point information of ongoing processes than gene expression. Due to reasons like destruction of ready proteins and their post-translational modification, gene expression levels do not indicate ready protein levels as well as one might expect [76; 77; 78]. On the other hand, proteomic data is harder to measure, especially for small proteins. Combining different levels of omic data has provided interesting insight [79; 80; 81], though it is experimentally demanding. While integrating multiple omics has several advances, most of them are related to the potential to identify and understand causality relationships behind the biological condition of interest instead of only correlations [82]. This is especially beneficial in studies involving complex and heterogeneous phenomena like cancer [83] or toxicology [84]. There are several databases providing multi-omic data, such as The Cancer Genome Atlas (TCGA), Cancer Cell Line Encyclopedia (CCLE), Alzheimer's Disease Neuroimaging Initiative (ADNI), and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC).

Pathway analysis naturally fits into integrative multi-omic analyses [60] as the interactions in a pathway can cross different omics (e.g. proteins might regulate transcriptomes). Despite that, there are only limited selection of pathway methods easily allowing combining different omics, as pointed out by Yan et al. [60]. However, some methodology is already available. For example, PARADIGM [85] is among the first multi-omic pathway methods and it is based on Bayesian approach, ActivePathways [86] is a recent method with finalised implementation, and MOSClip [87] combines different omics for pathway and survival analysis by utilising pathway topology. Some of the approaches (e.g. PARADIGM) utilise the detailed pathway

information and associate different omics with the corresponding details of the pathway structure, whereas some approaches (e.g. ActivePathways) map each omic to the same pathway genes so that each gene in a pathway has multiple values that can be further processed in different manners. Review by Subramanian et al. [88] summarises several aspect, including pathway analysis, of integrating different omic data.

While multi-omic approaches combine several omics, possibly including transcriptomics, there is also an interesting special case of transcriptomic data that should be introduced separately. In single cell RNA sequencing data (sc-RNAseq) RNA is extracted and measured from each cell separately, which results into very large datasets. It has been shown that cell populations are heterogeneous [89; 90] so fairly many cells are needed to get an overview of the cell population. Minimum of 50 cells per cell type has been suggested in the literature [91], but the heterogeneity of the tissue (i.e. number of cell types present) and presence of rare cell types affects the required cell count. This combined with the need of many individuals to make statistical conclusions about biological conditions causes the total number of cells to analyse to be very high. However, if the focus of a study is to compare cell populations to each other rather than biological conditions (e.g. case vs control samples study setting), the number of sample donors is usually lower. The most important difference to bulk RNAseq is that one cell does not express nearly as many genes as all the cells present in a sample combined. Therefore, in sc-RNAseq data there is considerable amount of zeros. As sc-RNAseq has become a popular experiment, number and size of publicly available datasets are growing rapidly [92].

There are several pathway methods designed for sc-RNAseq, such as PAGODA [93] which builds on SCDE [94] error model for single cell data, AUCell which is part of SCENIC R package [95] and provides activity scores, iDEA [96] which does simultaneous DE analysis and pathway analysis, UCell [97] which is a computationally light method based on Mann-Whitney U statistic, and function addModuleScore [98] from a popular sc-RNAseq tool kit Seurat [99]. Notably, none of these approaches utilise pathway topology. In addition, pathway methods providing deregulation scores based on for example input list of DEGs can be used similarly with sc-RNAseq data, but it is up to the user to define the input. It can be more challenging than with bulk RNAseq data including straightforward case vs control study setting as the definition of the sample groups does not follow immediately from the study design. As the average number of cells analysed per dataset is increasing [92], practical aspects related to implementation and running time are becoming more and more important.

3.3 Pathway methods

Table 2 includes several available pathway methods with implementation freely available at the time of writing (January 2020). Therefore, several methods have been excluded due to reasons like lack of implementation, implementation no longer available (broken links etc.), or the software is commercial or otherwise limited access. Also methods for subpathway analysis and those for other omic data than transcriptomics have been excluded. Many of the pathway methods listed in Table 2 are implemented in R packages `graphite` [100] and `ToPASeq` [101] in addition to their original implementation.

Table 2. List of several available pathway methods for transcriptomic data

Name	Speciality	Reference
ABP	For longitudinal analysis, uses GSVA for basic pathway analysis	[102]
ACST	Based on consistent subpathways	[103]
ActivePathways	Multi-omic method	[86]
BPA	Bayesian networks	[104]
CAMERA	Adjusts gene set test statistics with inter gene correlations	[105]
Cepa	Nodes are weighted based on pathway topology	[2]
CERNO	Fast ranking-based approach for gene sets	[106]
Clipper	Analyses also subpathways	[107]
DART	Prunes pathways in order to reduce noise	[108]
DAVID	Modified Fisher's exact test, multiple pathway databases	[109; 40]
EasyGO	Supports Affymetrix GeneChips of farm animals	[110]
EGSEA	Ensemble method for gene set analysis	[111]
EnrichNet	Combines analysis and visualisation	[112]
GANPA	Uses GSEA, but adds topology based weights	[113]
GGEA	Evaluates consistency of pathway nodes	[114]
GLOBALTEST	Evaluates significance of a gene set and focuses on visualisation	[115]
GOMiner	Enrichment of GO terms based on DEGs	[116]
GOstat	Provides sorted list of GO terms enriched with DEGs	[117]
GSA	Modified version of GSEA	[118]
GSEA	Utilises Kolmogorov-Smirnov-like rank statistic	[119]
GSVA	Utilises Kolmogorov-Smirnov-like rank statistic	[120]
iPANDA	Identifies biomarker pathways using co-expression and gene importance	[121]
iPAS	Sample-specific deregulation scores designed for cancer studies	[64]
mitch	Multi-contrast gene set enrichment	[122]
NEA	Evaluates the number of functional links between DEGs and gene sets	[123]
NetGen	GO term combination-based functional enrichment analysis	[124]
NetGSA	Models gene expression as a function of other genes in the network	[125; 126]
PADOG	Weights genes unique to the pathway	[127]
PAGODA	Designed for single-cell data	[93]
PARADIGM	Models pathways as factor graphs	[85]
PASI	Provides sample-specific deregulation or activity scores	[62]
Pathifier	Sample-specific scores based on principal component analysis	[61]
PathNet	Utilises connectivity information within and between pathways	[128]
PathOlogist	Calculates consistency and activity for each interaction	[129]
Pathway-express	Updated and maintained in R package ROntoTools [130]	[131]
PerPAS	Sample-specific pathway scores based on topology-weighted nodes	[132]

PWEA	Utilises topology and gene–gene correlations	[133]
pypath	Developed together with pathway resource Omnipath	[17]
SAM-GS	Gene set expansion of SAM analysis for single genes	[134]
Singscore	Sample-specific scores that are stable on small data	[135]
SPIA	Scores calculated utilising p-values and pathway topology	[1]
ssGSEA	Sample-level version of GSEA	[66]
TAPPA	Identifies phenotype-associated pathways utilising pathway topology	[136]
TBScore	Nodes are weighted based on number of significant down-stream nodes	[137]
TopoGSA	Focuses on topology and visualisation	[138]

3.3.1 Generations of pathway methods

The first pathway methods were published soon after microarray analyses became popular at early 2000 [139; 140; 141; 142]. While pathway methods are diverse and can be classified using multiple criteria, there are three main generations of methods, where newer generation attempts to address the issues of the previous one. However, within-generation variation of the methods is high especially in the second and the third generation. Reviews [12; 36] introduce the three generations of pathway methods together with their most outstanding limitations and [67; 143] provide a thorough reviews of the first two generations. In this summary, I utilise the terminology used in [12].

Methods for over representation analysis (ORA) represent the first generation of pathway methods. They mostly have similar working principle: investigate if pathways (from databases) include more differentially expressed genes than expected by random. The methods differ from each other by details like definition of differentially expressed genes and how they access and store the gene sets representing pathways, but the underlying tests are similar. Typically the tests are based on hypergeometric, chi-square, or binomial distribution. However, other hypotheses such as so called 'self-contained null-hypothesis', which assumes that a pathway contains no differentially expressed genes can also be tested. For example Garcia-Compos, Ackermann, and Maciejewski discuss different null-hypotheses in their review studies [36; 67; 144]. Early ORA approaches can be utilised only for group-level analysis.

As the first generation methods classify genes into two binary categories (DE or not), the level of differential expression is ignored and slightly, but systematically, altered pathways are likely to go unnoticed. The second generation of functional class scoring (FCS) methods aims to address these problems. First, a statistic value, like fold change or t-statistic, is assigned to each gene. Then these values can be scaled/transformed to increase robustness or to assure that both up- and down-regulated genes get to contribute. Finally, these gene-level values are summarised into pathway values, whose significance is then calculated. The first FCS methods were published in 2003 [145; 146].

The third generation of pathway methods utilises pathway topology. In its sim-

plest forms, the approach is otherwise very similar to FCS, but when summarising gene-level values into the pathway scores, different genes are weighted based on some topological value, like number of nodes connected to the gene or number of paths going through it. Other approaches include calculating values also for interactions and considering them as pathway units as well, nodes inheriting expression from other nodes according to the interactions, and utilising the whole pathway structure at once as in Bayesian graphs (though they are more commonly utilised when the goal is the identify pathway structures [147; 148; 149]). The first statements about pathway topology methods were published already in 2004 [150], but the first implementation came in 2007 [131]. The main issue with pathway topology methods is that the pathway topology is not always known and there has been statements that it could also change due to factors like age or disease [29].

3.3.2 Working principles of pathway methods

All pathway methods need to somehow access pathway information, map the measured input values into the pathways, and then summarise them into pathway scores. Many methods also somehow process the input data before and/or after it has been mapped into the pathways. How and which of these steps are done varies between and within method generations. Different steps of pathway analysis have been summarised also in the literature [36; 151].

Accessing pathways Pathway information, whether it includes pathway structures or not, can be obtained as an input from a user, as built-in knowledge, or by automatically connecting to some online pathway database via API. Especially in the last case when potentially very detailed information is available, the selected pathway method extracts only the level of information it utilises. For example, KEGG database includes interaction type 'indirect effect', and some methods might utilise only direct interactions, whereas others could utilise all of them.

Preprocessing the input data Common examples of data processing are scaling the data so that different genes and samples are comparable with each other, summarising samples within sample group in case of group-level analysis, extracting DEGs or calculating fold changes in case of FCS method, and filtering out lowly expressed genes (i.e. unreliable measurements).

Mapping input data to pathways Measured gene expression are mapped to the pathway nodes according to the gene ids associated with input transcriptomics and pathway nodes associated with gene products. If a pathway includes plenty of nodes not related to gene products (e.g. chemical compounds), or otherwise not measured, the final pathway scores are likely to be unreliable. This is a

challenge for all available pathway methods and some of them (e.g. PASI in original publication II) filter out pathways with too few measured nodes.

Processing values mapped to the pathways Pathway associated values can be further processed by for example weighting them according to some pathway topology related criteria [2], calculating values for interactions using node values [62], or updating the node values based on their neighbour nodes [1].

Calculating pathways scores The final pathway scores can be either values describing activity or deregulation (see Section 3.2.2) or simple statistical metric describing significance (p-value or FDR). The latter is common especially with ORA and FCS methods (see Section 3.3.1). To obtain activity or deregulation scores, the possibly processed values mapped into the pathway need to be somehow summarised. The simplest way to do it is to calculate their mean [85]. On the other hand, for example Pathifier [61] uses a very sophisticated pathway deregulation score defined as distance to control samples along principal curve in multidimesional space set by pathway nodes.

Notably, some methods (especially early ORA and FCS methods) might require some steps like processing of the input data, but instead of doing it internally, it is defined as a criteria for input data so that a user needs to do it before applying the method.

3.3.3 Evaluation of available methods

As pathway activity can not be directly measured, validating pathway analysis results is difficult. Several approaches have been proposed and utilised and they all have their own strengths and weaknesses. The commonly used strategies to validate pathway analysis results include

- Simulated data
- Knockout genes
- Literature review
- Target pathways
- Sampling the input data
- Similar datasets
- Sample classification
- Number of detections from real data as compared to dummy data

When simulating data to create a gold standard, measurements from selected genes are intentionally altered so that it is known that certain pathways including those genes should be detected [15]. The issue with simulated data is that assumptions of the person creating it affect the data and an end user can never be sure how a method would perform with real biological data. Another approach is to disable a gene all together (so called knockout gene), which causes pathways containing the gene to become altered. While this is among the better approaches, knockouts usually affect the expression of multiple other genes generating more real yet unknown pathway detections. Another issue with knockout approaches is that they are usually limited to animal models due to ethical reasons. While some gene knockdown approaches and modern techniques, like increasingly popular CRISPR method, have been applied on humans, the ethical aspects [152] heavily limit the sample sizes in such studies. Literature review as a validation strategy means investigating if the detected pathways are known to be relevant for the biological condition present in the data [132; 61]. The main weakness of this strategy is that the wide literature can link nearly all pathways to all biological conditions if the citations are selected to support the detections. A more reliable version of this approach is to investigate if a pathway directly related to the condition (i.e. target pathway), is among the top detections [131]. However, this is possible only when the condition is strongly associated with some pathway and the pathway is relevant (also) for the measured tissue. For example, it is not clear how well the Type I diabetes mellitus -pathway (KEGG id hsa04940), which is located in pancreas, can be detected from PBMC samples from T1D patients and healthy controls. Instead of defining the correct detections based on the literature, they can also be extracted ad hoc from the input data. This means sampling the data multiple times and stating that a good method detects the same pathways from the sample subsets [153]. However, this is possible only with large datasets containing very similar samples without further subgroups than the main ones in group-level analysis. Another strategy utilising reproducibility is to investigate if the same pathways are detected from similar datasets from separate studies. In this case the results are expected to differ more than with sampling the same data and it can be hard to define how similar results should be expected. This issue can be relived by utilising many similar yet separate datasets making the test more robust against one low quality or otherwise outlier dataset [132]. However, it can be difficult to find multiple similar yet different datasets, especially if the biological condition, sample preprocessing, or measurement technique is rare. In case of a sample-specific method, it can also be investigated if the samples can be classified biologically meaningfully (disease subtypes, survival time, case vs control) based on the obtained pathway scores [154; 132; 61]. Besides sample-specific pathway scores, also known sample groups are required. Finally, one option is to investigate whether a method finds more significant detections from real data than from dummy data without any real signal. For group-level methods this dummy data can be for

example randomly assigned DEGs [1]. Despite the lack of perfect validation, reasonably convincing evidence can be obtained by combining several approaches, such as knockout genes or target pathways and data sampling.

Besides the accuracy of the results, several other aspects related to the usability of the tested method are often evaluated. Typical topics to test include for example sensitivity to sample size, computational requirements such as running time and memory usage, effect of noise in input data, and false positive findings. These features are easier to evaluate than the correctness of the actual results. For evaluation purposes the input data can be modified by for example reducing sample size, adding noise, or mixing sample labels. Also different aspects of the pathways and their effect on the results can be tested. Typically this means evaluating how the method handles pathways of exceptional size or structure and does the utilisation of pathway topology improve the results.

Despite the challenges in validation, different pathway methods have been evaluated. There are several studies reviewing and comparing available pathway methods in theoretical level [34; 33; 155; 143; 156; 157]. Besides direct comparison of the available methods, also their ranking metrics have been evaluated [158]. Our publication I [159] includes empirical comparison of state-of-the-art methods covering also third generation tools utilising pathway structures. The study focuses on methods providing a list of pathways that behave differently between two sample groups. In the original publication I, we validated the tested methods by investigating if they find systematically similar pathways from similar data sets. Similar metric has been recently utilised by another study [106]. In order to prevent methods that claim all pathways as differentially expressed in all datasets appearing as optimal methods, we tested how prone the methods are to detect false positive findings. This was done by mixing case and control labels randomly multiple times; a good pathway method should not detect many pathways as significantly different between the artificial sample groups (false positive findings). About the same time another similar empirical comparison was published [160] and later on more [154; 15; 14; 161]. The comparisons use different evaluation criteria and their conclusions are not entirely uniform, but majority of them state that methods utilising pathway topology slightly outperform ORA and FCS approaches. Earlier empirical comparisons evaluated only generation one (ORA) and two (FCS) methods [162; 163; 164].

