UNIVERSITEIT GENT

FACULTY OF SCIENCES

Quantitative evaluation of network inference methods for single-cell cancer regulomes

Charlotte DE VOGELAERE

Master's dissertation submitted to obtain the degree of

Master of Science in Biochemistry and Biotechnology

Major Bioinformatics and Systems Biology

Academic year 2015-2016

Promoter: Prof. Dr. Ir. Katleen De Preter

Scientific supervisor: Ir. Robrecht Cannoodt

UGent

CENTRUM MEDISCHE GENETICA GENT

UNIVERSITEIT GENT

FACULTY OF SCIENCES

Quantitative evaluation of network inference methods for single-cell cancer regulomes

Charlotte DE VOGELAERE

Master's dissertation submitted to obtain the degree of

Master of Science in Biochemistry and Biotechnology

Major Bioinformatics and Systems Biology

Academic year 2015-2016

Promoter: Prof. Dr. Ir. Katleen De Preter

Scientific supervisor: Ir. Robrecht Cannoodt

UGent

CENTRUM MEDISCHE GENETICA GENT

Acknowledgements

In the first place, I would like to thank my promoter Prof. Dr. Ir. Katleen De Preter for allowing me to conduct research in the translational bioinformatics group at the CMGG and for guiding me in the right direction during our weekly meetings. Working with single-cell sequencing data was challenging, but I was very grateful for the opportunity to explore this state-of-the-art data type. I would also like to thank Ir. Robrecht Cannoodt, my supervisor, for his feedback and all his help, especially with the R programming language.

Further, I am grateful to my fellow *Bioinformatics and Systems Biology* students. The great atmosphere during out lectures and lunchbreaks really made the last two years the best years of my education at UGent.

Finally, I would also like to thank my parents for helping me realize that Medicine was not the most suitable choice for me, and for allowing me to switch to Biochemistry and Biotechnology, my brothers for being great roommates and Stefan, for always helping me look on the bright side of things.

Charlotte De Vogelaere, June 2016

Table of contents

Acknowledgements i

Table of contents iii

List of Abbreviations vi

Nederlandse samenvatting vii

English summary viii

Part 1: Introduction 9

1.1. **Oncogenesis** 9

1.1.1. How do oncogenic cells differ from healthy cells 9

1.1.2. Glioblastoma multiforme 10

1.1.3. Melanoma 12

1.2. Single-cell RNA-seq 13

1.2.1. Next generation sequencing for transcriptomics 13

1.2.2. Applications and advantages 15

1.2.3. Challenges 16

1.3. Network inference 17

1.3.1. Absolute value of Pearson's correlation 18

1.3.2. Algorithm for the Reconstruction of Gene Regulatory Networks 18

1.3.3. Context likelihood relatedness 18

1.3.4. Gene Network inference with Ensemble of Trees 19

1.3.5. Biological networks have scale-free properties 19

Part 2: Aims 21

2.1. Setting of the problem 21

2.2. Evaluation of network inference methods 21

Part 3: Results 23

3.1. Single cell RNA-seq datasets 23

3.1.1. Quality control 23

3.1.2. Mapping and normalization 24

3.1.3. Melanoma 25

3.2. Population vs. pooled vs. single cell 26

3.2.1. Pooled single-cell expression levels 26
3.2.2. Correlation 26
3.2.3. Dimensionality reduction 28
3.2.4. Conclusion 31
3.3. Network inference glioblastoma 32
3.3.1. Filtering of the expression matrix 32
3.3.2. Gold standard 32
3.3.3. Nonmalignant cells 33
3.3.4. Performance 33
3.3.5. Conclusion 36
3.4. Network inference melanoma 37
3.4.1. Filtering of the expression matrix 37
3.4.2. Gold standard 37
3.4.3. Nonmalignant cells 38
3.4.4. Performance 38
3.4.5. Hubs 40
3.4.6. Conclusion 44

Part 4: Discussion 45

4.1. Conclusion 45
4.2. Limitations 46
4.3. Future aspects of network inference in single-cells 47

Part 5: Samenvatting discussie 49

5.1. Conclusie 49
5.2. Limitaties 50
5.3. Toekomstige ontwikkelingen 51

Part 6: Materials and Methods 53

6.1. Glioblastoma data 53
6.1.1. Download data 53
6.1.2. Quality control 53
6.1.3. Mapping and normalization 53
6.1.4. Filter expression matrix 54
6.2. Melanoma data 54

6.2.1. Download data **54**
6.2.2. Filter expression matrix **54**
6.3. Control data **54**
6.3.1. Glioblastoma **54**
6.3.2. Melanoma **55**
6.4. Network inference **55**
6.4.1. Regulators **55**
6.4.2. Gold standard **56**
6.4.3. ARACNE and CLR **56**
6.4.4. GENIE3 ..... **57**
6.4.5. Absolute value of Pearson's correlation **57**
6.4.6. Hubs **57**
6.5. Pooled single-cell vs. population RNA-seq **58**
6.5.1. Pooling **58**
6.5.2. Correlation **58**
6.5.3. Comparison of the sequencing depth **58**
6.5.4. Dimensionality reduction **58**
References **61**
Attachments **65**
I. Supplementary tables **65**
II. Code **65**

List of Abbreviations

ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUC	Area under the curve
CLR	Context likelihood relatedness
ECM	Extracellular matrix
ERCC	External RNA controls consortium
GENIE3	Gene Network Inference with Ensemble of Trees
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRN	Gene regulatory network
GST	Gene set test
MDS	Multidimensional scaling
MGDB	Melanoma gene database
NI	Network inference
NIH	National Institute of Health
PINA	Protein Interaction Network Analysis
PR	precision-recall
QC	Quality control
ROC	Receiver Operating Characteristic
scLVM	single-cell Latent Variable Model
scRNA-seq	single-cell RNA-sequencing
SRA	Sequence Read Archive
TCGA	The Cancer Genome Atlas
TPM	Transcripts Per Million

Nederlandse samenvatting

Recente ontwikkelingen op het vlak van transcriptomics hebben de ontwikkeling van een *high-throughput* sequeneringsmethode die transcriptomen kan analyseren op *single-cell* niveau mogelijk gemaakt: *single-cell RNA-seq*. De eerste resultaten die gegenereerd werden aan de hand van deze methode, vooral in de ontwikkelingsbiologie en in immulogisch onderzoek, zijn veelbelovend. Daarbovenop is deze methode naar voor aan het treden in kankeronderzoek, waarbij vooral de nadruk gelegd wordt op het blootleggen van tumor heterogeniteit. Desondanks hebben onderzoekers nog altijd te kampen met enkele uitdagingen geassocieerd met dit nieuw datatype, op vlak van computationele methodes. Eén van de manieren om nieuwe biologische inzichten te verkrijgen op basis van expressiedata, is de inferentie van gen regulatorische netwerken. Momenteel bestaat er geen consensus over welke netwerk inferentie methode best gebruikt kan worden in *single-cell* data. De hoofddoelstelling van deze thesis is het evalueren van verschillende netwerk inferentie methodes aan de hand van een gouden standaard van kanker interacties in twee veelbestudeerde kankertypes: *melanoma* en *glioblastoma*. Ten eerste moest een gouden standaard van interacties in deze kankertypes opgesteld worden. Hiervoor werd een literatuurstudie uitgevoerd. Ten tweede werden de methodes geëvalueerd voor hun vermogen om deze interacties terug te vinden. Hierbij werd ook rekening gehouden met interacties die voorkomen in niet-kwaadaardige cellen. Tenslotte werd, voor de *melanoma* dataset, een analyse van de regulatoren uitgevoerd, waarbij regulatoren met een hoog aantal connecties geïdentificeerd werden als *hubs*. Deze regulatoren met veel connecties spelen mogelijks een belangrijke biologische rol.

Op basis van deze analyse kon geen enkele methode naar voor geschoven worden als consensusmethode voor netwerk inferentie in *single-cell RNA-seq*. Het evalueren van een *community-based* benadering, waarbij de resultaten van verschillende netwerk inferentie methodes gecombineerd worden, lijkt echter interessant. Het gebruik van een dergelijke benadering zou potentieel interessante kandidaat genen naar voor kunnen schuiven, die dan verder geanalyseerd kunnen worden *in vitro* en *in vivo*.

English summary

Recent developments in the field of transcriptomics have enabled the development of a high-throughput sequencing method that can analyze transcriptomes on a single cell level: single-cell RNA-seq. The first applications of this new technique have shown great promise in developmental and immunological research. Additionally, this method is also emerging in cancer research, focusing on uncovering tumor heterogeneity. However, researchers are still facing numerous computational challenges associated with this new datatype. One of the ways to obtain new biological insights based on expression data is the inference of gene regulatory networks. Currently, there is no consensus on which method to use for network inference in single-cell data. The main aim of this thesis is to evaluate network inference methods on their ability to find a set of gold standard interactions in two cancer types that have been extensively researched in the past: glioblastoma and melanoma. First, the gold standard needed to be constructed; this was done based on interactions described in literature. Secondly, the methods were scored using this gold standard, while taking into account interactions that occur in nonmalignant cells of the same type. Finally, for the melanoma dataset, the highly connected regulators were analyzed, as these were possible hubs in the network, meaning they could be of biological importance.

No single method showed a superior performance based on this analysis. However, it might be worth investigating a community-based approach. Combining the results from different network inference methods for the analysis of highly connected regulators might identify interesting candidate genes for further investigation *in vitro* and *in vivo*.

Part 1: Introduction

1.1. Oncogenesis

1.1.1. How do oncogenic cells differ from healthy cells

Cancer is a disease in which mechanisms involved in cellular growth and proliferation fail, causing cells to divide uncontrollably. It is suggested that the same fundamental capacities are acquired by most cancers, including resistance of cell death, replicative immortality and evasion of growth suppressors (Figure 1) (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011). This is facilitated by two enabling characteristics: genome instability and tumor-promoting inflammation (Hanahan & Weinberg 2011).

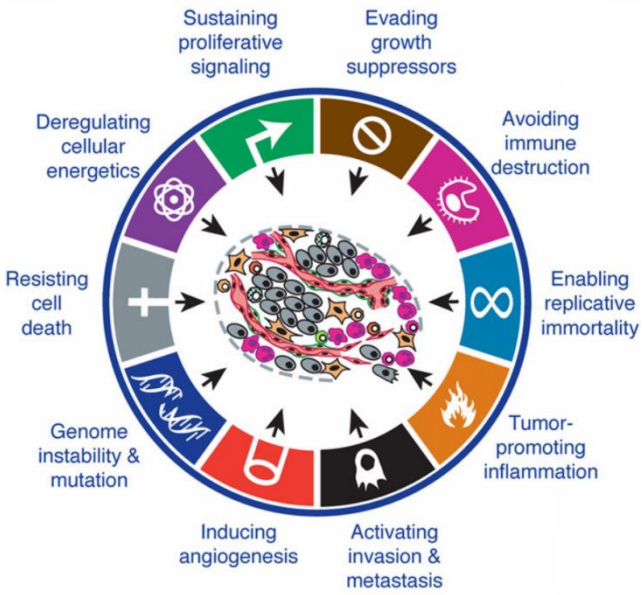

Figure 1: The hallmarks of cancer – Several features need to be acquired by cancer cells in order for them to evolve to a neoplastic state. Figure adapted from Hanahan & Weinberg 2011.

Microscopically, a lack of contact inhibition can be observed as well as the specific characteristics of rapidly growing cells, i.e. a high nucleus-to-cytoplasm ratio, prominent nucleoli, an increased frequency of mitotic cells, and relatively little specialized structure. Nonmalignant cells are localized to their original tissue by cell-cell adhesion signals and physical barriers, e.g. the basement membrane. This localization is also seen in benign tumors, while cells in malignant tumors, in contrast, have the ability to invade in the surrounding tissues, migrating to new sites and forming secondary tumors or metastases (Lodish et al. 2013).

Currently, tumors are regarded as complex tissues, focusing not only on the cancer cells themselves, but also on the interactions with their neighboring nonmalignant cells, including immune cells, blood vessels, the extracellular matrix (ECM), etc. (Figure 2) (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011). For instance, it is suggested that reciprocal signaling between stromal cells and cancer cells further

enhances cancer cell evolution, on the one hand, and, on the other hand, reprograms stromal cells to assist the cancer cells in various aspects of tumor formation (Hanahan & Weinberg 2011). Furthermore, a pan-cancer analysis among 12 cancer types reports that tumors often consist of multiple clones, each with a distinct genetic background. This intratumoral heterogeneity can be seen in all investigated cancer types. The amount of clones is reported to be predictive for the therapeutic outcome and to have implications for the development of therapeutic resistance (Andor et al. 2016).

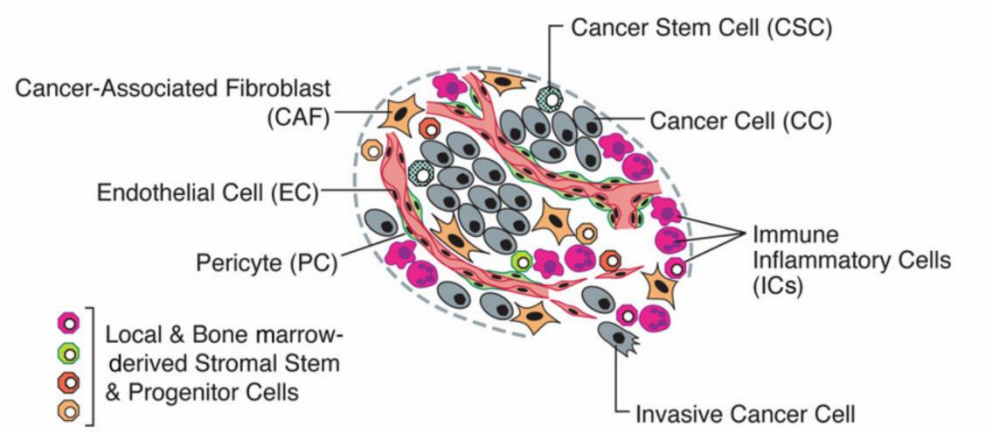

Figure 2: Cells of the tumor microenvironment – A tumor does not only consist of cancer cells, but contains a diverse cell types which all contribute to tumor biology, collectively enabling tumor growth. Figure taken from Hanahan & Weinberg 2011

In 2005, The Cancer Genome Atlas (TCGA) was launched by the National Institute of Health (NIH) in an attempt to create a better understanding of genomic alterations involved in different types of cancer in order to improve cancer prevention, early detection and treatment. To this end, high-throughput technologies, based on microarrays and next-generation sequencing methods, were applied to over 30 cancer types, studying single tumors as well as comparing diverse tumor types in pan-cancer analyses. The generated data is publicly available, enabling analysis by other researchers (Tomczak et al. 2015).

1.1.2. Glioblastoma multiforme

Glioblastoma multiforme or glioblastoma is the most common brain tumor in adults (Bleeker et al. 2012) and can arise *de novo* as a primary glioblastoma or secondary to a low-grade glioma. The former generally presents in older patients and is characterized by low survival rates and poor responses to therapy. Glioblastoma was the first cancer to be studied by TCGA (McLendon et al. 2008), and the initial findings were further researched in 2013 (Brennan et al. 2013), also including data generated by next-generation sequencing technology.

The TCGA pilot project (McLendon et al. 2008) mapped somatic mutations in 91 tumor samples to major pathways known to be involved in glioblastoma. This project

uncovered that most tumors showed biologically relevant deregulations in three core pathways: p53, retinoblastoma (RB) and receptor tyrosine kinase (RTK)/RAS/phosphoinositide-3 kinase (PI3K) (Figure 3). Based on this data, a molecular signature could be identified for two previously known (Phillips et al. 2006) and two newly identified glioblastoma subtypes (Verhaak et al. 2010). A list of 210 signature genes is available for each of these subtypes. The classical subtype is associated with high levels of epithelial growth factor receptor (EGFR) and also shows high levels of EGFR alteration combined with low levels of TP53 mutations. Expression of neuron markers is predominantly seen in the neuronal subtype, while the mesenchymal subtype shows expression of mesenchymal markers. The latter is also associated with lower expression levels of neurofibromin 1 (NF1) as a consequence of a deletion of the chromosomal region containing this gene. The proneural subtype, finally, is associated with alterations in platelet-derived growth factor receptor alpha (PDGFRA) and point mutations in isocitrate dehydrogenase 1 (IDH1). This subtype was generally associated with younger patients, longer survival and other characteristics associated with secondary glioblastoma. Data comparing treatment protocols suggests that these patients do not benefit from a more aggressive protocol. The expression patterns of the different subtypes suggest the presence of multiple stem cell-like populations, but this hypothesis requires further research. The classification of glioblastoma tumors has important consequences for the therapeutic strategies.

