Master Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Evaluating Directional and Association Methods on Single-cell RNA Sequencing Data

Bc. Eliška Dvořáková

Supervisor: Prof. Joe Song
Field of study: Open Informatics
Subfield: Bioinformatics
August 2020

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Dvořáková Eliška**	Personal ID number:	**456989**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Computer Science**		
Study program:	**Open Informatics**		
Specialisation:	**Bioinformatics**		

II. Master’s thesis details

Master’s thesis title in English:

Evaluating Directional and Association Methods on Single-cell RNA Sequencing Data

Master’s thesis title in Czech:

Porovnání směrových a asociativních metod na single-cell RNA sekvenčních datech

Guidelines:

Single-cell RNA sequencing (scRNA-seq) degrades data quality. Current methods for network inference face increased uncertainty from such data. To examine how directional and association inference methods work on scRNA-seq data, the thesis will study several methods that are either parametric or model-free without parametric model assumptions. They may include established methods such as conditional entropy (Cover and Thomas, 2012) and Kruskal-Wallis test (Kruskal and Wallis, 1952), and recent methods such as causal inference by stochastic complexity (Budhathoki and Vreeken, 2017) and function index (Kumar et al., 2018; Zhong and Song, 2019). The performance of the methods will be evaluated on simulated and real data at varying dropout rates, sample sizes, and discrete levels. The methods will be applied to discover directional interactions or association patterns across transcriptome and proteome in acute leukemia cells (Granja et al., 2019).

1. Get familiar with scRNA-seq and its characteristic features.
2. Review the methods for directional and association inference.
3. Examine the performance of these inference methods from scRNA-seq data, compare the methods mentioned ad 2, use AUROC and AUPR.
4. Discover directional or association patterns from the data published by (Granja et al., 2019).

Bibliography / sources:

[1] K. Budhathoki and J. Vreeken. MDL for causal inference on discrete data. In IEEE International Conference on Data Mining, pages 751–756, 2017.
[2] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley &amp; Sons, 2012.
[3] Granja, J.M., Klemm, S., McGinnis, L.M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol 37, 1458–1465 (2019) doi:10.1038/s41587-019-0332-7
[4] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. J Am Stat Assoc, 47(260):583–621, 1952.
[5] S. Kumar, H. Zhong, R. Sharma, Y. Li, and M. Song. Scrutinizing functional interaction networks from RNA-binding proteins to their targets in cancer. In IEEE International Conference on Bioinformatics and Biomedicine, pages 185–190, Madrid, Spain, 2018.
[6] Zhong, H., Song, M. Directional association test reveals high-quality putative cancer driver biomarkers including noncoding RNAs. BMC Med Genomics 12, 129 (2019) doi:10.1186/s12920-019-0565-9.

Name and workplace of master’s thesis supervisor:

prof. Joe Song, Department of Computer Science, New Mexico State University

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **11.02.2020** Deadline for master's thesis submission: **14.08.2020**

Assignment valid until: **30.09.2021**

prof. Joe Song
Supervisor’s signature

Head of department’s signature

prof. Mgr. Petr Páta, Ph.D.
Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student’s signature

Acknowledgements

I want to express my gratitude to my supervisor Prof. Joe M. Song for the knowledge he has passed to me in the last year, and also for the positive attitude that always lights my spirit to work harder.

Many thanks also belong to my family, friends and everyone that has supported me through my studies and life.

Declaration

I hereby declare that I have completed this thesis independently and that I have used only the sources (literature and web pages) listed in the enclosed bibliography.

In Prague, 14th August 2020

Abstract

This thesis aims to compare and evaluate directional and association methods performance on single-cell RNA sequencing (*scRNA-seq*) data. The scRNA-seq enables one to study biology at a single cell resolution.

Although this process of RNA sequencing opens up new possibilities, the data can be subject to technical distortions, such as a dropout where the loss of information ranges from 30 to 90 %. Thus methods that work reliably for the bulk RNA data sets may perform close to random guessing for the scRNA-seq. Therefore I present a comparison of multiple methods on both the simulated and the real data sets. The directional and non-directional studies are separated for tests using the simulated data to prevent influencing the results by methods that detect the inference inaccurately in only one of these studies. The best performing method is then used to discover new association patterns across transcriptome and proteome in acute leukaemia cells.

Secondly, I demonstrate the impact of data normalisation for association methods. Four current normalisation methods and a new approach proposed here are compared on real data. The functions are tested for a new artefact creation and the original artefact destruction. Examples of these pattern transformations are provided for each approach. The findings in this thesis suggest that the normalisation of the scRNA-seq data must be carefully handled to avoid introducing undesirable artefacts into the studying of relationships between genes.

Keywords: Single-cell RNA sequencing, Model-free directional dependency, scRNA-seq normalization, gene-protein co-expression, the dropout simulation

Supervisor: Prof. Joe Song

Abstrakt

Tato práce je zaměřena na porovnání a vyhodnocení směrových a asociativních metod na *single-cell RNA* sekvenčních (scRNA-seq) datech. ScRNA-seq umožňuje studovat biologii na jednobuněčné úrovni.

Přestože tento proces RNA sekvenování otevírá nové možnosti, data můžou podléhat technickému zkreslení jakým je výpadek informace s mírou ztráty pohybující se od 30 do 90 %. Proto metody, které fungovaly spolehlivě pro *bulk RNA* data, dávají výsledky blízké náhodnému odhadování pro scRNA-seq. Proto představuji porovnání několika metod na simulovaných i reálných datech. Směrové a nesměrové studie jsou rozděleny pro testy na simulovaných datech, aby se předešlo ovlivnění výsledků metodami, které predikují špatně jen pro jednu z těchto studií. Metoda s nejlepším vyhodnocením je následně použita pro objevení nových vzorů přes transkriptom a proteom v buňkách akutní leukémie.

Za druhé, demonstruji dopad normalizace dat na asociativní metody. Čtyři současné normalizační metody a nový přístup představený zde jsou porovnány na reálných datech. Funkce jsou testovány na tvorbu nových artefaktů a destrukci původních artefaktů. Příklady těchto vzorových přeměn jsou poskytnuté pro všechny postupy. Výsledky této práce naznačují, že normalizace scRNA-seq dat musí být pečlivě zpracována, aby se zabránilo zavádění nežádoucích artefaktů do studie vztahů mezi geny.

Klíčová slova: Single-cell RNA sekvenování, bezmodelová směrová závislost, normalizace scRNA-seq dat, koexprese genu a proteinu, simulace ztráty informace

Překlad názvu: Porovnání směrových a asociativních metod na single-cell RNA sekvenčních datech

Contents

Figures

Tables

Chapter 1
Introduction

Ribonucleic acid (RNA) sequencing is a powerful tool, which can be used to understand biological mechanisms. The idea of RNA sequencing is accurate quantification of messenger RNA (mRNA) expression levels across genes. The analysis of expression levels can reveal a gene to gene correlation of specific cell types, which helps us understand inner molecular mechanisms. A differential expression study can operate as treatment control in new medical procedures. *Single-cell RNA sequencing (scRNA-seq)* is a novel RNA sequencing method. One of the main characteristics of the scRNA-seq data is high dropout [8], which causes a lack of analysis methods accuracy that used to work for older RNA sequencing method, *bulk RNA sequencing* [9]. This work compares multiple methods in order to increase the accuracy of scRNA-seq data analysis.

The recent progress in next-generation sequencing techniques provides new possibilities in molecular biology. Especially technologies for genomics, transcriptomics, and proteomics focus on single-cell properties. The scRNA-seq was published for the first time by Tang et al. [10] in 2009. After this paper's release, scRNA-seq has been exploited and used to characterise biological mechanism of individual cells. Although a group of cells can share the donor or even the same tissue, cells show heterogeneous characteristics [11, 12]. The main advantage of scRNA-seq is the capability of distinguishing cells by their type and obtaining data that originates in only one type of cells, which is impossible with the bulk RNA sequencing. An example of the scRNA-seq application is the co-expression analysis, e.g. biological networks construction. An example of the scRNA-seq processing from cells to the gene expression analysis is shown in Figure 1.1.

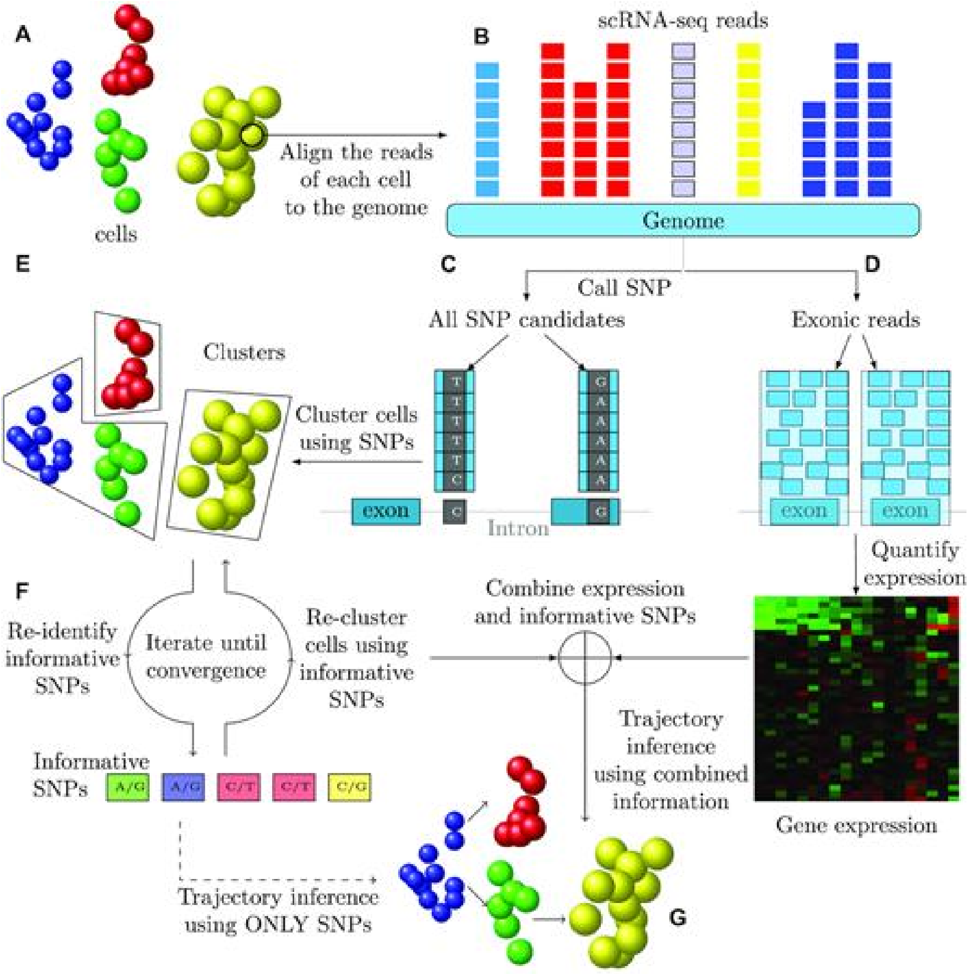


Figure 1.1: An example of the scRNA-seq and its futher analysis. Figure taken from [1].

Although the scRNA-seq allows studying molecular mechanisms with an unmatched resolution, the cell filtration brings data limitations. The library preparation techniques cause additional noise characterised by an observation of many zero values in the result data, also referred to as dropout [13]. The dropout of values can occur when a reverse transcription skips a cell [14]. The reverse transcription is a necessary step in the library preparation for the scRNA-seq. The result dropout rate equals to 30–90 % [8]. The high dropout rate causes the lack of accuracy of existing network inference software that used to work reliably for the bulk RNA sequencing. With no dropout consideration, old methods perform close to random guessing on scRNA-seq [9]. Due to dropout, new artefacts are introduced which affect the co-expression analysis.

Due to the lack of performance of existing inference methods, the accuracy of recent association methods is compared in this work. Four association methods are evaluated on both the simulated and the real data. The non-directional and the directional studies are performed separately to test each ability with focus, when using the simulated data.

To prevent the artefacts transformation due to dropout, four present normalisation methods are compared with a newly introduced normalisation function. All methods are evaluated on real data provided by [15].

1.1 Association methods

The central hypothesis of this thesis is that the *Functional index*, designed explicitly for the scRNA-seq, performs better on scRNA-seq data than other association methods. The *Functional index* is a recent association method which originates in the *FunChisq* and was introduced in [16, 17, 18].

The performance of the association methods is evaluated on both artificially generated and real data sets. The evaluation of methods is divided into two parts to test the ability to detect the inference separately from the ability to detect the direction of the dependency. In other words, this thesis aims to answer the following questions:

1. *How accurate does a method detect a dependency of two variables?*

2. *How accurate does a method detect the direction of dependency inference?*

For simplicity and correct evaluation, experiments based on binary decision making were designed. To answer the above questions, two different experiments were created. In the association test, a method only decides whether variables are dependent and therefore, half of the data consists of conditional inference, and the other half is non-dependent. This test studies the ability to detect an edge in the biological network.

Unlike the data set in the first experiment, the whole data set of the second test consists of dependent variables only, but the dependency is one-sided. The methods detect the direction of dependency inference. In this work, all the evaluated methods were designed to tell the correlation direction.

For both methods, the normalisation method gets a contingency table as input and returns *true* if $f : Y \to X$ is detected and *false* otherwise. In the association design X and Y are variables either independent ($f : X \not\to Y$ and $f : Y \not\to X$) or functional ($f : X \to Y$ and $f : Y \to X$), thus only a single direction of the inference is tested. Regardless of the tested direction, the outcome must be the same so testing a single direction is a correct and sufficient approach.

The second design shares the workflow of the contingency table processing. The normalisation methods get one contingency table as an input. However, since there is no assumption of the dependency type of the input table, only one direction is checked, and the methods have no information about the opposite direction inference. This design was chosen because when testing the real data sets, the type of inference is also hidden. An example of scRNA-seq data simulation is displayed in Figure 1.2.

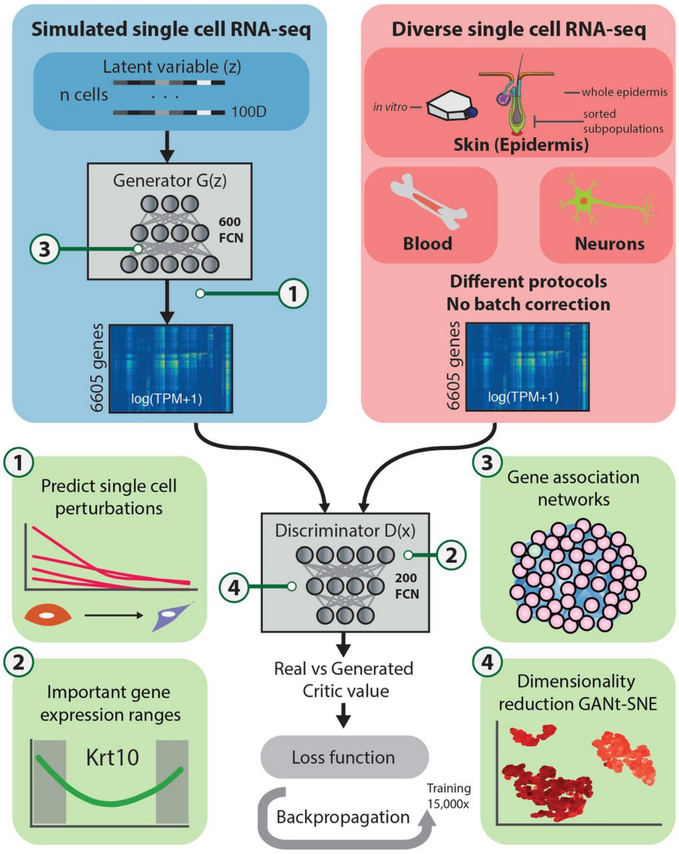


Figure 1.2: An example of the scRNA-seq data simulation. Figure taken from [2].

To prove the accuracy of the *Functional index*, the association method is compared with three other methods via *receiver-operating characteristic* (ROC) and *precision-recall* (PR), and calculated areas under the curves (AUROC and AUPR) on both the simulated and the real data. Then the *Functional index* is used to discover new gene-protein correlations.

In this work, four association methods (the *Kruskal-Wallis test*, the *Conditional entropy*, the *Causal inference by stochastic complexity (CISC)* and the *Functional index*) were compared on both the simulated and the real data. The *Functional index* was concluded to be the best performing association method.

The aim has been to propose a reliable method for biological inference networks construction given the data characteristics. There are several methods for scRNA-seq data. The simulated data generation proposed in this thesis allows concentrating the study to all parameters, both separately and all together. This feature enables us to recommend methods based on the data.

Although the scRNA-seq unfolds new fields of study, the imprecision in scRNA-seq data analysis by existing methods prohibits the reliable outcomes. The evaluation of new methods and their comparison helps overcome the

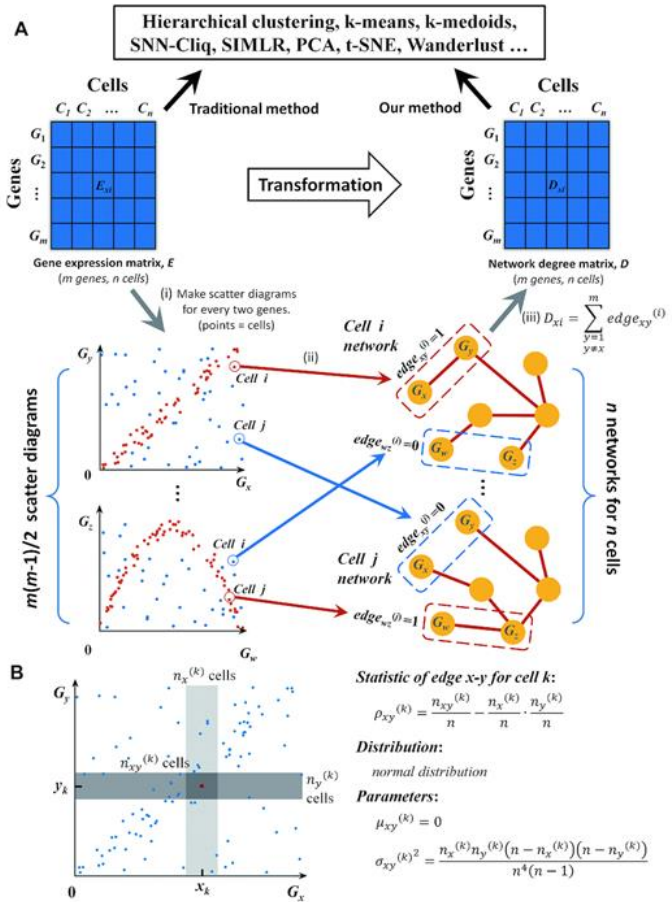

A
Hierarchical clustering, k-means, k-medoids, SNN-Cliq, SIMLR, PCA, t-SNE, Wanderlust ...
Traditional method
Our method
Cells
Genes
Transformation
Gene expression matrix, E (m genes, n cells)
Network degree matrix, D (m genes, n cells)
(i) Make scatter diagrams for every two genes. (points = cells)
(ii)
(iii) $D_{xi} = \sum_{y=1, y\neq x}^{m} edge_{xy}^{(i)}$
Cell i network
Cell j network
m(m-1)/2 scatter diagrams
n networks for n cells
B
Statistic of edge x-y for cell k:
$\rho_{xy}^{(k)} = \frac{n_{xy}^{(k)}}{n} - \frac{n_x^{(k)}}{n} \cdot \frac{n_y^{(k)}}{n}$
Distribution:
normal distribution
Parameters:
$\mu_{xy}^{(k)} = 0$
$\sigma_{xy}^{(k)2} = \frac{n_x^{(k)} n_y^{(k)} (n - n_x^{(k)})(n - n_y^{(k)})}{n^4 (n-1)}$

Figure 1.3: An example of a biological network construction from the scRNA-seq data. Figure taken from [3].

problems characteristic of scRNA-seq. The accuracy assessment of multiple methods provides recommendations for the association method end-users to speed up work with reliable outcomes. An example of biological network creation is shown in Figure 1.3.