3.4 Summary of pathway analysis

Pathway analysis is a regular step in studies utilising transcriptomic data and usually it aims to estimate pathway scores for wider set of pathways representing different biological processes. Its main benefits are robustness, data reduction, and insight to underlying biological phenomena, but it has also weaknesses. For example, results for tissue specific pathways can be difficult to interpret and effect of all interactions

(e.g. protein-protein interactions) is not visible in transcriptomic data, which can cause false conclusions in pathway analysis utilising pathway topology. On the other hand, methods not using pathway topology are often overly simplistic and they have been outperformed by topology methods in the literature (e.g. original publication I). There are many publicly available methods to do pathway analysis and they are very diverse. Some methods aim to estimate pathway activity and some its deregulation as compared to normal sample and the scores can be calculated to either sample groups or individual samples separately. Pathway method PASI (original publication II) utilises pathway structure and provides both score options, activity and deregulation, for sample specific pathway scores. When developing and introducing new methodology, special attention should be paid on validating it as usually true gold standards are not available.

4 Deconvolution

4.1 Introduction to deconvolution

If gene expression data is measured from heterogeneous tissue containing several types of cells, it can cause some difficulties. Diversity in samples' cell type compositions and cell type specific expression generate variability in the bulk expression, which can mask changes in one cell type. Analysis of single genes and bigger networks like pathways are both vulnerable to this noise from mixture of cell types. To address the issues caused by heterogeneous tissues, different cell types need to be separated either experimentally or computationally.

4.1.1 Different approaches to obtain cell type specific gene expression data

There are three main approaches to obtain cell type specific data. The first one is purified cell populations, which means that the experimental design includes purifying the samples so that they include only certain type of cells. More detailed experimental approach is single-cell analysis in which expression levels of different genes are measured separately for each cell. The third way is *deconvolution*, which means computationally extracting cell type specific information from samples originally containing different types of cells. Unlike the other two approaches related to experimental design, deconvolution is a computational approach applied on bulk data. All of these approaches have their own strengths and weaknesses briefly introduced below. Difficulties to analyse rare cell populations is one weakness common for all of them.

The main issues with purified cell populations are that the researcher needs to decide beforehand which cell types to look for, and not all the subtle cell types can be separated. On a positive side, the results are more robust as compared to single-cell data and do not have the uncertainty of computational estimates related to deconvolution. Common methods to isolate cell populations to extract the cell type frequencies include fluorescence-activated cell sorting (FACS) and laser-capture microdissection. The selection of isolation method of cell types can drastically affect the outcome [165]. Besides purified cell populations, this type of data can be called *sorted* or *enriched* cells.

While single-cell data provides the most detailed information among these three

approaches, the number of analysed cells should be high as the cell populations have been shown to be very heterogeneous [89; 90]. This demand of high number of cells often financially limits the number of biological replicates, which limits the biological conclusions as individuals are known to be heterogeneous as well [166; 167]. In addition, single-cell analysis is not easy to do for all cell types as some cells tend to either die during the processing or stick to other cells so that their separation for analysis is difficult [168]. Also for example fibrous and minute tissues are difficult to dissociate into single cells [169].

The main drawback of deconvolution is that it is likely to provide less accurate results as compared to experimental design based approaches. On a positive side, computational methods are free of charge (excluding commercial methods) and they can be applied on old data. The possibility to computationally extract cell type specific information from old data without re-doing the experiment in cell type specific manner (purified cell populations or single-cell) is especially valuable if the study of interest is difficult to reproduce. The source of difficulty can be for example long follow-up time of longitudinal data, rare disease, or samples that are technically challenging to extract.

4.1.2 Formal definition

First of all, it is important to notice that the term *deconvolution* used in the context of extracting cell type specific signal from bulk gene expression is not related to the generally used mathematical definition of convolution $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)\, d\tau$. Here the underlying model is almost always either directly or some modified version of

$$E = S \cdot C, \tag{1}$$

where matrix E is a bulk expression matrix with samples as columns and genes as rows, S is a matrix describing how strongly pure cell types (columns) express the genes (rows), and C is a cell type proportion matrix (i.e. columns of C sum into 1) that indicate proportions of cell types (rows) in samples (columns). Therefore, total expression of gene g in sample n is a linear combination of expression levels from different cell types t weighted by their corresponding proportions: $E_{gn} = \sum_{t \in T} S_{gt} \cdot C_{tn}$, where T is the set of all present cell types. Cell type proportions are also called cell type *fractions* or *compositions* in the literature and sometimes term *cell population* is used instead of cell type.

4.2 Objectives and requirements of deconvolution

Deconvolution can aim either to detect cell type proportions C or to extract cell type specific expression profiles S. Here these two approaches are called *composition* and

expression deconvolution, respectively (Figure 3). For composition deconvolution there are more methods available than for expression deconvolution, but in original publication III, we address this issue by introducing a novel robust method for expression deconvolution and evaluating it together with the other available methods. Methods that aim to infer both C and S are called *complete deconvolution* methods, whereas composition and expression deconvolution methods are *partial deconvolution*. Composition deconvolution includes also several methods that aim to estimate cell type abundances [170; 171] instead of cell type proportions. In these cases, predicted values are comparable over samples, but not over cell types. There is a public debate about pros and cons of cell type proportions and abundances [172; 173]. Another related goal is to estimate a score describing for example the purity of tumour samples (e.g. [174]). Detecting cell type specific DE genes (csDEGs) is a goal related to expression deconvolution and different types of methods for it are empirically compared in original publication IV.

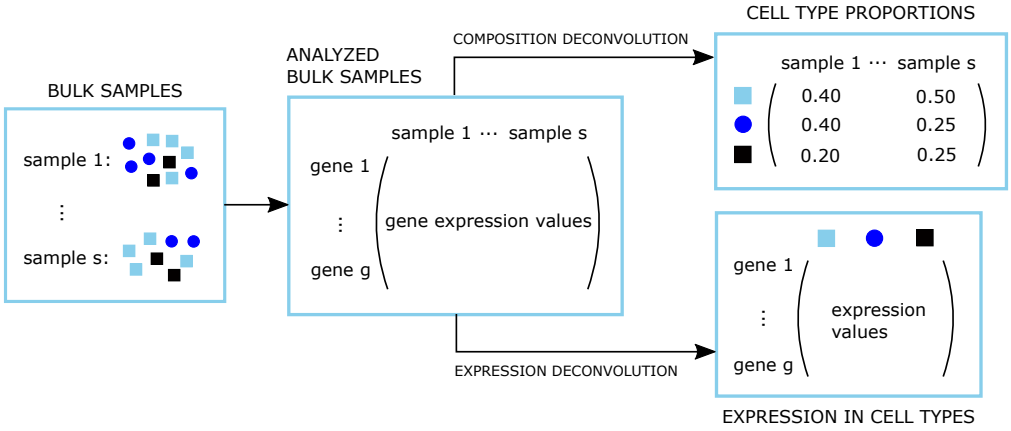

Figure 3. Overview of composition and expression deconvolution in a toy example of three cell types (cells notated with light blue squares, dark blue circles and black squares)

There are some special cases that contain deconvolving only two, or at most few, cell types. The most common example is cancer studies where the goal is to separate cancer cells from healthy tissue. Different characteristics of deconvolution in cancer research are discussed in more detail in Section 4.4.4. Another application for methods focusing on only two cell types is single cell doublets. In single cell techniques each cell is extracted into its own droplet or well, but errors happen in this process resulting into droplets/wells containing two or more cells. Originally these droplets/wells containing two or more cells, called doublets or multiplets, were just identified and excluded from the data, but recently approaches to identify and deconvolve them into usable single cell data have been proposed [175].

4.2.1 Signature matrix and marker genes

Most composition deconvolution methods require a *signature matrix* or a list of *marker genes* as an input (see Table 4). Some methods utilising such input offer a built-in version of it for some commonly analysed tissues like blood, or construct it internally. A signature matrix includes cell types assumed to be present in the bulk data as columns and signature genes as rows. The values are genes' expression levels in purified cell types. Signature genes should be stable over samples and tell the cell types apart from each other. The number of utilised signature genes is often (e.g. [176; 177; 178; 179; 180]) defined by minimising *a condition number* [181], which is formally defined as the product of the norm of the (signature) matrix and the norm of its inverse. The condition number reflects models' sensitivity to noise, so low number is associated with robust performance [182]. Typically this leads into several hundreds or thousands of signature genes [183; 169]. Genes that are exclusively expressed by only one cell type are called marker genes. Instead of expression values, a list of marker genes includes only names of genes that are enriched in cell types. The difference between marker genes and signature genes is vague, but marker genes are often more exclusive. However, also signature genes with big difference between the expression in the most expressing and in the second most expressing cell types have been reported to be favourable [184]. Notably, in other contexts than deconvolution, term 'gene signature' is usually used slightly differently, but it still typically characterises some phenomenon (rather than cell type). Here I call the bulk data to be deconvolved *target data* and the cell type specific data used to construct signature/markers *source data*.

A list of marker genes is easier to produce than a signature matrix, but it includes some caveats as well. The main issue with marker genes is that not all, especially closely related, cell types have clear transcriptomic markers [185], which often limits marker-based approaches to coarse cell types. Another issue is that the overlaps between markers from different studies or marker databases is often low [186] indicating that robust markers valid in most datasets are hard to find even for coarse cell types. Also, lowly expressed markers detected from RNAseq data (either single-cell or bulk from purified cell populations) might be unreliable if the target data to be deconvolved is old microarray data with higher noise level. Several studies [187; 188] have pointed out that marker candidates with medium to high expression are more robust than those with low or very high expression. Notably, surface markers that are used to separate cell populations from each other in cell sorting can not be directly used as marker genes for computational deconvolution as they are not necessarily expressed in RNA level [189].

Construction of signature matrix is a demanding yet important step in deconvolution [178] as it requires purified cell populations from samples similar to the target bulk data to be deconvolved. Vallania et al. demonstrated that the selection

of a signature matrix has even bigger impact on final results than the selection of the deconvolution method [190] and Cobos et al. stated that composition deconvolution methods utilising a signature matrix outperform those utilising marker genes [184]. In the same study they showed that technical, such as microarray platform, and biological, such as disease state, biases in signature significantly affect the deconvolution accuracy. Also other studies have reported similar observations. For example, signatures constructed using healthy samples are not necessarily accurate for cancer samples [191; 190], different normalisations yield different results [185], signature constructed using samples from adult donors causes bias when the data to be deconvolved is from babies [192], and the same immune cell types can have different expression profiles in different tissues [193; 194]. Furthermore, as cells interact with each other, the expression profile of a cell type can be affected by the other present cell types [98], making the matter more complicated. This phenomenon is an issue mainly for tissues involving very diverse cell types, like skin [195]. All this emphasises that the signature should be constructed with care and utilising source data similar to the target data to be deconvolved, if possible. As a signature matrix should represent all samples as well as possible, the selected signature genes should be stable over samples. In other words, a pure cell population should express a signature gene with similar intensity in all the samples. Therefore, an ideal signature matrix is a subset of S from model (1) including all its columns (cell types) and robust, stable, and cell type separating subset of its rows (genes).

Approaches to obtain a signature matrix or a marker gene list

There are three ways to get a signature matrix or a marker gene list:

- Using readily available one
- Using a computational method to construct one from source data
- Manually constructing one from source data

The last one is widely used and while the strategies vary from researcher to another, the process usually includes at least two main steps. The first one is some kind of identification of genes that are mostly expressed by one cell type, and the second one is about filtering out those genes that are not stable within samples of the same cell type. However, the process can be a lot more complicated [196].

There are ready signature matrices and marker gene lists publicly available. The signature matrices are usually constructed and published with a deconvolution method [171; 197; 178; 196], which makes the methods easy to use for those particular cell types. In many cases the built-in signature can also be extracted and used with other methods, if wanted. Typically this built-in information is for different cell types expected to be present in peripheral blood mononuclear cells (PBMC). For

target data from other tissues, those can not be used but the user needs to provide a signature matrix or marker gene list relevant to the tissue, similarly as with the deconvolution methods not providing built-in signature/markers. For marker genes, besides deconvolution methods providing them, there are also databases not related to deconvolution. These online resources include CellMarker [198], PanglaoDB [199], Blood Atlas [200] (included in The Human Protein Atlas [201]), and CTen [202]. Also few methods aiming to classify single cells, such as Garnett [203], SCINA [204], and DigitalCellSorter [205], contain cell type markers that can be utilised for deconvolution purposes.

While the most common approaches to get a signature matrix or marker gene list are to use readily available ones or to manually construct them from suitable source data, there are also few computational methods to do the task. Several composition deconvolution methods using a single cell data to internally construct a signature matrix allow the user to extract the constructed signature. Utilising this feature, methods like DWLS [179], SCDC [206], and BSEQ-sc [207] can be used to construct a signature matrix from single-cell data. Similarly CIBERSORTx [208] allows constructing the signature from either single cell data or bulk data from purified cell populations. Also methods to augment new cell types to an existing signature matrix have been published [209]. Some examples of methods to identify marker genes are Nano-dissection [210], which identifies genes with cell type specific expression, CellMapper [211], which searches for genes with similar expression pattern than given marker(s), and CellCODE [189], which is a deconvolution method.

Source data

Using single-cell data as source data has its own up and down sides as compared to bulk analysis of purified cell populations, which are more similar to the bulk target data. The challenges with single-cell data include 1) number of rare cells being too small to pool them for a reasonably reliable expression profiles, 2) sequencing depth affects the distribution of counts and it is different from that of a (target) bulk data, and 3) as single-cell studies often include only few donors, lack of individual heterogeneity can become a problem. However, not all cell types can be purified for the bulk approach either, and the same lack of donor individuals is often present in those studies as well. In addition, if the coming sc-RNAseq datasets are large as compared to the currently available ones, some of these issues could be relieved. Chen et al. have recently evaluated several challenges related to constructing a signature from single-cell data and they conclude that bulk-analysis based signatures are more reliable than single-cell based ones [193]. Also Lambrechts et al. argue that biases in sc-RNAseq may cause differences between cell types' expression profiles in bulk and single cell data [168]. However, single cell data has also several benefits, such as possibility to identify cell populations without distinct surface markers and pos-

sibility to identify and exclude cell subpopulations not expected to be present in the target data.

Notably, when a researcher is constructing a marker list from single cell data, they should be cautious of how the cells have been identified. In case the cell types of single cells have been defined based on only RNA expression markers (i.e. no additional surface protein markers or full expression profiles utilized), there is a risk of circular logic in using the data for deconvolution as this automatically leads to detecting the markers used in identification. However, currently the clustering-based cell type identification approaches are dominating the field so the issue is present mainly in old datasets. While surface protein markers are not perfectly reliable either, utilising also them adds to the confidence in the cell type identification [212].

There are several attempts to collect suitable source data into one database. Among them, the previously mentioned GEO and ArrayExpress are the widest and besides suitable source data, they also contain plenty of bulk datasets from mixture tissues. However, there are also more specialised databases available, especially for single-cell data, as summarised in Table 3. Besides online databases, several R packages also contain collected source datasets, namely deconvolution related R package CellMix [213] has some suitable source data and SingleR [214] provides access to several databases listed in Table 3 despite its main focus being at classifying single cells.

Table 3. Potential databases to search for source data. Databases marked with * require registration or request before access to the data, and those marked with ** provide links to other databases instead of actually hosting the data.

Name	Description	Reference
10x	single-cell data created with 10x platform	
ArrayExpress	general	[215]
BLUEPRINT*	human data, main focus on epigenomics	[216]
DICE	sorted human immune cells	[217]
GEO	general	[218; 219]
Human Cell Atlas	single-cell data from human	[220]
ImmGen	mouse data	[221]
JingleBells	single-cell data	[222]
PHANTOM5	mammal data	[223; 224]
Recount2	general	[225]
SCPortalen**	single-cell data	[226]
scRNASeqDB**	single-cell data from human	[227]
Single Cell Portal*	single-cell data	

4.2.2 Input for expression deconvolution methods

Complete deconvolution methods mainly utilise input similar to composition deconvolution methods, but expression deconvolution methods have generally different requirements. Most of them expect cell type proportion matrix C as an input. Sometimes it can be readily available, for example if FACS analysis has been done for the bulk samples to be deconvolved. However, that is an exception and most of the time it needs to be estimated, potentially with composition deconvolution. As estimates are never absolutely accurate, it is important that expression deconvolution methods tolerate some noise in C. In publication III, we demonstrated that most tested expression deconvolution methods endure minor noise in input reasonably well. On the other hand, in the same paper we showed that the number of samples in the input data has a considerable effect on the accuracy of the results, which is not the case for all composition deconvolution methods [188; 228].