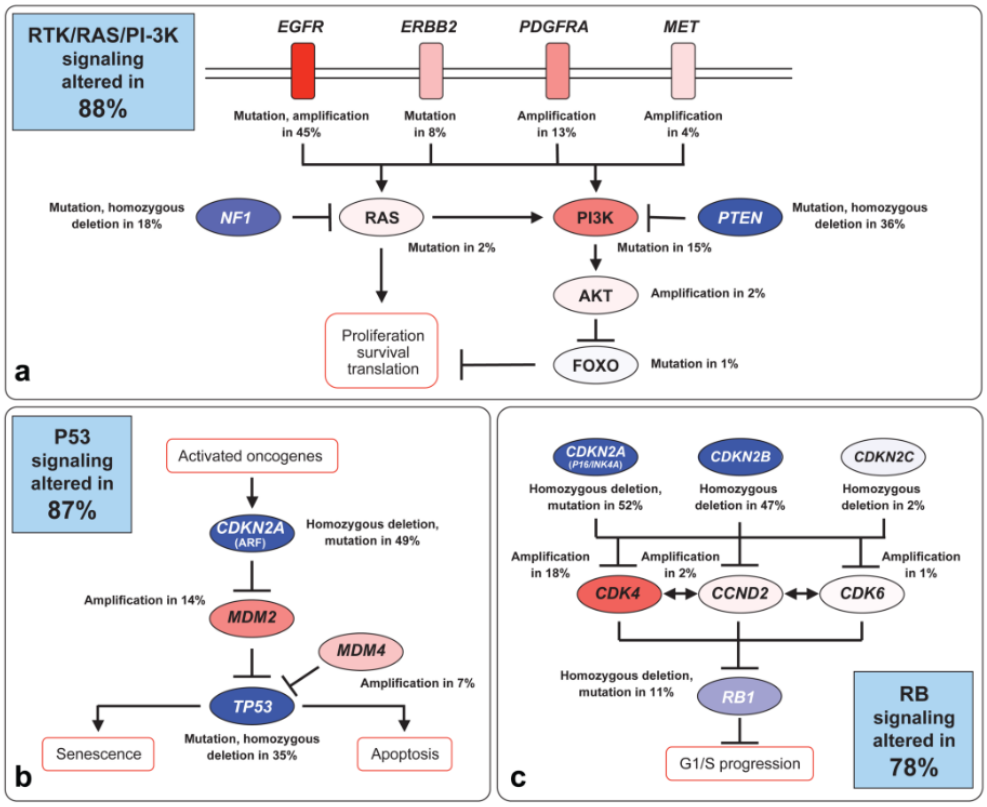

Figure 3: Three core pathways involved in glioblastoma pathogenesis - The TCGA pilot project showed that most of the glioblastoma samples showed alterations in RB, p53 and RTK/RAS/PI-3K signalling. Figure taken from McLendon et al. 2008.

Further, high frequency loss of tumor suppressor mitogen inducible gene 6 (Mig-6) is reported to be involved in gliomagenesis. Mig-6 is a regulator of EGFR transport to the late endosomes and lysosomes, thus supporting its degradation (Ying et al. 2010). Bcl2-Like-12 (Bcl2L12), a cell death regulator which represses p53 transactivation, was identified to play a role in therapeutic resistance of glioblastoma (Stegh et al. 2010). It is shown to be upregulated in tumors with wild type p53 and is less overexpressed in tumors with compromised p53.

More recently, single-cell RNA-seq (scRNA-seq) was used to uncover heterogeneity in glioblastoma, profiling cells from five patients (Patel et al. 2014). The dataset contains 672 single glioblastoma cells, five population controls, 192 single cells from two gliomasphere cell lines and six population samples from cell lines derived from three tumors and cultured under serum free and differentiated conditions.

1.1.3. Melanoma

Malignant melanoma or melanoma is a type of skin cancer that originates in the melanocytes, the cells producing the pigment melanin in the skin, eyes and hair. It shows a high incidence in Caucasians compared to other races. Despite having a low incidence compared to other types of skin cancer, the majority of the deaths associated with skin cancer are caused by melanoma. However, when diagnosed in an early stage, it is mostly curable (Cummins et al. 2006).

TCGA integrated data from multiple platforms from a cohort of 333 samples from 331 patients, resulting in the identification of four genomic melanoma subtypes which could have an impact on therapeutic decisions. The BRAF subtype is associated with BRAF hotspot mutations and is the most prevalent. It shows amplification of BRAF and MITF. The RAS subtype shows mutations in all three RAS family members (H, K and N) and is associated with increased MAP kinase (MAPK) activation and AKT3 expression. In the NF1 subtype, loss of function of NF1 is observed inhibiting downregulation of RAS activity, activating the MAP kinase pathway. In contrast, the triple wild type subtype shows no hot-spot mutations in BRAF, RAS and NF1. It does, however, show a significant amplification of KIT, MDM2, CDK4, CCND1 and TERT compared to the other subtypes (Watson et al. 2015).

To date, one dataset containing single melanoma cell RNA-seq data has been reported (Tirosh et al. 2016). This dataset contains 4645 cells, malignant as well as nonmalignant, from 19 melanoma patients with diverse pathogenic and therapeutic backgrounds. Each cell was marked as malignant, nonmalignant or unresolved, based on copy number variations (CNVs). The cells showing aneuploidy were classified as malignant. The nonmalignant cells were further categorized as T cells, B cells, macrophages, endothelial cells, cancer-associated fibroblasts (CAFs), and natural killer (NK) cells, based on the expression of marker genes.

1.2. Single-cell RNA-seq

1.2.1. Next generation sequencing for transcriptomics

In order to study the transcriptome - the assembly of all transcripts in the cell - in a high throughput fashion, several approaches have been developed. These approaches are aimed at quantifying expression levels under different conditions, assessing the transcriptional structure of genes and determining all classes of transcripts (Wang et al. 2009). In hybridization-based methods such as microarrays, the principle of hybridization of complementary sequences is used. The messenger RNA (mRNA) from a sample of interest is reverse transcribed to copy DNA (cDNA) and fluorescently labeled. An array of probes is constructed based on sequence knowledge of the organism under study and hybridization of the cDNA with the probes is quantified by exciting and measuring the fluorescent signal (Stears et al. 2003). Sequence-based methods, on the other hand, allow direct determination of the cDNA sequence, which creates several advantages over hybridization based methods. Mainly, it is no longer required to know the sequence of the organism in advance, enabling detection of unknown transcripts and splice variants, and transcriptome analysis of non-model organisms. Secondly, sequencing-based methods have very low background signal compared to the signal generated by cross-hybridization in hybridization-based methods. Further, there is no upper limit for quantification, meaning that the dynamic detection range is large. Finally, expression levels are much easier to compare between different experiments if normalization techniques are used (Wang et al. 2009). The development of high-throughput sequencing technologies and their subsequent application in transcriptomics, named RNA-seq, led to a revolution in this field in terms of mapping and quantifying transcripts.

Generally, RNA-seq protocols begin by isolating a fraction of RNA from the input cells of interest and reverse-transcribing this to a library of cDNA-fragments, which are amplified with PCR. The amplified fragments are then sequenced from one end (single-end sequencing) or both ends (paired-end sequencing), resulting in short reads of 30 - 400 base pairs (Wang et al. 2009; Stegle et al. 2015).

All raw reads generated by high-throughput sequencing methods should undergo quality checks before further analysis. One of the tools that can be used to this extent is FastQC (Andrews, 2010), which focusses on identifying problems with the sequencer or the library starting material. It evaluates the following criteria:

- Per base sequence quality: shows the range of quality for each position. Long reads might show inferior quality towards the end;
- Per sequence quality score: evaluates the general quality of the sequences within a run. If a substantial part of the sequences in a run show low scores, this might indicate a systematic problem;
- Per base sequence content: shows the proportion of each base at each position. The difference between the proportion of A and T or G and C is

expected to be under 10%, though a biased sequence composition might occur at the start of the read due to the method used for library production;

- Per sequence GC content: the GC distribution for all sequences in one file is plotted and compared to a normal distribution;
- Sequence length distribution: will raise a warning if all sequences are not the same length;
- Duplicate sequences: raises a warning if the amount of non-unique sequences exceeds a certain threshold. This module does, however, not differentiate between biological and technical duplications;
- Overrepresented sequences: the library is expected to be diverse; this module gives an error if one sequence occurs more than 1%. This could indicate a contamination;
- Kmer content: looks for relative enrichment of 7-mers. If the library is constructed using random primers, this will show a bias as the start of the library;
- Adapter content: will look for overrepresentation of adapters, indicating trimming is needed;
- Per tile sequence quality (for illumina data): assesses whether a certain area of the flow cell shows a loss in quality, which could indicate a technical bias related to the flow cell.

If the quality is deemed sufficient, the raw reads are either aligned to a reference genome or transcriptome, a process called mapping, or assembled *de novo*. Next, the number of mapped reads is quantified for each locus, generating counts. To remove technical biases caused by difference in length between transcripts and allow comparison of the counts across different samples, the counts should be normalized. Transcripts per million (TPM) is proposed as the measure of choice for transcript abundance (Wagner et al. 2012). Based on the normalized counts, an expression profile can be generated for each gene; the resulting gene expression value is an average of the levels across the population of input cells. Often, this is sufficient, but to solve certain biological problems, essential information is possibly masked when analyzing the average expression levels of a population of cells. In those cases, measurement of gene expression at single-cell resolution is required in order to correctly reflect tissue heterogeneity.

Until recently, research on single cells was done through low-throughput techniques such as single-cell qPCR or single-molecule RNA fluorescent in situ hybridization (RNA FISH). However, recent innovations such as automation of the methods to generate cDNA libraries from single cells (Stegle et al. 2015) allow RNA-seq transcriptome analysis at the single cell level in a high-throughput fashion using a protocol very similar to that of population analysis. In the first step, single cells are isolated, for instance by microfluidics. Next, the RNA molecules are processed in a similar way as the RNA fraction from a population of cells in population RNA-seq. The

main difference is the small amount of starting RNA, which necessitates an amplification step.

1.2.2. Applications and advantages

Single-cell RNA-seq (scRNA-seq) is most often applied in developmental research, where there are very few cells - each with a very distinct profile - or in cases where the tissue under study is expected to be very heterogeneous.

For example, differentiation of distal lung epithelium in mammalian lung tissue was studied by sequencing the RNA of single cells taken at four different time points, which lead to the discovery of formerly unknown cell-type markers (Treutlein et al. 2014). In early blood cell development, gene expression measured in single cells at four time points predicted the involvement of certain transcription factors. The authors were able to validate these predictions experimentally, suggesting that similar approaches could unravel regulatory networks in other developing organs (Moignard et al. 2015). The commitment of conventional dendritic cells (cDCs) to cDC1 or cDC2 is not yet fully understood, but single-cell mRNA sequencing analysis of DC development indicates that this occurs before the cells leave the bone marrow (Schlitzer et al. 2015). Clustering of the single-cells enabled reordering of the cells along the developmental continuum, showing the progression of gene expression.

Further, single-cell RNA-seq was used to study several populations of cells involved in the immune system. Sequencing of single dendritic cells at different time points, stimulated by three different pathogenic components, showed extensive heterogeneity between identically stimulated cells. This variability, suggested to play an important role in immune response plasticity, was less extensive when the cells were stimulated individually in their wells, suggesting the importance of paracrine signaling (Shalek et al. 2014). Th17 cells play a role in the adaptive immune system, but are also reported to be involved in pathogenesis of autoimmunity. Computational analysis of single-cell Th17 expression data and subsequent functional validation of candidate genes led to the detection of potential candidate genes involved in Th17 cell pathogenicity. These genes can be validated as therapeutic targets in autoimmune diseases caused by pathogenic Th17, meanwhile avoiding destruction of non-pathogenic, essential Th17 cells (Gaublomme et al. 2015).

Sequencing of single cells in cancer research uncovered a certain heterogeneity which could not be assessed previously using population based methods. This heterogeneity could have an impact on cancer diagnosis and treatment. Single cell expression analysis of 430 samples from five primary glioblastomas revealed intratumoral heterogeneity in relevant pathways. Based on population RNA-seq analysis, four of the tumors could be classified as one of the subtypes previously determined by TCGA. However, single-cell analysis showed that the tumors consisted of cells belonging to different subtypes, although the dominant subtype did correspond to the subtype determined by population RNA-seq analysis. In proneural tumors, the authors found that increased intratumoral heterogeneity is associated

with the clinical outcome; patients with high-heterogeneity tumors showed decreased survival. This finding also has an impact on targeted therapy, as the targeted molecules show high variability between cells (Patel et al. 2014). Experiments with single patient-derived xenograft (PDX) cells in lung adenocarcinoma (LUAD) confirm that single-cell RNA-seq experiments can identify possible factors involved in therapy resistance and metastases (Kim et al. 2015). Comparison of gene expression levels between population samples and pooled single-cell samples showed significant correlation, despite reported amplification bias in single-cell RNA-seq (Kim et al. 2015). The most recent large-scale single-cell RNA-seq study in cancer sequenced 4645 single cells from 19 metastatic melanoma patients, examining both malignant and non-malignant cells (stromal and immune cells) to explore the tumor microenvironment (Tirosh et al. 2016).

1.2.3. Challenges

Although the technique of single-cell RNA-seq shows great promise in unraveling the underlying biological mechanisms in oncogenesis, therapy resistance and metastatic potential of tumors, and the prediction of biomarkers and targets for targeted therapy, several challenges are still to be overcome to enable this technology to be used to its full potential (Prado et al. 2015).

The existing computational and statistical methods for transcript quantification and quality control, developed for population RNA-seq analysis, require improvements in three domains (Stegle et al. 2015). In the first place, a normalization method should be developed that takes into account differences in mRNA content between cells, as normalization methods applied to population RNA-seq data assume that the amount of RNA in each sample is the same. When using extrinsic spike-ins, e.g. the external RNA controls consortium (ERCC) RNA spike-in mix, a known amount of control RNA molecules is added to each well, allowing comparison of cells based on the amount of spiked-in material. Methods using unique molecular identifiers (UMIs), in contrast, label individual cDNA molecules during the reverse transcriptase phase (prior to amplification) with short random sequences. Counts are generated by quantifying the UMIs aligned to each position rather than the transcripts, avoiding amplification bias (Islam et al. 2014). It is also possible to normalize data in absence of UMIs and spike-ins by using population-based normalization methods such as TPM. Secondly, methods will need to deal with confounding factors, which can be of technical nature, e.g. batch effects and allelic dropout, or of biological nature, e.g. periodic processes in the cell such as the cell cycle. When analyzing population data, most biological confounding effects are cancelled out due to the fact that the average expression levels of each gene are calculated for a high amount of cells. Single-cell latent variable model (scLVM) is proposed as a method to remove variation caused by cell cycle or other confounders (Buettner et al. 2015). To our knowledge, no other methods have been developed. Thirdly, a high level of technical noise needs to be distinguished from genuine biological variability. A method based on external RNA

spike-ins, decomposing the variance in a technical and biological component, is suggested by Kim et al. (K.-T.T. Kim et al. 2015).

Other authors also pose that, in the area of experimental design, the number of cells that need to be sequenced in order to uncover covariant genes and the depth to which each of these cells should be sequenced, remain to be determined (Shapiro et al. 2013; Stegle et al. 2015).

Like in population RNA-seq, the gene expression levels can be used for gene regulatory network (GRN) inference. It is expected that the existing methods will be able to be used, once adaptations have been made to deal with the additional levels of technical noise and confounding factors (Stegle et al. 2015).

1.3. Network inference

The expression of all genes in a cell is strictly regulated by a limited number of regulators or transcription factors and the measured expression levels can be seen as the output of a gene regulatory network. By characterizing the interactions between regulators and genes through co-expression analysis, a model of the network of regulatory interactions can be reverse-engineered or inferred (D'haeseleer et al. 2000).