1.2 Normalisation methods

The existing normalisation methods implemented for the bulk RNA-seq introduce new or destroy original artefacts. Thus a new normalisation function is presented that prevents creating or destroying artefacts.

The *Up-down-sampling (UDSM)* normalisation method is presented in Chapter 4. Then the *UDSM* is compared to four other normalisation methods.

The normalisation methods are compared both visually and empirically. The co-expression graphs of the gene pairs are displayed and checked for the artefact creation or destruction. The empirical measurement is based on the *estimate* value calculated by the *Spearman* correlation test. The *Spearman* test was chosen because it has no assumption of the sample distribution. The normalisation methods are expected to modify the raw values but still stay close to the original patterns and estimate value.

To demonstrate the impact of the pattern destruction and formation, all values with at least one gene equal to zero are omitted. To prevent the evaluation of empty vectors, gene pairs are not included if their length is less than three after the zero filtration.

Two experiment designs were developed to check both original pattern exploitation and the new artefact creation. The study is designed to answer the following questions:

1. *Does a normalisation method exploit the dependency correctly?*
2. *Does a normalisation method introduces new artefacts?*

The first experiment contains selected pairs of genes with known co-expression dependency. The dependency types include a negative correlation, a positive correlation and also an independent pair of genes. In this study, the transformation of the raw values by normalisation methods is expected to improve the known dependency, so the estimate values can differ a lot as long as the new estimate value approaches the desired value.

The second study focuses on introducing new artefacts. The experiment is called permutation study because all the values are permuted row by row across cells. The permutation destroys all present dependency, and therefore the estimate value before and after normalisation should be close to zero. The data set for this study contains over 1000 gene pairs randomly picked from the real data [15]. An example of the scRNA-seq data normalisation is provided in Figure 1.4.

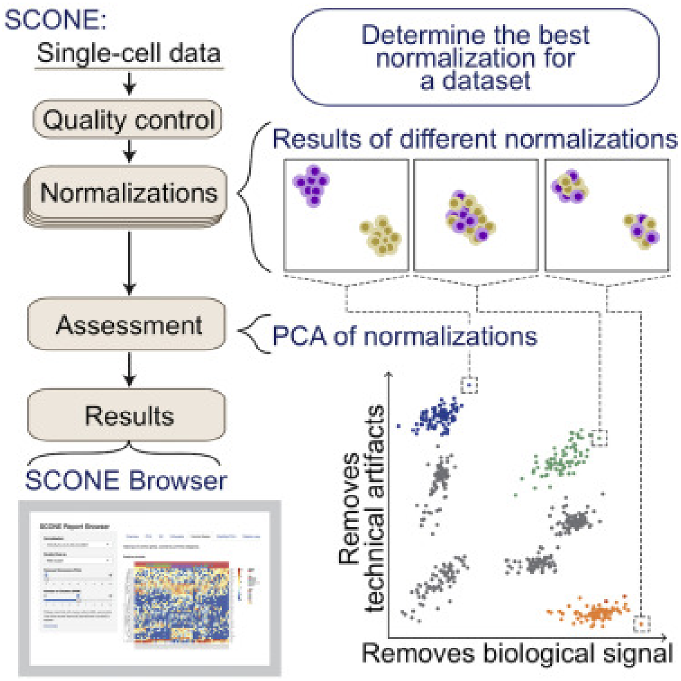

SCONE:
Single-cell data
Quality control
Normalizations
Assessment
Results
SCONE Browser
Determine the best normalization for a dataset
Results of different normalizations
PCA of normalizations
Removes technical artifacts
Removes biological signal

Figure 1.4: An example of the scRNA-seq data normalisation and the workflow of the analysis of the data. Figure taken from [4].

Five normalisation methods (including a newly presented method) were compared from the point of pattern destruction and creation on the real data. Two types of experiments were developed and evaluated.

The newly presented *Up-down-sampling (UDSM)* normalisation method was compared to four other normalisation methods *(Counts per million (CPM), Relative Log Expression (SF)*, 99^{th} *percentile (UQ)*, *Down-sampling (DSM)*). The experiments aim to test if any artefacts are created or destroyed during normalisation. Although the *UDSM* has been proven to perform better than the other normalisation methods, it still needs more testing of its parameters because it created few artefacts.

The correct normalisation enhances association analysis. The connection of the association and normalisation studies make the scRNA-seq examination more reliable and credible. Preventing the artefact formation reduces false-positive gene co-expression and suppressing the pattern destruction represses false-negatives.

The new normalisation method has proved to be preventing the new artefact creation and the pattern destruction. After further studies, the *UDSM* should improve the scRNA-seq analysis.

Chapter 2

Related work

The scRNA-seq is a relatively recent method which experienced massive development and modifications during the last decade. The idea of the RNA sequencing is accurate quantification of mRNA expression levels. The analysis of expression levels across genes provide valuable insights into biological mechanisms and can be beneficial in some medical treatments.

2.1 Single-cell RNA sequencing

The scRNA-seq belongs to the next-generation sequencing (NGS), which faces rapid progress in the past few years. The first description of the scRNA-seq was provided by Tang et al. [10] in 2009.

2.1.1 ScRNA-seq protocol

The scRNA-seq underwent several changes and development during the last years, and therefore there are multiple protocols for scRNA-seq [19]. All protocols require a minimum amount of starting material. One of the developments in scRNA-seq includes studies that increased the number of processed single cells in the assay. The original scRNA-seq study contained only one cell per experiment, but an improvement over the years ensures that current studies can contain up to 100,000 single cells [14].

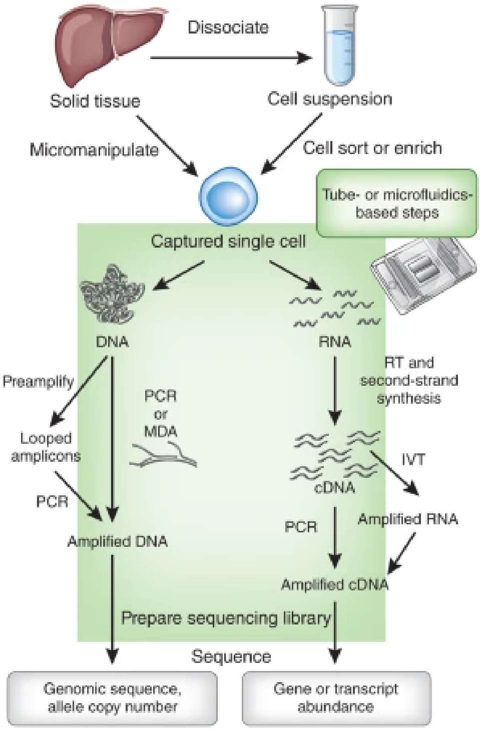


Figure 2.1: Visualization of the scRNA-seq process. Figure taken from [5].

The main steps of the scRNA-seq according to [14] are:

1. Capture the single cell material.
2. Convert transcribed RNA from cells to complementary DNA (cDNA) using the reverse transcription (RT).
3. Amplify the cDNA using the polymerase chain reaction (PCR) or the in vitro transcription (IVT).
    - When using the IVT, the resulting material is an amplified RNA and therefore is converted to cDNA again [5].
4. Prepare sequencing library.
5. Sequence.

The second step of converting the RNA to the cDNA is essential because the amount of the mRNA in a single cell is 1-5 % of its total RNA and degrades quickly [20]. The reverse transcription is required to convert the RNA to the cDNA, but it may produce positional bias during the process [21, 22]. Molecular barcodes can partially correct the positional bias [13], an example of a barcoding method is illustrated in Figure 2.2. However the correction comes with other sources of bias in the result data such as the PCR and sequencing errors [23].

To achieve the transfer from the RNA to the cDNA, adaptor sequences are added to all mRNA transcripts [14]. The PCR amplification operates exponentially and the IVT linearly. To obtain a sufficient amount of the cDNA material, multiple rounds of the IVT are needed. On the other hand, for PCR, adaptor sequences for both ends are required. Some genes can be preferred during the amplification which leads to additional bias in the result data [24]. The final read counts of some genes can be lower or completely dropped out [25].

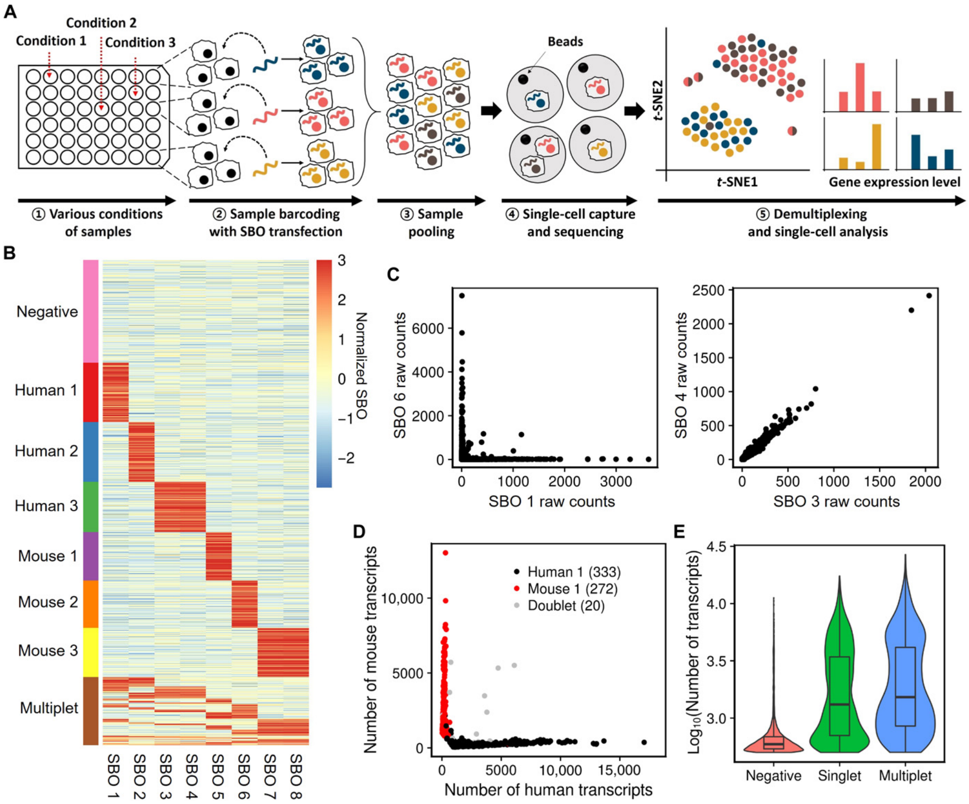


Figure 2.2: An illustration of a transient barcoding method. Figure taken from [6].

2.1.2 ScRNA-seq data characterisics

The first main characteristic of the scRNA-seq data is the amount of the input material. The amount of the analysed material is small and contains a few types of cells. This quality allows the research of molecular mechanisms and rare cell types that are difficult to cultivate.

The second characteristic is that during the essential part of the library preparation, the final data suffer from high information loss [26, 27]. Many zero observations in result data give the dropout with 30–90 % rate [8].

The next feature of the scRNA-seq is the creation of technical artefacts which can originate in cell-specific sequencing depth differences [28]. The

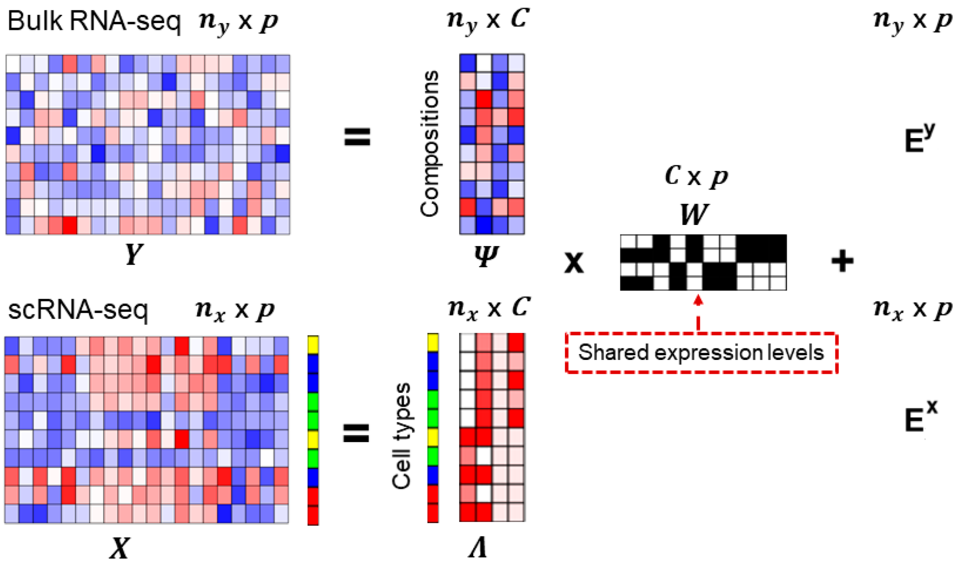


Figure 2.3: An example of the scRNA-seq and the bulk RNA-seq dissimilarity in the co-expression analysis. Figure taken from [7].

scRNA-seq data share a global pattern which is probably caused by the *cDNA* production [22].

2.1.3 ScRNA-seq vs. bulk RNA sequencing

Before the scRNA-seq became popular, the bulk RNA sequencing method preceded. The bulk RNA-seq processes a large population of cells within a tissue. The result data contains the average genetic content for each gene. The averaging across a large sample of cells discriminates the cell types with a rare type that is difficult to culture. The impact of these cells is abated or completely faded away [5]. The scRNA-seq fills the gap of averaging out some cell types because the scRNA-seq distinguishes cell types. Patal et al. [29] discovered heterogeneity within a tumour which would not be possible with the bulk RNA-seq.

Due to the divergent characteristics of the bulk RNA and the scRNA-seq, multiple methods with various focus perform close to random guessing [30]. Therefore a further research in order to increase the accuracy has been conducted. An example of how the dissimilarities of the scRNA-seq and bulk RNA-seq is shown in Figure 2.3.

2.2 ScRNA-seq normalisation

Since the bulk RNA-seq is the predecessor of the scRNA-seq, many normalisation methods were initially invented and tested on the bulk RNA-seq [31]. The scRNA-seq differ from the bulk data by not only the high dropout [26, 27]

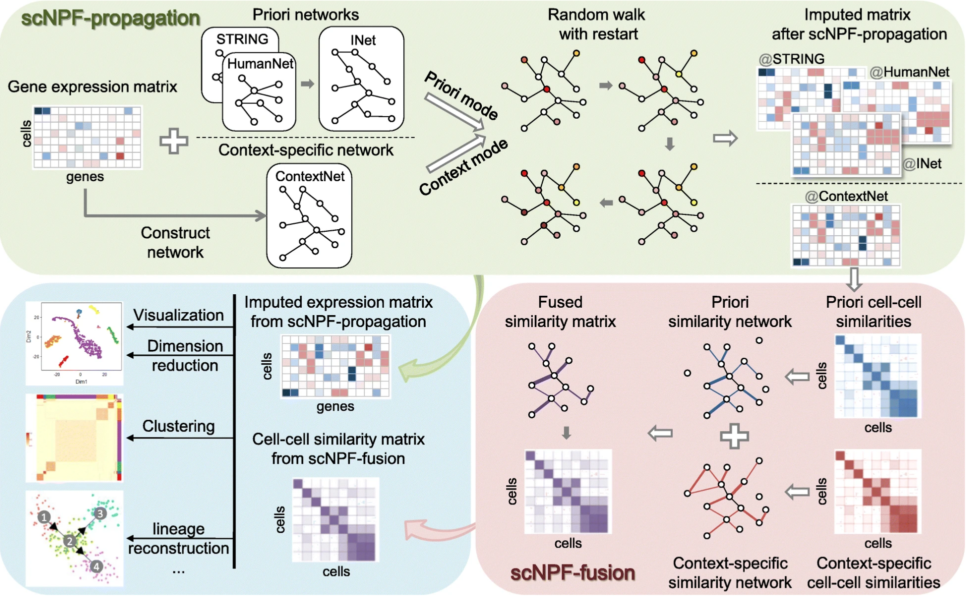


Figure 2.4: An example of scRNA-seq data analysis. Figure taken from [5].

but also by a technical noise and the bulk RNA-seq assumptions that do not apply on scRNA-seq [28]. Another problem are the technical artefacts created during the scRNA-seq process which can not be removed [28]. The goal of normalisation is to reduce or remove the technical artefacts and the batch effect influence [4]. Normalisation methods lose effectiveness when applied on the scRNA-seq data [32]. Since normalisation is an essential step during the data preprocessing, its improvement has become a theme of many recent studies [32, 25, 33, 34].

2.3 ScRNA-seq association analysis

Most of the standard association methods were designed before the invention of the scRNA-seq and were assessed on the bulk RNA. Thus the methods were not invented with the knowledge of the scRNA-seq characteristics. The difference between the bulk and the scRNA-seq causes the methods widely used for the bulk RNA analysis to perform close to random guessing [9]. Therefore new association tests are being developed, and their accuracy is being compared [35, 9]. An example of the association methods analysis is shown in Figure 2.4.

Chapter 3

Data sets

In this chapter, we look closer at the data that was used to evaluate the performance of the association methods. The experiments were performed both on the artificially generated and the real data. The initial testing is conducted on the simulated data set, then on the selection of genes from the real data provided by [15] for the final evaluation and the comparison of the normalisation methods. The real data set is also processed with the *Functional Index* association method to find new co-expressed protein-gene or gene-protein pairs.

3.1 Simulated data

The generated data for the association study is created in two steps. The first step is the data set creation with the *R* function *simulate_tables* from the *FunChisq* package. In the second step, the dropout is simulated.

The *simulate_tables* function allows the user to create a matrix representing the relationship of two variables with exclusively no inference, one-sided inference or functional inference. This feature facilitates the two-way experiment design applied in the methods evaluation. It also provides modifiable parameters for the matrix dimensions, number of samples and noise.

Because we aim to simulate not just any interactions but interactions specific for the scRNA-seq, the dropout is simulated. The original table is converted to the vectors of values for each variable. Then each vector is processed independently. According to the predefined dropout rate, a percentage of the vector values is set to zero.