4.3 Variation and validation in deconvolution

4.3.1 Sources of differences between bulk samples

In the basic deconvolution model $E = S \cdot C$ the only source of variation between samples is the different cell type compositions in matrix C. However, this is not the whole truth as cell types can also behave differently meaning that expression profile of pure cell type x might be different in sample i and sample j, which leads to individual differences in matrix S. Unfortunately estimating S (expression deconvolution) is difficult as shown in publication III, and estimating sample-specific S is even harder as discussed below. The source of differences in bulk expression (C, S, or both) varies from gene, condition, tissue, and study to another. T1D is an interesting example of uncertain source of variation as there is an open debate if the beta cells are fully absent from T1D patients' pancreas (C is altered) or if the beta cells are there to some extent, but they have lost their insulin production functionality (also S is altered) [229; 230; 231]. As beta cells are identified mainly by their insulin production, they can not be detected in composition deconvolution in either case as the signature matrix is invalid for the case samples. However, the issue can be avoided if the possible living, but dysfunctional beta cells would be systematically mistaken for one other cell type, say alpha cells. In this case, the case samples would have higher proportion of alpha cells as compared to controls without T1D. On the other hand, if the dysfunctional beta cells would be mistaken for several different cells or would go totally unnoticed, it would be impossible to distinguish the situation from dead beta cells.

Personalised S

Estimating S for each sample separately is a very demanding task and not many methods claim to do it. However, while our leave-one-out based attempts to develop a deconvolution method that estimates personalised S were not sufficiently accurate to our satisfaction, there are few other tools to obtain related goals. The closest one is CIBERSORTx [208] as it includes an option to estimate personalised S, but the size of the bulk data (number of genes and samples) is limited and analysing a typical modern dataset with the available online implementation would require requesting for more computational resources. DeMix [232] is another related tool, it provides personalised expression profiles for two cell types, tumor and healthy tissue. Unfortunately the link to the code has expired, so it is unsure if the method is available in practice. ISOpure [233] is an available method close to DeMix, it purifies expression of tumor samples from the effect of healthy tissue.

Cell type specific differentially expressed genes

A relaxed version of estimating a personalised S is to identify cell type specific DEGs (csDEGs), namely genes that are differentially expressed between two sample groups within one cell type. They may or may not be detectable from mixed bulk data, and if the cell type composition is different between sample groups, they get easily masked in the bulk data by genes that are strongly expressed by the cell type enriched in one sample group.

There are several expression deconvolution methods that allow cell type specific DE analysis [234; 235; 189; 207; 236; 237; 238; 239], and with a bit more effort, all expression deconvolution methods can be used for the task as follows:

Step 1 Split the bulk data E and cell type proportions C according to the sample groups and detect S for both groups separately

Step 2 Randomly split the samples into two groups of sizes equal to the real sample groups and estimate S for these random groups

Step 3 Repeat step 2 many times

Step 4 Calculate the differences in S between real sample groups (from step 1) and use the detected S from randomised sample groups to estimate the significance levels of the observed differences.

However, due to step 3, this unsophisticated approach is expected to be slow unless the utilised deconvolution method is especially fast. As demonstrated in publication IV, despite being computationally time consuming, the label sampling step 3 is important for the accuracy of the estimated csDEGs.

In the original publication IV, we compared nine approaches to detect csDEGs based on three inputs: a bulk expression matrix E, a cell type proportion matrix C, and a vector indicating sample groups. The tested methods involved four tools designed for the task (TOAST [239], csSAM [235], LRCDE [234], and CARseq [240]), two designed for similar task in methylation data (CellDMC [241] and TCA [242]), two expression deconvolution methods with accurate performance in publication III (Rodeo [243] and qprog [244]), and one general model without any focus on deconvolution (DESeq2 [245]). The results show that cell type proportion and individual heterogeneity in csGEPs are important factors defining how accurate estimates for csDEGs can be achieved. Among the tested methods, those designed for detecting cell type specific differences (either in gene expression or methylation) had the most accurate performance. In the paper, we also provide practical instructions how the end user can evaluate the level of individual heterogeneity and the possible presence of outlier samples.

4.3.2 Present cell types

When constructing the signature matrix or marker list, a researcher should decide which cell types to look for. The first decision is about defining which cell types are assumed to be present in the samples. The other task is drawing the line between cell types, which is ambiguous, and usually the researcher just needs to decide how detailed cell types they wants to use. For example, in PBMC T cells are present, but it is up to decision if they are considered as one group, or as two biggest subgroups CD4+ and CD8+ T cells, or possibly as dozen of small subgroups such as regulatory, naive, mature, memory, and activated cells. Aiming to detect fine subpopulations is called *deep deconvolution* [246]. Utilising smaller cell subpopulations provides more detailed picture of what is going on in the samples and are therefore clinically more interesting. Also, if a cell population is very heterogeneous due to different rates of clearly different subpopulations in different samples, finding a robust signature/markers for the superpopulation is difficult and using more stable subpopulations instead could be beneficial for the deconvolution accuracy. However, usually subpopulations resemble each other so the heterogeneity of superpopulation over the samples is not a big issue. Instead, closely related cell types are hard to distinguish from each other and especially finding multiple strong marker genes for each of them can be an issue. Notably, cell types with similar expression profiles are called *collinear* and collinearity is a known major difficulty for deconvolution as frequently stated in the literature [247; 172; 248; 178]. Another practical drawback with detailed subpopulations is that they are always more rare than their superpopulations and proportions of rare cell populations are harder to estimate than those of more dominating cell types [246; 244; 177; 178]. In fact, ability to detect also rare cell populations have become one measure to compare different deconvolution methods

[249]. Also the expression profiles of rare cell types are harder to estimate with expression deconvolution than those of abundant cells, as demonstrated in the original publication III.

4.3.3 Validation of deconvolution methods

Typically validation of bioinformatics tools is difficult and deconvolution is not an exception. For validation purposes, a bulk data together with its cell type proportions and/or expression levels of purified cell types should be known. There are several ways to simulate such situation and four commonly used ones are described below and summarised in Figure 4. Validation approaches 1-3 were used in publication III.

The first way is to not use real expression data, but simulate it computationally. This approach is very controlled and no unknown factors are present. Controllability makes it straightforward to draw the conclusions and avoids limitations like low number of biological replicates often present in measured data. On the negative side, it does not directly answer how well the tested method would perform with real data with all sort of uncertainty. Although, simulating some noise and bias into the data improves the realism.

The synthetic data becomes more realistic if it is constructed by combining measured expression levels from purified cell populations. While this approach is a step towards a more realistic situation, it still contains perfect linear relationship as described in model (1), which might not be the case with real data. Either single-cell data or data from purified cell populations can be used to construct an artificial bulk data with this approach. When using validation approaches two and three from Figure 4, it is important that the combined pure cell types are from different donors as otherwise the constructed bulk data does not contain individual heterogeneity of S present in a real bulk data. In the original publication IV, we tried to overcome this issue (and the insufficiently low number of sample donors present in publicly available data containing expression from pure cell populations) by generating samples following the same distribution as the measured samples. Thus, for artificial sample k, gene i, and cell type j we extracted an expression level from normal distribution with mean and standard deviation over measured samples for the given gene and cell type.

The third way is to combine purified cells in known proportions and then take the bulk measurements from the known mixture. This approach is very good, but in some of the most frequently used validation datasets the mixed cells are from totally different parts of a body. For example, there is a mixture containing rat's liver, brain, and lung tissue (GEO accession id GSE19830, also available in R package CellMix [213]). However, this type of datasets do not reflect real tissue data where the present cell types can be rather similar to each other. On a positive side, when the

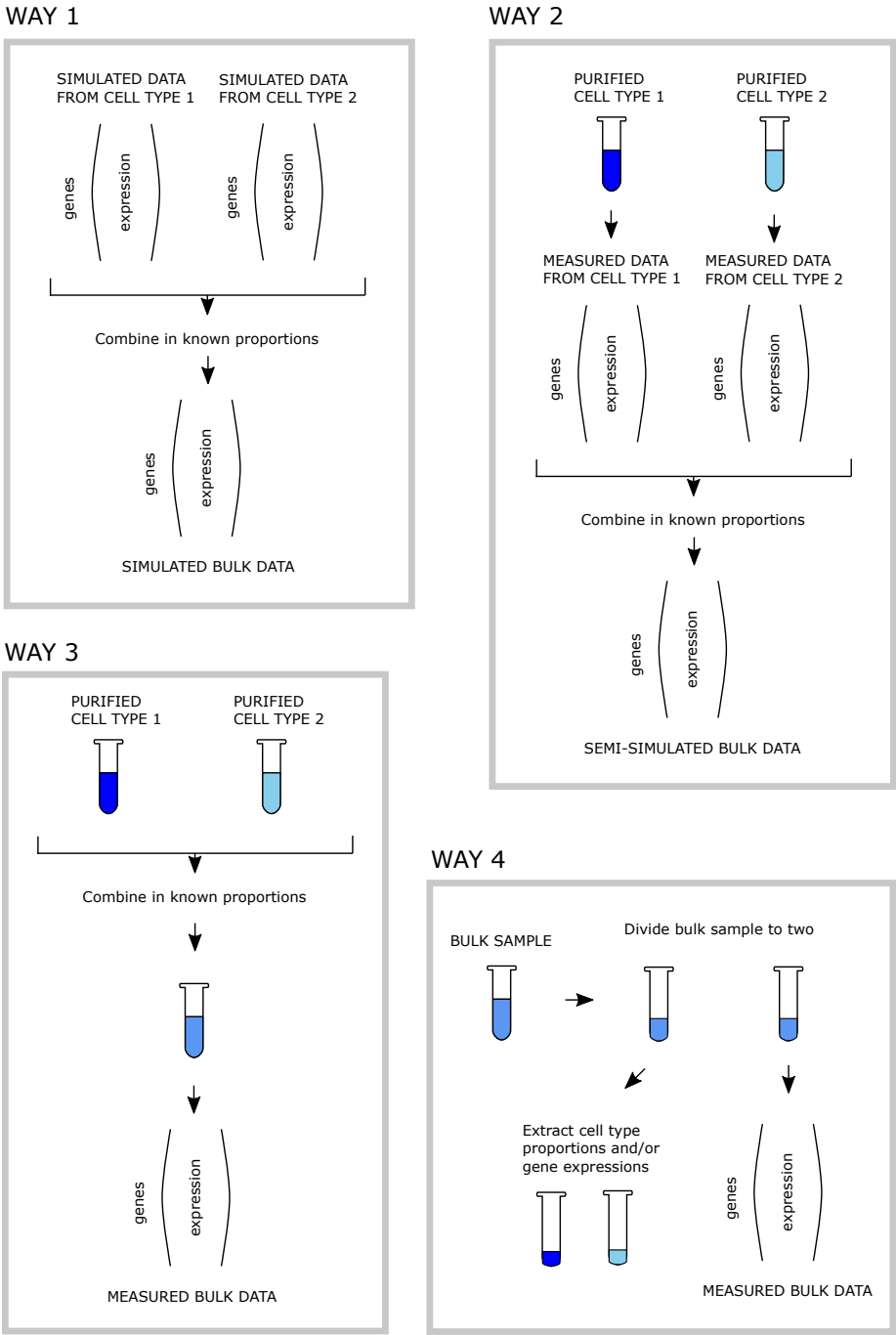

Figure 4. Four ways to obtain validation data for deconvolution methods. For the sake of simple visualisation, only two cell types (dark and light blue) and one bulk sample are included in the figure.

mixed cells are from different tissues, purity of cell types is guaranteed as they do not naturally appear together. Another variant of this approach is to combine RNA from different types of cells in known proportions instead of combining whole cells. From validation perspective the main difference between combining cells and combining RNA is that the proportion of cell types with low overall expression (like neutrophils [250; 251]) tends to get underestimated if cells are combined. Combining directly RNA avoids this issue, but it is not obvious if this factor should be present in the validation data or not. This challenge is further discussed in section 4.3.5. While mixing cells/RNA is closer to realistic data than mixing expression values (first two options), neither of these approaches contain unknown expression sources that are expected in real tissue data unless minor amounts of cells/RNA/data of other cell types is added to the bulk mixture as noise.

The most realistic approach in theory is to take a tissue sample, analyse part of it as a bulk mixture, and use the rest to extract proportions and/or expression profiles of purified cell populations. There are several techniques to obtain gold standard for cell type proportions from the remaining sample, such as FACS [252], DNA copy numbers [253], or DNA methylation [254]. This way bulk samples are realistic and known cell type proportions/expressions are from the same samples. The measured cell type proportions should be very close to those in the part of samples reserved for bulk analyses, if the cell types are quite uniformly distributed over the tissue samples. This is true for tissues like blood, but the assumption does not hold in case of for example solid tumours, where the extracted part for bulk analysis might contain different proportion of healthy tissue than the part left for defining the gold standard. In those cases, approach 3 might be a better choice.

4.3.4 Example of practical issues with validation

All validation approaches require data with known cell type proportions or expressions, but the public resources are limited. One suitable data set is by Linsley et al. [255] (GEO access GSE60424). It includes RNAseq from whole blood and from purified cell populations of neutrophils, monocytes, B cells, CD4+ T cells, CD8+ T cells, and NK cells from the same samples. Also cell counts (that can be converted to cell type proportions) of purified cell populations are provided, which makes the data a good validation set for many types of deconvolution methods. Based on model assumption (1), it should be possible to some extent reconstruct the bulk whole blood samples by combining the expression profiles of purified cell populations in the known proportions. However, when comparing correlations between reconstructed ($C \cdot S$, where also S is measured for each sample) and measured bulk samples, the same sample donor was not among the most dominating patterns (Figure 5).

If a method does not perform well with such validation data when measured whole blood samples are used as E, it is hard to say if it is due to

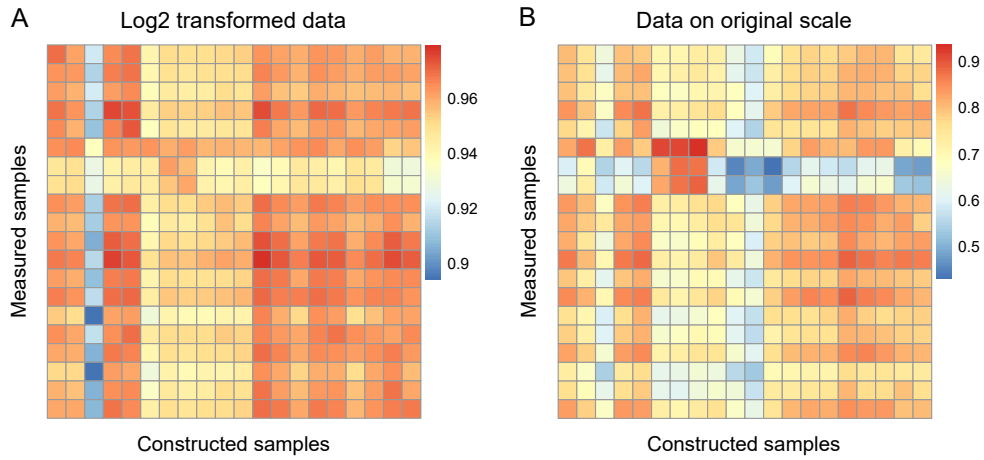

Figure 5. Pearson correlations of measured bulk samples (rows) and reconstructed bulk samples (columns) in (A) log and (B) original scale data. Importantly, logarithmic transformation has been done after reconstructing the bulk samples. Ideally a reconstructed bulk sample should correlate better with the corresponding measured bulk sample, i.e the diagonal should have higher correlation values than the the rest of the figure.

1. poor method
2. impure cell populations, technical noise, or other issue with the data processing
3. noise from rare unidentified cells
4. cell counts ignoring significant amount of cells causing inaccurate C
5. assumption of underlying model (1) not holding or
6. phenomena related to normalisation or other data processing currently not understood.