Robust reconstruction of GRNs from high-throughput gene expression data is an established problem in computational biology. A wide range of methods have been developed that, when applied to the same data, often result in divergent networks. During a transcriptional network inference challenge, the DREAM project (Marbach et al. 2012) assessed several of these methods, using microarray gene expression data. The aim of this challenge was to rank the methods based on their performance, evaluating their reliability and stimulating the development of more accurate methods. The performance was evaluated using a gold standard and classified the methods accordingly within four categories - regression, mutual information, Bayesian networks and correlation - based on the computational approach. Additionally, methods that did not belong to one of these categories were classified as 'other' and the category 'meta' contained methods that combined several of these approaches. The results indicate complementarity of the different methods, and the authors suggest that combining as little as three different methods should improve the performance drastically, since the advantages will amplify each other and the limitations will be cancelled out. None of the six categories was superior to the others. To our knowledge, no analysis of this extent has been carried out for RNA-seq or single-cell RNA-seq expression data.

Four of these methods will be discussed in more detail, selected based on their performance in the DREAM challenge and their popularity, including methods from various categories.

1.3.1. Absolute value of Pearson's correlation

The highest scoring method of the correlation-based methods was the absolute value of Pearson's correlation (Equation 1) (Marbach et al. 2012). The absolute value of the correlation coefficient between all transcription factors and genes is calculated.

Equation 1: Absolute value of Pearson's correlation coefficient – Pearson's correlation coefficient, r, for all transcription factors, x, and all target genes, y, with n the number of measurements of x and y. Equation from Marbach et al. 2012, supplemental data

$$r_{xy} = abs\left(\frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}\right)$$

The correlation coefficient is a value between -1 and 1, with 1 meaning the genes are perfectly correlated and -1 meaning they are perfectly anti-correlated. By using the absolute value, negative or inhibitory relations are also taken into account.

This method has as an advantage that it is very fast, but it does not take into account the difference between direct regulation of a gene by a transcription factor binding to its promotor, and indirect interactions through gene regulatory cascades.

1.3.2. Algorithm for the Reconstruction of Gene Regulatory Networks

Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE) reconstructs biological networks based on mutual information. Even though this method is widely used, it was outperformed by several other methods (Marbach et al. 2012).

Mutual information based methods consist of two steps. First, a mutual information matrix is calculated for all regulators and genes. Secondly, the edges are filtered. In ARACNE, all edges that can be explained by other interactions in the network are removed, suggesting that the remaining interactions are the direct interactions (Margolin et al. 2006). These remaining interactions are ranked according to their mutual information values.

1.3.3. Context likelihood relatedness

Context likelihood relatedness (CLR) was the highest ranking mutual information-based method (Marbach et al. 2012) CLR filters the edges by comparing the mutual information value to a set threshold, i.e. a background distribution of mutual information scores. This is expected to remove indirect interactions by removing edges where one transcription factor interacts weakly with a large number of genes, or one gene interacts weakly with many transcription factors (Faith et al. 2007).

1.3.4. Gene Network inference with Ensemble of Trees

Gene Network Inference with Ensemble of Trees (GENIE3) was the highest scoring method from the category 'other'. This method decomposes the problem into *p* regression problems, with *p* being the amount of genes in the network. The expression of each gene is assumed to be the function of the expression all other genes in the network, plus random noise. The importance of each transcription factor for the expression of each gene is predicted using a tree-based ensemble method, Random Forests. The ranking of all the edges is constructed from all possible regulatory links over all genes.

1.3.5. Biological networks have scale-free properties

Biological networks can be represented as a graph of directed, weighted interactions. All of the NI methods discussed above return a ranked list of weighed interactions which can be organized into network motifs, statistically significant subgraphs. These motifs, in turn, cluster into semi-independent modules. Finally, the ensemble of all modules will form the gene regulatory network (Figure 4) (Babu et al. 2004).

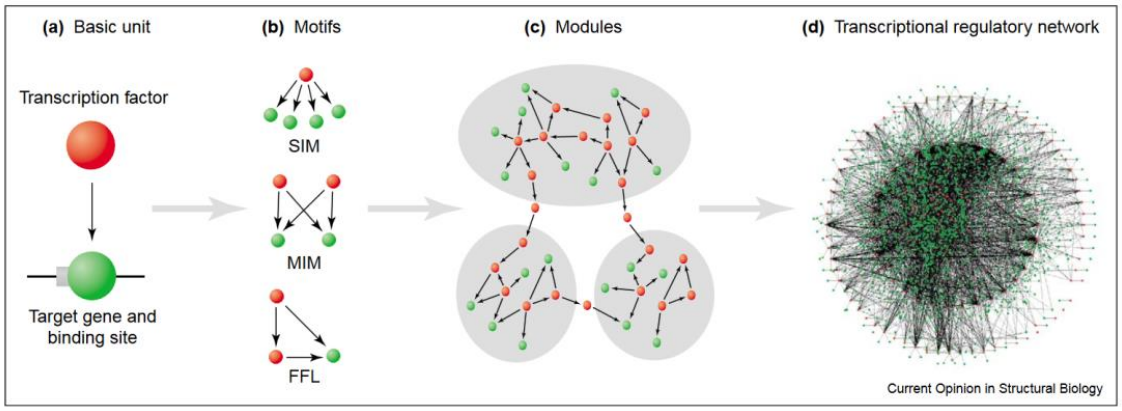

Figure 4: Structure of transcriptional regulatory networks at different levels – (a) the basic unit is the weighted, directed interaction from a transcription factor to a target gene, (b) interactions can be organized into motifs, (c) motifs cluster into modules and (d) the assembly of all modules forms the transcriptional regulatory network.
Figure taken from Babu et al. 2004

GRN are reported to approximate a scale-free topology (Barabási & Oltvai 2004; Barabási 2009), meaning that the degree distribution follows a power law. Certain nodes, called hubs, are highly connected and hold the network together. In biology, these hubs are more likely to be essential.

Part 2: Aims

2.1. Setting of the problem

Tumors are known to be very heterogeneous and this is thought to be one of the main causes of metastasis and therapy resistance. In order to fully understand the pathways altered in certain subclones, making it possible for them to withstand applied therapy, gene expression analysis on a population level is insufficient. Recent developments enable measuring gene expression at the single-cell level in a high-throughput way using NGS: scRNA-seq. Network inference from this data shows to be promising for the analysis of heterogeneous samples, such as tumor tissue, making it possible to identify subclones that were previously undetected.

Single-cell RNA-seq is an emerging technique and there are a large number of challenges that will need to be dealt with in data generation, pre-processing and analysis (Stegle et al. 2015). Currently, there is no consensus method for network inference. Additionally, no evaluation methods for single-cell networks exist, making it hard to differentiate between true biological variation and technical noise (J. Kim et al. 2015). There is no golden standard available to evaluate the networks, nor simulated data resembling single-cell data to assess their robustness.

2.2. Evaluation of network inference methods

The research in the translational bio-informatics lab at the Centre for Medical Genetics (CMGG) focusses on the application of techniques that uncover possible cancer drivers in pediatric tumors. Gene regulatory network inference from gene expression data is one of the methods used to achieve this in an unbiased way. The purchase of a Fluidigm C1 microfluidic platform, enabling scRNA-seq library construction in a high-throughput fashion, allows a plethora of new possible experiments including the search for cell or patient specific transcriptional perturbations linked to the cancer phenotype.

The main aim of the current thesis is to research whether network inference methods developed for the analysis of population RNA-seq can also be applied to scRNA-seq datasets, specifically containing cancer cells. A comparative evaluation of four widely used network inference methods based on three different computational methods will be performed. These methods will be scored for their ability to reconstruct GRN in these cells. To this end, in a first phase, the comparability of single-cell and population RNA-seq data will be evaluated. Secondly, a gold standard of interactions possibly involved in cancer will be constructed. Finally, the interactions inferred by the different methods will be compared to this gold standard, and to the interactions in nonmalignant cells, ultimately aiming to gain new insights in the impact of the specific features of this new datatype on the results of network inference. An overview of the aims is presented in Figure 5.

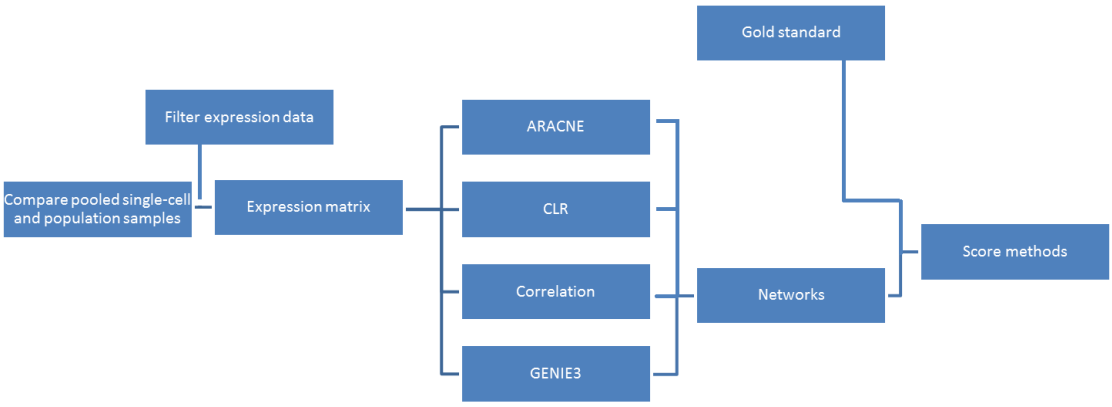

Figure 5: Overview of the aims – After comparison of the pooled single-cells and the population samples, depending on the results, the expression matrix will be filtered and the four selected NI methods will be applied. The inferred networks will be evaluated using a gold standard, ultimately giving the methods a score.

Part 3: Results

3.1. Single cell RNA-seq datasets

The four selected GRN inference methods were evaluated based on their ability to identify a set of predicted interactions in scRNA-seq data from two cancer types that have been extensively studied in the past: glioblastoma (Brennan et al. 2013; McLendon et al. 2008; Stegh et al. 2010; Verhaak et al. 2010; Ying et al. 2010) and melanoma (Cummins et al. 2006; Watson et al. 2015). A first step in this process was to find suitable datasets containing scRNA-seq samples from cancer patients. The raw reads of a scRNA-seq glioblastoma dataset (GSE57872) were available through the sequence read archive (SRA, accession number SRP042162). To convert these raw reads to expression levels, the reads were mapped to a reference transcriptome and the amount of reads mapped to each transcript were counted. These counts were then normalized to expression levels. For the single-cell melanoma dataset (GSE72056), only a supplementary file with the normalized gene expression levels was available. This file also contains information on the cell type.

3.1.1. Quality control

When dealing with raw sequencing data generated by high-throughput sequencing methods, it is important to verify whether it is of sufficient quality before continuing the analysis. To this end, FastQC was used. The runs were evaluated on a pass/fail basis for eight of the eleven features validated by FastQC; sequence length distribution, and k-mer and adapter content were excluded from the analysis. Additionally, the number and percentage of reads mapped and the amount of expressed genes were evaluated.

The dataset contains 875 samples in total, of which 672 are single glioblastoma cells. More than half of these single cells passed all quality control criteria (Figure 6).

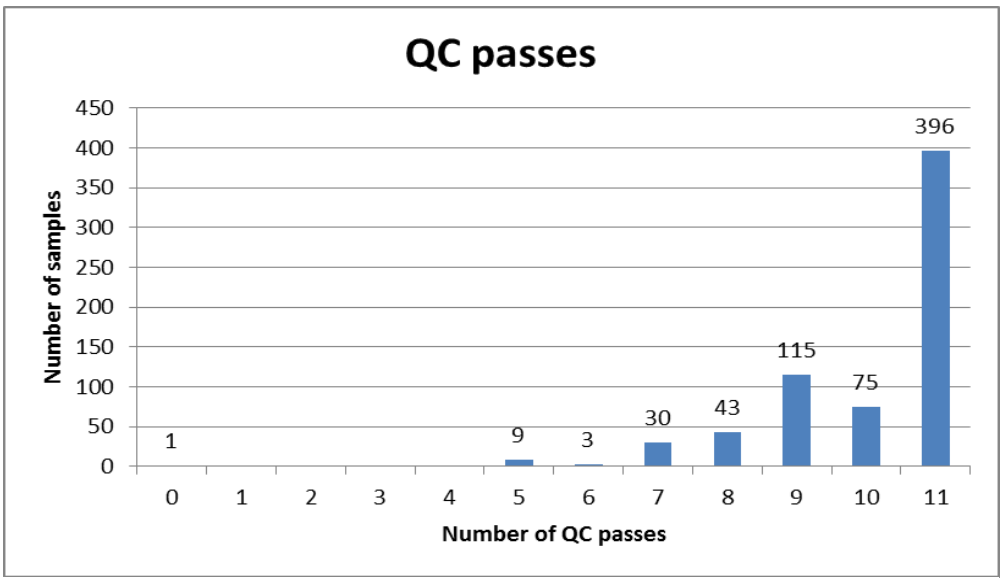

Figure 6: Number cells passing a certain number of QC criteria – About half of the samples pass all criteria

In order to verify if a certain criterion scores worse than the others, the amount of passing samples is counted for each of the criteria (Table 1). One criterion is passed by less than 80% of the samples: the amount of overrepresented sequences. A sequence present for more than 0.1% raises a warning and if a sequence makes up more than 1% of the total, this criterion is flagged as failed. A single overrepresented sequence could mean a contamination of the library.

Table 1: Number of samples that passed each criterion - The criteria that are passed by less than 80% of the samples are marked in red.

Criterion	Passes	Percentage (%)
Total sequences	669	99,55
Per base sequence quality	672	100
Per tile sequence quality	576	85,71
Per sequence quality scores	569	84,67
Per base sequence content	656	97,62
Per sequence %GC content	571	84,97
Per base n content	672	100
Overrepresented sequences	482	71,73
Number of reads mapped	660	98,21
Percentage mapped	599	89,14
Genes expressed	640	95,24

3.1.2. Mapping and normalization

After the quality of the data is confirmed, the raw reads should be mapped to a reference genome or transcriptome, or aligned *de novo*. Because this data originates from human cells, a reference genome and transcriptome are available. The tool Salmon will be used for the mapping and quantification of the transcripts. This tool requires indexing of the reference genome or transcriptome if used in quasi-mapping-based mode. For a read length of 75 base pairs, the length of the indices, k is ideally set to 31. Because the read length in the current dataset is 25 base pairs, a k of 11 is used. Depending on the read length and the number of errors permitted, mapping rates in Eukaryotic datasets are reported to be between 40 and 90% (Benjamin et al. 2014; Conesa et al. 2016; Mortazavi et al. 2008; Hatem et al. 2013). The mapping rates for the current dataset to the reference transcriptome are expected to be on the low end of this, due to short read length. The reads are mapped directly to the transcriptome, omitting problems with fragments overlapping splice junctions. Additionally, salmon is very robust to errors. However, when evaluating the mapping rates (Figure 7), much lower rates than expected were observed. This could, possibly be explained by the short read length causing fragments to match several loci, and thus be rejected by Salmon. Similar mapping rates are reported by the authors (Patel et al. 2014).

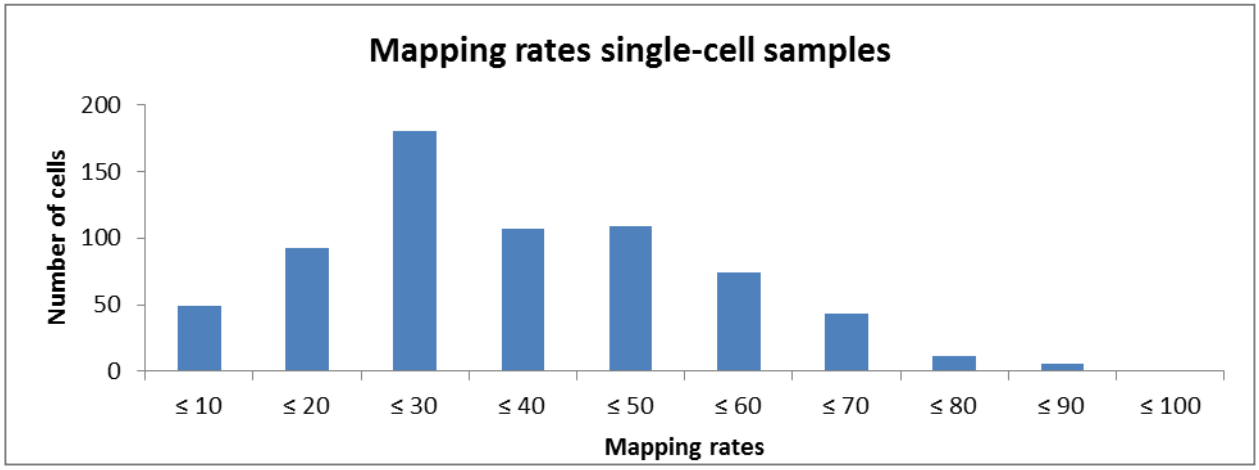

Figure 7: Mapping rates of single-cell samples – Most of the samples show a mapping rate of 21-30%, while rates between 40 and 90% are expected, depending on the data set and the allowed number of errors.