3.1.1 Impact of each parameter

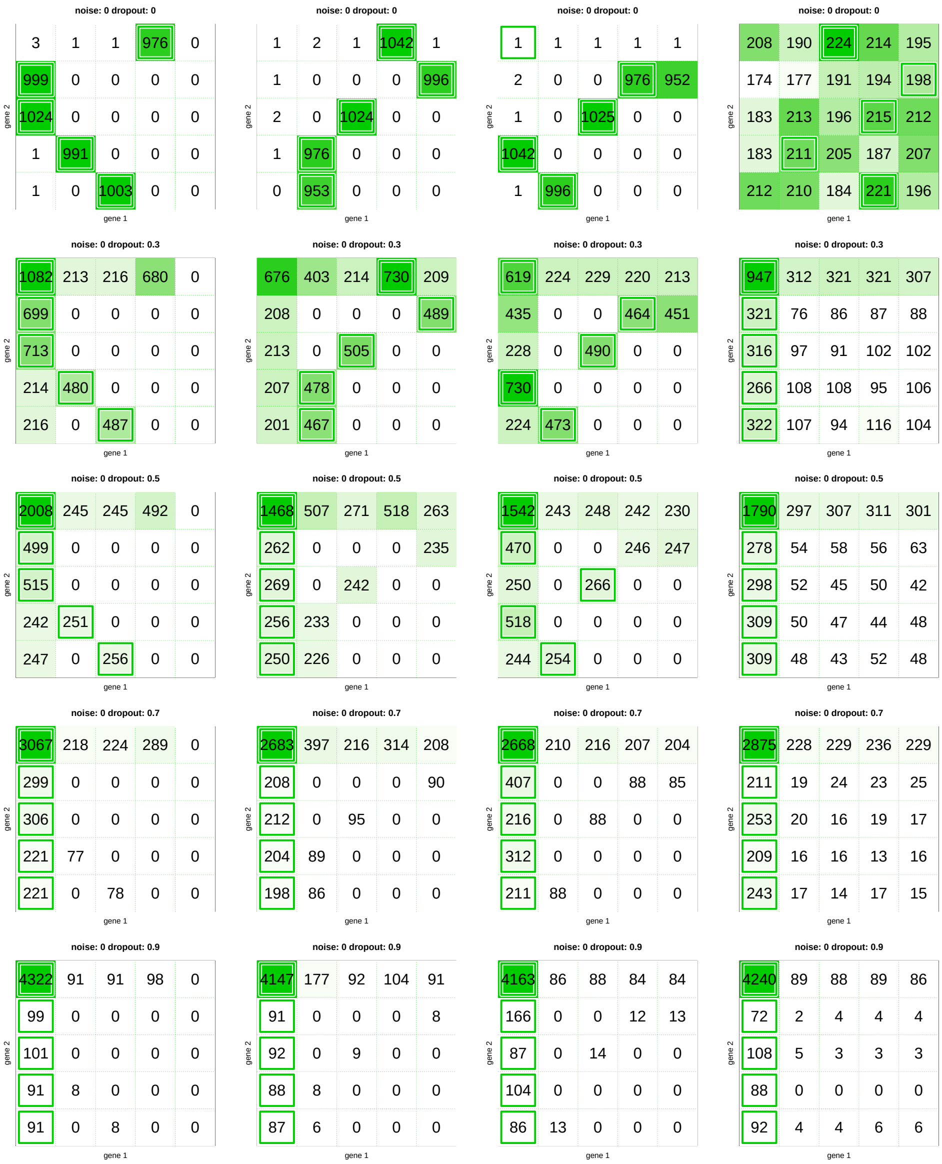
Due to all the modifiable parameters, it is possible to test the method robustness with a focus on each factor. The examples of all the previously mentioned parameters follow. All the following tables share these default settings if not mentioned otherwise:

- noise parameter: **0.2**
- dimension: **5x5**

- samples size: **5,000**
- type of inference: **Functional**

Dropout

Next Figure 3.1 illustrates the significant impact of the dropout on tables of all types of inference. The matrices are sorted from the top to the bottom by dropout (0–90 %) and from the left to the right by the dependency type (Functional, Many-to-one, One-to-many, Independent). The first row of tables shows an apparent difference between the types of inference. However, with the growing dropout, the difference is decreasing, especially when we compare the first line, where the dropout is set to 0 %, and the last line, where the dropout is set to 90 %.



noise: 0 dropout: 0

3	1	1	976	0
999	0	0	0	0
1024	0	0	0	0
1	991	0	0	0
1	0	1003	0	0

noise: 0 dropout: 0

1	2	1	1042	1
1	0	0	0	996
2	0	1024	0	0
1	976	0	0	0
0	953	0	0	0

noise: 0 dropout: 0

1	1	1	1	1
2	0	0	976	952
1	0	1025	0	0
1042	0	0	0	0
1	996	0	0	0

noise: 0 dropout: 0

208	190	224	214	195
174	177	191	194	198
183	213	196	215	212
183	211	205	187	207
212	210	184	221	196

noise: 0 dropout: 0.3

1082	213	216	680	0
699	0	0	0	0
713	0	0	0	0
214	480	0	0	0
216	0	487	0	0

noise: 0 dropout: 0.3

676	403	214	730	209
208	0	0	0	489
213	0	505	0	0
207	478	0	0	0
201	467	0	0	0

noise: 0 dropout: 0.3

619	224	229	220	213
435	0	0	464	451
228	0	490	0	0
730	0	0	0	0
224	473	0	0	0

noise: 0 dropout: 0.3

947	312	321	321	307
321	76	86	87	88
316	97	91	102	102
266	108	108	95	106
322	107	94	116	104

noise: 0 dropout: 0.5

2008	245	245	492	0
499	0	0	0	0
515	0	0	0	0
242	251	0	0	0
247	0	256	0	0

noise: 0 dropout: 0.5

1468	507	271	518	263
262	0	0	0	235
269	0	242	0	0
256	233	0	0	0
250	226	0	0	0

noise: 0 dropout: 0.5

1542	243	248	242	230
470	0	0	246	247
250	0	266	0	0
518	0	0	0	0
244	254	0	0	0

noise: 0 dropout: 0.5

1790	297	307	311	301
278	54	58	56	63
298	52	45	50	42
309	50	47	44	48
309	48	43	52	48

noise: 0 dropout: 0.7

3067	218	224	289	0
299	0	0	0	0
306	0	0	0	0
221	77	0	0	0
221	0	78	0	0

noise: 0 dropout: 0.7

2683	397	216	314	208
208	0	0	0	90
212	0	95	0	0
204	89	0	0	0
198	86	0	0	0

noise: 0 dropout: 0.7

2668	210	216	207	204
407	0	0	88	85
216	0	88	0	0
312	0	0	0	0
211	88	0	0	0

noise: 0 dropout: 0.7

2875	228	229	236	229
211	19	24	23	25
253	20	16	19	17
209	16	16	13	16
243	17	14	17	15

noise: 0 dropout: 0.9

4322	91	91	98	0
99	0	0	0	0
101	0	0	0	0
91	8	0	0	0
91	0	8	0	0

noise: 0 dropout: 0.9

4147	177	92	104	91
91	0	0	0	8
92	0	9	0	0
88	8	0	0	0
87	6	0	0	0

noise: 0 dropout: 0.9

4163	86	88	84	84
166	0	0	12	13
87	0	14	0	0
104	0	0	0	0
86	13	0	0	0

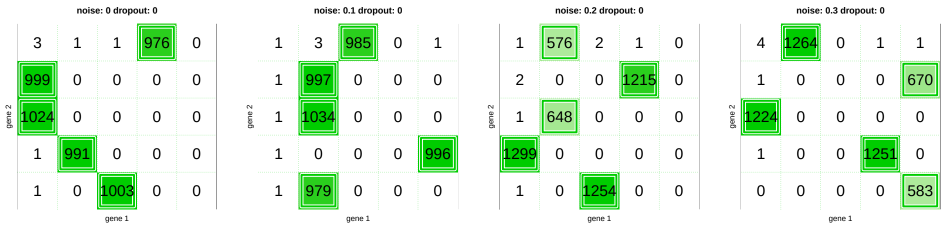
noise: 0 dropout: 0.9

4240	89	88	89	86
72	2	4	4	4
108	5	3	3	3
88	0	0	0	0
92	4	4	6	6

Figure 3.1: Generated tables of all the dependency types with zero noise and various dropouts. The types of inference from the left to the right: Functional, Many-to-one, One-to-many, Independent. Dropout from the top to the bottom: 0 %, 30 %, 50 %, 70 %, 90 %.

3.1.2 Noise

The following graphs show the effect of various noise levels.


noise: 0 dropout: 0

3	1	1	976	0
999	0	0	0	0
1024	0	0	0	0
1	991	0	0	0
1	0	1003	0	0

noise: 0.1 dropout: 0

1	3	985	0	1
1	997	0	0	0
1	1034	0	0	0
1	0	0	0	996
1	979	0	0	0

noise: 0.2 dropout: 0

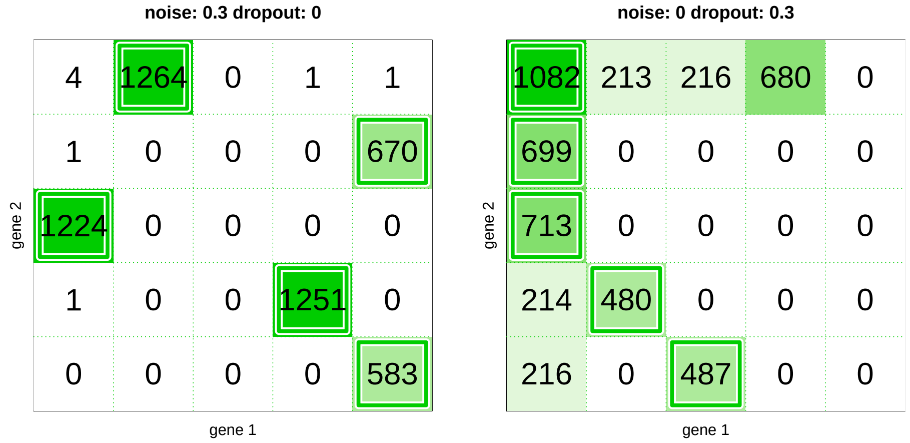
1	576	2	1	0
2	0	0	1215	0
1	648	0	0	0
1299	0	0	0	0
1	0	1254	0	0

noise: 0.3 dropout: 0

4	1264	0	1	1
1	0	0	0	670
1224	0	0	0	0
1	0	0	1251	0
0	0	0	0	583

Figure 3.2: Generated functional tables of various noise with a zero dropout. The noise parameter from the left to the right: 0.0 , 0.1, 0.2, 0.3.

Figure 3.3 shows functional tables with various noise and zero dropout. The impact of the noise is less noticeable than the impact of the dropout. That's because the noise produces random numbers and dropout produces zeros, which shifts a lot of values in only one direction and introduces new dependencies. To demonstrate the different impact of the noise and dropout, following Figure 3.4 puts two functional tables side by side – one with the noise parameter set to 0.3 and zero dropout and the second one with zero noise and 30% dropout.


noise: 0.3 dropout: 0

4	1264	0	1	1
1	0	0	0	670
1224	0	0	0	0
1	0	0	1251	0
0	0	0	0	583

noise: 0 dropout: 0.3

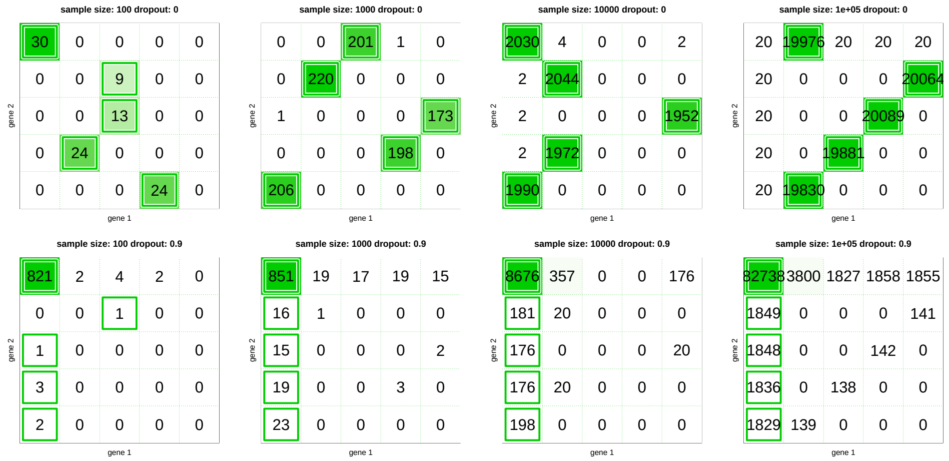
1082	213	216	680	0
699	0	0	0	0
713	0	0	0	0
214	480	0	0	0
216	0	487	0	0

Figure 3.3: The comparison of two generated functional tables, one with a zero dropout but the noise parameter set to 0.3 (left) and the second one with a 30% dropout and zero noise (right).

Sample size

Demonstrating all aspects of how the various sample size impacts the inference of two variables, eight contingency tables are provided as an example. The set is sorted by sample size from the left to the right: 100; 1,000; 10,000; 100,000). The first line of matrices is without the dropout effect. The second line contains the same functional tables but with the simulated 90%

dropout. Even without the dropout, the tables with greater sample size are distinguishable visually because the dependence is visibly more robust. The greater number of samples also resists the dropout rate more. Even with a very high dropout rate of 90 %, the pattern is still evident.



sample size: 100 dropout: 0

30	0	0	0	0
0	0	9	0	0
0	0	13	0	0
0	24	0	0	0
0	0	0	24	0

sample size: 1000 dropout: 0

0	0	201	1	0
0	220	0	0	0
1	0	0	0	173
0	0	0	198	0
206	0	0	0	0

sample size: 10000 dropout: 0

2030	4	0	0	2
2	2044	0	0	0
2	0	0	0	1952
2	1972	0	0	0
1990	0	0	0	0

sample size: 1e+05 dropout: 0

20	19976	20	20	20
20	0	0	0	20064
20	0	0	20089	0
20	0	19881	0	0
20	19830	0	0	0

sample size: 100 dropout: 0.9

821	2	4	2	0
0	0	1	0	0
1	0	0	0	0
3	0	0	0	0
2	0	0	0	0

sample size: 1000 dropout: 0.9

851	19	17	19	15
16	1	0	0	0
15	0	0	0	2
19	0	0	3	0
23	0	0	0	0

sample size: 10000 dropout: 0.9

8676	357	0	0	176
181	20	0	0	0
176	0	0	0	20
176	20	0	0	0
198	0	0	0	0

sample size: 1e+05 dropout: 0.9

82738	3800	1827	1858	1855
1849	0	0	0	141
1848	0	0	142	0
1836	0	138	0	0
1829	139	0	0	0

Figure 3.4: Generated functional tables of various sample size with a 90% dropout. Sample size from the left to the right: 100; 1,000; 10,000; 100,000.

3.1.3 Details of the data generated for the association experiments

We simulate two primary types of data sets, which both include 200 edges. The first type consists of both the independent and the functional inference with a 50:50 ratio. This type aims to test whether the association method is capable of detecting the dependency existence. The second type contains matrices with only one-sided dependence. The generated matrix is saved with its transpose. Due to the transposition, we know that only one direction is dependent and that the set includes 50 % of a correctly detected inference and the rest 50 % is independent for the tested direction.

The advantage of simulated data is the known ground truth. It allows evaluating the methods with only pairwise dependency that precludes the results from being affected by indirect dependencies. E.g. if a biological network forms a circle, all the vertices are more or less dependent on each other. Then it is harder to resolve the ground truth because even if a pair of vertices isn't neighbours, they still depend on each other.

3.2 Real data

This work uses a data set, which was presented at the end of the last year by Granja et al. [15]. The data collection includes the scRNA-seq, the

Antibody-Derived Tag sequencing (scADT-seq), and the Assay of Transposase-Accessible Chromatin using sequencing (scATAC-seq). The set contains 16 samples, where 10 were obtained from the individuals diagnosed with a mixed-phenotype acute leukaemia (MPAL), and the rest is from the healthy individuals (HI). The youngest MPAL donor was 22 years old and the oldest was 72. Also, both the female and the male donors were included, which shows the diversity of the data.

The real data are used for both the evaluation and a new association detection, but we only use the scRNA-seq and scADT-seq. The scRNA-seq data consists of matrices, where the rows are the genes and the columns correspond to the cells. The values of a specific gene in a given cell represent the number of reads of the particular gene in the given cell. The scADT-seq shares the same logic, but instead of genes, the expression levels of proteins are present.

Because the dropout is the main focus of this work studies, brief statistics of the scRNA-seq data follow. The overview is mainly focused on demonstrating a high percentage of zero values and a high amount of values per a matrix. The dropout metric cannot be used here because all that can be seen are the final tables, and it would only be a guess to state the dropout rate.

Name	**State**	**Age**	**Genes**	**Cells**	**Zero %**
GSM4138872	HI	18-55	20287	6270	*91 %*
GSM4138873	HI	18-55	20287	6332	*91 %*
GSM4138874	HI	18-55	20287	2424	*85 %*
GSM4138875	HI	18-55	20287	5752	*89 %*
GSM4138876	HI	18-55	20287	7544	*91 %*
GSM4138877	HI	18-55	20287	7260	*91 %*
GSM4138878	MPAL	36	20287	196	*90 %*
GSM4138879	MPAL	36	20287	1539	*92 %*
GSM4138880	MPAL	65	20287	5885	*88 %*
GSM4138881	MPAL	22	20287	510	*85 %*
GSM4138882	MPAL	22	20287	325	*85 %*
GSM4138883	MPAL	46	20287	1579	*86 %*
GSM4138884	MPAL	46	20287	1908	*86 %*
GSM4138885	MPAL	71	20287	4161	*90 %*
GSM4138886	MPAL	72	20287	465	*91 %*
GSM4138887	MPAL	72	20287	1488	*88 %*

Table 3.1: The statistics of the real data.

Table 3.1 demonstrates the high percentage of zeros in all tables and the high amount of data. The average of the zero percentage per a matrix is 89 %, and the mean of the data values count is 68,009,632, which corresponds to the average of 3352 cells per a sample. To test the gene-gene co-expression, it takes

$$2\binom{n}{2} = 2\frac{n\,(n-1)}{2} = n(n-1)$$

comparisons, where n is the number of genes. It corresponds to 411,542,082 association tests.

Figure 3.5 demonstrates how rapidly can the high dropout affect the co-expression analysis. The graph shows a co-expression of two genes CD34 and ABCG2. According to [36], the CD34 and the ABCG2 genes should be positively correlated. Unfortunately, the dropout shifted all points to the axes, so the final inference statistics are negative, which would suggest a negative correlation.