4.3.5 Factors affecting the linear model assumption

The previous example leads us to issues with the assumed underlying model (1). Currently there is no consensus about the effect of normalisation on preserving the linear relationship (1), yet its importance has been highlighted in several studies [256; 257; 188]. Especially for microarray data, conflicting statements about the best normalisation and background correction have been made [256; 232; 258; 259; 260]. For RNAseq data transcript per million (TPM) normalisation has been stated to be the best normalisation from deconvolution perspective [261] and study [185] introduces an entirely new normalisation developed specifically for deconvolution purposes. The same study also addresses the issue of different overall expression between cell types (see the last paragraph of this subsection).

Besides normalisation, also log transformation can affect the linearity assumption (1). As compared to normalisation questions, the available literature is more in line about the effect of log transformation, and most studies state that original scale preserves the linearity better than log transformation in case of both microarray [262; 238; 187; 249] and RNAseq data [263; 261; 184]. However, there are several studies whose conclusions challenge the superiority of original scale [264; 237; 188]. Notably, multiple studies supporting the usage of original scale still suggest that the normalisation of data can be done on log scale as long as it is returned back to original scale before deconvolution [265; 266; 252]. Impact of both normalisation and log transformation have been thoroughly investigated at [261] and reviews [249; 184] provide a nice and up-to-date summary of the discussion about both topics.

The linear combination naturally does not hold if the input and conclusions are made on different levels; it is important to not confuse cell counts and RNA. Bulk transcriptomic data is measured by investigating the total RNA present in the samples and signature matrices are also built on measuring RNA. However, cell type proportion matrix C (input for expression deconvolution or validation of composition deconvolution) is often defined based on cell counts. This is an issue as not all cell types express equally much. For example, neutrophils have low overall expression as compared to leukocytes [185]. Therefore, if a sample contains 60% neutrophils and 40% leukocytes (simplified example), only 20% of its total RNA might be from neutrophils and 80% from leukocytes. Due to this phenomenon, the proportions of cell types with low overall expression tend to be underestimated. There are two approaches to avoid the issue. Either all inputs should be about RNA rather than cell counts, and also the deconvolution results should be interpreted in the level of RNA. The other option is to use methods that consider the variation in total expression rate of cell types. There are some such composition deconvolution methods [267; 197] and one complete method [268], but they require background information that is laborious to obtain, which makes utilising these features difficult. Zaitsev et al. demonstrate the issues related to different cell sizes and total RNA contents and suggest possible solutions utilising ERCC spike-ins (i.e. artificial transcripts added to the sample in known quantity) in experimental level [267].

4.4 Deconvolution methods

Here I discuss several methodology related topics, such as different statistical approaches often utilised in deconvolution methods, methods without input requirements, the available studies reviewing and comparing deconvolution methods, and deconvolution methods for certain type of data. Table 4 lists available methods to deconvolve transcriptomic data and provides references for further reading.

Table 4. List of several available deconvolution methods. Column 'Type' indicates the goal of the method. It has following options: estimate cell type proportions or abundances (composition), cell type specific expression profiles (expression), or cell type specific differentially expressed genes (csDEG). Term 'complete' indicates both composition and expression deconvolution simultaneously. If multiple options are listed exclusively (or), the type of the output depends on the given input. The main input besides the bulk expression to be deconvolved (column 'Input') can be either marker gene list (M), cell type proportion matrix (C), signature matrix (S), single cell RNAseq data (sc-RNAseq), any type of source data to construct signature (source), number of cell types (#T), or none. Notably, many of the listed methods contain built-in signatures matrices/marker lists making the user input optional for those tissue types. An asterisk after the programming language of the implementation indicates that the code can no longer be accessed due to broken link or some other issue. Many of the listed methods provide some additional information and have some special features and/or requirements not listed here.

Name	Type	Input excl. bulk	Implementation	Reference
ABIS	composition	none	online	[185]
ADAPTS	complete	source	R	[209]
BayICE	composition	S	R	[269]
Bisque	composition	sc-RNAseq	R	[270]
BRETIGEA	composition	M	R	[271]
BSEQ-sc	composition and csDEG	sc-RNAseq	R	[207]
CAM	composition	none	Java-R	[252]
CARseq	csDEG	C	R	[240]
CDSeq	complete	#T	Matlab, Octave, R	[268]
CellCODE	composition and csDEG	S	R	[189]
CellPred	composition	S	online	[272]
CellR	complete	sc-RNAseq	R	[273]
CIBERSORT	composition	S	online	[178]
CIBERSORTx	complete	S or source	online	[208]
Clarke et al.	composition	S	R*	[257]
collapseRows	composition	S	R	[274]
COMPMIX	csDEG	unknown	R*	[236]
contamDE	tumor purity	none	R	[275]
csSAM	expression and csDEG	C	R	[235]
DCQ	composition	S	R, Java, and online	[276]
Deblender	complete	M	Matlab	[228]
DECODER	complete	#T	Matlab	[277]
Decon-cell	composition	none	R and online	[278]
Deconf	complete	M	R	[188]
DeconRNASeq	composition	S	R	[177]
DECONVOLUTE	composition	S	Java*	[279]
deconvSeq	composition	S	R	[169]
DeMix	complete	S or none	R*	[232]
DeMixT	complete	partial S	R	[280]
DSA	expression	M	R	[263]
Dsection	expression and csDEG	C	R and online	[237]
dtangle	composition	S	R	[281]
DWLS	composition	S or sc-RNAseq	R	[179]
DynamicDA	composition	none	Matlab	[282]
Enumerarteblood	composition	S	R	[283]
EPIC	composition	S	R	[197]
ESTIMATE	tumour purity	S	R	[174]
FARDEEP	composition	S	R	[284]
GEDIT	composition	S	online	[285]
ImmuCC	composition	none	online	[259; 286]
ImmQuant	composition	M	Java-R	[287]
ImSig	composition	none	R	[196]
ISOpure	complete	none	R	[233]

LinDeconSeq	composition	S	R	[180]
LinSeed	composition	none	R	[267]
LRCDE	expression and csDEG	C	R	[234]
lsfit	composition	S	R	[288]
MCP-counter	composition	M	R	[171]
MMAD	composition or expression	S or C	Matlab	[260]
MuSiC	composition	sc-RNAseq	R	[248]
MySort	composition	S	R	[176]
PERT	composition	S	Octave	[289]
PSEA	expression and csDEG	M	R	[238]
qprog	composition	S	R	[244]
quanTIseq	composition	none	docker image	[290]
RAD	complete	#T	Python	[291]
Reinartz-Finkernagel	purified cancer	S	Python	[292]
Rodeo	expression	C	R	[243]
SCDC	composition	sc-RNAseq	R	[206]
semi-CAM	composition	M	R	[293]
SMC	composition	#T	Matlab	[294]
SPEC	composition	M	R	[295]
ssNMF	complete	M	R	[296]
TEMT	expression	C	Python	[297]
TIMER	composition	none	R and online	[254]
TOAST	csDEG	C	R	[239]
UNDO	complete	none	R	[298]
VoCAL	composition	S	R	[183]
xCell	composition	none	R and online	[170]

4.4.1 Working principles of deconvolution methods

While technical aspects about machine learning and optimisation are out of the scope of this introduction, here I briefly introduce several techniques that are utilised in different deconvolution methods. More comprehensive reviews are available by Mohammadi et al. [299] and Avila Cobos et al. [249]. Here I follow the classification into the regression and probabilistic models used in [299].

Different regression approaches are common among deconvolution methods. In linear regression, a dependent variable (e.g. bulk gene expression) is modelled as a linear function of explanatory variables (e.g. cell type proportions in composition deconvolution or csGEPs of pure cell types in expression deconvolution) and the aim is to define the coefficients for the linear equation so that the linear curve fits the observed data as well as possible. These types of methods are very heterogeneous and they differ from each other in data preprocessing, how the objective function to be optimised is defined, by the possible additional constrains in the optimisation problem, and by the final implementation of how to solve the selected optimisation problem. The objective functions can be for example ordinary least squares with or without regularisation term such as elastic net. It is a commonly used regularisation in machine learning and it has been utilised also in the context of deconvolution [276]. Common additional constrains include non-negativity of coefficients [289] and sum-to-one constrain [244], which forces the coefficient to represent cell type proportions. However, these features can be enforced also after the optimisation

problem has been solved [288]. Support vector regression is a subtype of regression where small deviation from the regression line is not penalised at all in the optimisation problem. This robustness is favourable in deconvolution problems [182; 178] due to the heterogeneous and erroneous nature of biological data.

Another major group of methods is probabilistic models. Two examples are Bayesian models (e.g. DSection, COMPMIX) and latent dirichlet allocation models (e.g. CDSeq, ISOpure, ISOLATE). In Bayesian models a likelihood function is maximised, but the form of the function is dependent on used parameters, hyperparameters, and definitions of a priori and a posteriori functions. Latent dirichlet allocation is mainly applied in natural language processing and it assumes few unobserved background elements (e.g. cell types in deconvolution or topics in natural language processing) to explain similarities among observations (e.g. bulk expression in deconvolution or written text in natural language processing).

Besides sophisticated modelling, the cell types proportions and especially abundances can be estimated by using simple marker-based approaches [171; 260]. In this type of methods, the expression levels of different marker genes are somehow summarised within cell type, and these summary values are then interpreted to reflect the over-samples variation of that cell type. Different filtering, clustering, scaling, and identification of new markers can be done prior to summarising the markers representing the same cell type.

One important subgroup of such marker-based methods is unsupervised deconvolution methods (e.g. CAM, LinSeed) aiming to first identify gene clusters with high internal correlation (or other measure of similar behaviour) over samples from the bulk data. These gene clusters are then interpreted to represent marker genes of different cell types. Then a simplex with marker clusters as vertices is created and used to further estimate cell type abundances/proportions and/or cell type specific expression profiles in the bulk samples. However, the methods utilising such strategy are very heterogeneous and differ from each other by aspects like normalisation of the data and selection criteria for marker genes.

4.4.2 Unsupervised and semi-supervised methods

Some methods do not require any input from a user besides the bulk expression data to be deconvolved [267; 188]. These type of methods are called reference free or unsupervised methods. Methods that use built-in signature matrix or marker gene list, or pre-trained machine learning model are not considered as reference free despite appearing as such for the end-user. This type of methods are typically easy to use due to their minimal input requirements, but interpreting the output can be challenging as different extracted output profiles are not directly linked to cell types. Due to the lack of preliminary information, unsupervised methods might detect some other source of variation than cell type composition [296]. As shown in the original pub-

lication III for expression deconvolution and in several other studies [296; 293] for composition deconvolution, unsupervised methods are less accurate than supervised methods especially when the bulk data contains related cell types. However, despite their limitations, unsupervised methods can be useful for studying poorly known and very diverse cell populations like tumour subclones [300].

As the name suggests, semi-supervised methods require some cell type specific input, but it does not need to cover all the present cell types. Notably, the term semi-supervised has been used vaguely in the literature and some studies [296; 184] address composition methods utilising a marker gene list instead of a signature matrix as semi-supervised. There is one recent semi-supervised method semi-CAM [293] that allows markers for any number of cell types present in the data. The other available semi-supervised methods are more restricted in a sense that they allow missing input for only one cell type. This can mean either utilising a signature matrix that covers all but one of the present cell types (e.g. ISOLATE [301], DeMix [232], and DeMixT [280]) or providing cell type proportion also for unknown content (e.g. EPIC [197] and BayICE [269]).

4.4.3 Comparisons and reviews of available deconvolution methods

Several theoretical comparisons and reviews of different deconvolution methods are available [302; 246; 303; 299; 249], introducing different algorithms and methods together with their scopes and required inputs. Empirical comparisons are more rare and most of them are related to validating new methods [178], including our comparison of expression deconvolution methods in publication III. However, recently few empirical third-party comparisons of composition deconvolution methods have been published [304; 184]. Additionally, study [303] includes a brief comparison of tumour purity estimates besides the theoretical review. For expression deconvolution methods there is no third party empirical comparison available at the time of writing, and our publication III introducing Rodeo and evaluating it together with similar methods is the only recent study involving empirical comparison. In publication IV deconvolution methods to detect csDEGs are evaluated from practical point of view, but it is not third-party comparison either as our own method Rodeo from publication III is one of the tested methods, and also the evaluated aspect is more specialised than general expression deconvolution. Due to expression deconvolution being less studied than composition deconvolution, the conclusions and summary statements about it have less support and narrower scope.

All the evaluation studies suffer from the limitations of the available test data. The main underlying issue is that as many cell types need to be analysed separately, the number of biological replicates is typically low. While the small sample size is an issue for especially expression deconvolution, as demonstrated in publication

III, also many composition methods and particularly unsupervised ones have been shown to benefit from larger sample size as well [188; 294; 246; 232]. Low number of sample donors does not always lead to low sample size in the test data, if multiple mixtures of cells/RNA are made from the material from the same few donors. For example, in the dataset GSE19830 (briefly introduced in section 4.3.3) the samples are constructed by combining different types of cells from one donor in known proportions. However, this approach suffers from the lack of individual heterogeneity and provides too easy test data. In the original publication III, we demonstrated the issues related to over-simplistic test data. Difficulties to obtain accurate results from data with low sample size and low number of donors causing too simple test data highlight the need for large and realistic dataset containing purified cell populations from many sample donors. An ideal empirical comparison would 1) be carried out by a third-party not affiliated with any of the tested methods, 2) utilise such realistic validation data (preferably multiple unrelated datasets with different characteristics), 3) following validation approaches 3 and 4 in Section 4.3.3, and 4) test wide set of state-of-the-art methods for both composition and expression deconvolution.

4.4.4 Deconvolution methods for cancer studies

Cell type proportions have been associated with several important cancer related questions like prognosis, survival, and response to treatment in the literature [305; 306; 307; 308], so the wide range of deconvolution methods developed and demonstrated specifically for cancer studies is not unexpected. When deconvolving bulk data from cancer tissue, the goal is typically to separate at least cancer tissue, healthy tissue, and immune cells. Weather different types of immune cells are treated separately or not varies from study and method to another. Cancer studies have several aspects that make them particularly challenging for deconvolution. The two main obstacles requiring special attention are 1) heterogeneity of tumours and 2) altered immune cells (see also Section 4.3.1). Both issues have been at least somehow addressed by the available methods. The heterogeneity of tumour tissues makes it very difficult to construct even remotely reliable signature expression profile suitable for all samples in the study, whereas the issue of altered immune cells is less tricky, but requires attention regardless. It has been shown that expression profiles of immune cells measured from healthy individuals do not correspond to those measured from cancer patients [190]. Therefore, it is important to use data from cancer patients similar to the sample donors of the study when constructing the input signature/markers. Public resources like The Cancer Genome Atlas TCGA are important assets for this.

Many expression and composition deconvolution methods have been developed to asses cancer research [275; 236; 228; 232; 280; 197; 174; 233; 302; 298; 284]. Due to the difficulties related to defining typical expression profile for tumour, semi-supervised methods allowing some unknown source of RNA are well suited for de-

convolution in cancer studies. Also, some models, such as the one implemented in EPIC, assume that the input data includes many heavily varying genes, which is characteristic for cancer data. Besides classical composition (e.g. [197; 286]) and expression (e.g. [263]) cancer deconvolution methods, there are also several ambitious complete deconvolution methods designed for cancer studies [298; 232; 280; 228; 233; 291] and relaxed versions of composition and expression deconvolution methods. As an example of a sophisticated complete method designed for cancer studies, ISOpure [233] takes the complete deconvolution as far as attempts to estimate personalised cancer profiles. Relaxed composition deconvolution methods aim to estimate the purity level of tumour samples [275; 174] and relaxed expression methods aim to purify the tumour tissue expression from the effect of other components like immune cells [292]. Some methods developed for cancer studies assume only limited number of cell types (like two for tumour and healthy tissue, or also a third one for immune cells), which makes them unsuited for other studies containing more cell types. However, exceptions without any limitations and wider built-in signature/markers, like quanTIseq (10 cell types) and TIMER (6 immune cell types), exist.