For all genes that have at least two samples with 5 counts, the expression is calculated as the amount of transcripts per million (TPM) and log transformed.

3.1.3. Melanoma

To date, only one dataset containing single melanoma cell RNA-seq data is reported, containing malignant as well as nonmalignant cells from 19 malignant melanoma patients. For the current analysis, the cells that were classified as malignant and that did not have a nonmalignant cell type category assigned to them (*non-malignant cell type* = 0) were selected. 1252 malignant cells (Table 2) were isolated from different patients. The tumors showed varying numbers of malignant cells.

Table 2: Number of malignant cells for each melanoma patient – In total there are 1252 malignant samples from 14 different patients.

Patient	Number of malignant cells
53	16
59	54
60	9
65	4
71	54
78	120
79	468
80	125
81	133
82	32
84	14
88	115
89	98
94	10
	1252

3.2. Population vs. pooled vs. single cell

To obtain better insights in the features of scRNA-seq data and its comparability to its population counterpart, a comparison of population and single-cell glioblastoma expression data was carried out. The glioblastoma dataset under analysis did not only contain single glioblastoma patient data, but also a population sample for each patient, consisting of 2000 – 10000 cells. Gene expression levels in these samples were compared to levels in single-cells and pooled single-cells. Expression levels in the pooled single cells are expected to be comparable to those in the population samples, because certain biological confounders are expected to be levelled out the same way as they are in population RNA-seq. Prior to these analyses, no samples were excluded.

3.2.1. Pooled single-cell expression levels

First, the single-cell reads were concatenated, creating a file with all forward reads and a file with all reverse reads. Next, these reads were mapped to the human reference transcriptome and expression levels were quantified in the same way as described above for single-cell reads (*supra* 3.1.2 Mapping and normalization).

Only protein coding genes that were expressed in at least one of the groups, i.e. single-cell, pooled or population, were regarded for further analysis. This led to the inclusion of 20740 genes, of which 19339 were expressed in all three groups (Figure 8).

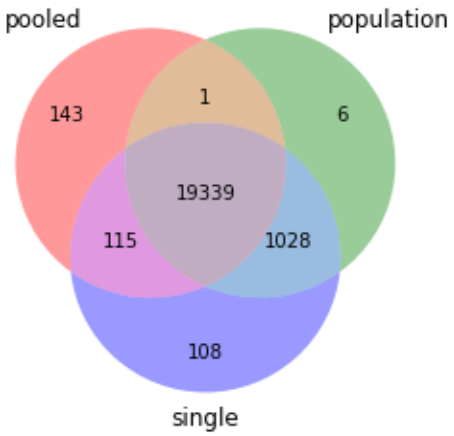

Figure 8: Comparison of the genes present in each of the expression matrices – 19339 genes are expressed in all three groups

3.2.2. Correlation

To compare expression levels of the different genes in pooled and in population samples, the correlation was plotted (Figure 9) and calculated. If the results in patient MGH31 are disregarded, the shape of all the plots shows a similar trend, i.e. a linear relation between the expression values, but with a slight bend towards the x-

axis for the lower expression values. This means that the expression of each gene, on average, is higher in pooled than in population samples.

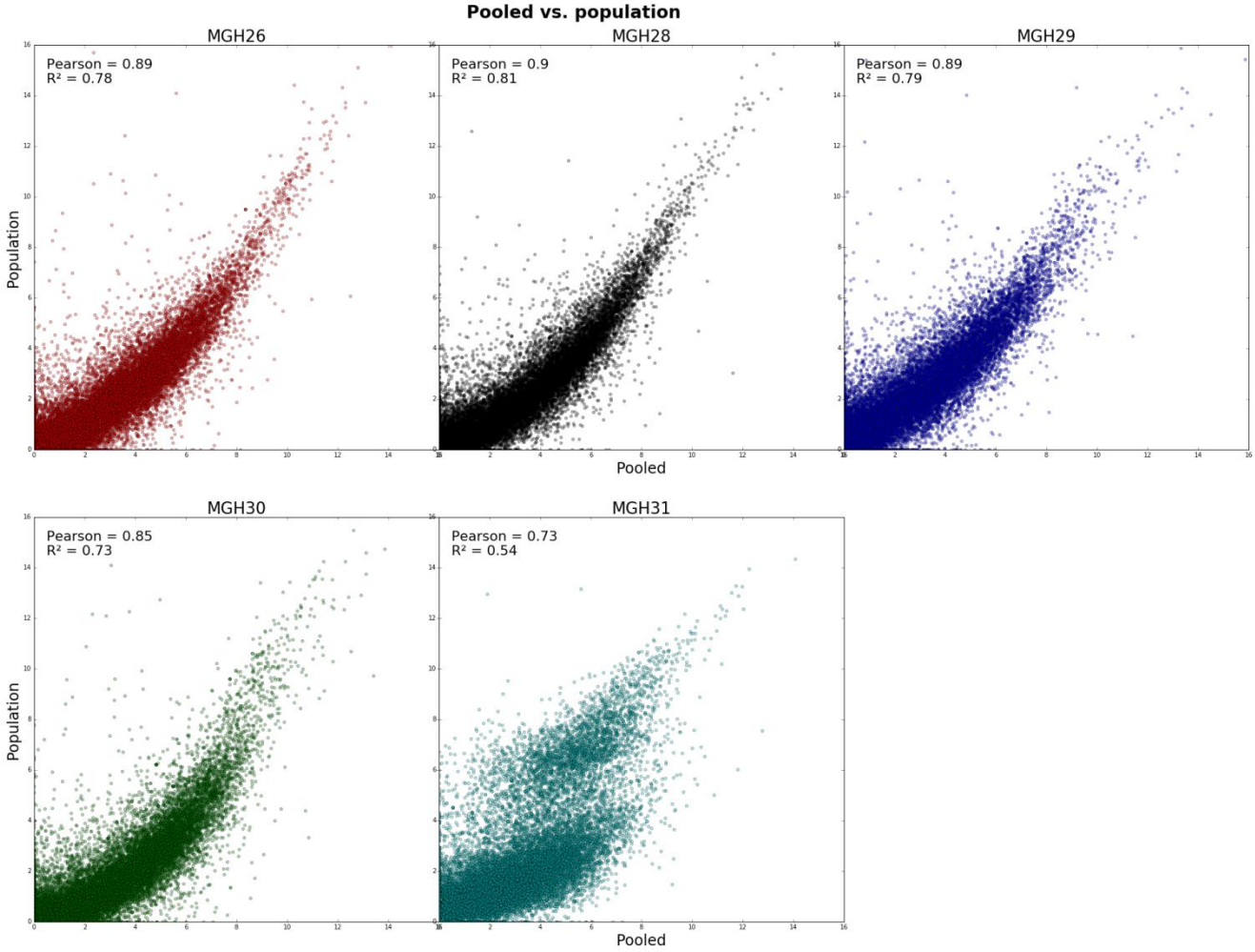

Figure 9: Correlation of expression levels in pooled single cells vs. in population samples, per patient – For MGH26, MGH28, MGH29 and MGH30 there is a linear relation between the expression values but for lower values, the expression in the pooled samples is slightly higher. Correlation for MGH31 is poor.

This trend is confirmed when studying the genes that are not expressed in one or both conditions (Table 3). Almost 1.5 times more genes are not expressed in population samples compared to pooled samples.

Table 3: Evaluation of the expression of all genes for the different patients in a binary discrete way: expressed (+) or not-expressed (-) – The population samples have about 1.5 times more non-expressed genes that are expressed by the pooled samples than vice versa.

	population +	population -	total
pooled +	89257	6353	95610
pooled -	4811	3279	8090
total	94068	9632	103700

To find a possible explanation for the aberrant results in patient MGH31, the average mapping rates per patient were analyzed. The mapping rate for this patient is slightly under 12%, while the others all have a rate higher than 15% (Supplementary table 1). Possibly, sequence data from this patient has a lower quality.

3.2.3. Dimensionality reduction

Of the 20740 genes that were present in at least one of the matrices, the top 50% genes with the highest variance were selected for further analysis, because it is most likely that these genes are of biological importance. The dimensions of the dataset were reduced from 682 samples and 10370 genes to three dimensions in the gene direction (Figure 10), based on a distance matrix containing correlation distances between all genes. This was done using classical Torgerson multidimensional scaling (MDS).

Except for the data of patient MGH31 – which had also shown deviating results when studying correlation - population, pooled and single-cell data grouped together per patient. However, a subset of single-cell samples (Figure 10, circled in red) is observed that does not group together with the other samples from the same patient. This subset consists of all samples within the sphere with center (0.20, -0.3, 0.28) and radius 0.29. In total, a group of 47 samples, containing samples from each of the patients, is selected.

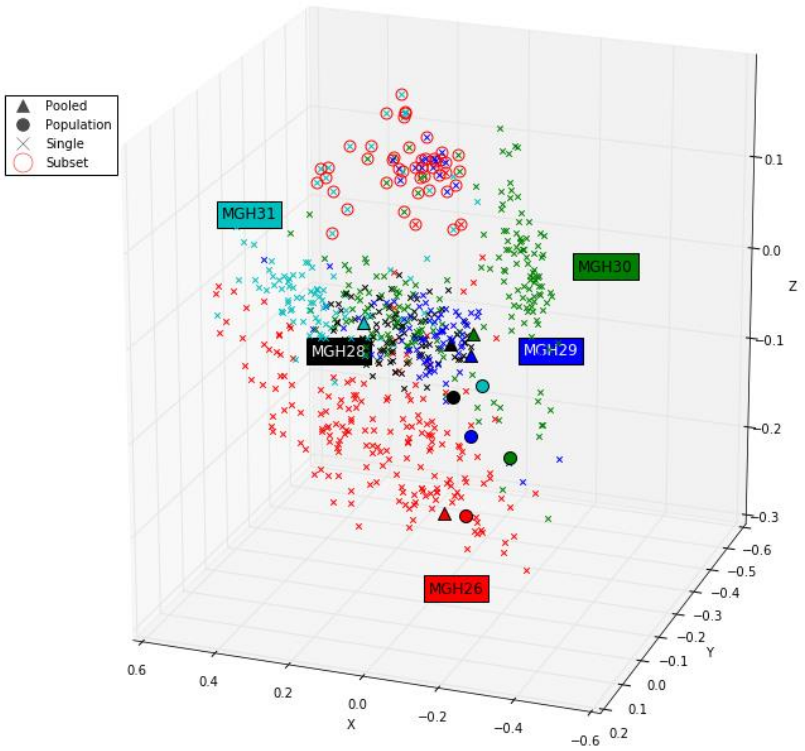

Figure 10: Classical MDS in 3 dimensions, showing single-cell, pooled and population samples from all five patients – Most of the samples cluster together per patient, however, samples circled in red belong to a possibly interesting subset

To analyze why these samples do not group together with other single-cell samples from the same patient, a differential gene expression analysis is executed. First, the dataset is divided into six groups, i.e. the five patients and the subset. Next, differential expression between the subset and each of the patients is calculated for all 10370 genes. It does, however, need to be taken into account that, strictly speaking, the different samples cannot be seen as replicates of the same patient. Assuming they are replicates, will cause the difference to seem more significant than it is in reality. 1977 genes with a p-value of <0.01, which was corrected for multiple testing using the method of Benjamin-Hochberg, and a log fold change of >4 (Figure 11) in at least one of the patient-subset analyses are retained for further analysis.

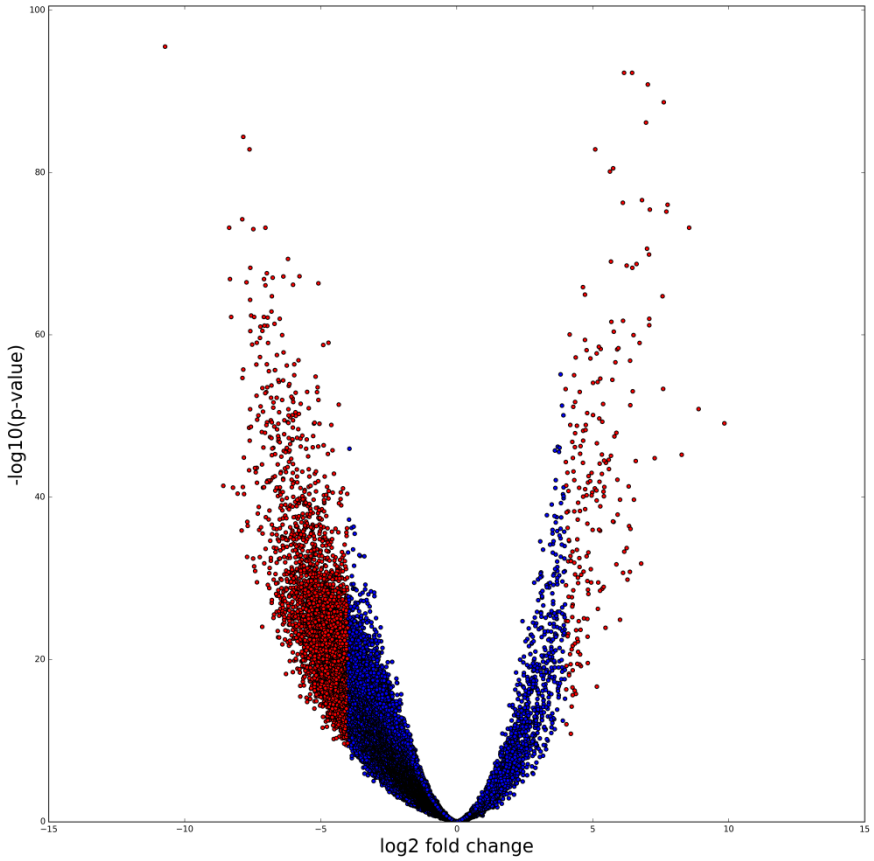

Figure 11: Volcano plot showing the log2 of the fold change of the gene expression versus - log10 of the p-value – samples in red, with p-value < 0.01 and a log fold change of > 4 are considered significant and retained for further analysis

Due to the large amount (1885) of genes that are downregulated in the subset as opposed to the patients, these samples are suspected to be of inferior quality. This is tested by comparing, among others, the mapping rate and number of genes expressed between the two groups (Figure 12). To assess whether the difference between the two groups is significant, Mann–Whitney U tests ($\alpha < 0.05$) were performed for all features (Supplementary table 2), as the data does not have a normal distribution. These tests confirm that the quality of the subset differs significantly from the quality of the other samples.