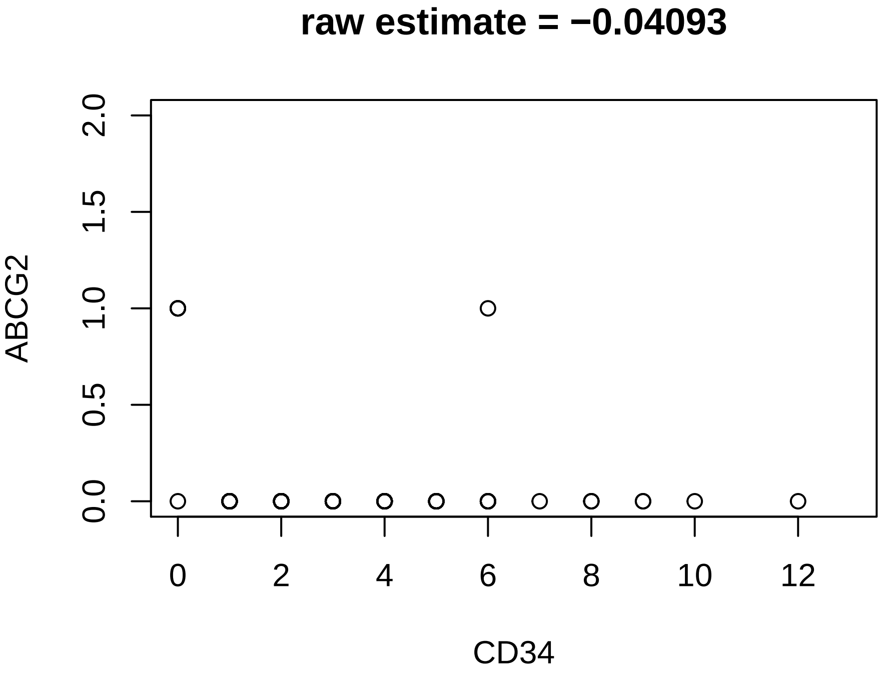


Figure 3.5: The illustration of the dropout effect on the CD34 and the ABCG2 genes co-expression.

Chapter 4

Methods

In this chapter, all methods used in this work are explained in detail. The first part is dedicated to the normalisation methods, including a presentation of a new normalisation method. The second part contains all the association methods.

4.1 Normalisation methods

Four current normalisation methods are compared using the scRNA-seq data. Three of them are based on the normalisation factors, one of which is the library size that corresponds to the total number of reads in one cell and is calculated with the formula

$$c_i = \sum_j (n_{ij})$$

where c_i is the library size of $i-th$ cell and n_{ij} is number of reads of $j-th$ gene in $i-th$ cell.

Another important statistic is the *length of* $j-th$ *gene* and is calculated by

$$l_j = \sum_i (n_{ij})$$

The length of the gene corresponds to the sum of the gene expression levels across all cells.

The last method recreates the data by the means of a multinomial distribution based random generator, and the probability is calculated from the previously mentioned statistics.

A new normalisation method is presented, which is based on the technique of using a multinomial distribution to recreate the data, with some additional steps.

To demonstrate the impact of each normalisation, an example of the gene to gene inference is provided from the real data set provided by [15]. The gene pair selected for the demonstration is the CDA and the CCM2. The co-expression graph is displayed side by side with the estimate statistics calculated by the *Spearman* correlation method. For better visual understanding, the data values with at least one coordinate equal to zero or both equal to one are omitted. However, the statistics are calculated with all values before the zero filtration. An example of the CDA and the CCM2 gene pair co-expression graph before any normalisation is shown in Figure 4.1.

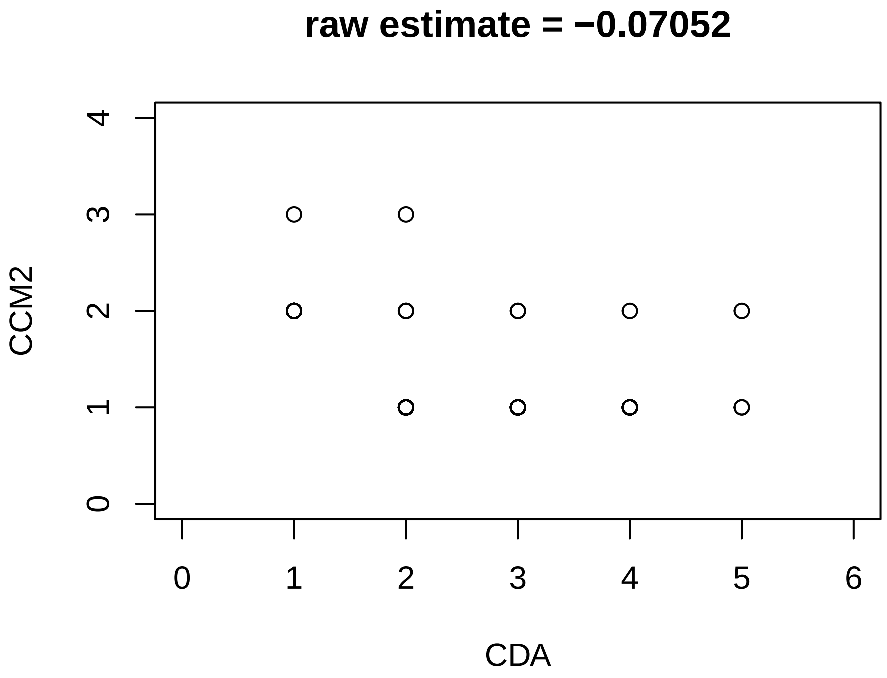


Figure 4.1: The data before any normalisation.

4.1.1 Counts per million (CPM)

The normalisation method *Counts per million* [37] normalises the read counts by the library size. The result counts are scaled by 1,000,000.

$$CMP = \frac{n_i}{\sum_j (n_j)} \cdot 1,000,000$$

For a better understanding, the code in the *R* programming language is provided:

```
calc\_cpm <- function (expr_mat)
{
  norm\_factor <- colSums( expr\_mat )
  return( t( t( expr\_mat ) / norm\_factor ) * 10^6 )
}
```

The following graph shows the inference of two genes after the *Counts per million* normalisation method. Using only the normalisation factors affects the inference by creating or destroying artefacts.

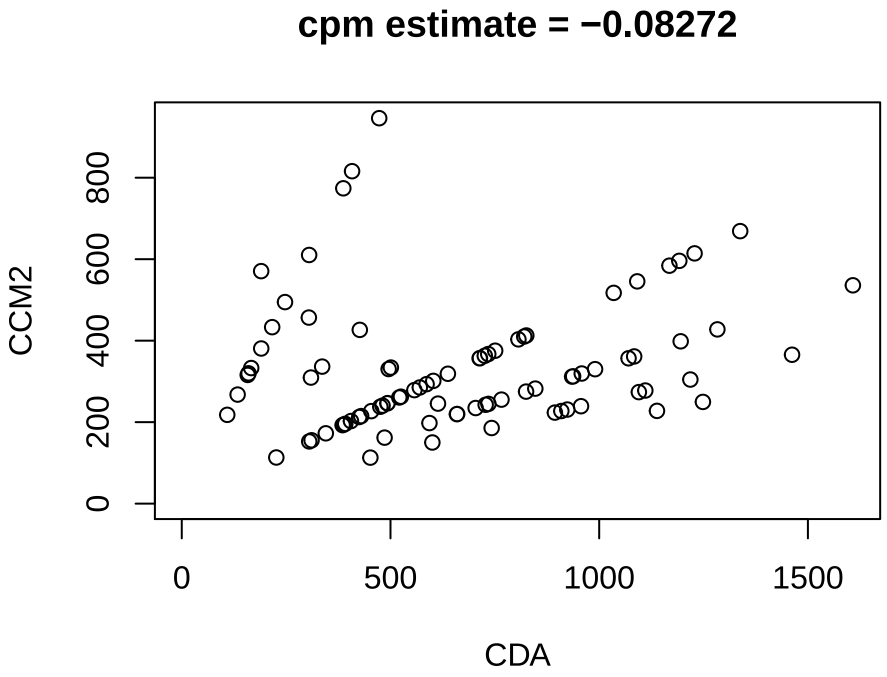


Figure 4.2: The data normalised by the *Counts per million* normalisation method.

4.1.2 Relative log expression (SF)

The *Relative log expression* method [37] is based on the size factor. The normalisation factor used for scaling is not linear but geometrical. At first a vector gm of the geometrical mean of each row is calculated by

$$gm_i = e^{\frac{\sum_j log(n_{ij})}{s_i}}; n_{ij} \neq 0,$$

where gm_i is the geometrical mean of the $i-th$ row, s_i is the length of the row, n_{ij} is the number of reads of the $j-th$ gene in the $i-th$ cell.

The size factor sz is then obtained using the gm vector. The size factor is a vector whose size is the same as the size of the column. The $j-th$ element of the size factor is calculated as follows

$$sz_j = median\left(\frac{n_{.j}}{gm}\right),$$

where sz_j is the $j-th$ element of the size factor $n_{.j}$ is the $j-th$ row, the zero values are omitted as in gm vector.

The normalisation is obtained by dividing the each row by the size factor

$$SF_i = \frac{n_{i.}}{sz}; \forall i \in 1, 2, ..., s_i; i \in \mathbb{N}.$$

The code in *R* is provided for better understanding:

```
calc_sf <- function (expr_mat) {
  gm <- function(cnts) {
    exp( mean( log( cnts[!(cnts==0)] ) ) )
  }
  geomeans <- apply( expr_mat, 1, gm)
  SF <- function(cnts) {
    tmp = ((cnts / geomeans )
        [(is.finite(geomeans) & geomeans > 0) ] )
    median( tmp[tmp>0] )
  }
  norm_factor <- apply( expr_mat, 2, SF)
  keep_cols = norm_factor > 0 # prevents division by 0
  expr_mat[,keep_cols] =
    t( t( expr_mat[,keep_cols] ) / norm_factor )
  return(expr_mat)
}
```

The *SF* normalisation method also creates new patterns. However, the artefacts affect the result estimate value less than in the case of the previous *CPM* method.

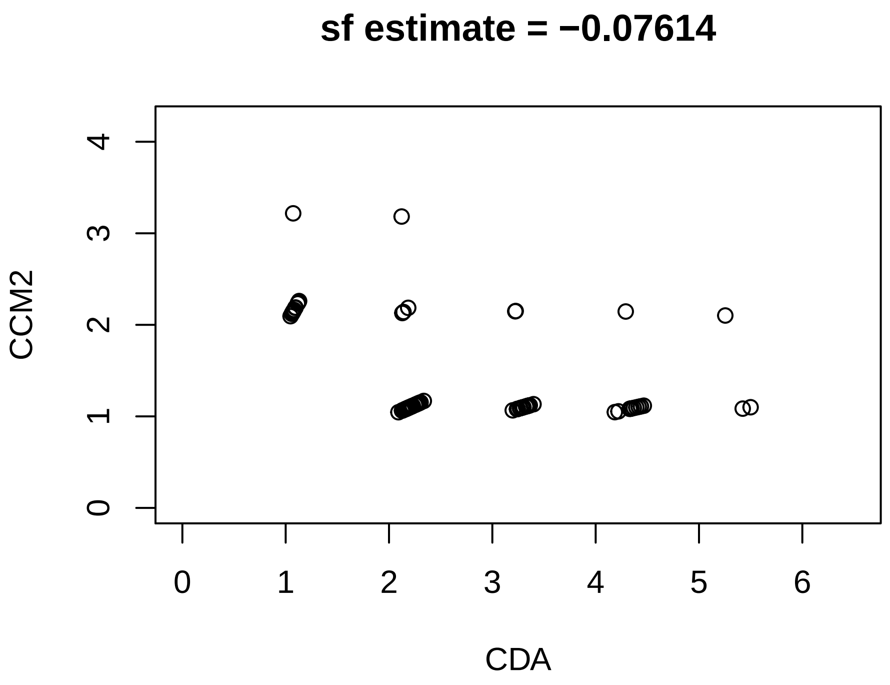


Figure 4.3: The data normalised by the *Relative log expression* normalisation method.

4.1.3 99^{th} percentile (UQ)

The 99^{th} *percentile* normalisation [37] creates the normalisation factor with the percentile statistics.

Firstly, a vector uq containing the 99^{th} percentile of each column is created. Then the vector is divided by its median. The normalisation factor is obtained and used to scale each row

$$uq_j = p_{99}\left(n_{.j}\right),$$

where $p_{99}\left(n_{.j}\right)$ is the 99^{th} percentile of the $j-th$ cell,

$$P_{99} = \frac{n_{i.}}{\frac{ug}{median(uq)}}; \forall i \in 1, 2, ..., s_i; i \in \mathbb{N},$$

where P_{99} is the final normalisation matrix, $n_{i.}$ is the $i-th$ row.

```
calc_uq <- function (expr_mat, quantile=0.99)
{
  UQ <- function(x) { quantile( x[ x > 0 ], quantile )}
  keep_cols = colSums(expr_mat) > 0 #omits 0 devision
  non_zero = expr_mat[,keep_cols]
  uq <- unlist( apply( non_zero, 2, UQ ) )

  norm_factor <- uq / median(uq)
  result = ( t( t( non_zero ) / norm_factor) )
  expr_mat[,keep_cols] = result
  return(expr_mat)
}
```

Figure 4.4 shows the co-expression graph of the example gene pair after the *UQ* normalisation method. We can see the newly created artefact with a similar significance as after applying the *CPM* normalisation method.

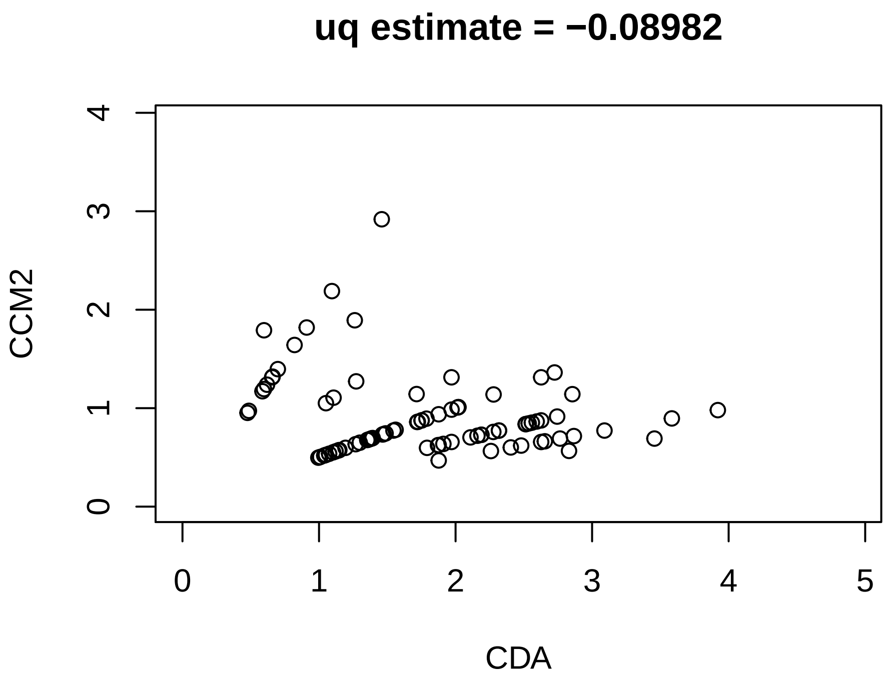


Figure 4.4: The data normalised by the 99^{th} *percentile* normalisation method.

4.1.4 Down-sampling (DSM)

The *Down-sampling (DSM)* normalisation method [37] differs from the previous ones in the main idea. This method is not based on scaling by the normalisation factor like the previous methods but on recreating the whole set again with a multinomial distribution based random generator, the result data of which then replace the original. In our implementation the *rbinom* function is used from the *stats* package. The new observation is generated with the probability of the minimal *library size* divided by the *library size* of the current cell.

Since this normalisation method is based on regenerating the data set instead of the normalisation factor, only the code in *R* is provided for better understanding:

```
Down_Sample_Matrix <- function(expr_mat) {
  keep_cols = colSums(expr_mat) > 0 # prevents 0 probability
  non_zero = expr_mat[,keep_cols]
  min_lib_size <- min(colSums(non_zero))

  down_sample <- function(x) {
    prob <- min_lib_size/sum(x)
    return(
      sapply(x, function(y) { rbinom(1, y, prob) } )
    ) }

  down_sampled_mat <- apply(non_zero, 2, down_sample)
  expr_mat[,keep_cols] = down_sampled_mat
  return(expr_mat)
}
```

A new created artefact are not present when using the *DSM* normalisation function, which is demonstrated in Figure 4.5. That's because the multinomial random generator generates only integers. It takes the current value and the probability of the cell library size, and based on this information it generates a new value. The result value is, therefore, an integer lower or equal to the original value. Unfortunately, the co-expression space is too small to keep the original pattern.

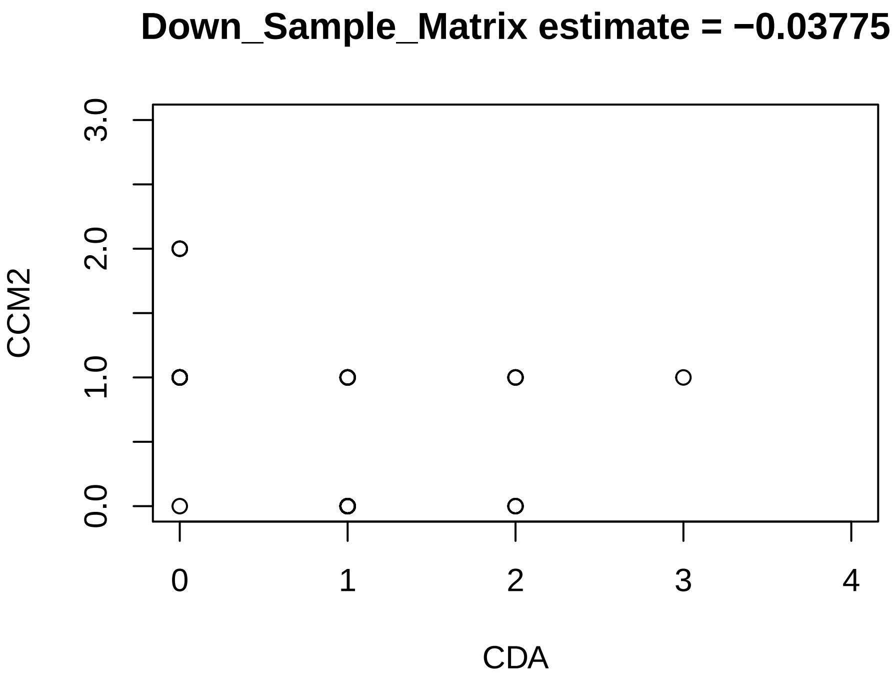


Figure 4.5: The data normalised by the *Down-sampling* normalisation method.

4.1.5 Up-down-sampling (UDSM)

To improve the normalisation, I present the *Up-down-sampling (UDSM)* normalisation method. In order to prevent new artefacts formation, the *Down-sampling* normalisation is chosen as the method foundation. However, some additional steps are added to preprocess the data set and scale the resulting space.

Firstly, noise of uniform distribution with the resulting interval of a length one is added to the data to separate the same values. To prevent negative values, the zero values are only increased. Secondly, the final space is stretched by a multiplication parameter m. The last step is the random generation performed equally as in the *Down-sampling* normalisation.