4.4.5 Deconvolution methods for DNA methylation data

Several deconvolution methods have been developed also for other data types than transcriptomics. Among those, DNA methylation data has the widest selection of methods implemented. DNA methylation means that a methyl group is added into a cytosine, which is one of the four bases in DNA, and it is one type of epigenomics. Methylation can either activate or inactivate genes [309] and, as different type of cells have different functions, it is not surprising that they can have distinct methylation profiles as well. Most methods to deconvolve methylation data correspond to composition deconvolution of transcriptomic data, and they aim to estimate cell type composition of bulk samples. Also, similar to the methods for transcriptomic data, there are unsupervised and supervised methods, and the supervised ones are likely to outperform the unsupervised ones [310]. The supervised methods typically utilise reference cell type specific DNA methylation profiles (csDMP) of present cell types, similar to a signature matrix.

Deconvolution methods for methylation data have been reviewed [195; 311] and compared empirically [312; 313] by a third party. Table 5 summarises some of the available methods. The summary does not include more general implementations for matrix decomposition, like SVA [314], ISVA [315], and RUV [316] despite them being utilised in DNA methylation deconvolution as unsupervised methods. Differences between methods for transcriptomic and methylation data are not dramatic. For example, R package MethylCIBERSORT includes functions related to processing the methylation data, but the actual deconvolution is done with the same on-

Table 5. Summary of several available deconvolution methods for DNA methylation data. Notably, while EPISCORE and TCA do not require cell type specific DNA methylation profiles as input, EPISCORE expects single cell RNAseq data (i.e. it combines transcriptomic and methylation data), and TCA expects cell type proportions.

Name	Type	csDMP as input	Citation
BCheterogeneity	composition	yes	[317]
eDEC	complete	both options	[318]
EpiDISH	complete	yes	[319; 241]
EPISCORE	composition	no	[320]
FaST-LMM-EWASher	composition	no	[321]
HIRE	complete	no	[322]
Houseman's CP	composition	yes	[323]
MeDeCom	complete	both options	[324]
MethylCIBERSORT	composition	yes	[325; 326]
MethylResolver	composition	yes	[327]
ReFACTOR	composition	no	[328]
RefFreeEWAS	complete	no	[329; 310]
TCA	expression	no	[242]

line tool CIBERSORT used for transcriptomic data. Another example is deconvolution method TOAST, which has been initially developed and validated for both data types. Also, approaches like quadratic programming are utilised in both applications [311; 299]. In the original publication IV, we show that for estimating csDEGs methods designed for methylation data had performance equal to those designed for transcriptomics. In their study [325], Chakravarthy et al. show that cell type proportion estimates obtained from methylation data are more accurate than those obtained from transcriptomic data of the same samples, likely due to the lower noise level of methylation data. Notably, in the literature of deconvolution methods for other data types than transcriptomics, terms reference-based and reference free are commonly used instead of supervised and unsupervised.

4.5 Deconvolution summary

Computational deconvolution is an affordable approach to obtain cell type specific information also from old datasets. Composition deconvolution aims to detect different cell types' proportions or abundances in bulk samples and in expression deconvolution their pure expression profiles are estimated. Methods doing both tasks simultaneously are called complete methods. In contrast to the rather well studied composition deconvolution, for expression deconvolution there are fewer methods and summary studies available. However, in this thesis I have addressed this gap by developing a new method to estimate cell type specific gene expression profiles from

bulk data and thoroughly evaluating it with the other available similar tools (publication III), and by empirically comparing different methods to estimate csDEGs (publication IV), which is a task related to expression deconvolution.

Typically composition and expression deconvolution methods require a signature matrix and a cell type proportion matrix, respectively, as an input. Also other inputs are possible and unsupervised methods have been developed as well, but they have been shown to provide less accurate results than supervised methods. Constructing a signature matrix is a demanding yet crucial step in composition deconvolution and it should be built using preferably purified bulk csGEPs as close to the target data to be deconvolved as possible. On the other hand, individual heterogeneity of csGEPs causes issues particularly for expression deconvolution, whereas both types of partial deconvolution struggle with rare cell types, collinear cell types, and small sample size.

5 Discussion

5.1 Challenges

Here I discuss the general issues related to developing and applying the bioinformatics methods that I discovered over the course of this dissertation. The scientific challenges specific to pathway analysis and deconvolution have been introduced in the corresponding parts of this thesis.

One of the major practical issues in method development is to get researchers without a computational background to use the most suitable, possibly new, tool instead of the one they are already familiar with. This requires as a minimum easy installation, intuitive usage, and clear documentation and instructions. A graphical user interface could also increase attractiveness of a method. While these requirements sound reasonable, they are not always granted. During the empirical comparison of pathway methods (original publication I), I had to exclude around half of the originally planned methods as I was unable to use them due to reasons like missing documentation, expired links to software, no-longer-available dependences, and bugs in the source code. On a positive note, pathway analysis is a commonly used step in diverse biological studies involving transcriptomic data (and several other data types as well), so pathway analysis methods, though sometimes suboptimal ones, are regularly used by researchers.

While the issue with pathway analysis is about using the most suitable rather than familiar methods, deconvolution still struggles with becoming widely utilized. The literature about deconvolution consists mainly of method-oriented articles, pure application studies that only use deconvolution to investigate a biological phenomenon are more rare and often contain a co-author who has been involved in deconvolution method development. This indicates that deconvolution methods either 1) are not widely known among the target users, 2) do not solve an interesting issue, 3) provide too inaccurate results for practical purposes, or 4) are too difficult to use.

5.2 Limitations of this thesis

Besides obvious limitations related to leaving some related topics out of this thesis, there are several other drawbacks as well. First of all, no new gene expression data has been introduced, but all the projects have been built on publicly available data. Deconvolution projects III and IV especially would have benefited from well known

in-house data with RNAseq bulk samples and purified cell population expressions from the same samples (validation approach 4 in Figure 4). As mentioned before, the publicly available cell type specific data contain too few sample donors for expression deconvolution, which is also an issue.

The second major drawback of this thesis is the lack of graphical user interfaces for the developed methods. While the current implementations come with clear user manuals, they still require the end user to import their data into an R session. This can be a deal breaker for the potential users not comfortable with any level of programming.

5.3 Impact and applications

In this thesis work I have compared different computational methods from a practical perspective and developed robust new tools. The main target audience of these studies are researchers who analyse gene expression data but are not method oriented. I aimed to provide them assistance with selecting a suitable method for their purposes and offer novel tools that are easy to use and provide accurate results with realistic noisy data. Our studies also reveal how accurate the results can be expected to be with different study settings, which is an important yet often overlooked step of test design. In the original publication IV in particular, we also give practical instructions on how the end user can evaluate the accuracy of their findings.

Other method developers may also benefit from our work as they can further build on it. Different validation strategies and semi-simulated datasets with gold standards available can be especially valuable for them. Factors affecting the tested methods' performance are of interest for researchers both developing/comparing methods and those simply choosing which one to apply.

All the articles related to this thesis are open access and the developed software are published under GPL-2 licence. These decisions towards open science allow everyone to utilise our work free of charge. This also includes researchers beyond academia.

5.4 Further research on pathway analysis and deconvolution

Currently our R package PASI (available at `https://github.com/elolab`) includes the basic pathway analysis tool (original publication II). In the future, I aim to add tools for pathway analysis of sc-RNAseq data and longitudinal data. Other planned upgrades are the utilisation of pathway databases other than KEGG and the possibility to use proteomic data as complementary to gene expression data. The latter is especially interesting as KEGG pathways include information about interaction types such as protein-protein interaction (for example modification and binding) or

gene expression interaction (interaction between transcription factor and gene product).

Our next planned deconvolution work is related to generating a signature matrix to be used as an input for composition deconvolution. While there are already a few tools to generate it, none of them utilise the target data to be deconvolved, inclusion of which I believe considerably improves the final deconvolution results. Also, it would be interesting to test if sophisticated expression deconvolution methods designed for cancer studies could be used for more general purposes. The main obstacle to this would be that cancer-focused methods typically expect only two cell types ('tumour' and 'healthy'), which is not the case for most of the tissues. This can be overcome by estimating an expression profile for one cell type t at a time and using all the other cell types as another group. As tumours are more heterogeneous than healthy tissue, it is intuitive to set one cell type t (probably the most abundant one) as 'healthy' and all the other cells in various proportions as 'tumour'. One advantage of using this approach is that using a mixture of different cells as one heterogeneous cell group ('tumour') allows for unknown content in the cell mix. After all, real tissue samples always include some unidentified cells, making the assumption that present cell types would all be known unrealistic. Among cancer oriented methods, there are several attempts to obtain personalised or 'cleaned' results for one cell type. Ideally, if it would be possible to obtain very accurate personalised estimates for the most abundant cell type, an iterative algorithm could be used to obtain personalised S:

1. Estimate personalised expression profiles for the most abundant cell type t
2. Subtract it from bulk matrix E and re-scale C without the cell type t
3. Iterate from step 1 for the new most abundant cell type.

With this approach, each cell type would be the most abundant one (i.e. the easiest one to analyse) at time, eventually providing accurate personalised results for all cell types.

5.5 Conclusions

In this thesis I have evaluated and developed methods to conduct pathway analysis and expression deconvolution. The main goal was to help other researchers applying bioinformatic tools by providing new robust and accurate methods and assisting with selecting a suitable method from those available. Pathway analysis and deconvolution have been introduced in this summary part of the thesis, and specific aspects have been studied in more detail in the original publications I-IV. Our results show that group-level pathway methods utilising pathway structure outperform those not using it, but all the available methods struggle with data containing only minor differences between the sample groups. We also introduced a new pathway method

PASI, which provides sample-level pathway scores. PASI also performs reasonably well with challenging data containing only minor differences between the sample groups and it has favourable performance compared to the other available methods for similar analyses.

Deconvolution is a difficult type of analysis and in publication III we have evaluated expression deconvolution methods. All of the evaluated methods were sensitive to sample size and, similar to composition deconvolution, rare cell types were harder to analyse than abundant ones. In this thesis, we also introduced Rodeo, which is a robust expression deconvolution method designed to tolerate outlier samples in the data. It together with two methods originally implemented for composition deconvolution, namely cs-qprog and cs-lsfit, had the most accurate performance among the tested methods. For detecting csDEGs, methods designed for the purpose are more accurate than general models or expression deconvolution methods. Methods designed to identify cell type specific differential methylation can also be used as their performance is comparable to that of methods' designed for gene expression data. According to our results, identifying csGEPs and identifying csDEGs are both difficult tasks and the user attempting to do these should consider if sufficiently accurate results can be obtained from their particular data. Besides the previously mentioned sample size and cell type abundance, residuals (see publication IV) can also be used to evaluate how challenging the data is to analyse.

List of References

[1] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2008.

[2] Zuguang Gu, Jialin Liu, Kunming Cao, Junfeng Zhang, and Jin Wang. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC systems biology*, 6(1):56, 2012.

[3] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[4] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644, 2014.

[5] Muhammad Farooq Rai, Eric D Tycksen, Linda J Sandell, and Robert H Brophy. Advantages of rna-seq compared to rna microarrays for transcriptome profiling of anterior cruciate ligament tears. *Journal of Orthopaedic Research®*, 36(1):484–497, 2018.

[6] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1): 1–25, 2004.

[7] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[8] Laura L Elo, Sanna Filén, Riitta Lahesmaa, and Tero Aittokallio. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(3):423–431, 2008.

[9] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422, 2010.

[10] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[11] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3):83–92, 2004.

[12] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[13] Pau Creixell, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, et al. Pathway and network analysis of cancer genomes. *Nature methods*, 12(7):615, 2015.

[14] Ivana Ihnatova, Vlad Popovici, and Eva Budinska. A critical comparison of topology-based pathway analysis methods. *PloS one*, 13(1):e0191154, 2018.

[15] Jing Ma, Ali Shojaie, and George Michailidis. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*, 20(1):546, 2019.

[16] Alicia Amadoz, M Hidalgo, Cankut Cubuk, José Carbonell-Caballero, and Joaquín Dopazo. A comparison of mechanistic signaling pathway activity analysis methods. *Brief Bioinform*, 2018.

[17] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966, 2016.

[18] Gordon Fehringer, Geoffrey Liu, Laurent Briollais, Paul Brennan, Christopher I Amos, Margaret R Spitz, Heike Bickeböller, H Erich Wichmann, Angela Risch, and Rayjean J Hung. Comparison of pathway analysis approaches using lung cancer gwas data sets. *PloS one*, 7(2):e31816, 2012.

[19] Hongsheng Gui, Miaoxin Li, Pak C Sham, and Stacey S Cherny. Comparisons of seven algorithms for pathway analysis using the wtccc crohn's disease dataset. *BMC research notes*, 4(1): 386, 2011.

[20] Marina Evangelou, Augusto Rendon, Willem H Ouwehand, Lorenz Wernisch, and Frank Dudbridge. Comparison of methods for competitive tests of pathway analysis. *PloS one*, 7(7): e41018, 2012.

[21] Peter Holmans. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In *Advances in genetics*, volume 72, pages 141–179. Elsevier, 2010.

[22] Heather B Mayes, Sangyun Lee, Andrew D White, Gregory A Voth, and Jessica MJ Swanson. Multiscale kinetic modeling reveals an ensemble of cl–/h+ exchange pathways in clc-ec1 antiporter. *Journal of the American Chemical Society*, 140(5):1793–1804, 2018.

[23] Calin Stoicov, Reza Saffari, Xun Cai, Chhaya Hasyagar, and JeanMarie Houghton. Molecular biology of gastric cancer: Helicobacter infection and gastric adenocarcinoma: bacterial and host factors responsible for altered growth signaling. *Gene*, 341:1–17, 2004.

[24] Marco S Nobile, Daniela Besozzi, Paolo Cazzaniga, Giancarlo Mauri, and Dario Pescini. cupsoda: a cuda-powered simulator of mass-action kinetics. In *International Conference on Parallel Computing Technologies*, pages 344–357. Springer, 2013.

[25] Wolfgang Wiechert and Stephan Noack. Mechanistic pathway modeling for industrial biotechnology: challenging but worthwhile. *Current opinion in biotechnology*, 22(5):604–610, 2011.

[26] Fei Hua, Melanie G Cornejo, Michael H Cardone, Cynthia L Stokes, and Douglas A Lauffenburger. Effects of bcl-2 levels on fas signaling-induced caspase-3 activation: molecular genetic tests of computational model predictions. *The Journal of Immunology*, 175(2):985–995, 2005.

[27] JA Marchand, ME Neugebauer, MC Ing, C-I Lin, JG Pelton, and MCY Chang. Discovery of a pathway for terminal-alkyne amino acid biosynthesis. *Nature*, 567(7748):420–424, 2019.

[28] Blaz Zupan, Janez Demsar, Ivan Bratko, Peter Juvan, John A Halter, Adam Kuspa, and Gad Shaulsky. Genepath: a system for automated construction of genetic networks from mutant data. *Bioinformatics*, 19(3):383–389, 2003.

[29] Marieke Lydia Kuijjer, Matthew George Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *iScience*, 14:226–240, 2019.

[30] Teneale A Stewart, Kunsala TDS Yapa, and Gregory R Monteith. Altered calcium signaling in cancer cells. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(10):2502–2511, 2015.

[31] Peter Burbelo, Anton Wellstein, and Richard G Pestell. Altered rho gtpase signaling pathways in breast cancer cells. *Breast cancer research and treatment*, 84(1):43–48, 2004.

[32] Gary D Bader, Michael P Cary, and Chris Sander. Pathguide: a pathway resource list. *Nucleic acids research*, 34(suppl_1):D504–D506, 2006.

[33] Vijay K Ramanan, Li Shen, Jason H Moore, and Andrew J Saykin. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics*, 28(7): 323–332, 2012.

[34] Lv Jin, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, and Shao-Qi Rao. Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, 12(5):210–220, 2014.

[35] Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Math-

ivanan, Christian Pecquet, et al. Netpath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):1–9, 2010.

[36] Miguel A García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Pathway analysis: state of the art. *Frontiers in physiology*, 6:383, 2015.

[37] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report: The Computer Software Journal for Scient*, 2(3):117–120, 2001.

[38] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34 (suppl_1):D535–D539, 2006.

[39] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2008.

[40] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.

[41] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861, 2012.

[42] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363, 2014.

[43] M Kanehisa and S Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[44] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1): D674–D679, 2008.

[45] Dexter Pratt, Jing Chen, David Welker, Ricardo Rivas, Rudolf Pillich, Vladimir Rynkov, Keiichiro Ono, Carol Miello, Lyndon Hicks, Sandor Szalma, et al. Ndex, the network data exchange. *Cell systems*, 1(4):302–305, 2015.