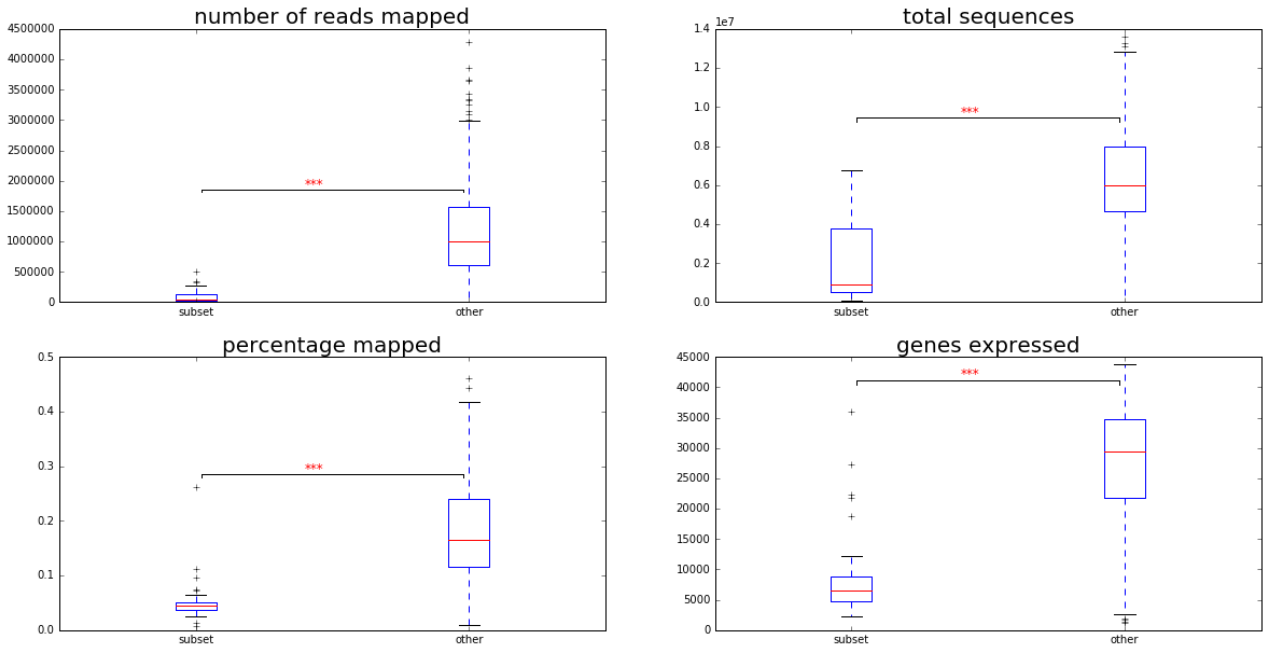

Figure 12: Comparison of several quality control values between the subset and the other single-cell samples – levels of significance: p-value < 0.05: *, p-value < 0.01: **, p-value < 0.001: ***

The analysis is repeated only including the single-cell samples that passed more than 8 of the quality control criteria evaluated (*supra* 3.1.1 Quality control). Again, only the genes that are expressed in at least one of the groups are included.

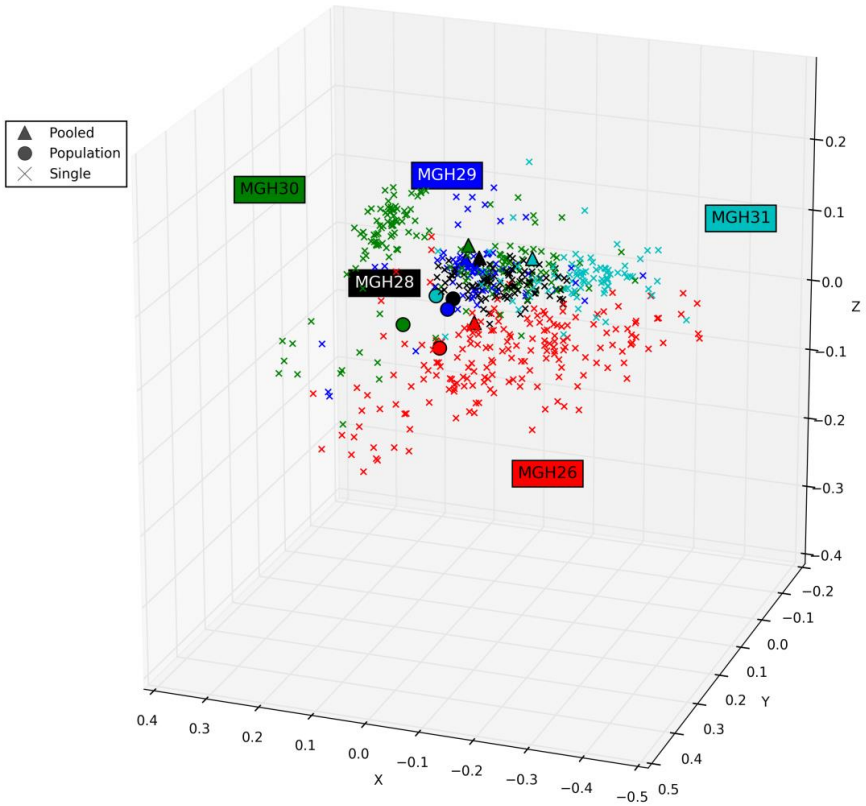

Figure 13: MDS after exclusion of the low quality samples – exclusion of the low quality samples improves clustering per patient

After exclusion of the low quality samples, the dimensionality reduction plot (Figure 13) shows a much better correlation for the single-cell samples of the different patients.

3.2.4. Conclusion

When comparing TPM-normalized expression levels in pooled and population samples, on average the genes show higher expression in the pooled samples, especially for genes with lower expression rates. Also, about 1.5 times more genes that show expression in pooled samples are not seen in population than vice versa. This could be explained by a difference in sequencing depth between the population samples and the single-cell samples, which compose the pooled samples (Supplementary table 3). A smaller sequencing depth causes the sensitivity to be lower for lowly expressed genes.

The data was also visualized in three dimensions using MDS. Initially, no genes or samples were excluded from the analysis, leading to the grouping of a subset of single-cell samples. Evaluation of the mapping rate and the number of genes expressed by all samples, uncovered that there were significant differences in quality between this subset and the other samples. Exclusion of samples that passed less than 8 of the evaluated quality control criteria, led to a better clustering of the samples.

These findings motivated the development of a filtering method of the expression matrices before network inference. The inferior correlation in patient MGH31 (*supra,* 3.2.2 Correlation) is expected to be caused by low quality of the samples; if the considered filtering method is applied, these samples will also be removed.

3.3. Network inference glioblastoma

In one of the previous steps, expression matrices have been generated from two single-cell cancer datasets. In the next step, the selected network inference methods were evaluated for their ability to reconstruct GRN in the cancer cells based on single-cell expression data. This requires a gold standard of expected interactions is these cancer types, which will be constructed by integrating data from different pathway databases. The interactions in nonmalignant cells, however, also need to be taken into account when evaluating the identified interactions. To reduce computational complexity, a list of transcription factors, which should be regarded as possible regulators, was passed to the algorithm.

3.3.1. Filtering of the expression matrix

As most network inference methods are computationally intensive, the size of the expression matrix was reduced in such a way that it only included samples of interest that showed sufficient quality, i.e. passed at least 8 of the aforementioned quality control criteria, and genes that code for proteins and are expressed in at least 10% of the samples. Further selection of the genes is done based on the variance of expression across all samples, considering genes with a higher variance have a higher chance to be of biological importance. The top 8000 genes with the highest variance are selected.

3.3.2. Gold standard

As a gold standard on glioblastoma cells does not exist, important genes and interactions were inferred from the literature. The TCGA pilot paper on glioblastoma (McLendon et al. 2008) reported three main pathways involved in glioblastoma: the p53 and retinoblastoma tumor suppressor pathway, dysregulation of cell growth via mutations in RTK genes and activation of the PI3K pathway. The human pathways *PI3K-Akt signaling pathway, cell cycle* and *p53 signaling pathway* were downloaded from the KEGG database and the interactions with type protein-protein interaction (PPrel) and gene expression interaction (GErel) were written to a file. From the KEGG disease database, interactions predicted to be involved in glioma were also added to the list of predicted interactions. Finally, stringDB was searched for proteins interacting with Mig-6 (Ying et al. 2010) Bcl2-L12 (Stegh et al. 2010) PTEN and interactions between genes predicted to be involved in glioblastoma by PathCards (Belinky et al. 2015) This resulted in a total of 383 interactions, containing 189 different genes. Additionally, four clinically relevant subtypes of glioblastoma were identified using TCGA data (Verhaak et al. 2010) . A list of 210 signature genes for each of these types is available, but only 679 of these genes could be mapped to an ensembl identifier. Finally, the KEGG pathway microRNAs in cancer (hsa05206) was consulted. Certain microRNAs are upregulated in glioblastoma, inhibiting tumor suppressor gene activity and the downregulation of other microRNAs may cause

oncogene activation. 8 genes that are possibly upregulated and 7 genes that are possibly downregulated were added to the list of relevant genes.

Of these in 855 unique genes, three did not occur in the glioblastoma expression matrix; these genes were omitted from further analysis. When comparing the expected genes to the 8000 filtered glioblastoma genes (Figure 14), about half of the expected genes are included. The other half of these genes did not show sufficient variance over the different samples.

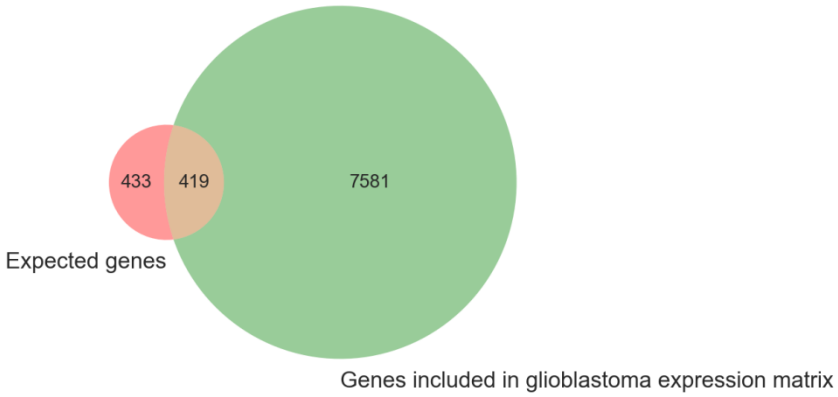

Figure 14: Number of expected genes included in filtered expression matrix – about half of the expected genes are included in the filtered gene list.

3.3.3. Nonmalignant cells

To avoid qualifying normal astrocyte interactions as false positives because they do not occur in the list of important interactions in glioblastoma, a control matrix containing expression levels of the filtered glioblastoma genes in astrocytes was subjected to the same network inference methods. This way, a list of control interactions was constructed.

3.3.4. Performance

As a first exploratory analysis, each of the ranked lists of predicted interactions was validated against the combined lists of control interactions and interactions involved in glioblastoma, plotting a bar if the inferred interaction occurred in the list of expected interactions (Figure 15). All interactions are expected to show a match with either a control interaction or a glioblastoma specific interaction. However, the matching rates are lower than expected.

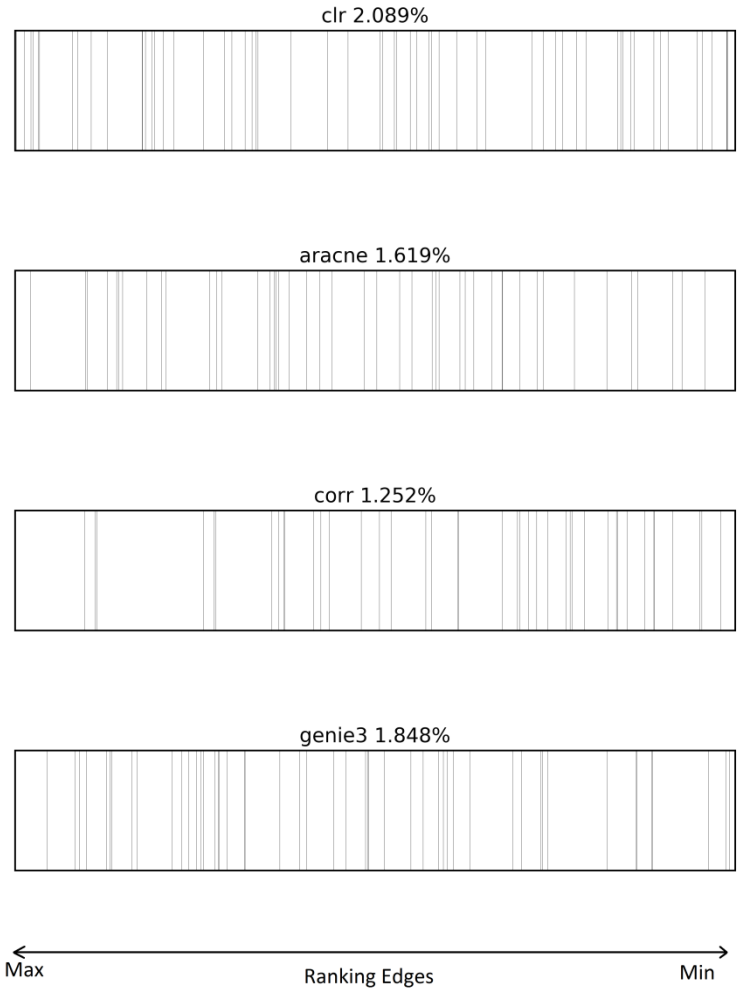

Figure 15: Barcode plot; a stripe is plotted if an interaction is found in the combined list of expected and control interactions – the edges are ranked from high to low, with the highest weight on the left hand side of the plot. The percentage indicates the fraction of interactions that were in the combined list.

Next, the occurrence of interactions involved in glioblastoma was evaluated in the ranked interactions inferred from the cancer expression data and compared to those inferred from control expression data (data not shown). The same was done for the genes involved in these interactions (Figure 16). As some of these interactions also exist in nonmalignant cells, they are expected to be positive in that dataset but to occur at random. In malignant cells, on the other hand, they should rank highly. These hypotheses are not confirmed by the data. Only 11 of the interactions in cancer and control data is also in the list of expected interactions (data not shown) and the pattern generated by the evaluation of the genes involved in these interactions in cancer data is more random than expected.

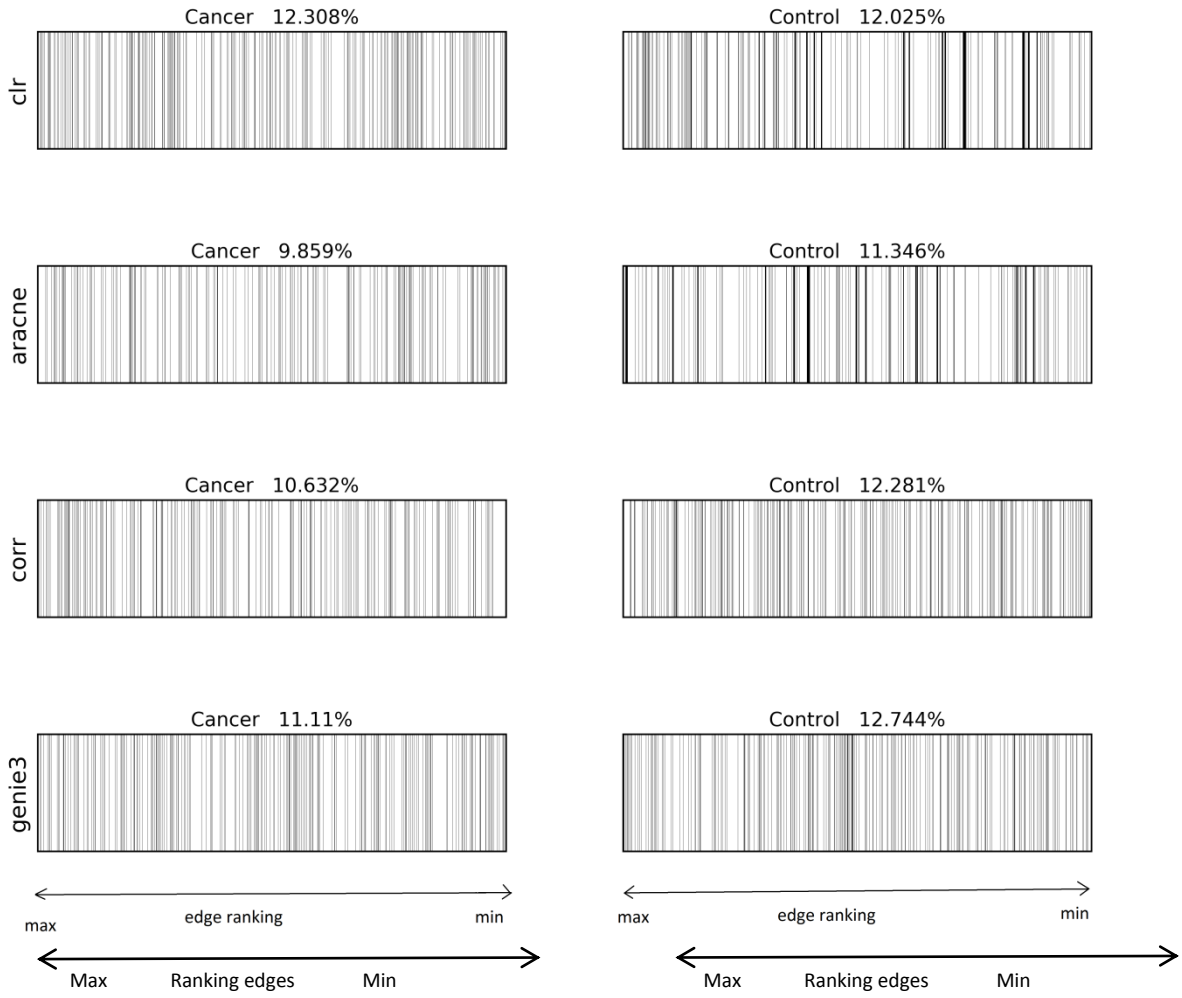

Figure 16: Barcode plot showing the ranks of expected genes in cancer; these are compared in cancer and control data – the edges are ranked from high to low, with the highest weight on the left hand side of the plot. clr = context likelihood relatedness, aracne = Algorithm for the Reconstruction of Accurate Cellular Networks, corr = correlation, ctrl = control.