The code in *R* follows to provide a better understanding:

```
Add_Noise <- function(expr_mat, l = 0.5){
  print("adding noise")
  expr_mat[expr_mat==0] =
    runif(sum(expr_mat == 0),min = 0, max = (l))
  expr_mat[expr_mat!=0] =
    expr_mat[expr_mat!=0] +
      runif(sum(expr_mat != 0),min = -l, max = (l))
  print("noise added")
  return(expr_mat)
}

udsm <- function(expr_mat,m=1000) {

  min_lib_size <- min(colSums(expr_mat))

  up_down_sample <- function(x) {
    prob <- min_lib_size/sum(x)
    return(
      sapply(x, function(y) {
        if(y <= 0){
          return(0)
        }
        if(prob >= 1){
          prob = 1
        }
        return(rbinom(1, round(y*m ), prob))

          } )
    ) }

  up_down_sampled_mat =
    apply(expr_mat, 2, up_down_sample)
  expr_mat = up_down_sampled_mat
  return(expr_mat)
}
```

The example of the data modification by the *Up-down-sampling* normalisation method follows in Figure 4.6. The example proves that *UDSM* does not introduce new artefacts, but at the same time holds the original patterns.

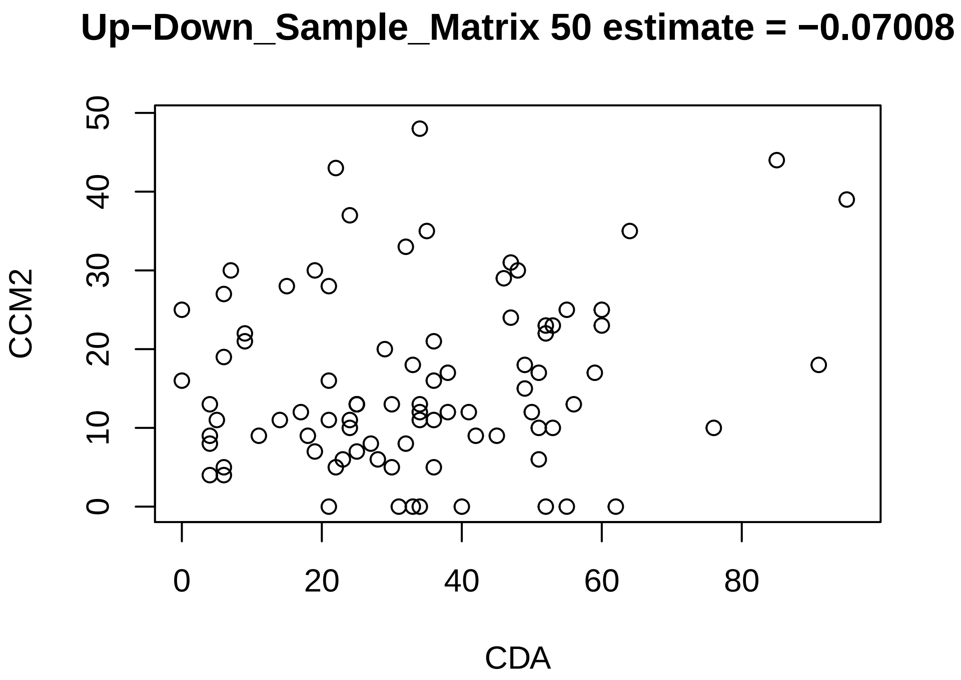


Figure 4.6: The data normalised by the *Up-Down-sampling* normalisation method.

4.2 Association methods

In this work, four different association methods are compared. All included methods can predict the direction of the directional association. It is crucial to mention that all the methods have the same settings for both directional and non-directional experiments.

4.2.1 Kruskal-Wallis test

The *Kruskal-Wallis (KW)* test, also called *Kruskal-Wallis rank sum test*, is the oldest method used in our comparison. It was firstly introduced in [38] more than half a century ago. The zero hypothesis for this test is:

- *The samples are from the same population.*

The base is the H test which is calculated by

$$H = \frac{12}{N\ (N+1)} \sum_{i=1}^{c} \frac{R_i^2}{n_i} - 3\,(N+1)\,,$$

where C is the number of samples, n_i the number of observations in the $i-th$ sample, N is the total sum of all observations n_i and R_i is the sum of the ranks of the $i-th$ sample.

If the samples come from the same population, the H value is expected to have $\chi^2(C-1)$ distribution. Thus the zero hypothesis is rejected for the large H values. The samples with a small number of observations $n_i \leq 5$ are to be handled differently as an exception. But since the scRNA-seq data is considered “BigData”, n_i is always large enough for this test. In *R*, the *kruskal.test* function from the *stats* package is used in our implementation.

4.2.2 Conditional entropy

[39] defines the *Conditional entropy* $H(Y|X)$ as:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathscr{X}} p(x)\ H(Y|X=x) \\ &= -\sum_{x\in\mathscr{X}} p\,(x) \sum_{y\in\mathscr{Y}} p\,(y|x)\ \ log\,(\ p\,(y|x)\ ) \\ &= -\sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} p\,(x,y)\ \ log\,(\ p\,(y|x)\ ) \\ &= -E_{p(x,y)}\ log\,(\ p(Y|X)\ )\,, \end{aligned}$$

where X and Y are random variables, $p(x,y)$ is the joint probability mass function, $p(x)$ is the marginal probability mass function of X, $p(y|x)$ is the conditional probability function of y when x is observed.

As evident from the definition formula, the *Conditional entropy* of $H(Y|X)$ is not equal to $H(X|Y)$. This means that the *Conditional entropy* is able to test whether Y is a function of X or vice versa and resolve both the dependency existence and its direction.

4.2.3 Causal inference by stochastic complexity (CISC)

Budhathoki and Vreeken [39] presents a new association method, which is founded on the Minimum Description Length (MDL) principle. The causal inference of two variables is calculated with the algorithmic Markov condition and the provable mini-max guarantees the optimality. They also defined an indicator for the directional association based on the stochastic complexity.

The stochastic complexity relative to $\mathcal{M}_m$ is calculated by

$$\mathcal{S}\,(X;\mathcal{M}_m) = n\ log\ n\ \ -\sum_{j=1}^{m} h_j\ log\ h_j\ \ + logR\,(\mathcal{M}_m, n)$$

$$R\,(\mathcal{M}_m, n) = \sum_{h_1+\cdots+h_m=n} \frac{n!}{h_1! \ldots h_m!} \prod_{j=1}^{m} \left(\frac{h_j}{n}\right)^{h_j},$$

where $\mathcal{M}_m$ is a multinomial model class, h_j is the number of times an outcome j is seen in X, and $logR(\mathcal{M}_m, n)$ is the parametric complexity of the model class $\mathcal{M}_m$, n is the size of the data set.

The total stochastic complexity from X to Y is calculated as follows:

$$\mathcal{S}_{X\to Y} = \mathcal{S}\left(X; \mathcal{M}_m\right) + \mathcal{S}\left(Y|X; \mathcal{M}_m\right)$$

The smaller the $\mathcal{S}_{X\to Y}$ statistics, the stronger the correlation. The statistics in our code are divided by the sample size multiplied by 10. Then the statistics are treated as the *p-value*. I emphasize that the *p-value* and $\mathcal{S}_{X\to Y}$ are not equal.

4.2.4 Functional index

Kumar et al. define the *Functional index*, a new method for dependency analysis, in [17]. The method originates in the *Functional chi-square* test. The *Functional chi-square* statistics χ_f^2 is defined by [16] as follows:

$$\chi_f^2\left(X \to Y\right) = \left[\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}}{s}\right)^2}{\frac{n_{i\cdot}}{s}}\right] - \left[\sum_{j=1}^{s} \frac{\left(n_{\cdot j} - \frac{n}{s}\right)^2}{\frac{n}{s}}\right],$$

where X and Y are random variables with the same number of samples, n is the number of samples, r is the number of samples of X, s is the number of samples of Y.

The *Functional index* is an improvement to the *Functional chi-square* test. It's defined by the following formula

$$\xi_f = \frac{\chi_f^2\left(X \to Y\right)}{n(s-1) - \chi^2(Y)}$$

The *Functional index* $\xi_f \in [0;1]$ The correlation grows with the grow of the ξ_f and therefore the $1 - \xi_f$ is treated a the *p-value* in the evaluation. I emphasize that *p-value* and ξ_f are not equal. [17] also provide the recommended thresholds for the dependency detection, which I follow when using the *Functional index* to detect new dependencies in the real data. To state the inference the statistics must satisfy the following condition: $\xi_f \geq 0.48$ and *p-value*≤ 0.05.

Chapter 5

Results

All the results that were achieved are presented in this chapter and described in detail. The first part comprises the normalisation methods evaluation. The second part displays the results of the association study.

5.1 Normalisation methods

The dropout creates artefacts that distort a negative correlation to the positive correlation and exaggerate the positive correlation. The most straightforward strategy is to remove all zeros when computing any statistics on relationships, so all zeros are omitted when comparing the normalisation methods.

The normalisation methods were compared empirically and visually on a real data set, provided by [15]. The estimate values of the *Spearman* correlation test are used for empirical evaluation. The value after normalisation is expected to be close to the original value before. The estimate value was selected to capture the transformation of the negative to the positive correlation or vice versa.

5.1.1 Examples on genes with known co-expression

The first study is focused on testing if and how the normalisation methods modify the existing correlation. Firstly, the dependency of a pair of genes is tested by the *Spearman* correlation test before normalisation. Then all the normalisation methods are applied separately, and the estimate value is calculated again. Then we compare how much the statistics changed.

Gene pairs with known correlations were selected to test the correctness of the raw data statistics. The chosen pairs include negatively correlated, positively correlated and independent genes.

Gene X	Gene Y	Type of correlation	H/L	Source
PHTF1	BCL11B	negative	L	[40]
PHTF1	BCL11B	positive	H	[40]
PHTF1	FEM1B	positive	L	[40]
PHTF1	APAF1	positive	L	[40]
CD34	ABCB1	positive	L	[36]
CD34	ABCG2	positive	L	[36]
CD34	ABCC1	independent	L	[36]
CD34	LRP1	independent	L	[36]

Table 5.1: The selected genes with a known dependency.

In following Figures, one example for each type of dependency is presented. The cases are shown in this order: negative dependency, positive dependency and independent gene pairs. The gene pair co-expression graph of the raw values is displayed first. Co-expression graphs after normalisation by present methods follow. The last Figure demonstrates the normalisation of the newly presented method.

Negative correlation

The negative correlation is illustrated by the PHTF1 and the BCL11B gene pair for the individuals diagnosed with an acute leukaemia, the estimate values are therefore expected to be negative. The ground truth is taken from [40]. The PCR was used to perform the co-expression study.

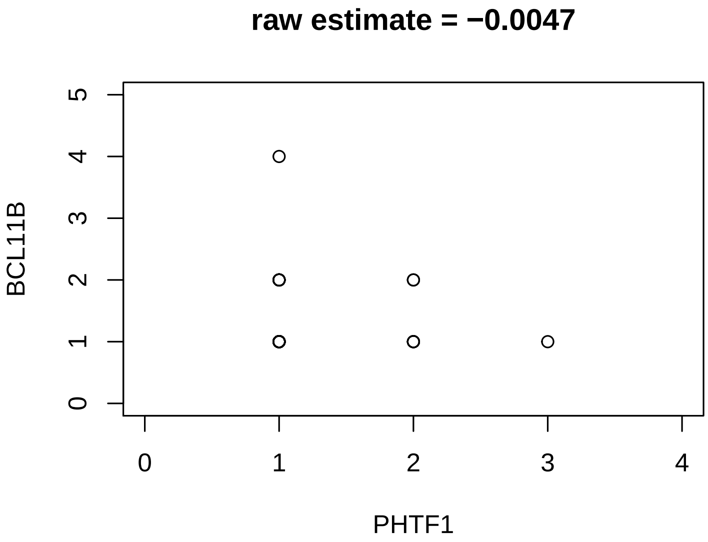


Figure 5.1: An example of the negative correlation of the PHTF1 and the BCL11B gene pair before any normalisation.

The first Figure 5.1 shows the estimate value and the co-expression graph before any normalisation. Although negative estimate values are expected, only a weak negative correlation is present. The normalisation methods are therefore assumed to increase the negative inference. The PCR was used to perform the co-expression study.

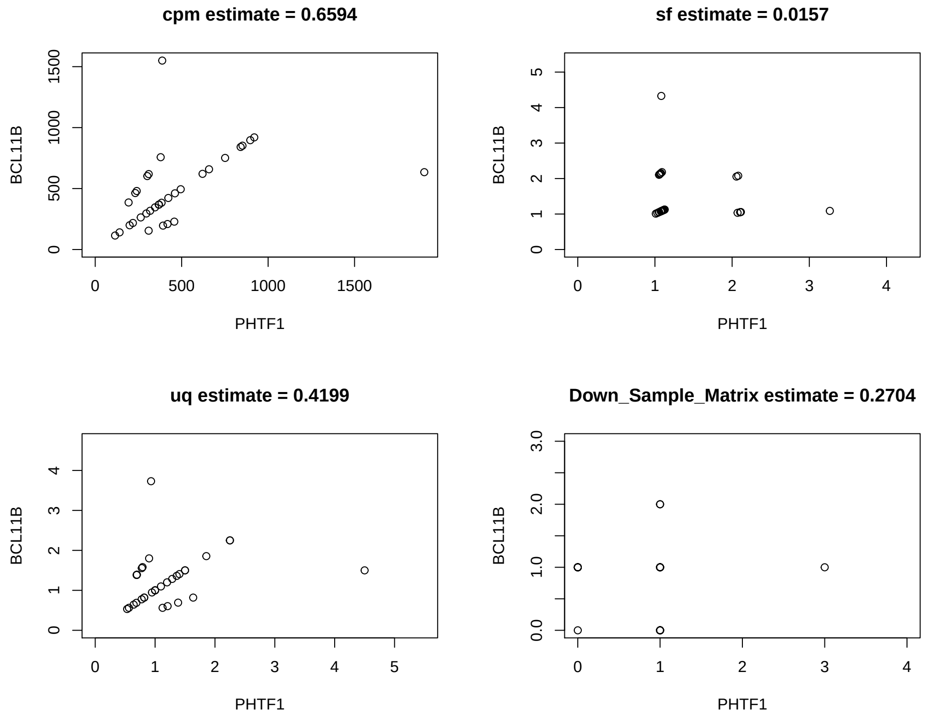


Figure 5.2: An example of the negative correlation of the PHTF1 and the BCL11B gene pair after the normalisation by four the normalisation methods.

The effect of the present normalisation methods is illustrated in Figure 5.2. All current normalisation methods have affected the estimate value, but the change has led to increasing the positive correlation. The *CPM* and *UQ* methods introduced new strong positive artefact, which were the origin of such high positive relationship. The *SF* method also created a new artefact, while however affecting the correlation statistic less than the two previous methods.

The *DSM* normalisation destroyed the original pattern, which also resulted in increasing the positive correlation.

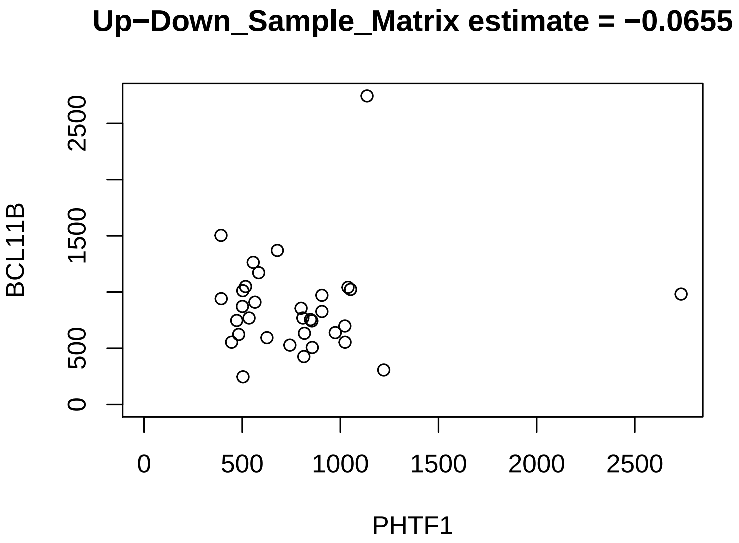


Figure 5.3: An example of the negative correlation of the PHTF1 and the BCL11B gene pair after the newly presented method normalisation.

Figure 5.3 shows the transformation of the data by the newly presented method. The *UDSM* method was the only method that increased the negative dependency. The *UDSM* also shows no sign of a new artefact creation. Therefore the *UDSM* normalisation was the most successful.

Positive correlation

An example of the positive correlation is demonstrated by the CD34 and the ABCG2 gene pair. Thus the estimate value must be positive. [36] introduces the ground truth.

Figure 5.4 shows the co-expression graph before any normalisation.

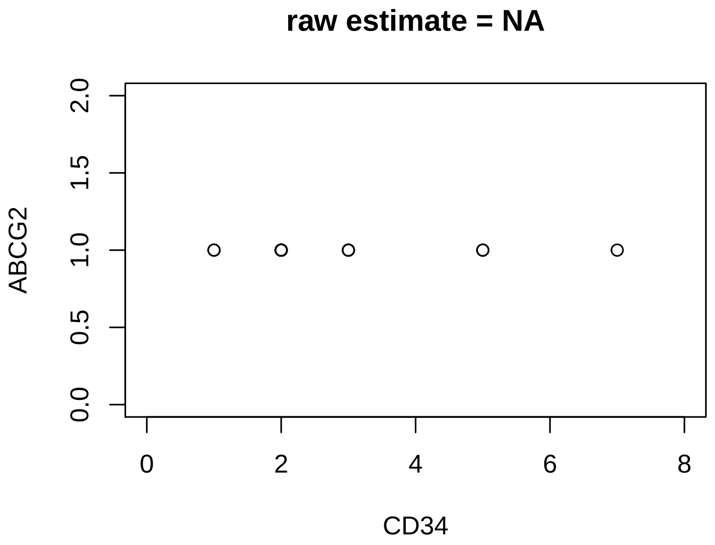


Figure 5.4: An example of the positive correlation of the CD34 and the ABCG2 gene pair before any normalisation.