[46] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A Fulcher, Timothy A Holland, Ingrid M Keseler, Anamika Kothari, Aya Kubo, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, 42(D1):D459–D471, 2014.

[47] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9):2129–2141, 2003.

[48] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010.

[49] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[50] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl_1):D428–D432, 2005.

[51] Alexander R Pico, Thomas Kelder, Martijn P Van Iersel, Kristina Hanspers, Bruce R Conklin, and Chris Evelo. Wikipathways: pathway editing for the people. *PLoS biology*, 6(7):e184, 2008.

[52] ML Green and PD Karp. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, 34(13):3687–3697, 2006.

[53] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The

systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[54] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[55] Lena Strömbäck and Patrick Lambrix. Representations of molecular pathways: an evaluation of sbml, psi mi and biopax. *Bioinformatics*, 21(24):4401–4407, 2005.

[56] Anna Bauer-Mehren, Laura I Furlong, and Ferran Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology*, 5(1):290, 2009.

[57] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019.

[58] Thomas Kelder, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS Biol*, 8(8):e1000472, 2010.

[59] Junjie Su, Byung-Jun Yoon, and Edward R Dougherty. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PloS one*, 4(12):e8161, 2009.

[60] Jingwen Yan, Shannon L Risacher, Li Shen, and Andrew J Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6):1370–1381, 2018.

[61] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

[62] Maria K Jaakkola, Aidan J McGlinchey, Riku Klén, and Laura L Elo. Pasi: A novel pathway method to identify delicate group effects. *PloS one*, 13(7):e0199991, 2018.

[63] Chengyu Liu, Rainer Lehtonen, and Sampsa Hautaniemi. Perpas: topology-based single sample pathway analysis method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 15(3):1022–1027, 2018.

[64] TaeJin Ahn, Eunjin Lee, Nam Huh, and Taesung Park. Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics*, 30(17):i422–i429, 2014.

[65] Gunes Gundem and Nuria Lopez-Bigas. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. *Genome medicine*, 4(3):28, 2012.

[66] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.

[67] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, 10(1):47, 2009.

[68] Raphael BM Aggio, Katya Ruggiero, and Silas Granato Villas-Bôas. Pathway activity profiling (papi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics*, 26(23): 2969–2976, 2010.

[69] Seungyoon Nam, Hae Ryung Chang, Kyung-Tae Kim, Myeong-Cherl Kook, Dongwan Hong, ChangHyuk Kwon, Hae Rim Jung, Hee Seo Park, Garth Powis, Han Liang, et al. Pathome: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene*, 33(41): 4941, 2014.

[70] Thair Judeh, Cole Johnson, Anuj Kumar, and Dongxiao Zhu. Teak: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic acids research*, 41(3):1425–1437, 2012.

[71] Chunquan Li, Xia Li, Yingbo Miao, Qianghu Wang, Wei Jiang, Chun Xu, Jing Li, Junwei Han, Fan Zhang, Binsheng Gong, et al. Subpathwayminer: a software package for flexible identification of pathways. *Nucleic acids research*, 37(19):e131–e131, 2009.

[72] Aristidis G Vrahatis, Panos Balomenos, Athanasios K Tsakalidis, and Anastasios Bezerianos. Desubs: an r package for flexible identification of differentially expressed subpathways using rna-seq experiments. *Bioinformatics*, 32(24):3844–3846, 2016.

[73] Chenchen Feng, Jian Zhang, Xuecang Li, Bo Ai, Junwei Han, Qiuyu Wang, Taiming Wei, Yong Xu, Meng Li, Shang Li, et al. Subpathway-corsp: Identification of metabolic subpathways via integrating expression correlations and topological features between metabolites and genes of interest within pathways. *Scientific reports*, 6:33262, 2016.

[74] Xia Li, Chunquan Li, Desi Shang, Jing Li, Junwei Han, Yingbo Miao, Yan Wang, Qianghu Wang, Wei Li, Chao Wu, et al. The implications of relationships between human diseases and metabolic subpathways. *PloS one*, 6(6):e21131, 2011.

[75] Chunquan Li, Junwei Han, Qianlan Yao, Chendan Zou, Yanjun Xu, Chunlong Zhang, Desi Shang, Lingyun Zhou, Chaoxia Zou, Zeguo Sun, et al. Subpathway-gm: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic acids research*, 41(9):e101–e101, 2013.

[76] Guoan Chen, Tarek G Gharib, Chiang-Ching Huang, Jeremy MG Taylor, David E Misek, Sharon LR Kardia, Thomas J Giordano, Mark D Iannettoni, Mark B Orringer, Samir M Hanash, et al. Discordant protein and mrna expression in lung adenocarcinomas. *Molecular & cellular proteomics*, 1(4):304–313, 2002.

[77] Trey Ideker, Vesteinn Thorsson, Jeffrey A Ranish, Rowan Christmas, Jeremy Buhler, Jimmy K Eng, Roger Bumgarner, David R Goodlett, Ruedi Aebersold, and Leroy Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292 (5518):929–934, 2001.

[78] Timothy J Griffin, Steven P Gygi, Trey Ideker, Beate Rist, Jimmy Eng, Leroy Hood, and Ruedi Aebersold. Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, 1(4):323–333, 2002.

[79] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, 2014.

[80] Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C Chambers, Lisa J Zimmerman, Kent F Shaddox, Sangtae Kim, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, 2014.

[81] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[82] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[83] Miaolong Lu and Xianquan Zhan. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA Journal*, 9(1):77–102, 2018.

[84] Sebastian Canzler, Jana Schor, Wibke Busch, Kristin Schubert, Ulrike E Rolle-Kampczyk, Hervé Seitz, Hennicke Kamp, Martin von Bergen, Roland Buesen, and Jörg Hackermüller. Prospects and challenges of multi-omics data integration in toxicology. *Archives of Toxicology*, pages 1–18, 2020.

[85] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[86] Marta Paczkowska, Jonathan Barenboim, Nardnisa Sintupisut, Natalie S Fox, Helen Zhu, Diala Abd-Rabbo, Miles W Mee, Paul C Boutros, and Jüri Reimand. Integrative pathway enrichment analysis of multivariate omics data. *Nature communications*, 11(1):1–16, 2020.

[87] Paolo Martini, Monica Chiogna, Enrica Calura, and Chiara Romualdi. Mosclip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic acids research*, 47(14):e80–e80, 2019.

[88] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[89] Josien C van Wolfswinkel, Daniel E Wagner, and Peter W Reddien. Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell stem cell*, 15(3): 326–339, 2014.

[90] Emir Hodzic. Single-cell analysis: Advances and future perspectives. *Bosnian journal of basic medical sciences*, 16(4):313, 2016.

[91] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[92] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[93] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3):241, 2016.

[94] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740, 2014.

[95] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11): 1083–1086, 2017.

[96] Ying Ma, Shiquan Sun, Xuequn Shang, Evan T Keller, Mengjie Chen, and Xiang Zhou. Integrative differential expression and gene set enrichment analysis using summary statistics for scrna-seq studies. *Nature communications*, 11(1):1–13, 2020.

[97] Massimo Andreatta and Santiago J Carmona. Ucell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal*, 19:3796–3798, 2021.

[98] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282): 189–196, 2016.

[99] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[100] G Sales, E Calura, and C Romualdi. graphite: Graph interaction from pathway topological environment. r package version 1.16.0. 2015.

[101] Ivana Ihnatova and Eva Budinska. Topaseq: an r package for topology-based pathway analysis of microarray and rna-seq data. *BMC bioinformatics*, 16(1):350, 2015.

[102] Shaoyan Sun, Xiangtian Yu, Fengnan Sun, Ying Tang, Juan Zhao, and Tao Zeng. Dynamically characterizing individual clinical change by the steady state of disease-associated pathway. *BMC bioinformatics*, 20(25):1–12, 2019.

[103] Jakub Mieczkowski, Karolina Swiatek-Machado, and Bozena Kaminska. Identification of pathway deregulation–gene expression based analysis of consistent signal transduction. *PloS one*, 7 (7):e41541, 2012.

[104] Senol Isci, Cengizhan Ozturk, Jon Jones, and Hasan H Otu. Pathway analysis of high-throughput biological data within a bayesian network framework. *Bioinformatics*, 27(12):1667–1674, 2011.

[105] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

[106] Joanna Zyla, Michal Marczyk, Teresa Domaszewska, Stefan HE Kaufmann, Joanna Polanska, and January Weiner. Gene set enrichment for reproducible science: comparison of cerno and eight other algorithms. *Bioinformatics*, 2019.

[107] Paolo Martini, Gabriele Sales, M Sofia Massa, Monica Chiogna, and Chiara Romualdi. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research*, 41(1):e19–e19, 2012.

[108] Yan Jiao, Katherine Lawler, Gargi S Patel, Arnie Purushotham, Annette F Jones, Anita Grigoriadis, Andrew Tutt, Tony Ng, and Andrew E Teschendorff. Dart: Denoising algorithm based on relevance network topology improves molecular pathway activity inference. *BMC bioinformatics*, 12(1):403, 2011.

[109] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

[110] Xin Zhou and Zhen Su. Easygo: Gene ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC genomics*, 8(1):246, 2007.

[111] Monther Alhamdoosh, Milica Ng, Nicholas J Wilson, Julie M Sheridan, Huy Huynh, Michael J Wilson, and Matthew E Ritchie. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, 33(3):414–424, 2017.

[112] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, Reinhard Schneider, and Alfonso Valencia. Enrichnet: network-based gene set enrichment analysis. *Bioinformatics*, 28(18):i451–i457, 2012.

[113] Zhaoyuan Fang, Weidong Tian, and Hongbin Ji. A network-based gene-weighting approach for pathway analysis. *Cell research*, 22(3):565, 2012.

[114] Ludwig Geistlinger, Gergely Csaba, Robert Küffner, Nicola Mulder, and Ralf Zimmer. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, 27 (13):i366–i373, 2011.

[115] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[116] Barry R Zeeberg, Weimin Feng, Geoffrey Wang, May D Wang, Anthony T Fojo, Margot Sunshine, Sudarshan Narasimhan, David W Kane, William C Reinhold, Samir Lababidi, et al. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome biology*, 4(4):R28, 2003.

[117] Tim Beißbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[118] Bradley Efron, Robert Tibshirani, et al. On testing the significance of sets of genes. *The annals of applied statistics*, 1(1):107–129, 2007.

[119] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[120] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):7, 2013.

[121] Ivan V Ozerov, Ksenia V Lezhnina, Evgeny Izumchenko, Artem V Artemov, Sergey Medintsev, Quentin Vanhaelen, Alexander Aliper, Jan Vijg, Andreyan N Osipov, Ivan Labat, et al. In silico pathway activation network decomposition analysis (ipanda) as a method for biomarker development. *Nature communications*, 7:13427, 2016.

[122] Antony Kaspi and Mark Ziemann. mitch: multi-contrast pathway enrichment for multi-omics and single-cell profiling data. *BMC genomics*, 21(1):1–17, 2020.

[123] Andrey Alexeyenko, Woojoo Lee, Maria Pernemalm, Justin Guegan, Philippe Dessen, Vladimir Lazar, Janne Lehtiö, and Yudi Pawitan. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC bioinformatics*, 13(1):226, 2012.

[124] Duanchen Sun, Yinliang Liu, Xiang-Sun Zhang, and Ling-Yun Wu. Netgen: a novel network-based probabilistic generative model for gene set functional enrichment analysis. *BMC systems biology*, 11(4):75, 2017.

[125] Ali Shojaie and George Michailidis. Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, 16(3):407–426, 2009.

[126] Ali Shojaie and George Michailidis. Network enrichment analysis in complex experiments. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[127] Adi Laurentiu Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC bioinformatics*, 13(1):136, 2012.

[128] Bhaskar Dutta, Anders Wallqvist, and Jaques Reifman. Pathnet: a tool for pathway analysis using topological information. *Source code for biology and medicine*, 7(1):10, 2012.

[129] Sharon I Greenblum, Sol Efroni, Carl F Schaefer, and Ken H Buetow. The pathologist: an automated tool for pathway-centric analysis. *BMC bioinformatics*, 12(1):133, 2011.

[130] Sahar Ansari, Calin Voichita, Michele Donato, Rebecca Tagett, and Sorin Draghici. A novel pathway analysis approach based on the unexplained disregulation of genes. *Proceedings of the IEEE*, 105(3):482–495, 2016.

[131] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome research*, 17(10):1537–1545, 2007.

[132] Chengyu Liu, Rainer Lehtonen, and Sampsa Hautaniemi. Perpas: topology-based single sample pathway analysis method. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(3):1022–1027, 2017.

[133] Jui-Hung Hung, Troy W Whitfield, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome biology*, 11(2):R23, 2010.

[134] Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Improving gene set analysis of microarray data by sam-gs. *BMC bioinformatics*, 8(1):242, 2007.

[135] Momeneh Foroutan, Dharmesh D Bhuva, Ruqian Lyu, Kristy Horan, Joseph Cursons, and Melissa J Davis. Single sample scoring of molecular phenotypes. *BMC bioinformatics*, 19 (1):404, 2018.

[136] Shouguo Gao and Xujing Wang. Tappa: topological analysis of pathway phenotype association. *Bioinformatics*, 23(22):3100–3102, 2007.

[137] Maysson Al-Haj Ibrahim, Sabah Jassim, Michael Anthony Cawthorne, and Kenneth Langlands. A topology-based score for pathway enrichment. *Journal of Computational Biology*, 19(5):563–573, 2012.

[138] Enrico Glaab, Anaïs Baudot, Natalio Krasnogor, and Alfonso Valencia. Topogsa: network topological gene set analysis. *Bioinformatics*, 26(9):1271–1272, 2010.

[139] Mark D Robinson, Jörg Grigull, Naveed Mohammad, and Timothy R Hughes. Funspec: a web-based cluster interpreter for yeast. *BMC bioinformatics*, 3(1):35, 2002.

[140] Kam D Dahlquist, Nathan Salomonis, Karen Vranizan, Steven C Lawlor, and Bruce R Conklin. Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nature genetics*, 31(1):19, 2002.

[141] Purvesh Khatri, Sorin Draghici, G Charles Ostermeier, and Stephen A Krawetz. Profiling gene expression using onto-express. *Genomics*, 79(2):266–270, 2002.

[142] Cristian I Castillo-Davis and Daniel L Hartl. Genemerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, 2003.

[143] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in bioinformatics*, 9(3):189–197, 2008.

[144] Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics*, 15(4):504–518, 2014.

[145] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[146] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267, 2003.

[147] Hui Zhou and Tian Zheng. Bayesian hierarchical graph-structured model for pathway analysis using gene expression data. *Statistical applications in genetics and molecular biology*, 12(3): 393–412, 2013.

[148] Seiya Imoto, Sunyong Kim, Takao Goto, Sachiyo Aburatani, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of bioinformatics and computational biology*, 1(02): 231–252, 2003.

[149] Melike Korucuoglu, Senol Isci, Arzucan Ozgur, and Hasan H Otu. Bayesian pathway analysis of cancer microarray data. *PloS one*, 9(7):e102803, 2014.

[150] Jörg Rahnenführer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, 3(1):1–29, 2004.

[151] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

[152] Rebecca A Lea and Kathy K Niakan. Human germline genome editing. *Nature cell biology*, 21 (12):1479–1489, 2019.

[153] Chenggang Yu, Hyung Jun Woo, Xueping Yu, Tatsuya Oyama, Anders Wallqvist, and Jaques Reifman. A strategy for evaluating pathway analysis methods. *BMC bioinformatics*, 18(1): 1–11, 2017.

[154] Sangsoo Lim, Sangseon Lee, Inuk Jung, Sungmin Rhee, and Sun Kim. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*, 2018.

[155] Muhammad Faiz Misman, Safaai Deris, Suhairul ZM Hashim, R Jumali, and Mohd Saberi Mohamad. Pathway-based microarray analysis for defining statistical significant phenotype-related pathways: a review of common approaches. In *2009 International Conference on Information Management and Engineering*, pages 496–500. IEEE, 2009.

[156] Joaquin Dopazo. Formulating and testing hypotheses in functional genomics. *Artificial intelligence in medicine*, 45(2-3):97–107, 2009.