Finally, the performance of each of the methods was evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUROC) and the area under the precision-recall curve (AUPR) (Figure 17). Ideally, these areas approach 1 as closely as possible. For this dataset, none of the methods score better than a random method would.

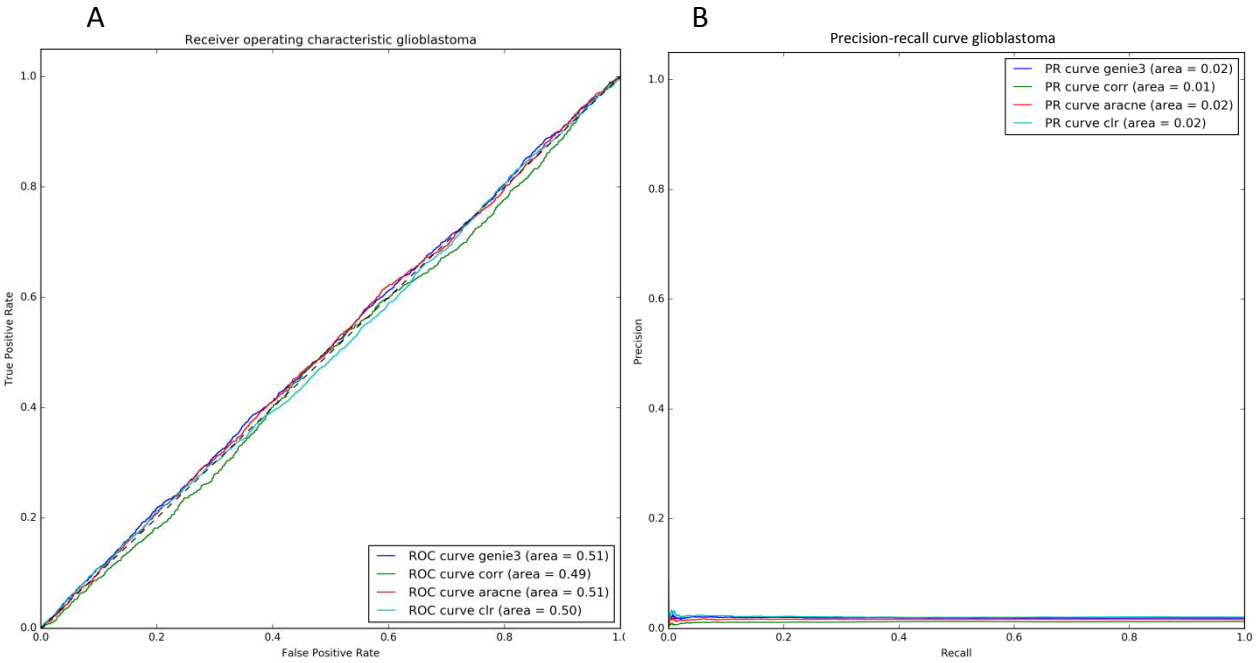

Figure 17: Receiver operating characteristics (A) and precision-recall curve (B) for all the different network inference methods for analyzing single-cell glioblastoma data – None of the methods score better than random (0,50).

3.3.5. Conclusion

Analysis of NI methods through this dataset has not enabled the designation of a single method that seems promising for the inference of GRN in scRNA-seq data. On the contrary, none of the methods achieved a better score than a random method would, based on the AUROC scores. This could be explained in a number of ways.

Due to the short raw reads and, possibly, subsequent low mapping rates, it is possible that expression values of several genes that are expressed *in vivo*, are not included in the expression matrix. Also, the values in the matrix might not reflect the true expression levels. Since GRN are inferred from the expression levels, this might have an impact on the amount of false negative interactions.

Alternatively, the lack of existence of a true gold standard for glioblastoma, resulting in the construction of a list of putative cancer interactions based on interactions known in literature, might have led to these results. Though the interactions reported in literature are highly confident, not all data sources used to create this list have the same level of curation. The use of interaction prediction databases, such as

stringDB, which integrates several prediction tools, will also include interactions of lower confidentiality.

It also needs to be taken into account that the control data originated from microarray experiments due to the absence of control scRNA-seq experiments. The comparison of their derived GRNs will be more accepted if the raw data has the same source.

3.4. Network inference melanoma

3.4.1. Filtering of the expression matrix

As mentioned before (*supra,* 3.1.3 Melanoma), only malignant samples were included in this analysis. No information on sample quality or mapping rates was available. The genes were filtered in a similar way as the glioblastoma genes: first, only the genes that have a gene type '*protein_coding*' were selected. Next, the variance of the genes that are expressed in at least 10% of the samples was calculated. Due to the amount of samples, the top 6000 genes with the highest variance were selected for network inference.

3.4.2. Gold standard

As for glioblastoma, no melanoma golden standard is available. The Melanoma Gene Database (MGDB) (Zhang et al. 2015), a manually curated database containing information on 527 genes reported in literature to be involved in melanoma, was consulted. First, the list of genes, obtained from PubMed abstracts by the authors, was downloaded. A distinction was made between 422 coding and 105 non-coding genes. For further analysis, only the coding genes were obtained. Secondly, the file containing interaction information for these genes downloaded. This information had been obtained by the authors through the Protein Interaction Network Analysis (PINA) (Cowley et al. 2012) platform. Of the 422 genes, roughly one third had a highly variant expression and was included in the filtered expression matrix (Figure 18).

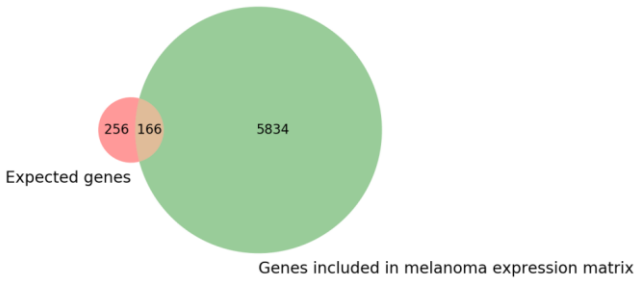

Figure 18: Number of expected genes included in filtered expression matrix – about half of the expected genes are included in the filtered gene list of genes included in the melanoma expression matrix

3.4.3. Nonmalignant cells

Microarray expression data from 7 melanocytes was filtered to include only the 6000 genes that were filtered in the melanoma single cell data. Only 4730 of these genes were found in the control dataset, probably caused by incomplete mapping between the control dataset's gene symbols and the melanoma ensembl identifiers.

3.4.4. Performance

The initial analyses executed for the inferred networks in melanoma data were the same as for glioblastoma data. First, the ranked interactions were verified against the combined list of predicted and control interactions. The number of inferred interactions that could be found in the combined list was higher in melanoma than in glioblastoma, but in terms of percentage, the increase seems insignificant (data not shown).

The ranks of interactions (data not shown) and genes (Figure 19) likely to be involved in melanoma were compared in cancer and control interactions. Here too, we would expect the interactions to occur at random in the control interactions and to rank highly in the cancer dataset, but the results are comparable to those in glioblastoma.

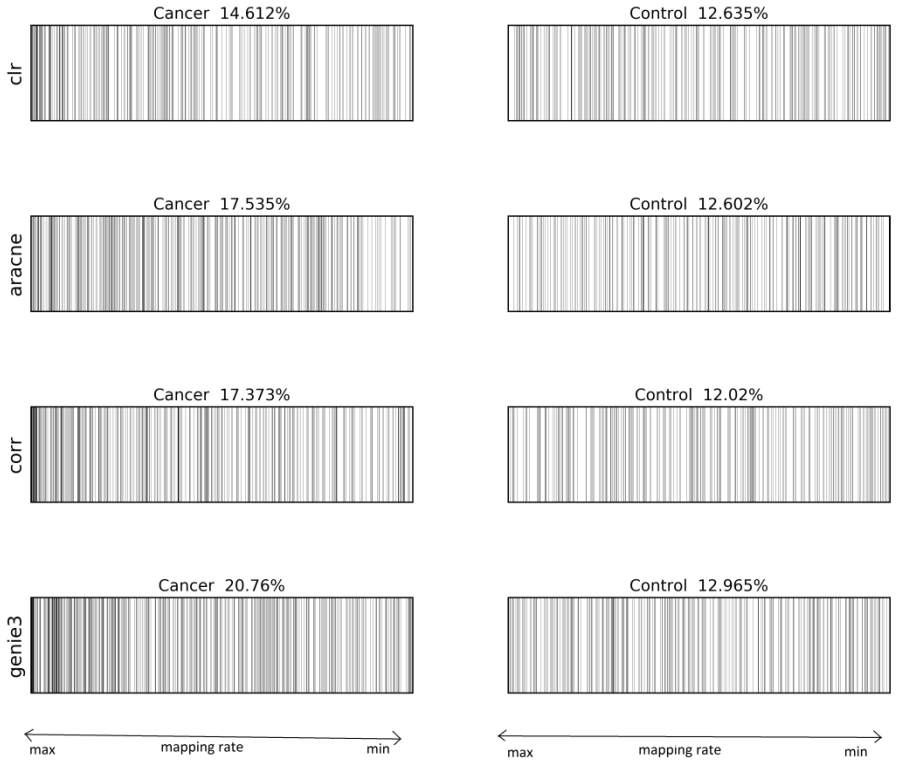

Figure 19: Comparison - in cancer versus control - of the ranks of genes expected to be important for melanoma - clr = context likelihood relatedness, aracne = Algorithm for the Reconstruction of Accurate Cellular Networks, corr = correlation, ctrl = control.

The performance of each of the methods was evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUROC) and the area under the precision-recall curve (AUPR) (Figure 20).

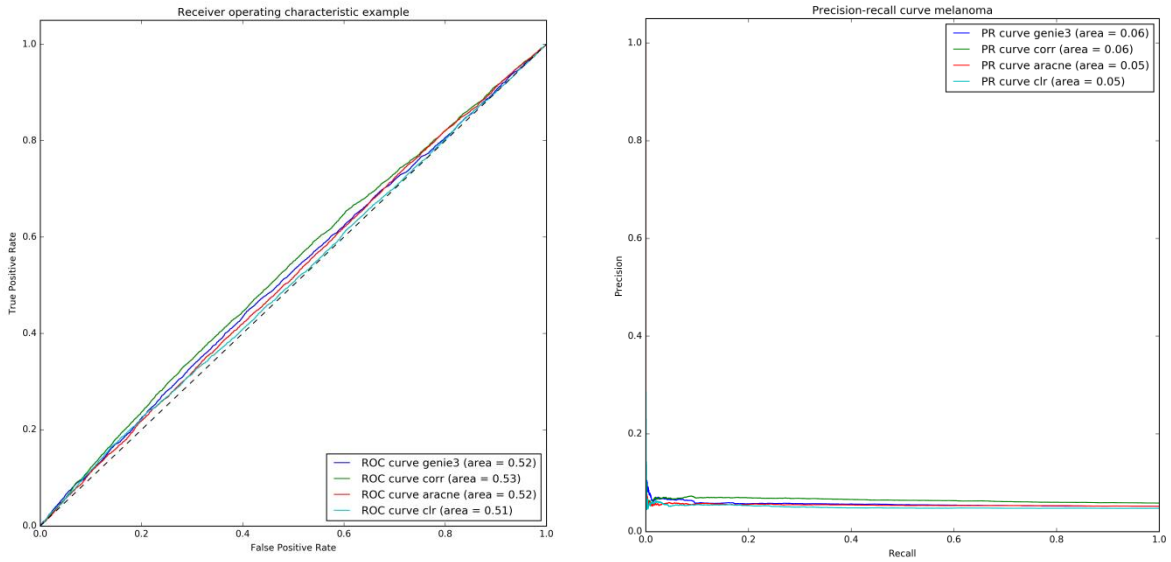

Figure 20: Receiver operating characteristics (A) and precision-recall curve (B) for all the different network inference methods for analyzing single-cell melanoma data – None of the methods score better than random (0.50)

3.4.5. Hubs

Biological networks are reported to have scale-free properties, meaning that more nodes than expected have a high degree and the node connectivity follows a power-scale distribution. As a result, to evaluate the importance of a gene in a biological network, it is not sufficient to evaluate the weight of its edges, but it is also important to evaluate its connectivity. Highly connected genes, the so called hubs, are more likely to be essential.

The degree distribution was verified for each of the methods (Figure 21). These plots show that there are a few highly connected nodes, and a high amount of genes with a very low connectivity.

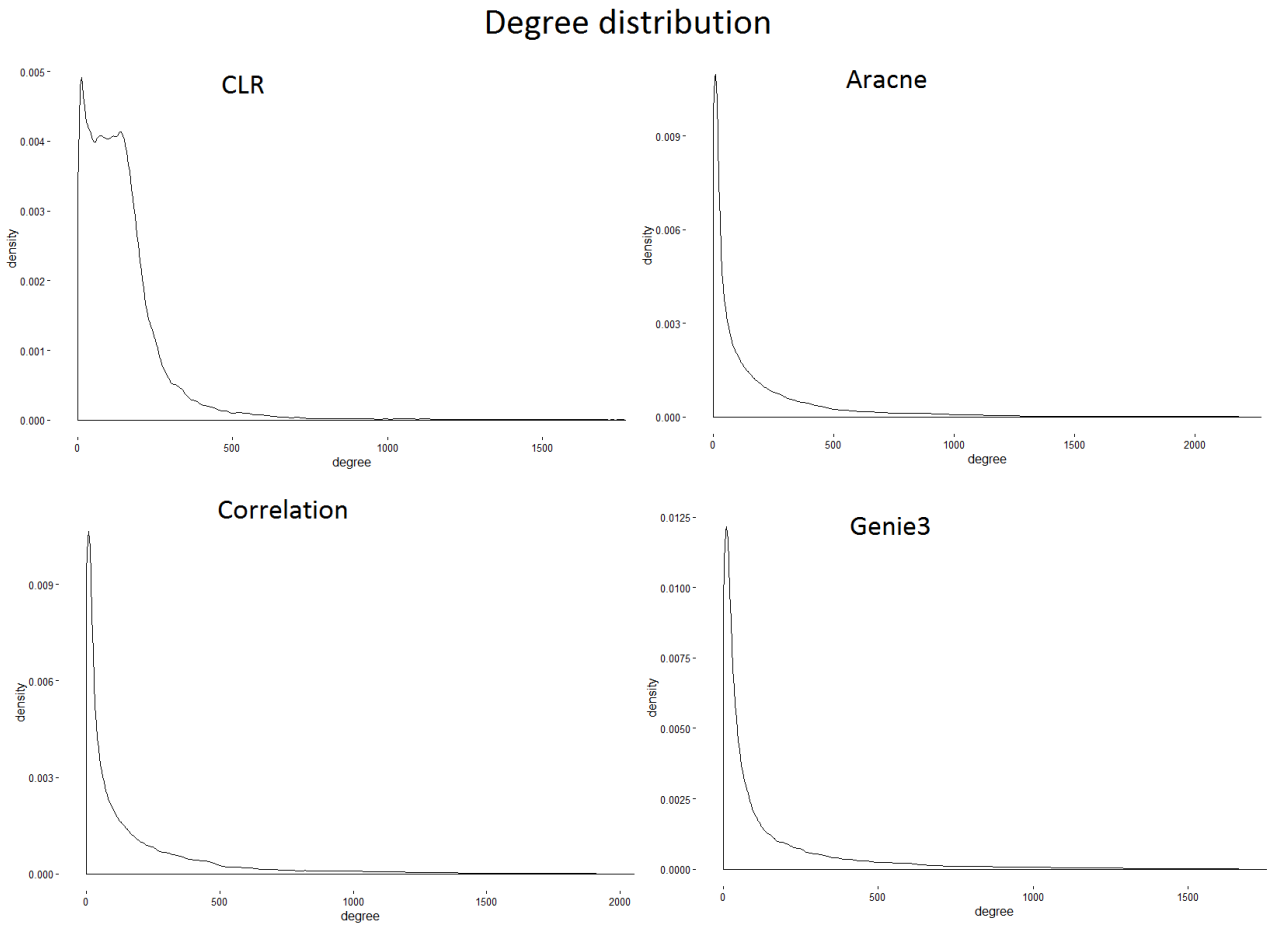

Figure 21: Degree density for the networks inferred by the different NI methods – all degrees are not normally distributed but have a small amount of highly linked nodes and a large amount of nodes that are lowly linked.