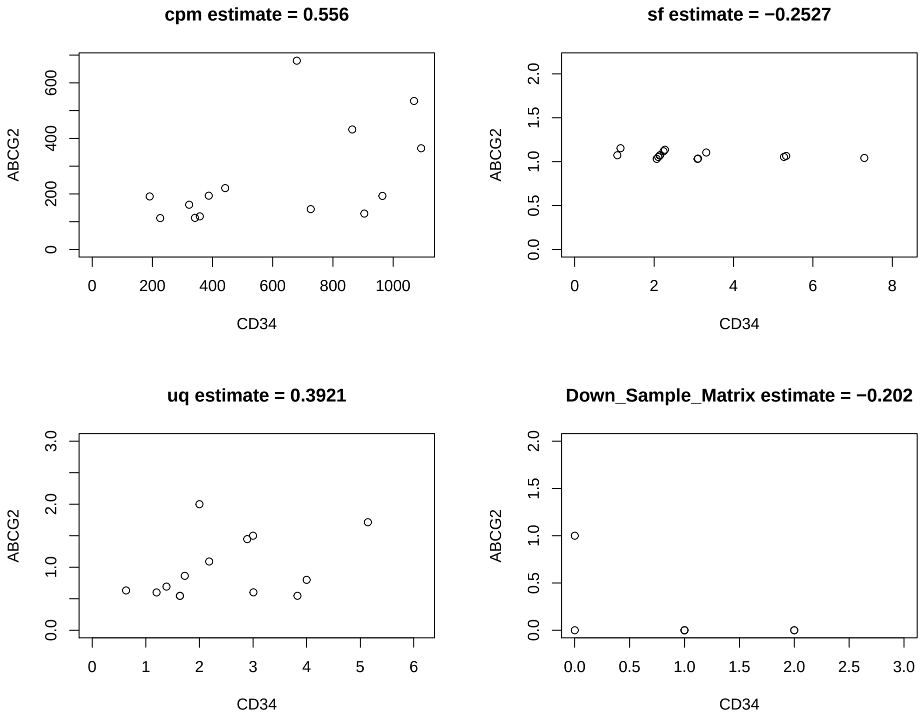


Figure 5.5: An example of the positive correlation of the CD34 and the ABCG2 gene pair after the normalisation by the four normalisation methods.

The normalisation effect of all present methods is displayed in Figure 5.5. The two methods *SF* and *DMS* have influenced the estimate value towards the negative correlation, although this gene pair is known to have a positive dependency. The *DSM* also destroyed the pattern noticeable in the raw data. The *CPM* and the *UQ* normalisation improved the correlation statistics. The *CPM* performed the best.

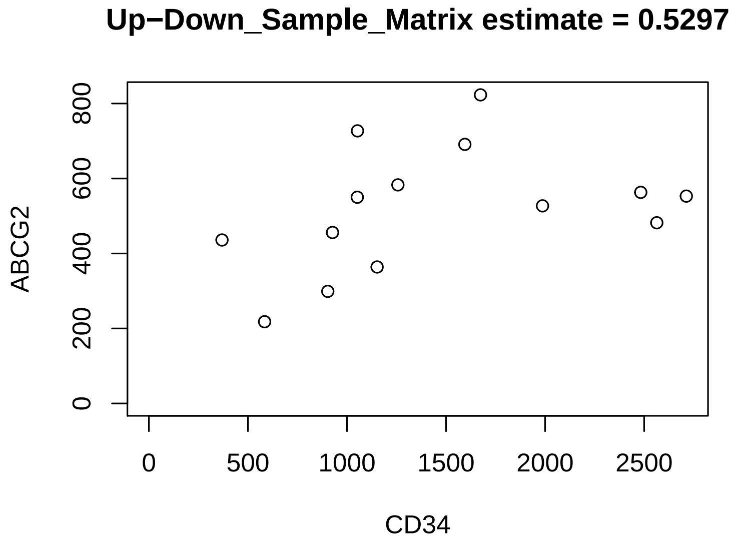


Figure 5.6: An example of the positive correlation of the CD34 and the ABCG2 gene pair after the newly presented method normalisation.

The *UDSM* normalisation improved the positive correlation, which is demonstrated in Figure 5.6. The estimate value approaches the best performance of the current methods.

Independent pair of genes

The independent pair of genes is illustrated by the CD34 and the ABCC1 gene pair for the individuals diagnosed with an acute leukaemia, the estimate values are therefore expected to be zero. The ground truth is taken from [40]. The PCR was used to perform the co-expression study.

The estimate value before any normalisation is already close to zero as shown in Figure 5.7.

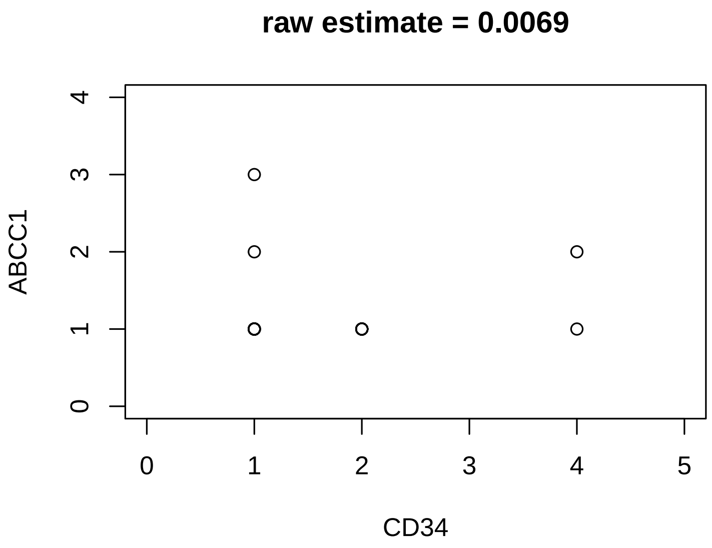


Figure 5.7: An example of the independent dependency is the CD34 and the ABCC1 gene pair before any normalisation.

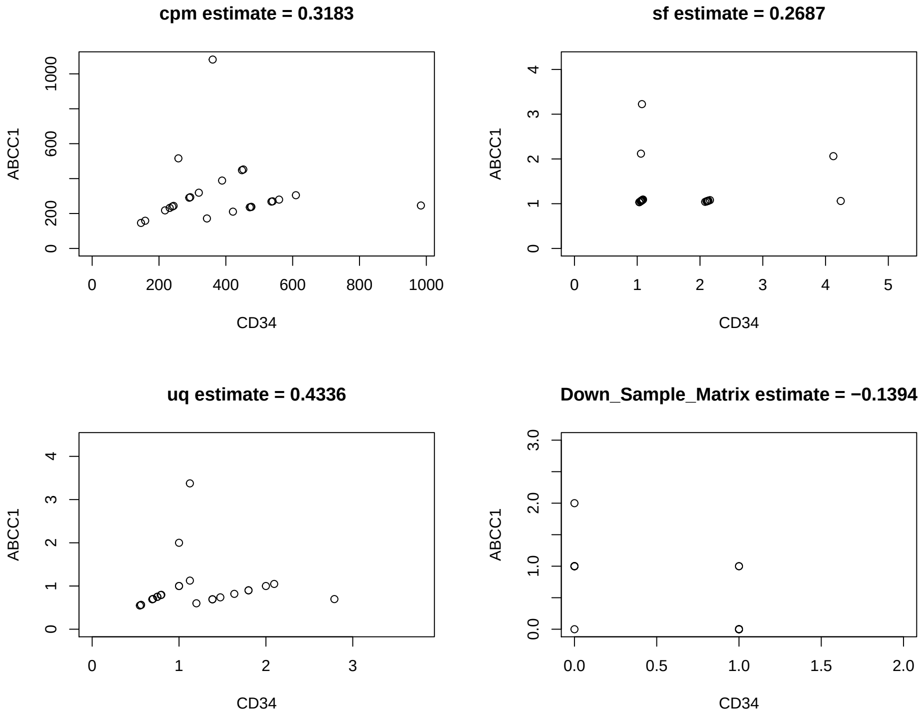


Figure 5.8: An example of the independent correlation is the CD34 and the ABCC1 gene pair after the normalisation by the four current normalisation methods.

Figure 5.8 demonstrates the performance of how the current methods have increased the correlation in either a positive or a negative direction. The *CPM* and *UQ* have also created a new artefact that affected the result significantly.

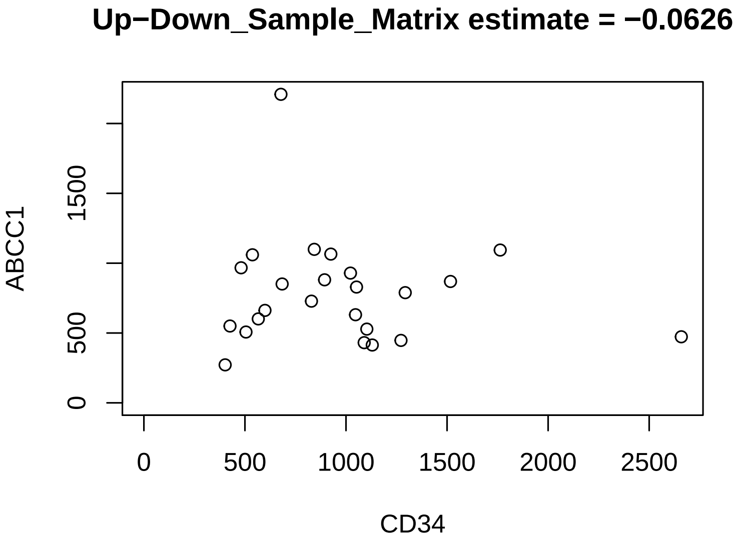


Figure 5.9: An example of the independent gene pair the CD34 and the ABCC1 after the newly presented method normalisation.

The *UDSM* method performed the best, which is demonstrated in Figure 5.9. The *UDSM* has also increased the correlation estimate value, but the increment is insignificant. Also, the *UDSM* estimate value is closest to zero compared to the other methods.

5.1.2 Permutation study

To check if the normalisation methods introduce a new artefact, we present a permutation test on the real data as a negative control study. The permutation study was designed with the following steps. Firstly, the data are permuted, for each row across the cells independently. Then a thousand of gene pairs is randomly selected and tested by the *Spearman* correlation test. The *Spearman* correlation test has no assumption of the distribution, which is the reason for its use in this study. When all the cells are permuted randomly across the genes, all the positive or negative correlation should be destroyed. Therefore the estimate values are expected to be close to zero. To prevent the influence of the dropout on the estimate value, we omit all zero values.

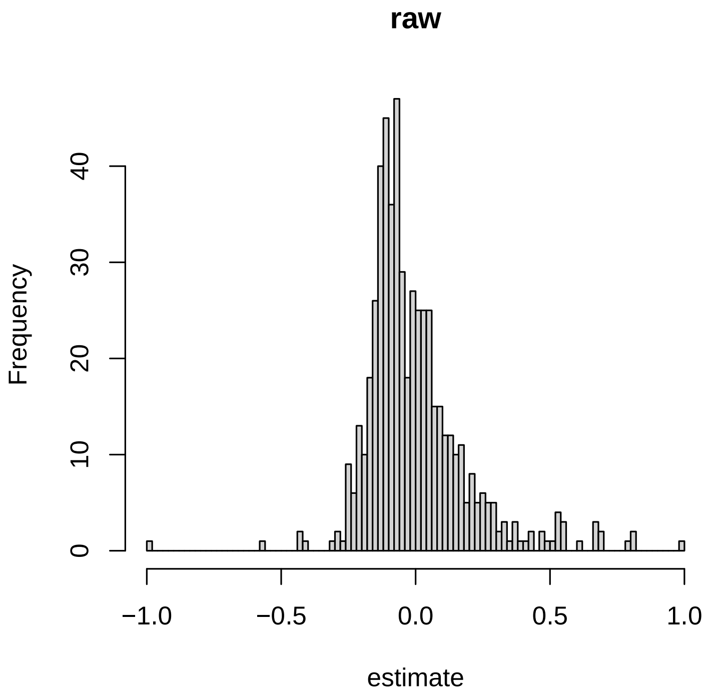


Figure 5.10: The permutation study estimate value histogram of the raw data with permuted cells.

Figures 5.10 5.11 5.12 show the results of the permutation study. We can see in Figure 5.10 that the estimate value of the raw permuted data set is close to zero for most the permuted pairs, which means that the main idea of the permutation study is correct.

The next important thing to notice are the estimate value histograms of all three normalisation methods based on the library size in Figure 5.11. The estimate values of Counts per million, Relative log expression and 99^{th} *percentile* normalisation methods is equal to 1 for most of the values, which means that all of the three methods introduce a new positive correlation. Although the permutation study shows the *Down-sample (DSM)* to be the best performing method, it is not surprising because it was previously proved that it destroys the original pattern.

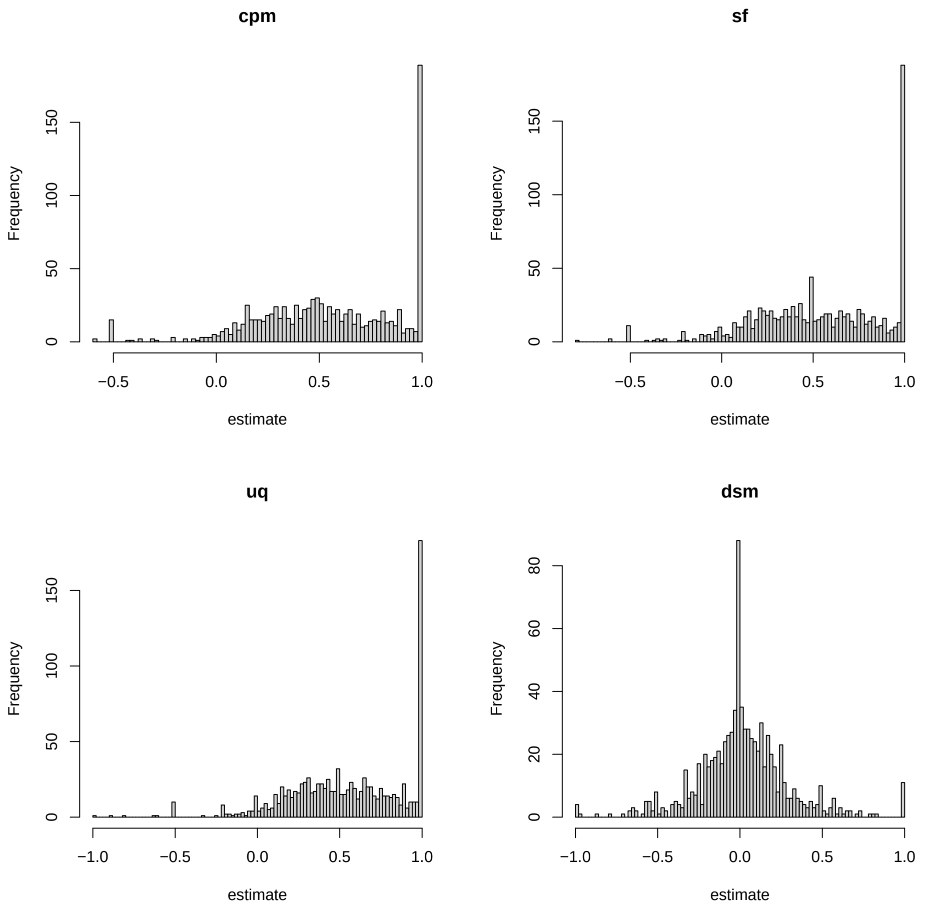


Figure 5.11: The permutation study estimate value histograms of all the current normalisation methods.

The new normalisation method presented here performs with the most of the estimate values falling in the [-0.5, 0.5] interval, and following the *DSM* method, the results are the second-best (Figure 5.12). This proves that the *UDSM* method does not introduce new patterns in most cases that would cause a positive correlation in comparison to the previous methods. However, some artefacts are created which is illustrated by the peaks for absolute estimate value greater than 0.5, but no trend was captured, unlike in the case of the library size normalisation methods. The spikes are randomly distributed, which can be a side effect of the additional noise or the cell permutation. Even though the histogram proves that the *UDSM* doesn't show a tendency of introducing new patterns of the same type, there was a new pattern introduced. The decomposition of peaks illustrates a random pattern which is probably connected to the randomised noise or the cell permutation. This means that a new study should occur with a focus on the parameterf of the *UDSM* method in the future.

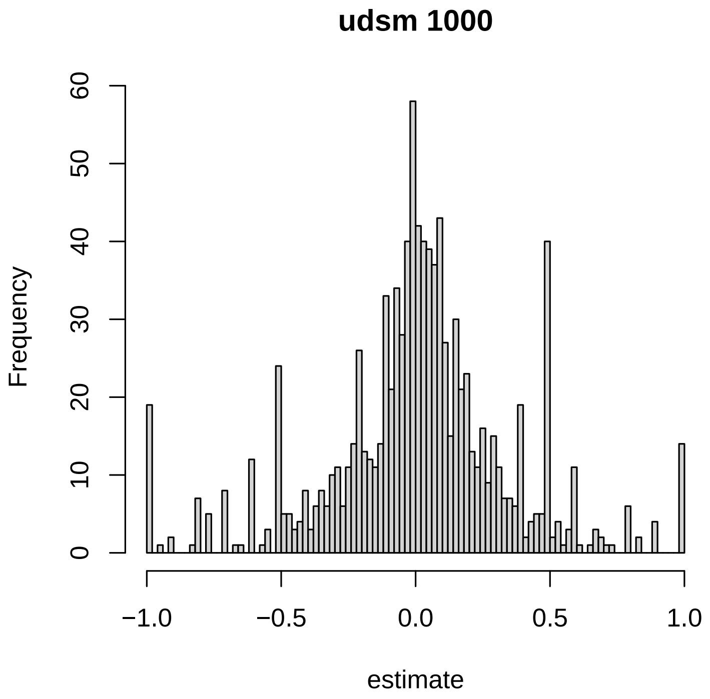


Figure 5.12: The permutation study estimate value histogram for the *Up-down-sampling* normalisation method.

5.2 Association methods

The association experiments are divided by the data type: simulated data and real data. To measure the performance of the methods, we use receiver-operating characteristic (ROC) and precision-recall (PR) curves and the areas under the curves (AUROC and AUPR).

5.2.1 Simulated data

The evaluation of the association methods on the simulated data sets is divided into two main parts: The non-directional study and the directional study. The non-directional study focuses on how accurately is a method capable of detecting the existence of dependency. The data set generated for both studies contains 200 contingency tables. The half of the data consists of samples where X is a function of Y $(f : X \to Y)$ and Y is a function of X $(f : Y \to X)$. The other half of the tables are contingency tables of independent variables. The directional study tests how accurate can the methods predict the direction of the inference. Therefore the data set for the directional study contains contingency tables where X is a function of Y and Y is not a function X, where both directions are tested.

The experiments were run for varying parameters. The whole experiment settings consist of the combination of dropout rate levels 0 %, 30 %, 50 %, 70 %, 90 %; noise parameters: 0.1, 0.2, 0.3 and sample sizes: 100; 1,000; 10,000. The combination of all metrics allows us to detect which parameter impacts the performance of the methods the most. The dropout rate influences the performance most significantly; even a low dropout rate causes a visible change in the accuracy. The sample size also causes a noticeable effect but with the a positive tendency along with the size growth. The noise level has almost no impact on the association experiment. Although a small range of noise levels was tested, the impact of the dropout is noticeable even at 30 %, while at 0.3 the noise level parameter is not. Thus the noise level impact is the lowest in comparison to the other parameters.

Next Figures 5.13 5.14 5.15 show the evaluation of all the methods between each other by AUROC and AUPR metrics in the non-directional study. We can see that the *Functional index* overcame or performed at least the same as all the other methods. The second best is the *CISC* and the *Conditional entropy*. Even though the *Kruskal-Wallis* test shows the worst performance over all settings, its accuracy increases with a high number of the sample size.