[157] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Calin Voichita, and Sorin Draghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, 4:278, 2013.

[158] Joanna Zyla, Michal Marczyk, January Weiner, and Joanna Polanska. Ranking metrics in gene set enrichment analysis: do they matter? *BMC bioinformatics*, 18(1):256, 2017.

[159] Maria K Jaakkola and Laura L Elo. Empirical comparison of structure-based pathway methods. *Briefings in bioinformatics*, 17(2):336–345, 2015.

[160] Michaela Bayerlová, Klaus Jung, Frank Kramer, Florian Klemm, Annalen Bleckmann, and Tim Beißbarth. Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, 16(1):334, 2015.

[161] Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, and Sorin Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome biology*, 20(1):1–15, 2019.

[162] Joanna Zyla, Michal Marczyk, and Joanna Polanska. Reproducibility of finding enriched gene sets in biological data analysis. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 146–154. Springer, 2017.

[163] Sarah Song and Michael A Black. Microarray-based gene set analysis: a comparison of current methods. *BMC bioinformatics*, 9(1):502, 2008.

[164] Qi Liu, Irina Dinu, Adeniyi J Adewale, John D Potter, and Yutaka Yasui. Comparative evaluation of gene-set analysis methods. *BMC bioinformatics*, 8(1):431, 2007.

[165] S Debey, U Schoenbeck, M Hellmich, BS Gathof, R Pillai, T Zander, and JL Schultze. Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. *The pharmacogenomics journal*, 4(3):193, 2004.

[166] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

[167] André Nicolau Aquime Gonçalves, Melissa Lever, Pedro Russo, Bruno Gomes-Correia, Alysson H Urbanski, Gabriele Pollara, Mahdad Noursadeghi, Vinicius Maracaja-Coutinho, and Helder I Nakaya. Assessing the impact of sample heterogeneity on transcriptome analysis of human diseases using mdp webtool. *Frontiers in genetics*, 10:971, 2019.

[168] Diether Lambrechts, Els Wauters, Bram Boeckx, Sara Aibar, David Nittner, Oliver Burton, Ayse Bassez, Herbert Decaluwé, Andreas Pircher, Kathleen Van den Eynde, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature medicine*, 24(8):1277–1289, 2018.

[169] Rose Du, Vince Carey, and Scott T Weiss. deconvseq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics*, 35(24):5095–5102, 2019.

[170] Dvir Aran, Zicheng Hu, and Atul J Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. *Genome biology*, 18(1):220, 2017.

[171] Etienne Becht, Nicolas A Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, Wolf H Fridman, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology*, 17(1):218, 2016.

[172] Bo Li, Jun S Liu, and X Shirley Liu. Revisit linear regression-based deconvolution methods for tumor gene expression data. *Genome biology*, 18(1):127, 2017.

[173] Aaron M Newman, Andrew J Gentles, Chih Long Liu, Maximilian Diehn, and Ash A Alizadeh. Data normalization considerations for digital tumor dissection. *Genome biology*, 18(1):1–6, 2017.

[174] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4(1):1–11, 2013.

[175] Erica AK DePasquale, Daniel J Schnell, Pieter-Jan Van Camp, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, Harinder Singh, and Nathan Salomonis. Doubletdecon: Deconvoluting doublets from single-cell rna-sequencing data. *Cell reports*, 29(6):1718–1727, 2019.

[176] Shu-Hwa Chen, Wen-Yu Kuo, Sheng-Yao Su, Wei-Chun Chung, Jen-Ming Ho, Henry Horng-Shing Lu, and Chung-Yen Lin. A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. *BMC bioinformatics*, 19(4):154, 2018.

[177] Ting Gong and Joseph D Szustakowski. Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

[178] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[179] Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature communications*, 10(1): 1–9, 2019.

[180] Huamei Li, Amit Sharma, Wenglong Ming, Xiao Sun, and Hongde Liu. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC genomics*, 21(1):1–15, 2020.

[181] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

[182] Yen-Jung Chiu, Yi-Hsuan Hsieh, and Yen-Hua Huang. Improved cell composition deconvolution method of bulk gene expression profiles to quantify subsets of immune cells. *BMC medical genomics*, 12(8):169, 2019.

[183] Yael Steuerman and Irit Gat-Viks. Exploiting gene-expression deconvolution to probe the genetics of the immune system. *PLoS computational biology*, 12(4), 2016.

[184] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, 11(1):1–14, 2020.

[185] Gianni Monaco, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carre, Nicolas Burdin, Lucian Visan, Michele Ceccarelli, Michael Poidinger, et al. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. *Cell reports*, 26(6):1627–1640, 2019.

[186] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

[187] Alexandre Kuhn, Azad Kumar, Alexandra Beilina, Allissa Dillman, Mark R Cookson, and Andrew B Singleton. Cell population-specific expression analysis of human cerebellum. *BMC genomics*, 13(1):610, 2012.

[188] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F Black, Joachim Selbig, Shreemanta K Parida, Stefan HE Kaufmann, and Marc Jacobsen. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC bioinformatics*, 11(1):27, 2010.

[189] Maria Chikina, Elena Zaslavsky, and Stuart C Sealfon. Cellcode: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, 31 (10):1584–1591, 2015.

[190] Francesco Vallania, Andrew Tam, Shane Lofgren, Steven Schaffert, Tej D Azad, Erika Bongen, Winston Haynes, Meia Alsup, Michael Alonso, Mark Davis, et al. Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature communications*, 9(1):4735, 2018.

[191] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin MacBeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell rna-seq data. *Nature communications*, 8(1):2032, 2017.

[192] Paul Yousefi, Karen Huen, Hong Quach, Girish Motwani, Alan Hubbard, Brenda Eskenazi, and Nina Holland. Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environmental and molecular mutagenesis*, 56(9):751–758, 2015.

[193] Ziyi Chen, Chengyang Ji, Qin Shen, Wei Liu, F Xiao-Feng Qin, and Aiping Wu. Tissue-specific deconvolution of immune cell composition by integrating bulk and single-cell transcriptomes. *Bioinformatics*, 36(3):819–827, 2020.

[194] Elvira Mass, Ivan Ballesteros, Matthias Farlik, Florian Halbritter, Patrick Günther, Lucile Crozet, Christian E Jacome-Galarza, Kristian Händler, Johanna Klughammer, Yasuhiro Kobayashi, et al. Specification of tissue-resident macrophages during organogenesis. *Science*, 353(6304), 2016.

[195] Andrew E Teschendorff and Shijie C Zheng. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768, 2017.

[196] Ajit J Nirmal, Tim Regan, Barbara B Shih, David A Hume, Andrew H Sims, and Tom C Freeman. Immune cell gene signatures for profiling the microenvironment of solid tumors. *Cancer immunology research*, 6(11):1388–1400, 2018.

[197] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6, 2017.

[198] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2019.

[199] Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019, 2019.

[200] Mathias Uhlen, Max J Karlsson, Wen Zhong, Abdellah Tebani, Christian Pou, Jaromir Mikes, Tadepally Lakshmikanth, Björn Forsström, Fredrik Edfors, Jacob Odeberg, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science*, 366(6472), 2019.

[201] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[202] Jason E Shoemaker, Tiago JS Lopes, Samik Ghosh, Yukiko Matsuoka, Yoshihiro Kawaoka, and Hiroaki Kitano. Cten: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC genomics*, 13(1):460, 2012.

[203] Hannah A Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nature methods*, 16(10):983–986, 2019.

[204] Ze Zhang, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, Wei Guo, Eric W Stawiski, Zora Modrusan, et al. Scina: Semi-supervised analysis of single cells in silico. *Genes*, 10(7):531, 2019.

[205] Sergii Domanskyi, Anthony Szedlak, Nathaniel T Hawkins, Jiayin Wang, Giovanni Paternostro, and Carlo Piermarocchi. Polled digital cell sorter (p-dcs): Automatic identification of hematological cell types from single cell rna-sequencing clusters. *BMC bioinformatics*, 20(1):369, 2019.

[206] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in Bioinformatics*, 2020.

[207] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[208] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

[209] Samuel A Danziger, David L Gibbs, Ilya Shmulevich, Mark McConnell, Matthew WB Trotter, Frank Schmitz, David J Reiss, and Alexander V Ratushny. Adapts: Automated deconvolution augmentation of profiles for tissue specific cells. *PloS one*, 14(11):e0224693, 2019.

[210] Wenjun Ju, Casey S Greene, Felix Eichinger, Viji Nair, Jeffrey B Hodgin, Markus Bitzer, Youngsuk Lee, Qian Zhu, Masami Kehata, Min Li, et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome research*, 23(11):1862–1873, 2013.

[211] Bradlee D Nelms, Levi Waldron, Luis A Barrera, Andrew W Weflen, Jeremy A Goettel, Guoji Guo, Robert K Montgomery, Marian R Neutra, David T Breault, Scott B Snapper, et al. Cellmapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types. *Genome biology*, 17(1):201, 2016.

[212] Qiuyu Lian, Hongyi Xin, Jianzhu Ma, Liza Konnikova, Wei Chen, Jin Gu, and Kong Chen. Artificial-cell-type aware cell type classification in cite-seq. *Bioinformatics*, 36(Supplement_1): i542–i550, 2020.

[213] Renaud Gaujoux and Cathal Seoighe. Cellmix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, 29(17):2211–2212, 2013.

[214] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung

single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2): 163–172, 2019.

[215] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, et al. Arrayexpress update—simplifying data submissions. *Nucleic acids research*, 43(D1):D1113–D1116, 2014.

[216] Alison Abbott. Europe to map the human epigenome. *Nature*, 477(7366):518, 2011.

[217] Benjamin J Schmiedel, Divya Singh, Ariel Madrigal, Alan G Valdovino-Gonzalez, Brandie M White, Jose Zapardiel-Gonzalo, Brendan Ha, Gokmen Altay, Jason A Greenbaum, Graham McVicker, et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, 175(6):1701–1715, 2018.

[218] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[219] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1): D991–D995, 2012.

[220] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

[221] Tracy SP Heng, Michio W Painter, Kutlu Elpek, Veronika Lukacs-Kornek, Nora Mauermann, Shannon J Turley, Daphne Koller, Francis S Kim, Amy J Wagers, Natasha Asinovski, et al. The immunological genome project: networks of gene expression in immune cells. *Nature immunology*, 9(10):1091, 2008.

[222] Hadas Ner-Gaon, Ariel Melchior, Nili Golan, Yael Ben-Haim, and Tal Shay. Jinglebells: a repository of immune-related single-cell rna–sequencing datasets. *The Journal of Immunology*, 198(9):3375–3379, 2017.

[223] Marina Lizio, Imad Abugessaisa, Shuhei Noguchi, Atsushi Kondo, Akira Hasegawa, Chung Chau Hon, Michiel De Hoon, Jessica Severin, Shinya Oki, Yoshihide Hayashizaki, et al. Update of the fantom web resource: expansion to provide additional transcriptome atlases. *Nucleic acids research*, 47(D1):D752–D758, 2019.

[224] Marina Lizio, Jayson Harshbarger, Hisashi Shimoji, Jessica Severin, Takeya Kasukawa, Serkan Sahin, Imad Abugessaisa, Shiro Fukuda, Fumi Hori, Sachi Ishikawa-Kato, et al. Gateways to the fantom5 promoter level mammalian expression atlas. *Genome biology*, 16(1):1–14, 2015.

[225] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible rna-seq analysis using recount2. *Nature biotechnology*, 35(4):319–321, 2017.

[226] Imad Abugessaisa, Shuhei Noguchi, Michael Böttcher, Akira Hasegawa, Tsukasa Kouno, Sachi Kato, Yuhki Tada, Hiroki Ura, Kuniya Abe, Jay W Shin, et al. Scportalen: human and mouse single-cell centric database. *Nucleic acids research*, 46(D1):D781–D787, 2017.

[227] Yuan Cao, Junjie Zhu, Guangchun Han, Peilin Jia, and Zhongming Zhao. scrnaseqdb: a database for gene expression profiling in human single cell by rna-seq. *bioRxiv*, page 104810, 2017.

[228] Konstantina Dimitrakopoulou, Elisabeth Wik, Lars A Akslen, and Inge Jonassen. Deblender: a semi-/unsupervised multi-operational computational method for complete deconvolution of expression data from heterogeneous samples. *BMC bioinformatics*, 19(1):408, 2018.

[229] Richard A Oram, Emily K Sims, and Carmella Evans-Molina. Beta cells in type 1 diabetes: mass and function; sleeping or dead? *Diabetologia*, 62(4):567–577, 2019.

[230] Carol J Lam, Daniel R Jacobson, Matthew M Rankin, Aaron R Cox, and Jake A Kushner. β cells persist in t1d pancreata without evidence of ongoing β-cell turnover or neogenesis. *The Journal of Clinical Endocrinology & Metabolism*, 102(8):2647–2659, 2017.

[231] JJ Meier, A Bhushan, AE Butler, RA Rizza, and PC Butler. Sustained beta cell apoptosis in patients with long-standing type 1 diabetes: indirect evidence for islet regeneration? *Diabetologia*, 48(11):2221–2228, 2005.

[232] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B Suraokar, Lixia Diao, Ignacio I Wistuba, and Wenyi Wang. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[233] Gerald Quon, Syed Haider, Amit G Deshwar, Ang Cui, Paul C Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome medicine*, 5(3):29, 2013.

[234] Edmund R Glass and Mikhail G Dozmorov. Improving sensitivity of linear regression-based cell type-specific differential expression deconvolution with per-gene vs. global significance threshold. In *BMC bioinformatics*, volume 17, page 334. BioMed Central, 2016.

[235] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4):287, 2010.

[236] Debashis Ghosh. Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, 20(11):1663–1669, 2004.

[237] Timo Erkkilä, Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, Ilya Shmulevich, and Harri Lähdesmäki. Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26(20):2571–2577, 2010.

[238] Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard LM Faull, and Ruth Luthi-Carter. Population-specific expression analysis (psea) reveals molecular changes in diseased brain. *Nature methods*, 8(11):945, 2011.

[239] Ziyi Li and Hao Wu. Toast: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome biology*, 20(1):190, 2019.

[240] Chong Jin, Mengjie Chen, Dan-Yu Lin, and Wei Sun. Cell-type-aware analysis of rna-seq data. *Nature Computational Science*, 1(4):253–261, 2021.

[241] Shijie C Zheng, Charles E Breeze, Stephan Beck, and Andrew E Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods*, 15 (12):1059–1066, 2018.

[242] Elior Rahmani, Regev Schweiger, Brooke Rhead, Lindsey A Criswell, Lisa F Barcellos, Eleazar Eskin, Saharon Rosset, Sriram Sankararaman, and Eran Halperin. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications*, 10 (1):1–11, 2019.

[243] Maria K Jaakkola and Laura L Elo. Computational deconvolution to estimate cell type-specific gene expression from bulk data. *NAR Genomics and Bioinformatics*, 3(1):lqaa110, 2021.

[244] Ting Gong, Nicole Hartmann, Isaac S Kohane, Volker Brinkmann, Frank Staedtler, Martin Letzkus, Sandrine Bongiovanni, and Joseph D Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PloS one*, 6(11), 2011.

[245] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.

[246] Shai S Shen-Orr and Renaud Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology*, 25(5):571–578, 2013.

[247] Kai Kang, Qian Meng, Igor Shats, David M Umbach, Melissa Li, Yuanyuan Li, Xiaoling Li, and Leping Li. A novel computational complete deconvolution method using rna-seq data. *bioRxiv*, page 496596, 2018.

[248] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):1–9, 2019.

[249] Francisco Avila Cobos, Jo Vandesompele, Pieter Mestdagh, and Katleen De Preter. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34 (11):1969–1979, 2018.

[250] S Stejskal, I Koutna, and Z Rucka. Isolation of granulocytes: which transcriptome do we analyse—neutrophils or eosinophils. *Folia Biol.(Praha)*, 56:252–255, 2010.

[251] Mark T Quinn, Frank R DeLeo, and Gary M Bokoch. *Neutrophil methods and protocols*, volume 412. Springer, 2007.

[252] Niya Wang, Eric P Hoffman, Lulu Chen, Li Chen, Zhen Zhang, Chunyu Liu, Guoqiang Yu, David M Herrington, Robert Clarke, and Yue Wang. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Scientific reports*, 6:18909, 2016.