Next, the ranked interactions are evaluated one by one, verifying the degree of the regulators with each step. Regulators that rapidly gain a high degree are more likely to be of biological importance as these nodes have a lot of connections with a high weight. For each of the regulators, this change in degree with descending weight is plotted and the area under the curve is calculated. For example, in the case of CLR (Figure 22), most regulators show a linear increase in degree. However, a few regulators have a large amount of high weighted edges and show a more logarithmic increase in degree; these are possible hubs.

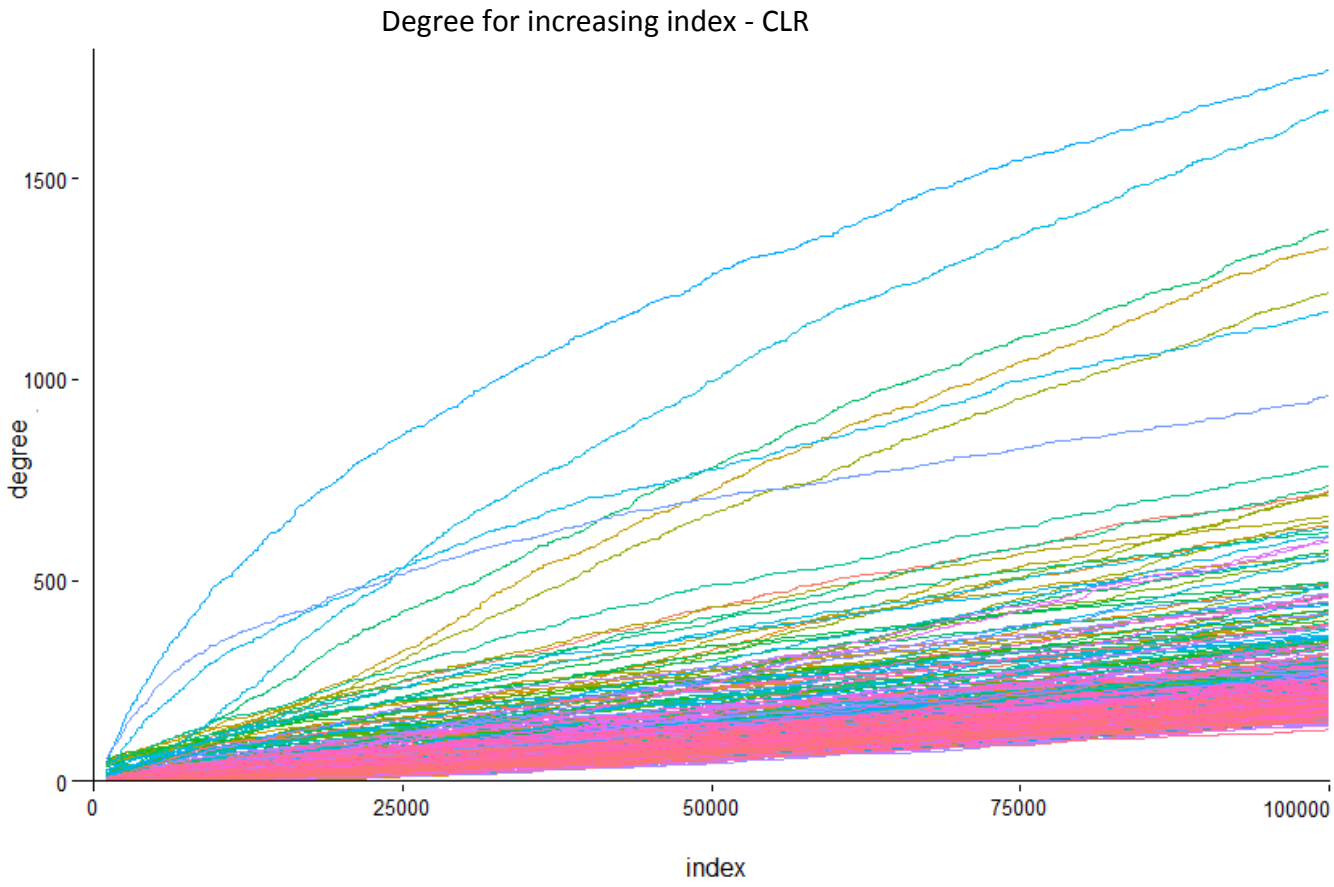

Figure 22: Plot showing the degree in function of the index – most regulators show a linear relation but some regulators have a large amount of high weight edges; these are possible hubs.

Secondly, a gene set test (GST) is performed to test whether the set of regulators is ranked more highly as opposed to a randomly chosen set of genes. For each of the methods, the log transformed p-value of the GST for each regulator is plotted against its area under the change-in-degree curve. The 10 most significant regulators are labeled (Figure 23).

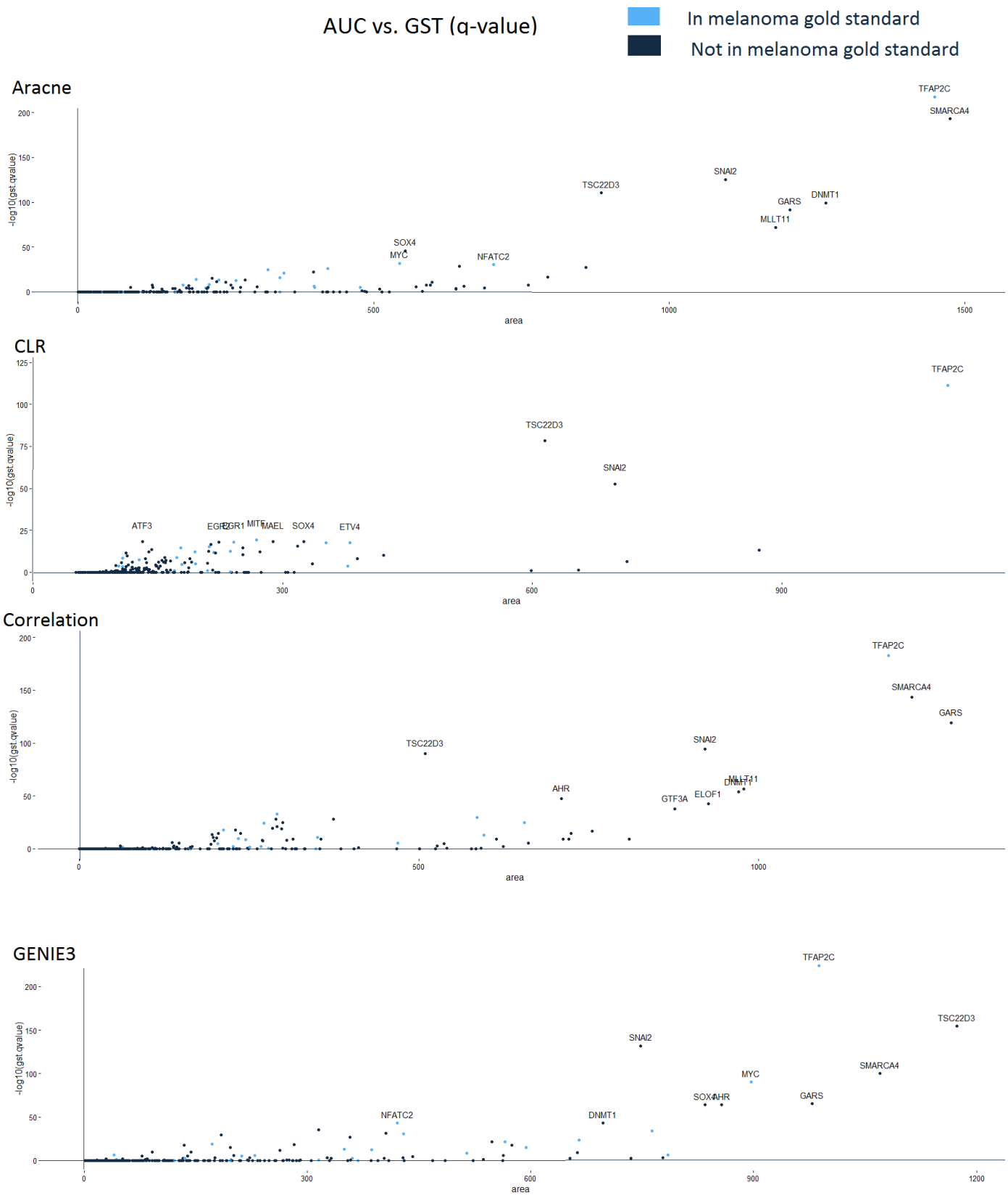

Figure 23: the log transformed p-value, corrected for multiple testing, is plotted against the area under the degree curve for each of the methods – the top 10 genes with the most significant p-value are labeled and genes which are in the melanoma gold standard are colored light blue.

These plots show that the regulators with the highest area under the curve are not necessarily the regulators show the most significant results for the gene set test. Finally, the 10 most significant regulators are compared for each NI method in the melanoma and the control data, and evaluated against the gold standard for melanoma (Table 4).

Table 4: Top 10 most significant regulators for each method – For each of the methods, the top 10 most significant regulators compared between melanoma and control. The regulators that occur in the gold standard, are marked with a *

	ARACNE		**Correlation**		**GENIE3**		**CLR**	
	Melanoma	Control	Melanoma	Control	Melanoma	Control	Melanoma	Control
1	TFAP2C *	HIF1A	TFAP2C *	DRAP1	TFAP2C *	ID1	TFAP2C *	ZNF195
2	SMARCA4	RELA	SMARCA4	PCBP2	TSC22D3	NOC4L	TSC22D3	ZNF207
3	SNAI2	PHB2	GARS	AEBP1	SNAI2	TSC22D2	SNAI2	ZNF672
4	TSC22D3	PIAS4	SNAI2	HMGA1	SMARCA4	ZFP36L2	MITF *	CENPT
5	DNMT1	PCBP2	TSC22D3	PHB2	MYC *	GATAD2A	ATF3	SFPQ
6	GARS	FOXJ3	MLLT11	CREB3L2	GARS	E2F6	MAEL	NFKBIB
7	MLLT11	ARID5B	DNMT1	KAT5	SOX4	ARNT2	SOX4	DDB2
8	SOX4	ZNF451	AHR	HMG20B	AHR	KAT5	EGR1 *	MBD4
9	MYC *	TSC22D1	ELOF1	RELA	DNMT1	TFB1M	EGR2	NFE2L1
10	NFATC2 *	ZNF593	GTF3A	CERS4	NFATC2 *	ZBTB48	ETV4 *	ZNF174

For the melanoma data, 19 different regulators are seen in total, of which three, SNAI2, TFAP2C and TSC22D3, are reported by all four methods. These highly connected genes can be seen as hubs. Transcription factor AP-2 gamma (TFAP2C) is most significant in all methods. Silencing of TFAP2C by micro RNAs (miRNAs) is reported to be involved in oncogenesis of melanoma (Penna et al. 2011).

Apart from TFAP2C, five other hubs can also be found in the gold standard, i.e. EGR1, ETV4, MITF, MYC and NFATC2 (indicated with *, table x). However, it is possible that the gold standard is incomplete. SNAI2, for instance, which is ranked third by three of the methods, is reported to play a role in the epithelial-to-mesenchymal transition (EMT) in healthy melanocyte development. It is also suggested that it influences melanoma progression by enhancing motility and invasiveness (Shirley et al. 2012). The same publication also reported an association between expression of SNAI2 and MITF, a gene that was present in the current gold standard. MITF activates SNAI2 *in vivo*.

Finally, the regulators were sorted by descending AUC and ascending GST q-value. The ranks of each of the 422 gold standard melanoma genes were evaluated for both criteria (Figure 24). The melanoma genes clearly rank higher, showing that these genes are more likely to be highly connected.

Figure 24: The ranks of each of the 422 gold standard melanoma genes evaluated for AUC and GST - The melanoma genes clearly rank higher, showing that these genes are more likely to be highly connected. AUC = area under the curve, GST = gene set test.

3.4.6. Conclusion

Similar to the glioblastoma dataset, analysis of the ranked interactions has not led to the anticipated result. Analysis of transcription factor connectivity, however, showed that genes that are involved in melanoma are more highly connected. This means that these genes, hubs, are more likely to be involved in important interactions.

Of the 19 most highly connected regulators, only six are reported in the gold standard. However, it is possible that the gold standard is incomplete. Several of these regulators have also been identified to play a role in oncogenesis of other cancer types and might be interesting to investigate *in vitro* and *in vivo*.

Part 4: Discussion

4.1. Conclusion

Due to recent developments in the field of microfluidics, it has become possible to analyze transcriptomes on a single-cell level in a high-throughput fashion: scRNA-seq. This technique has already shown promising results in developmental and immunological research and is also emerging in cancer research, particularly in uncovering aspects of tumor heterogeneity. To date, there is no consensus on which computational methods to use for the analysis of this new data type. For some steps of the analytical process, it is suggested that tools developed for the analysis of population RNA-seq can be used, although adaptations will be required. For other steps, new tools will need to be developed.

One of the ways to obtain biological insights from expression data is through reconstructing gene regulatory networks from this data. It has been reported that network inference (NI) methods developed for population RNAseq can be used for single-cell data, but that adjustments will be required to deal with the higher levels of technical noise and with confounding factors. The main aim of this thesis was to evaluate selected NI methods for their ability to correctly infer gene regulatory networks (GRN) using scRNA-seq data.

Since scRNA-seq has only been developed recently and not much has been reported about the comparability of this data to its population counterpart, the first part of the analysis consisted of a comparison between average single-cell expression levels and population expression levels. Plots of the correlation between expression levels in the pooled single-cells and the population cells showed that, especially for lowly expressed genes, the expression values in pooled samples were slightly higher. Also, the amount of non-expressed genes was lower in these samples. A possible explanation for this was the difference in sequencing depth, with the depth being lower in the population samples. Next, the samples were visualized in 3D using dimensionality reduction. This analysis led to the detection of a group of outliers. The quality of these samples was markedly lower than the other samples; their exclusion led to better clustering results and motivated the exclusion of low quality samples for further analysis.

The second part of the current thesis consisted of the scoring of the performance of four selected NI methods, using data from two publicly available single cancer cell RNA-seq experiments. First, a gold standard of interactions reported in these two cancer types, glioblastoma and melanoma, needed to be constructed. For the glioblastoma dataset, this was done integrating data from different pathway and interaction prediction databases, based on a literature study. Analysis of the networks inferred by the different methods, did not show the expected results: few of the gold standard interactions were found by any of the methods and ROC analysis showed that none of the methods scored better than random. For the melanoma gold standard, all coding genes from the highly curated melanoma gene

database (MGDB) were used. Again, none of the NI methods showed to be superior in a ROC analysis. Finally, the connectedness of the regulators was evaluated for each of the networks. The interactions were ranked based on the predicted edge weights, and a gene set test (GST) was used to evaluate whether certain regulators were ranked more highly than expected by chance. For each of the regulators, the degree was also evaluated with ascending interaction weight. Regulators for which the degree rises more rapidly within the first - higher weighted - interactions, are expected to be biologically more important, as they have more connections with a higher weight. For each method, the top ten genes with the most significant p-value for the GST were compared. In total, this resulted in 19 different genes, of which six were also reported in the gold standard, although a superficial literature search shows that many of the other genes are reported in cancer, and some of them even in melanoma. These genes are possible candidates for further investigation. It would also be interesting to evaluate the effect on the network of the removal of one of these genes *in silico*, and compare those to knock-out or knock-down experiments *in vitro* or *in vivo*. The results of these experiments could help to improve the accuracy of the model, while the results *in silico* could help to make predictions on the effect of certain drugs that inhibit a specific gene.