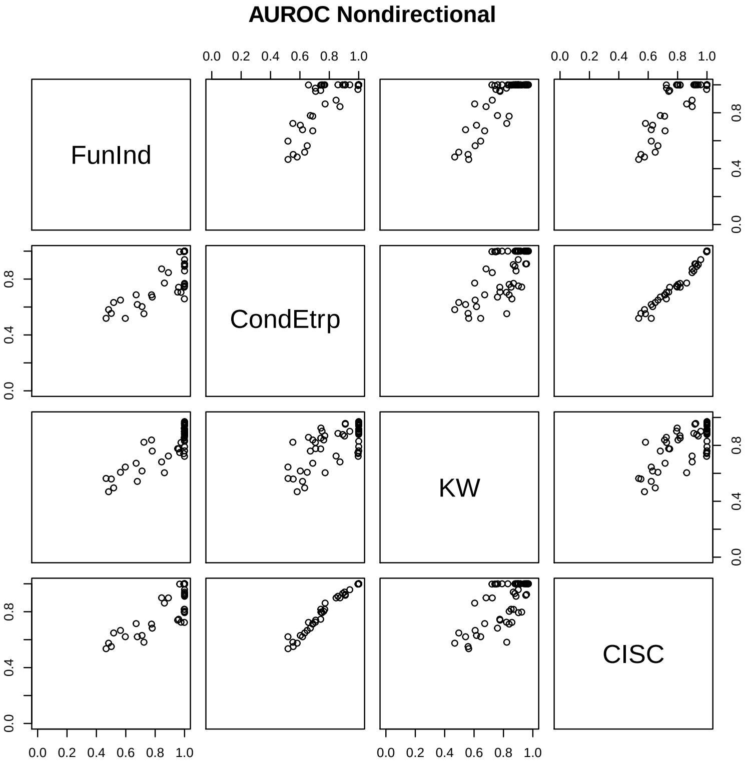


Figure 5.13: The performance evaluation of the association methods in the non-directional study (AUROC).

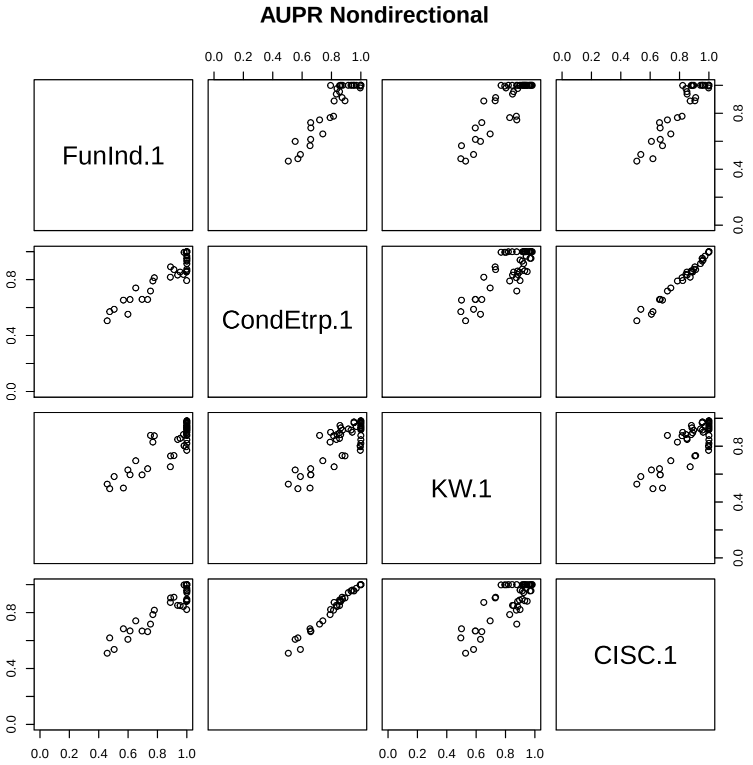


Figure 5.14: The performance evaluation of the association methods in the non-directional study (AUPR).

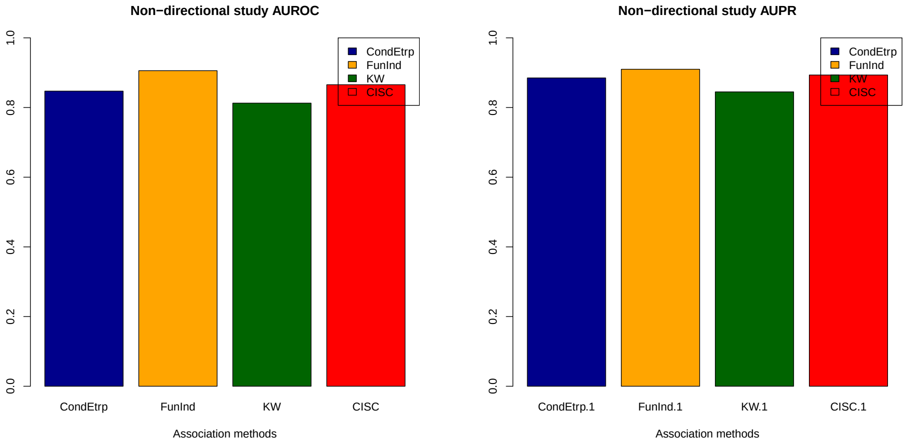


Figure 5.15: The bar plots of the association study evaluation.

The evaluation of the association methods in the directional study is illustrated in Figures 5.16 5.17 5.18. Even though the *CISC* and the *Kruskal-Wallis* methods are designed to predict the dependency directions, their performance is close to random in the study outlined in this thesis. The reason for this behaviour can be the base idea of the directional test. In the analysis presented here, no assumptions on dependency are made before the experiment, so the methods answer the question "*is X a function of Y*", regardless of the study type instead of answering the question "*is* $(f : X \rightarrow Y)$ *stronger than* $(f : Y \rightarrow X)$*?*" The *Functional index* performs slightly more accurately using the AUROC metric. *Conditional entropy's* accuracy was lightly better than the accuracy of the *Functional index* using the AUPR metric. But when the patterns of the strongest correlations are printed for both methods, we see that patterns picked by the *Functional index* are more reliable than the patterns rated the highest by the *Conditional entropy.* An example of these patterns is demonstrated in Figure 5.19.

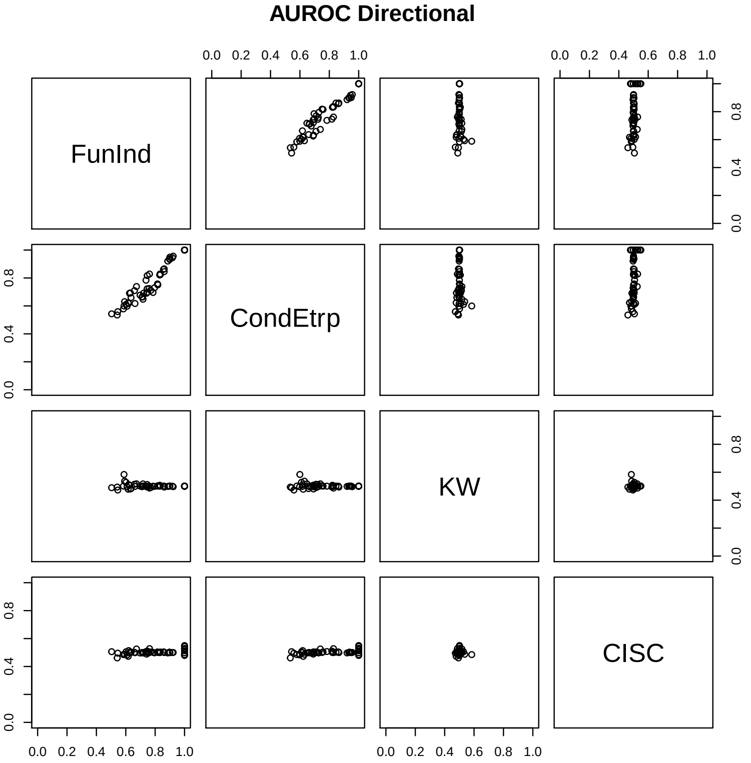


Figure 5.16: The performance of the association methods in the directional study (AUROC).

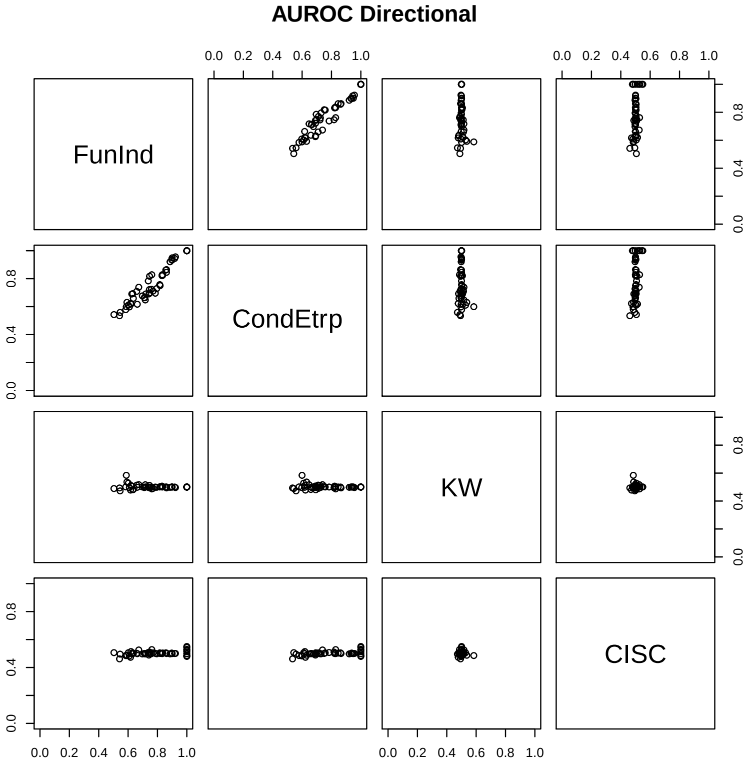


Figure 5.17: The performance evaluation of the association methods in the directional study (AUPR).

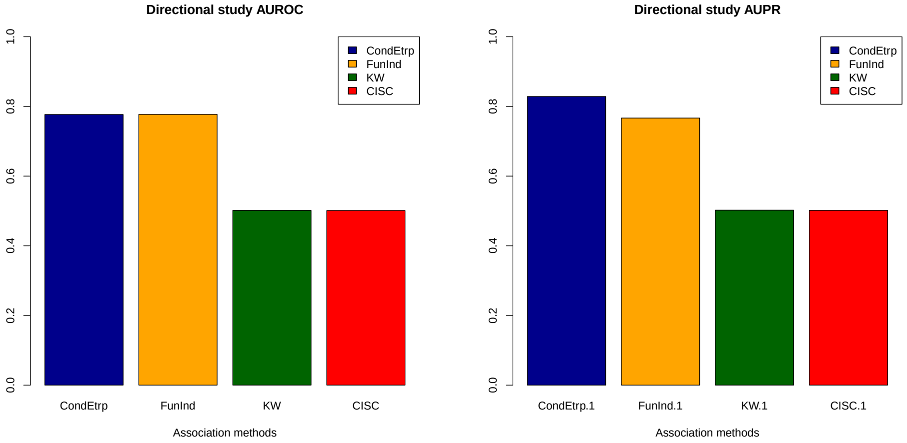
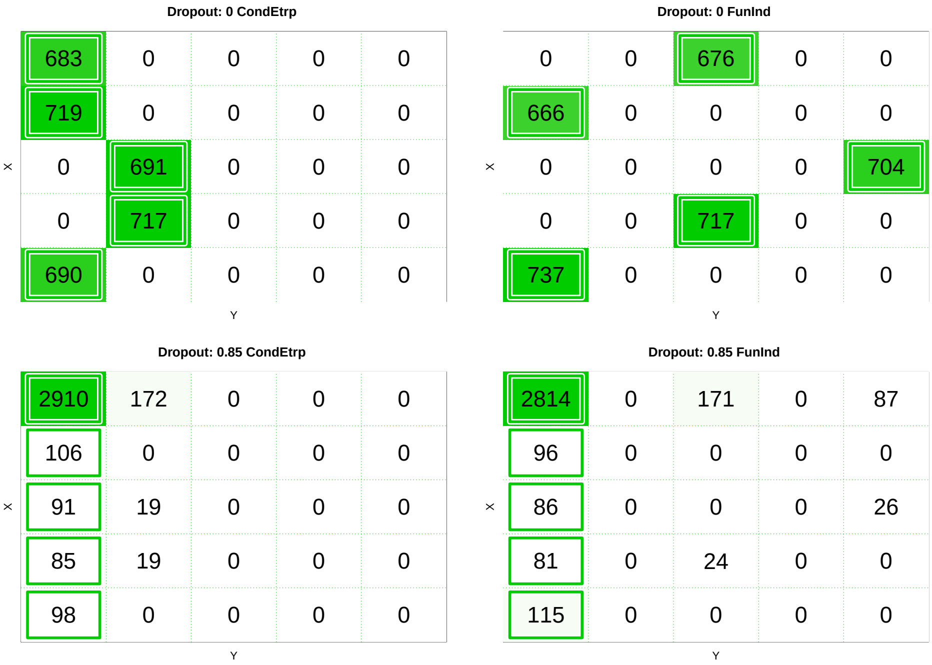


Figure 5.18: The bar plots of the directional study evaluation.

Figure 5.19 shows the patterns rated as the best by the *Functional index* and the *Conditional entropy*. The setting of the experiment was the same for all selected contingency tables by both methods. The environment has been defined by a noise parameter 0.2, a 85% dropout rate and a sample size of 3,500 to mimic the real data set settings as much as possible. On the left, there is a contingency table with the highest rating given by *Conditional entropy*. On the right, there is the most correlated pair of variables selected by the *Functional index*. The first row contains tables without dropout to demonstrate the original pattern. The second row consists of the tables from the first row with simulated 85% dropout.

The *Conditional entropy* selected the contingency table with the most of the values with at least one of the coordinate equal to zero. But since zero values in at least one coordinate characterise the dropout, this pattern increases the false positive rate. Compared to that, the *Functional index* rated the highest the tables with a strong correlation pattern, which prevents false positive ratings.



Dropout: 0 CondEtrp

683	0	0	0	0
719	0	0	0	0
0	691	0	0	0
0	717	0	0	0
690	0	0	0	0

Dropout: 0 FunInd

0	0	676	0	0
666	0	0	0	0
0	0	0	0	704
0	0	717	0	0
737	0	0	0	0

Dropout: 0.85 CondEtrp

2910	172	0	0	0
106	0	0	0	0
91	19	0	0	0
85	19	0	0	0
98	0	0	0	0

Dropout: 0.85 FunInd

2814	0	171	0	87
96	0	0	0	0
86	0	0	0	26
81	0	24	0	0
115	0	0	0	0

Figure 5.19: The patterns rated the highest by the *Conditional Entropy* (left) and the *Functional index* (right).

5.2.2 Real data

The real data experiments were firstly run on the genes from mouse cerebellar development data provided by [41]. These results were already published in [35], which has also proven that the *Functional index* performs the best.

The real data experiment in this work is based on the examples from literature shown in Table 5.1. Since the *UDSM* normalisation method performed the best in previous normalisation experiments, the data have been normalised with the *UDSM* first. The test of the *UDSM* was performed with all the zero values omitted, so this study is also evaluated with all the values with at least one coordinate equal to zero omitted to prevent an unexpected impact caused by the normalisation. Thus an additional rule is presented to preclude evaluation of the empty sample, the size of which is at least 3. The results are presented in the following graphs in Figure 5.21.

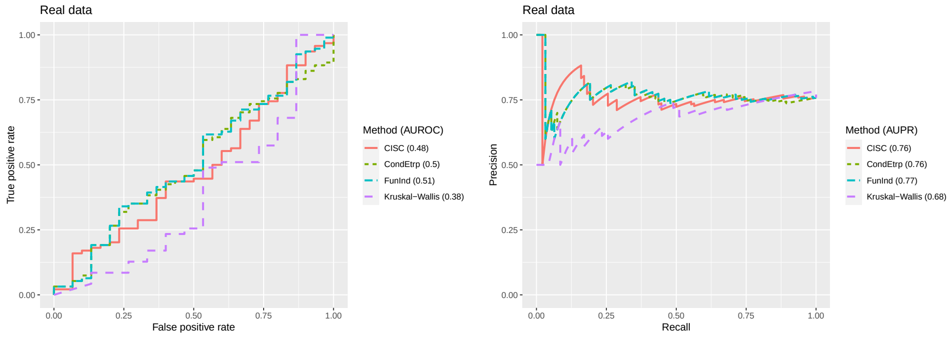


Figure 5.20: Results of the real data study.

The results of the real data study are illustrated in Figure 5.21. Although the *Functional index* performed slightly better than the other methods, all the functions achieved the accuracy close to random guessing using the AUROC metric. The reason for this might be the high zero rate in the real data, as mentioned in 3.1. The gene pairs found in the literature appear to have low expression levels in the used real data, so further testing with more gene pairs is encouraged.

5.2.3 Protein-gene dependency discovery

All real data samples of individuals diagnosed with an acute leukaemia were processed by the *Functional index* association method to discover protein-gene or gene-protein dependencies.

The experiment follows these steps

1. Normalisation of transcriptomic data (scRNA-seq) by *UDSM*.
2. Genes and proteins data are connected by the cell names
3. The *Functional index* detects inference.

The results are displayed in Table 5.2. The pairs detected in more than one sample are listed. However, no dependency was discovered more than twice. We can see that most of the relationships come from the *GSM4138879* sample, the sample size of which is greater than 1,000 and lower than 2,000. Chapter 3.1 includes the statistics of the real data samples. The samples *GSM4138883* and *GSM4138887* have a similar sample size lower then 2000 cells and are also listed as a source of the dependency multiple times. Only a single dependence originates from a sample *GSM4138880* with a size greater than 2000, which is surprising.

X	Y	Type	Average ξ_f	Average *p-value*	Samples
CD46	CD3	$G \rightarrow P$	0.76	0.03	84, 86
CEP70	CD3	$G \rightarrow P$	0.73	0.03	86, 87
RCN2	CD7	$G \rightarrow P$	0.71	0.04	79, 87
CLU	CD34	$G \rightarrow P$	0.70	0.03	79, 82
ADD1	CD34	$G \rightarrow P$	0.69	0.04	79, 83
CD34	SSBP2	$P \rightarrow G$	0.67	0.02	81, 82
TCEA2	CD45RA	$G \rightarrow P$	0.64	0.01	79, 83
ARHGAP11B	CD45RA	$G \rightarrow P$	0.63	0.03	79, 87
ARNT	CD34	$G \rightarrow P$	0.62	0.04	79, 83
SHOX2	CD7	$G \rightarrow P$	0.61	0.03	79, 87
SIGLEC8	CD7	$G \rightarrow P$	0.61	0.01	79, 87
ARL13B	CD34	$G \rightarrow P$	0.61	0.03	79, 83
DNAJC5B	CD7	$G \rightarrow P$	0.60	0.03	79, 87
CACNA2D2	CD34	$G \rightarrow P$	0.59	0.03	79, 83
ASNA1	CD45RA	$G \rightarrow P$	0.58	0.03	79, 80
MAGI2	CD33	$G \rightarrow P$	0.57	0.03	79, 87

Table 5.2: The discovered protein-gene or gene-protein dependencies.