[253] Yasin Şenbabaoğlu, Ron S Gejman, Andrew G Winer, Ming Liu, Eliezer M Van Allen, Guillermo de Velasco, Diana Miao, Irina Ostrovnaya, Esther Drill, Augustin Luna, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger rna signatures. *Genome biology*, 17(1):231, 2016.

[254] Bo Li, Eric Severson, Jean-Christophe Pignon, Haoquan Zhao, Taiwen Li, Jesse Novak, Peng Jiang, Hui Shen, Jon C Aster, Scott Rodig, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome biology*, 17(1):174, 2016.

[255] Peter S Linsley, Cate Speake, Elizabeth Whalen, and Damien Chaussabel. Copy number loss of the interferon gene cluster in melanomas is linked to reduced t cell infiltrate and poor patient prognosis. *PLoS One*, 9(10):e109760, 2014.

[256] Martin Hoffmann, Dirk Pohlers, Dirk Koczan, Hans-Jürgen Thiesen, Stefan Wölfl, and Raimund W Kinne. Robust computational reconstitution–a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC bioinformatics*, 7(1):369, 2006.

[257] Jennifer Clarke, Pearl Seo, and Bertrand Clarke. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, 26(8):1043–1049, 2010.

[258] Rafael A Irizarry, Zhijin Wu, and Harris A Jaffee. Comparison of affymetrix genechip expression measures. *Bioinformatics*, 22(7):789–794, 2006.

[259] Ziyi Chen, Anfei Huang, Jiya Sun, Taijiao Jiang, F Xiao-Feng Qin, and Aiping Wu. Inference of immune cell composition on the expression profiles of mouse tissue. *Scientific reports*, 7:40508, 2017.

[260] David A Liebner, Kun Huang, and Jeffrey D Parvin. Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, 30(5):682–689, 2013.

[261] Haijing Jin, Ying-Wooi Wan, and Zhandong Liu. Comprehensive evaluation of rna-seq quantification methods for linearity. *BMC bioinformatics*, 18(4):117, 2017.

[262] Yi Zhong and Zhandong Liu. Gene expression deconvolution in linear space. *Nature methods*, 9 (1):8, 2012.

[263] Yi Zhong, Ying-Wooi Wan, Kaifang Pang, Lionel ML Chow, and Zhandong Liu. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC bioinformatics*, 14(1): 89, 2013.

[264] Casey P Shannon, Robert Balshaw, Raymond T Ng, Janet E Wilson-McManus, Paul Keown, Robert McMaster, Bruce M McManus, David Landsberg, Nicole M Isbel, Greg Knoll, et al. Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PloS one*, 9(4):e95224, 2014.

[265] Catalina V Anghel, Gerald Quon, Syed Haider, Francis Nguyen, Amit G Deshwar, Quaid D Morris, and Paul C Boutros. Isopurer: an r implementation of a computational purification algorithm of mixed tumour profiles. *BMC bioinformatics*, 16(1):156, 2015.

[266] Harri Lähdesmäki, Valerie Dunmire, Olli Yli-Harja, Wei Zhang, et al. In silico microdissection of microarray data from heterogeneous cell populations. *BMC bioinformatics*, 6(1):54, 2005.

[267] Konstantin Zaitsev, Monika Bambouskova, Amanda Swain, and Maxim N Artyomov. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, 10(1):2209, 2019.

[268] Kai Kang, Qian Meng, Igor Shats, David M Umbach, Melissa Li, Yuanyuan Li, Xiaoling Li, and Leping Li. Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLoS Computational Biology*, 15(12), 2019.

[269] An-Shun Tai, George Tseng, and Wen-Ping Hsieh. Bayice: A hierarchical bayesian deconvolution model with stochastic search variable selection. *BioRxiv*, page 732743, 2019.

[270] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature Communications*, 11(1):1–11, 2020.

[271] Andrew T McKenzie, Minghui Wang, Mads E Hauberg, John F Fullard, Alexey Kozlenkov, Alexandra Keenan, Yasmin L Hurd, Stella Dracheva, Patrizia Casaccia, Panos Roussos, et al. Brain cell type specific gene expression and co-expression network architectures. *Scientific reports*, 8(1):1–19, 2018.

[272] Yipeng Wang, Xiao-Qin Xia, Zhenyu Jia, Anne Sawyers, Huazhen Yao, Jessica Wang-Rodriquez, Dan Mercola, and Michael McClelland. In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer research*, 70(16):6448–6455, 2010.

[273] Abolfazl Doostparast Torshizi, Jubao Duan, and Kai Wang. A computational method for direct imputation of cell type-specific expression profiles and cellular compositions from bulk-tissue rna-seq in brain disorders. *NAR Genomics and Bioinformatics*, 3(2):lqab056, 2021.

[274] Jeremy A Miller, Chaochao Cai, Peter Langfelder, Daniel H Geschwind, Sunil M Kurian, Daniel R Salomon, and Steve Horvath. Strategies for aggregating gene expression data: the collapserows r function. *BMC bioinformatics*, 12(1):322, 2011.

[275] Qi Shen, Jiyuan Hu, Ning Jiang, Xiaohua Hu, Zewei Luo, and Hong Zhang. contamde: differential expression analysis of rna-seq data for contaminated tumor samples. *Bioinformatics*, 32 (5):705–712, 2015.

[276] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, Irit Gat-Viks, et al. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular systems biology*, 10(2), 2014.

[277] Xianlu Laura Peng, Richard A Moffitt, Robert J Torphy, Keith E Volmar, and Jen Jen Yeh. De novo compartment deconvolution and weight estimation of tumor samples using decoder. *Nature communications*, 10(1):1–11, 2019.

[278] Raúl Aguirre-Gamboa, Niek de Klein, Jennifer di Tommaso, Annique Claringbould, Monique GP van der Wijst, Dylan de Vries, Harm Brugge, Roy Oelen, Urmo Võsa, Maria M Zorro, et al. Deconvolution of bulk blood eqtl effects into immune cell subpopulations. *BMC Bioinformatics*, 21(1):1–23, 2020.

[279] Peng Lu, Aleksey Nakorchevskiy, and Edward M Marcotte. Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences*, 100(18):10370–10375, 2003.

[280] Zeya Wang, Shaolong Cao, Jeffrey S Morris, Jaeil Ahn, Rongjie Liu, Svitlana Tyekucheva, Fan Gao, Bo Li, Wei Lu, Ximing Tang, et al. Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *iScience*, 9:451–460, 2018.

[281] Gregory J Hunt, Saskia Freytag, Melanie Bahlo, and Johann A Gagnon-Bartsch. dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, 35(12):2093–2099, 2019.

[282] Noa Bossel Ben-Moshe, Shelly Hen-Avivi, Natalia Levitin, Dror Yehezkel, Marije Oosting, Leo AB Joosten, Mihai G Netea, and Roi Avraham. Predicting bacterial infection outcomes using single cell rna-sequencing analysis of human immune cells. *Nature communications*, 10 (1):1–16, 2019.

[283] Casey P Shannon, Robert Balshaw, Virginia Chen, Zsuzsanna Hollander, Mustafa Toma, Bruce M McManus, J Mark FitzGerald, Don D Sin, Raymond T Ng, and Scott J Tebbutt. Enumerateblood–an r package to estimate the cellular composition of whole blood from affymetrix gene st gene expression profiles. *BMC genomics*, 18(1):43, 2017.

[284] Yuning Hao, Ming Yan, Blake R Heath, Yu L Lei, and Yuying Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS computational biology*, 15(5):e1006976, 2019.

[285] Brian B Nadel, David Lopez, Dennis J Montoya, Feiyang Ma, Hannah Waddel, Misha M Khan, Serghei Mangul, and Matteo Pellegrini. The gene expression deconvolution interactive tool (gedit): Accurate cell type quantification from gene expression data. *GigaScience*, 10(2): giab002, 2021.

[286] Ziyi Chen, Lijun Quan, Anfei Huang, Qiang Zhao, Yao Yuan, Xuye Yuan, Qin Shen, Jingzhe Shang, Yinyin Ben, F Qin, et al. seq-immucc: cell-centric view of tissue transcriptome measuring cellular compositions of immune microenvironment from mouse rna-seq data. *Frontiers in Immunology*, 9:1286, 2018.

[287] Amit Frishberg, Avital Brodt, Yael Steuerman, and Irit Gat-Viks. Immquant: a user-friendly tool for inferring immune cell-type composition from gene-expression data. *Bioinformatics*, 32(24): 3842–3843, 2016.

[288] Alexander R Abbas, Kristen Wolslegel, Dhaya Seshasayee, Zora Modrusan, and Hilary F Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PloS one*, 4(7), 2009.

[289] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W Zandstra. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS computational biology*, 8(12), 2012.

[290] Francesca Finotello, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, Zuzana Loncova, Wilfried Posch, Doris Wilflingseder, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data. *Genome medicine*, 11(1):1–20, 2019.

[291] Yifeng Tao, Haoyun Lei, Xuecong Fu, Adrian V Lee, Jian Ma, and Russell Schwartz. Robust and accurate deconvolution of tumor populations uncovers evolutionary mechanisms of breast cancer metastasis. *Bioinformatics*, 36(Supplement_1):i407–i416, 2020.

[292] Silke Reinartz, Florian Finkernagel, Till Adhikary, Verena Rohnalter, Tim Schumann, Yvonne Schober, W Andreas Nockher, Andrea Nist, Thorsten Stiewe, Julia M Jansen, et al. A transcriptome-based global map of signaling pathways in the ovarian cancer microenvironment associated with clinical outcome. *Genome biology*, 17(1):108, 2016.

[293] Li Dong, Avinash Kollipara, Toni Darville, Fei Zou, and Xiaojing Zheng. Semi-cam: A semi-supervised deconvolution method for bulk transcriptomic data with partial marker gene information. *Scientific reports*, 10(1):1–12, 2020.

[294] Oyetunji E Ogundijo and Xiaodong Wang. A sequential monte carlo approach to gene expression deconvolution. *PloS one*, 12(10):e0186167, 2017.

[295] Christopher R Bolen, Mohamed Uduman, and Steven H Kleinstein. Cell subset prediction for blood genomic studies. *BMC bioinformatics*, 12(1):258, 2011.

[296] Renaud Gaujoux and Cathal Seoighe. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infection, Genetics and Evolution*, 12(5):913–921, 2012.

[297] Yi Li and Xiaohui Xie. A mixture model for expression deconvolution from rna-seq in heterogeneous tissues. *BMC bioinformatics*, 14(5):S11, 2013.

[298] Niya Wang, Ting Gong, Robert Clarke, Lulu Chen, Ie-Ming Shih, Zhen Zhang, Douglas A Levine, Jianhua Xuan, and Yue Wang. Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, 31(1):137–139, 2014.

[299] Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proceedings of the IEEE*, 105(2):340–366, 2016.

[300] Fangzheng Xie, Mingyuan Zhou, Yanxun Xu, et al. Baycount: A bayesian decomposition method for inferring tumor heterogeneity using rna-seq counts. *The Annals of Applied Statistics*, 12(3):1605–1627, 2018.

[301] Gerald Quon and Quaid Morris. Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–2889, 2009.

[302] Francesca Finotello and Zlatko Trajanoski. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy*, 67(7):1031–1040, 2018.

[303] Vinod Kumar Yadav and Subhajyoti De. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Briefings in bioinformatics*, 16(2):232–241, 2014.

[304] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445, 2019.

[305] Etienne Becht, Nicolas A Giraldo, Marie-Caroline Dieu-Nosjean, Catherine Sautès-Fridman, and Wolf Herman Fridman. Cancer immune contexture and immunotherapy. *Current opinion in immunology*, 39:7–13, 2016.

[306] Andrew J Gentles, Aaron M Newman, Chih Long Liu, Scott V Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S Nair, Yue Xu, Amanda Khuong, Chuong D Hoang, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine*, 21(8): 938–945, 2015.

[307] Wolf Herman Fridman, Franck Pagès, Catherine Sautès-Fridman, and Jérôme Galon. The immune contexture in human tumours: impact on clinical outcome. *Nature Reviews Cancer*, 12(4): 298–306, 2012.

[308] Wolf H Fridman, Laurence Zitvogel, Catherine Sautès-Fridman, and Guido Kroemer. The immune contexture in cancer prognosis and treatment. *Nature reviews Clinical oncology*, 14(12): 717, 2017.

[309] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[310] E Andres Houseman, Molly L Kile, David C Christiani, Tan A Ince, Karl T Kelsey, and Carmen J Marsit. Reference-free deconvolution of dna methylation data and mediation by cell composition effects. *BMC bioinformatics*, 17(1):259, 2016.

[311] Alexander J Titus, Rachel M Gallimore, Lucas A Salas, and Brock C Christensen. Cell-type deconvolution from dna methylation: a review of recent applications. *Human molecular genetics*, 26(R2):R216–R224, 2017.

[312] Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia MT Greenwood. An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies. *Genome biology*, 17(1):84, 2016.

[313] Clémentine Decamps, Florian Privé, Raphael Bacher, Daniel Jost, Arthur Waguet, Eugene Andres Houseman, Eugene Lurie, Pavlo Lutsik, Aleksandar Milosavljevic, Michael Scherer, et al. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free dna methylation deconvolution software. *BMC bioinformatics*, 21(1):1–15, 2020.

[314] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[315] Andrew E Teschendorff, Joanna Zhuang, and Martin Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011.

[316] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[317] Yanhua Wen, Yanjun Wei, Shumei Zhang, Song Li, Hongbo Liu, Fang Wang, Yue Zhao, Dongwei Zhang, and Yan Zhang. Cell subpopulation deconvolution reveals breast cancer heterogeneity based on dna methylation signature. *Briefings in bioinformatics*, 18(3):426–440, 2017.

[318] Vitor Onuchic, Ryan J Hartmaier, David N Boone, Michael L Samuels, Ronak Y Patel, Wendy M White, Vesna D Garovic, Steffi Oesterreich, Matt E Roth, Adrian V Lee, et al. Epigenomic deconvolution of breast tumors reveals metabolic coupling between constituent cell types. *Cell reports*, 17(8):2075–2086, 2016.

[319] Andrew E Teschendorff, Charles E Breeze, Shijie C Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC bioinformatics*, 18(1):105, 2017.

[320] Andrew E Teschendorff, Tianyu Zhu, Charles E Breeze, and Stephan Beck. Episcore: cell type deconvolution of bulk tissue dna methylomes from single-cell rna-seq data. *Genome biology*, 21 (1):1–33, 2020.

[321] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 11(3):309–311, 2014.

[322] Xiangyu Luo, Can Yang, and Yingying Wei. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature communications*, 10(1):1–12, 2019.

[323] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

[324] Pavlo Lutsik, Martin Slawski, Gilles Gasparoni, Nikita Vedeneev, Matthias Hein, and Jörn Walter. Medecom: discovery and quantification of latent components of heterogeneous methylomes. *Genome biology*, 18(1):1–20, 2017.

[325] Ankur Chakravarthy, Andrew Furness, Kroopa Joshi, Ehsan Ghorani, Kirsty Ford, Matthew J Ward, Emma V King, Matt Lechner, Teresa Marafioti, Sergio A Quezada, et al. Pan-cancer deconvolution of tumour composition using dna methylation. *Nature communications*, 9(1): 1–13, 2018.

[326] Ankur Chakravarthy and Daniel D De Carvalho. Using epigenetic data to estimate immune composition in admixed samples. In *Methods in Enzymology*, volume 636, pages 77–92. Elsevier, 2020.

[327] Douglas Arneson, Xia Yang, and Kai Wang. Methylresolver—a method for deconvoluting bulk dna methylation profiles into known and unknown cell contents. *Communications biology*, 3(1): 1–13, 2020.

[328] Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, et al. Sparse pca corrects for cell type heterogeneity in epigenome-wide association studies. *Nature methods*, 13(5):443, 2016.

[329] Eugene Andres Houseman, John Molitor, and Carmen J Marsit. Reference-free cell mixture adjustments in analysis of dna methylation data. *Bioinformatics*, 30(10):1431–1439, 2014.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama Oy, Turku, Finland 2022

ISBN 978-951-29-8845-7 (PRINT)
ISBN 978-951-29-8846-4 (PDF)
ISSN 0082-7002 (PRINT)
ISSN 2343-3175 (ONLINE)