Table 5.2 shows the gene-protein correlated pairs discovered using the real data. One record contains the name of the correlated gene and the protein in the order of dependency direction. Then the direction type is displayed, either $G \rightarrow P$ or $P \rightarrow G$, where the G corresponds to the gene and the P to the protein. The next two numbers are the average ξ_f and the average *p-value* statistics. The names of the samples, where the correlation was found, are shown in the last column. The sample names are shortcuts, which correspond to the last two digits of the sample name. All the dependencies are sorted by the average ξ_f value.

Figure 5.21 shows the gene-protein correlation with the highest ξ_f values. The example includes two graphs, the graph on the left corresponds to $CD46 \rightarrow CD3$ inference before any normalisation and the graph on the right contains data after the *UDSM* normalisation. The added noise for zeros seems to influence the of the association study and therefore further study on the zero handling will be conducted in my future work.

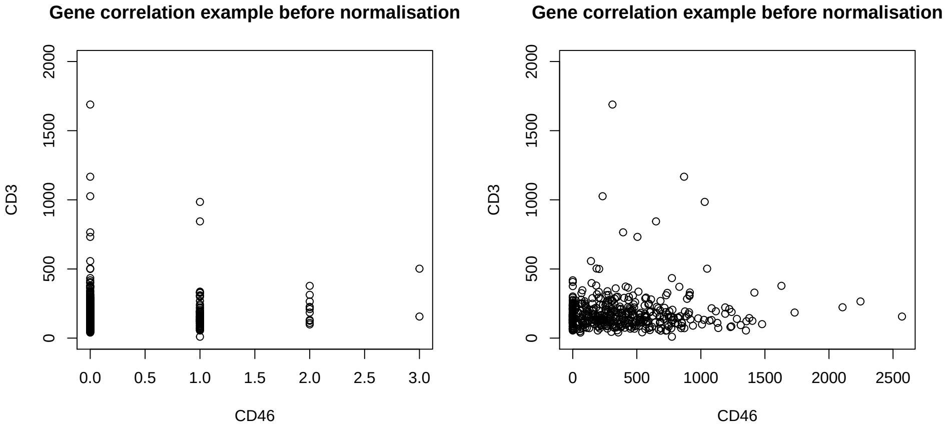


Figure 5.21: The discovered pattern rated the best ξ_f from the *GSM4138884* sample. The pattern before the normalisation is shown on the left and the pattern after the normalisation is shown on the right.

Chapter 6

Discussion

In this chapter the main findings are described.

6.1 Design of the Up-down-sampling normalisation method

Firstly, the correctness of the new *Up-down-sampling* method, presented in this work, is discussed. In the first step, new extra noise is added to the original data. The additional noise prevents a new artefact creation. A small distortion, in the beginning, detaches points with the same coordinates. This separation prevents originally identical values from creating a smooth diagonal after the data normalisation. The smaller the initial values are, the more significant impact the distortion has. The strong correlation requires a number of different original values that form a pattern. As long as the initial distortion is kept relatively small, the overall pattern is not destroyed.

The extensive number of zeros is characteristic for the scRNA-seq data and has a significant influence on the inference study, as mentioned in Chapter 1. Therefore the data manipulation that affects the lower values the most is a crucial advantage of decreasing the effect of the zero values. It also prevents the creation of a diagonal pattern from many original points with equal coordinates. The distortion breaks the smooth diagonal directly proportional to the length parameter.

To prove that the initial distortion is appropriately designed, its detailed description follows. The *runif* function from the *stats* package creates the noise distribution. The noise is generated for the whole data set, separately for the zero values and the rest. The new values are uniformly distributed within an open interval where the minimum is set to $o-0.5$ and the maximum is set to $o+0.5$, where o is the original value. E.g. all original values $o=1$ are evenly distributed into an interval $(0.5, 1.5)$. The zero values are an exception because it is necessary to prevent the negative values, so the permissible minimum of the resulting interval equals zero.

The length of the distortion interval l is one of the parameters of the presented normalisation method. In the case of the study presented in this work, the length parameter l is set to one. The parameter was chosen to be

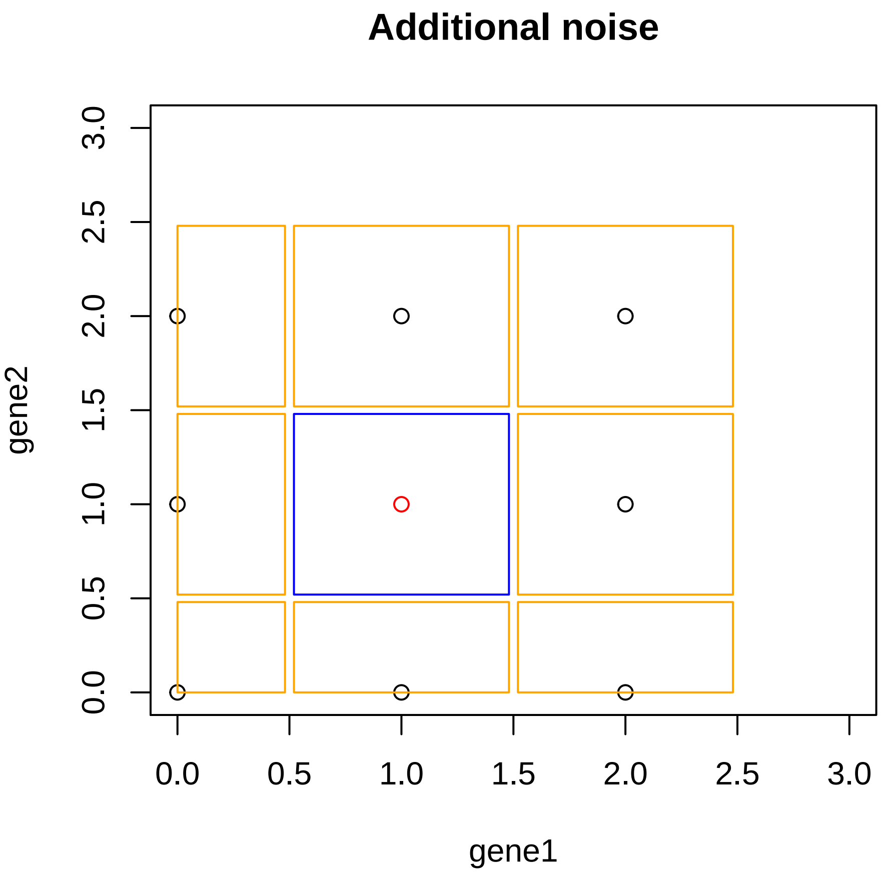


Figure 6.1: The areas where the shifted values are located. The blue square demonstrates the location of the distorted values, the original coordinates of which were [1, 1].

as large as possible but still reversible by rounding of the shifted values. We emphasize that this setting is an initial compromise and worth testing the performance for other values. This parameter reflects the accuracy of the RNA sequencing. If the error is assumed to be greater than one, the length parameter can reflect that.

Figure 6.1 illustrates the final distortion. The previous example of the distortion of ones is demonstrated by the blue square, which restricts the area of the shifted values. The figure shows a graph of a relationship between two genes (rows). The orange squares display areas of other shifted values. The target areas are equal and separable, which means that the shifted values are still the closest to the original integer values.

Before generating the data with the random generator based on the multinomial distribution, the target space for the generator is stretched. Each value is multiplied by 1000. The scaling is necessary for the patterns to have enough space to be shaped. The need for this step is demonstrated by the *Down-sampling* in Chapter 5. The multiplication parameter m is also an initial configuration which needs to be studied more.

6.2 Possible improvement

The *UDSM* normalisation method has two modifiable parameters: the length of the noise distortion l and the multiplication parameter m. The impact of the parameters is not known in detail and needs to be tested further.

6.3 Results of the normalisation methods evaluation

The newly presented normalisation method outperformed the other methods in the permutation study and also in the gene examples correlation study. The correlation study shows on examples that the *UDSM* improves the estimate value and is not prone to destroying and creating an artefact.

All normalisation methods based on the library size create new artefacts that increase the positive correlation, which has been proved in the example study and the permutation study. The *CPM* creates new patterns with the most significant impact on the estimate value. The *UQ* also affects the statistics towards a positive correlation, but less notably than the *CPM*. Although the *SR* method also impacts the estimate value, the effect is the least significant of all the library size normalisation methods. However, the effect of creating new artefacts is still relevant.

The *DSM* normalisation method prevents creating new artefacts but destroys the original pattern, which was demonstrated in the example study. The permutation study also shows the signs of the *DSM* erasing the dependency patterns.

The method presented in this work, *UDSM* has been proved to improve the estimate value in the example study and performed the second best in the permutation study. The study also shows that the *UDSM* method has created a few artefacts, which means that a further study of the method parameters must follow.

6.4 Results of the association study

The association study evaluates four association methods. All methods are able to detect the direction of the inference. The study using the simulated data has been divided into two parts, where one is dedicated to testing the ability of detecting the existence of inference. The second study evaluates the accuracy of detecting the relationship direction.

The non-directional study shows a dominance of the *Functional index*. Also, the *Kruskal-Wallis* method performed reliably, but only for the large sample sizes, and therefore has the highest error rate. The *Conditional entropy* and the *Causal inference by stochastic complexity* share similar accuracy.

The directional study is more interesting because of the performance of the *Causal inference by stochastic complexity* and the *Kruskal-Wallis* test. Although the *CISC* was explicitly designed to detect the inference direction, its accuracy is equal to random guessing. The reason might be the design of our experiment. In the directional experiment presented here, all methods answer a simple question $f : X \to Y$ with *false* or *true*. There is no assumption of the inference of the variables, and both directions are tested independently, so there is information only about a single direction. Also, the *Kruskal-Wallis* test performs close to random guessing, but its accuracy grows with the growing sample size.

The *Conditional entropy* and the *Functional index* perform similarly. Therefore the patterns rated the highest by both methods were examined in detail. The contingency tables rated the highest by the *Functional index* reveal a more reliable inference than the patterns selected by the *Conditional entropy*. Therefore I claim the *Functional index* is the best performing method.

We published the evaluation of the association methods on real data in [35], where the *Functional index* outperformed the other methods with the lowest error rate. Another study on the real data was performed in this work. Gene pair examples were selected with the ground truth from the literature [36, 40]. The *Functional index* detected dependencies with slightly better accuracy, but the size of gene examples set should be increased for the test in the future. The chosen data set provided by [15] contains very sparse matrices which make the analysis more difficult.

The *Functional index* was used to discover new dependencies. Sixteen dependencies were found in two samples out of ten. Some samples contain only a few cells which makes the analysis more difficult.

Chapter 7

Conclusions

This thesis has studied how noise in the single-cell RNA sequencing (scRNA-seq) affects data analysis, specifically from the point of view of dropout and normalisation. The noise is created during the necessary library preparation process and is thus always present. The information dropout is characterised by many zero observations in the result scRNA-seq data, with a rate of 30–90 %. Although the scRNA-seq brings new possibilities for studying, it also decreases the reliability of analytical methods developed using the bulk RNA-seq data. Three association methods (*Causal inference by stochastic complexity*, *Kruskal-Wallis test*, *Conditional entropy*), on both directional and non-directional tests, were evaluated and compared. A recent method *Functional index* was suggested and compared with the previously mentioned methods. The experiments were performed on both real and artificially generated data sets. The simulated data allows us to test each parameter separately with a precise ground truth. On the other hand, examinations based on real data are essential for everyday use. The *Functional index* performed better on both the directional and the non-directional analysis and both the artificial and the real data. A test using real data provided by [15] was performed in this work. Although the *Functional index* performed the best, the AUROC and AUPR statistics are low for all the methods. However, we have already published the results of an experiment using different real data [35], which has shown more promising results of the *Functional index*. The *Functional index* has also been used to discover protein-gene or gene-protein correlations. Sixteen correlated pairs were detected in two different samples.

Secondly, I have studied and demonstrated how the normalisation during data preprocessing affects the results of the association method. The standard normalisation methods were examined and compared using the real gene-gene relationships. I have found out that in the case of the association analysis, it can create and also destruct the original patterns we are trying to find. Four standard normalisation methods (*Counts per million*, *Relative log expression*, 99^{th} *percentile*, *Down-sampling*) were assessed to see how the artefacts are modified after normalisation. Several examples of the artefact creation and destruction have been shown, and a new solution has been proposed.

The new *Up-down-sampling (UDSM)* method, which was presented in this work, has been shown to improve the association methods and minimize the distortion of gene-gene relationships. The *UDSM* introduced less new artefacts in the permutation study than all the library size based methods and improved the estimate correlation statistics in the example study. Although the *UDSM* has performed well in the presented experiment, it also has shown signs of unexpected behaviour in the permutation study, so a further examination of the method's parameters is planned for my future work.

Bibliography

[1] Jun Ding, Chieh Lin, and Ziv Bar-Joseph. Cell lineage inference from SNP and scRNA-Seq data. *Nucleic Acids Research*, 47(10):e56–e56, 03 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz146. URL `https://doi.org/10.1093/nar/gkz146`.

[2] Arsham Ghahramani, Fiona M Watt, and Nicholas M Luscombe. Generative adversarial networks simulate gene expression and predict perturbations in single cells. *BioRxiv*, page 262501, 2018.

[3] Hao Dai, Lin Li, Tao Zeng, and Luonan Chen. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Research*, 47(11):e62–e62, 03 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz172. URL `https://doi.org/10.1093/nar/gkz172`.

[4] Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell systems*, 8(4):315–328, 2019.

[5] Tal Nawy. Single-cell sequencing. *Nature methods*, 11(1):18–18, 2014.

[6] Dongju Shin, Wookjae Lee, Ji Hyun Lee, and Duhee Bang. Multiplexed single-cell rna-seq via transient barcoding for simultaneous expression profiling of various drug perturbations. *Science advances*, 5(5):eaav2249, 2019.

[7] Xifang Sun, Shiquan Sun, and Sheng Yang. An efficient and flexible method for deconvoluting bulk rna-seq data with single-cell rna-seq data. *Cells*, 8(10):1161, 2019.

[8] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, 2019.

[9] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):1–21, 2018.

[10] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[11] Sui Huang. Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 136(23):3853–3862, 2009.

[12] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[13] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11 (2):163, 2014.

[14] 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.

[15] Jeffrey M Granja, Sandy Klemm, Lisa M McGinnis, Arwa S Kathiria, Anja Mezger, M Ryan Corces, Benjamin Parks, Eric Gars, Michaela Liedtke, Grace XY Zheng, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology*, 37(12):1458–1465, 2019.

[16] Hua Zhong and Mingzhou Song. Directional association test reveals high-quality putative cancer driver biomarkers including noncoding rnas. *BMC Medical Genomics*, 12(7):1–10, 2019.

[17] Sajal Kumar, Hua Zhong, Ruby Sharma, Yiyi Li, and Mingzhou Song. Scrutinizing functional interaction networks from rna-binding proteins to their targets in cancer. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 185–190. IEEE, 2018.

[18] Hua Zhong and Mingzhou Song. A fast exact functional test for directional association and cancer biology applications. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):818–826, 2018.

[19] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell rna-sequencing experiments. *Nature methods*, 14(4):381–387, 2017.

[20] Wah Chin Boon, Karolina Petkovic-Duran, Yonggang Zhu, Richard Manasseh, Malcolm K Horne, and Tim D Aumann. Increasing cdna yields from single-cell quantities of mrna in standard laboratory reverse transcriptase reactions using acoustic microstreaming. *JoVE (Journal of Visualized Experiments)*, (53):e3144, 2011.

[21] Daniel Hebenstreit. Methods, challenges and potentials of single cell rna-seq. *Biology*, 1(3):658–667, 2012.

[22] Nathan Archer, Mark D Walsh, Vahid Shahrezaei, and Daniel Hebenstreit. Modeling enzyme processivity reveals that rna-seq libraries are biased in characteristic and correctable ways. *Cell systems*, 3(5):467–479, 2016.

[23] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[24] Karlynn E Neu, Qingming Tang, Patrick C Wilson, and Aly A Khan. Single-cell genomics: approaches and utility in immunology. *Trends in immunology*, 38(2):140–149, 2017.

[25] Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Dudoit, and Lieven Clement. Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome biology*, 19 (1):1–17, 2018.

[26] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740–742, 2014.

[27] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):1–13, 2015.

[28] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

[29] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190): 1396–1401, 2014.

[30] Hailun Wang, Pak Sham, Tiejun Tong, and Herbert Pang. Pathway-based single-cell rna-seq classification, clustering, and construction of gene-gene interactions networks using random forests. *IEEE Journal of Biomedical and Health Informatics*, 24(6):1814–1822, 2019.

[31] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17 (1):63, 2016.

[32] Nicholas Lytal, Di Ran, and Lingling An. Normalization methods on single-cell rna-seq data: An empirical survey. *Frontiers in Genetics*, 11:41, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00041. URL https://www.frontiersin.org/article/10.3389/fgene.2020.00041.

[33] Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. Assessment of batch-correction methods for scrna-seq data with a new test metric. *BioRxiv*, page 200345, 2017.

[34] Antonio Scialdone, Kedar N Natarajan, Luis R Saraiva, Valentina Proserpio, Sarah A Teichmann, Oliver Stegle, John C Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[35] Eliška Dvořáková, Sajal Kumar, Jiří Kléma, Filip Železný, Karel Drbal, and Mingzhou Song. Evaluating model-free directional dependency methods on single-cell rna sequencing data with severe dropout. In *Proceedings of the 2019 6th International Conference on Bioinformatics Research and Applications*, ICBRA '19, page 55–62, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450372183. doi: 10.1145/3383783.3383793. URL https://doi.org/10.1145/3383783.3383793.

[36] Marry M van den Heuvel-Eibrink, Bronno van der Holt, Alan K Burnett, Wolfgang U Knauf, Martin F Fey, Gregor EG Verhoef, Edo Vellenga, Gert J Ossenkoppele, Bob Löwenberg, and Pieter Sonneveld. Cd34-related coexpression of mdr1 and bcrp indicates a clinically resistant phenotype in patients with acute myeloid leukemia (aml) of older age. *Annals of hematology*, 86(5):329–337, 2007.

[37] Vladimir Kiselev, Tallulah Andrews, Jennifer Westoby, Davis McCarthy, Maren Buttner, and Martin Hemberg. Analysis of single cell rna-seq data., 2 2018. Section 7.7.

[38] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47 (260):583–621, 1952.

[39] Kailash Budhathoki and Jilles Vreeken. Mdl for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 751–756. IEEE, 2017.

[40] Xin Huang, Suxia Geng, Jianyu Weng, Zesheng Lu, Lingji Zeng, Minming Li, Chengxin Deng, Xiuli Wu, Yangqiu Li, and Xin Du. Analysis of the expression of phtf1 and related genes in acute lymphoblastic leukemia. *Cancer cell international*, 15(1):93, 2015.

[41] Robert A Carter, Laure Bihannic, Celeste Rosencrance, Jennifer L Hadley, Yiai Tong, Timothy N Phoenix, Sivaraman Natarajan, John Easton, Paul A Northcott, and Charles Gawad. A single-cell transcriptional atlas of the developing murine cerebellum. *Current Biology*, 28 (18):2910–2920, 2018.