4.2. Limitations

The fact that not one single NI method could be suggested for further analysis in single-cell data could have a number of possible causes. If, in the future, NI methods would be evaluated for use in single-cell data, the following remarks should be taken into account:

For the glioblastoma dataset, the raw data was obtained from the SRA and preprocessed; this process includes mapping the raw reads, generating counts and normalizing these counts. Mapping of the 25 base pair paired-end raw reads showed mapping rates that were lower than expected. A possible cause was the length of the reads; shorter reads are statistically more likely to have more than one match with the transcriptome, causing them to be excluded by the used mapping tool. Mapping rates of around 30% indicate that 70% of the reads is not used to generate the expression matrix. The effect of this on the inference of networks is difficult to assess. Secondly, normalization of the counts was done using TPM, the method of choice for sequencing depth correction in population based data. This method is based on the assumption that the amount of RNA in each sample is approximately the same, which does not hold true for scRNA-seq data. A suggested method to deal with this is the use of extrinsic spike-ins to estimate the amount of RNA in each sample. As these spike-ins had not been added to the raw data, it was not possible to correct for the difference in RNA content between the samples.

One of the main sources of variability between single-cell and population RNA-seq is biological variability, caused by periodic processes in the cell, such as the cell cycle. In a population of cells, this variability is masked by the fact that expression levels of

the genes are in fact the average expression levels over the total population of cells. Though this biological confounder has been reported, only one method, scLVM, has been developed that allegedly removes these effects. Due to lack of evidence, this method was not used in the current analysis.

For the melanoma data, an expression matrix containing the normalized counts was obtained, as the raw data had not yet been made available. The mapping of 30 base pair paired-end reads was done using bowtie, allowing single-base mutations. No mapping rates were available. Normalization was done using TPM. If possible in the future, it would be interesting to preprocess the melanoma data using the same workflow as was used for the glioblastoma data. This way, results will be more comparable and possible computational bias associated with the preprocessing method, will be eliminated.

In terms of the method used to score the NI methods, there are also several improvements which can be made. Firstly, in the construction of the gold standard for the glioblastoma data, interaction prediction databases were used. As these interactions are predicted with different levels of confidence, these are not suited for a gold standard. Including these genes could lead to a lot of seemingly false negative predictions, giving the methods a worse score than they have in reality. This was taken into account when constructing the gold standard for melanoma. The authors manually extracted the genes from PubMed abstracts, drastically increasing levels of confidence. Interacting partners for these genes were inferred from PINA, a protein-protein interaction database containing data from six manually curated interaction databases. However, the higher confidence levels of the gold standard did not have an influence on the score. This could indicate that the scoring method, AUROC, should be reevaluated.

Another issue could be that the importance of a regulator should not only be measured by the weight of its edges, but also by the number of edges. To this end, an analysis of possible hubs was executed for the melanoma data, ranking the regulators by their degree, while also taking into account the weight of the edges.

Additionally, none of the methods takes into account the origin (patient) of each of the samples. Including this information in the analysis and combining this with metadata on the progress of the disease in this particular patient, could lead to new biological insights.

Even though CLR identified the highest number of gold standard genes, it is not possible to recognize this as the best method based on this data. If, however, a way is found to overcome all challenges mentioned above, the method to identify hubs shows great promise as a tool for prioritizing genes.

4.3. Future aspects of network inference in single-cells

The main priority for the near future of scRNA-seq data analysis should be the development of appropriate tools to deal with specific issues of this new data type, for instance cell cycle effects and differences in RNA content between the samples.

Once the problems concerning generation of correct expression values have been tackled, more specific methods can be developed to obtain new biological insights based on this data.

Network inference in single cells, specifically, could focus on predicting which pathways are responsible for the cell's phenotype. Uncovering subclones in heterogeneous tissue samples such as tumors could lead to a better understanding of driver genes playing a role in therapy resistance or the development of metastases. These *in silico* uncovered genes could be validated *in vitro* and *in vivo*, ultimately leading to the discovery of a more effective therapy.

Similar to the results of the DREAM challenge, no single method performs optimally based on this analysis. The solution proposed by the authors is a community-based method, combining information from different methods to generate the most robust networks; this is an approach worth investigating.

One of the main problems of scRNA-seq is its cost. Currently, however, the field is evolving towards the development of higher-throughput methods, analyzing 1000-5000 cells but with a much lower coverage using methods like, for instance, drop-seq. Increasing the amount of samples could have beneficial effects on the statistical power of the tests, although the lower coverage might present researchers with new computational challenges. Additionally, the lower cost and high speed of these methods could increase the amount of experiments, making this data more accessible, which could benefit the progress in the field.

Part 5: Samenvatting discussie

5.1. Conclusie

Dankzij recente ontwikkelingen in het veld van de *microfluidics* is het mogelijk geworden om het transcriptoom te analyseren van één enkele cell, waarbij meerdere cellen in parallel kunnen geanalyseerd worden. Deze techniek wordt *single-cell RNA-seq* genoemd (scRNA-seq). In ontwikkelingsbiologisch en immunologisch onderzoek heeft het gebruik van deze methode al geleid tot enkele veelbelovende resultaten en ook in kanker onderzoek is deze manier van transcriptoom analyse aan het opkomen, meer specifiek om aspecten van tumor heterogeniteit bloot te leggen. Tot op de dag van vandaag is er geen consensus over de computationele methodes die gebruikt dienen te worden bij de analyse van dit nieuwe data type. Voor bepaalde stappen binnen het analytische proces wordt gesuggereerd dat methodes die ontwikkeld werden voor de analyse van de RNA-seq data van een groep cellen, *population* RNA-seq, zouden moeten gebruikt kunnen worden, hoewel aanpassingen nodig zullen zijn. Voor andere stappen zullen nieuwe methodes ontwikkeld moeten worden.

Eén van de manieren waarop nieuwe biologische inzichten verkregen kunnen worden aan de hand van expressie data, is door het reconstrueren van gen regulatorische netwerken (GRNs) aan de hand van deze data. Bepaalde auteurs hebben gerapporteerd dat netwerk inferentie (NI) methodes voor population RNA-seq ook gebruikt kunnen worden voor *single-cell* data, mits enkele aanpassingen die rekening houden met verhoogde niveaus van technische ruis en bepaalde verstorende factoren. De hoofddoelstelling van deze thesis was het vergelijken van het vermogen van verschillende geselecteerde NI methodes om de juiste netwerken te infereren op basis van *single-cell* data.

Aangezien scRNA-seq een vrij recent ontwikkelde techniek is, bestaan er weinig publicaties over de vergelijking van deze data met population RNA-seq. Het eerste deel van deze thesis bestond uit een vergelijking van de gemiddelde *single-cell* expressie waarden met de *population* expressie waarden. Plots van de correlatie tussen deze waarden toonden aan dat de gemiddelde expressie waarden in de single-cells enigermate hoger waren dan die in de *population* stalen; dit gold voornamelijk voor de genen die laag tot expressie komen. Het aantal genen dat helemaal niet tot expressie kwam was ook hoger in de population stalen. Een mogelijke verklaring hiervoor is het verschil in *sequencing depth*. Vervolgens werden de stalen gevisualiseerd in drie dimensies met behulp van dimensionaliteitsreductie. Hierbij werd een groep van *outliers* gevonden die van lagere kwaliteit bleken te zijn. Het uitsluiten van deze stalen leidde tot betere clustering resultaten en toonde aan dat de stalen van lage kwaliteit uitgesloten moesten worden bij verdere analyses.

Het tweede deel van deze thesis bestond uit het evalueren van de vier geselecteerde NI methodes. Hierbij werd scRNA-seq data gebruikt van twee verschillende kanker types, nl. *melanoma* en *glioblastoma*; deze data was publiekelijk beschikbaar. Ten

eerste moest een gouden standaard van gekende interacties in deze twee kanker types opgesteld worden. Voor *glioblastoma* werd hiervoor data van verschillende *pathway* en interactie-predictie databanken geïntegreerd. ROC analyse toonde aan dat geen enkele van de NI methodes erin slaagde om de juiste interacties te voorspellen. Voor melanoma werden de genen van MGDB gebruikt. Opnieuw kon geen enkel van de methodes de interacties op een significante manier voorspellen. Een evaluatie van de hubs in het netwerk, de regulatoren met een hoge connectiviteit, konden echter wel enkele belangrijke genen voorspeld worden. Voor elke methode werden de beste tien *hubs* vergeleken. Bij het opstellen van de top tien werd niet alleen rekening gehouden met het aantal connecties, maar ook met hun voorspelde belang. In totaal bevatten deze lijsten 19 verschillende genen, waarvan zes ook in de gouden standaard opgenomen waren. Een oppervlakkige literatuurstudie toonde wel aan dat voor veel van de 13 andere genen aangetoond was dat ze een rol speelden in oncogenese, en sommige zelfs meer specifiek in melanogenese.

5.2. Limitaties

Geen enkele van de NI methodes kon naar voor geschoven worden als de beste methode voor gebruik in single-cell analyse. Indien in de toekomst opnieuw een evaluatie van NI methodes zou gebeuren, moet voor het verwerken van de datasets rekening gehouden worden met de factoren die hieronder opgesomd worden:

- Bij het mappen van de *glioblastoma reads* op het humaan transcriptoom, werd een zeer lage graad van mapping geobserveerd. Vermoedelijk komt dit doordat ze *reads* zeer kort zijn, nl. 25 basenparen, waardoor ze mogelijks op meer dan één plaats in het transcriptoom mappen. De gebruikte tool zal dergelijke fragmenten verwerpen. Het is moeilijk om het effect hiervan op de NI in te schatten.
- Normalizatie van de *counts* van de *glioblastoma* data werd gedaan door middel van TPM. Hierbij wordt er van uit gegaan dat de hoeveelheid RNA in elk staal bij benadering hetzelfde is; deze veronderstelling klopt niet voor scRNA-seq data. Het gebruik van *spike-ins* zou hiervoor kunnen corrigeren, maar in deze dataset waren geen *spike-ins* aanwezig.
- Een deel van de variabiliteit tussen single-cell en population RNA-seq is biologische variabiliteit veroorzaakt door periodieke processen in de cel, zoals bijvoorbeeld de cel cyclus. Deze processen hebben geen invloed op de expressieniveaus bij *population* RNA-seq omdat hierbij het gemiddelde niveau over een groep van cellen wordt genomen. De methode scLVM zou deze effecten verwijderen, maar doordat deze methode nog maar recent ontwikkeld is en zijn effectiviteit nog niet uitgebreid bewezen is, werd ervoor gekozen om deze niet te gebruiken.

- Voor de *melanoma* data waren er ten tijde van de analyse nog geen *raw reads* ter beschikking. Hierdoor moest gewerkt worden met data die gepreprocessed was door de auteurs. Indien deze *raw* data in de toekomst ter beschikking komt, zou het interessant zijn om zelf de preprocessing uit te voeren, om technische verschillen tussen de datasets uit te sluiten.

Verder zijn er op vlak van de methodologie ook nog verschillende verbeteringen mogelijk:

- Het gebruik van interactie-predictie databanken voor het construeren van een gouden standaard is niet doeltreffend. Interacties in dergelijke databanken hebben een verschillend niveau van betrouwbaarheid, wat zou kunnen leiden tot veel schijnbaar vals negatieve resultaten. Een aantal van deze interacties zullen nl. onjuist zijn, waardoor het onmogelijk is voor de NI methodes om ze te voorspellen.
- Met het bovenstaande werd rekening gehouden bij het construeren van de gouden standaard voor de *melanoma* dataset. Desalniettemin was er geen significant verschil bij het scoren van de verschillende NI methodes; geen enkele methode scoorde beter dan random. Dit resultaat geeft aan dat de gebruikte score methode, AUROC, misschien geherevalueerd moet worden voor gebruik in deze context.
- Tot nu toe werd geen rekening gehouden met de herkomst van elk staal. Het integreren van patient data, gaande van biometrische kenmerken tot het verloop van de ziekte, zou tot nieuwe biologische inzichten kunnen leiden.

Tenslotte heeft de analyse van de *melanoma* data aangetoond dat het belang van een regulator misschien niet enkel af hangt van het gewicht van zijn interacties, maar ook van het aantal interacties. Deze laatste methode lijkt veelbelovend voor het naar voor schuiven van nieuwe kandidaat regulatoren die verder onderzocht kunnen worden *in vitro* en *in* vivo.

5.3. Toekomstige ontwikkelingen

De prioriteit voor de analyse van scRNA-seq data zou moeten liggen bij het ontwikkelen van geschikte methodes die kunnen omgaan met specifieke aspecten van dit nieuwe data type, zoals bijvoorbeeld effecten van de cel cyclus en verschillen in RNA inhoud tussen de stalen. Eens de problemen omtrent het genereren van een correcte expressie matrix opgelost zijn, kunnen meer specifieke methodes ontwikkeld worden die helpen bij het verkrijgen van nieuwe biologische inzichten op basis van deze data.

Het infereren van netwerken op basis van single-cell data zou erop gericht kunnen zijn om te voorspellen welke pathways verantwoordelijk zijn voor het fenotype van een bepaalde cel. Het blootleggen van subklonen in heterogene weefsel stalen zoals tumoren zou kunnen leiden tot een beter inzicht in de driver genen die een rol spelen in therapie resistentie, of het ontwikkelen van metastasen. Deze *in silico*

ontdekte genen zouden gevalideerd kunnen worden *in vitro* en *in vivo*. Het ultieme doel zou zijn om een meer effectieve behandeling te ontwikkelen.

Het feit dat geen enkel van de methodes, volgens experimenten die hier uitgevoerd werden, optimaal presteerde is een resultaat dat ook gezien werd door de auteurs van de DREAM challenge. Zij stelden een meta-methode voor waarbij NI data van verschillend methodes geïntegreerd wordt. Een dergelijke benadering zou ook in dit geval nagegaan kunnen worden.

Huidige trends wijzen er op dat er binnenkort scRNAseq methoden, bvb. drop-seq, in gebruik zullen genomen worden waarbij er veel meer cellen (1000 à 5000) gesequeneerd worden maar aan een veel lagere *coverage*. Een groter aantal cellen per dataset zou de statistische *power* vergroten. De lagere *coverage* zal onderzoekers mogelijks voor nieuwe computationele uitdagingen stellen. Dergelijke methodes zijn goedkoper dan de bestaande methodes en werken snel, wat ervoor zou kunnen zorgen dat het aantal experimenten snel stijgt, met als gevolg dat meer onderzoekers zich kunnen toeleggen op scRNA-seq, wat de vooruitgang in dit veld ten goede zou komen.

Part 6: Materials and Methods

6.1. Glioblastoma data

6.1.1. Download data

The raw sequencing data from a publicly available dataset of 5 patients with glioblastoma multiforme (GSE57872) (Patel et al. 2014) was downloaded from NCBI's sequence read archive (SRA), using the SRA accession number SRP042161. This dataset contains the expression profiles of 576 single glioblastoma cells, 192 cells from glioblastoma cell lines and a population glioblastoma sample for each patient, containing 2000–10000 cells. The reads were generated using the SMART-seq protocol; this resulted in paired-end reads of 25 base pairs. The metadata for this dataset was composed combining the information on the runs from the SRA run selector and the metadata associated with the GEO samples. The latter was accessed by using the getGEO function from the Bioconductor package GEOquery.

6.1.2. Quality control

The quality of the sequencing data was evaluated using FastQC. This tool requires that the input data is in fastQ, BAM or SAM format. The downloaded runs, which were in the sra format, were converted to fastQ using the sra toolkit fastq-dump tool. The option --split-files was added to split forward and reverse reads.

For each of the runs, 8 of the criteria were evaluated using a fail/pass system and all results were combined in a single table. When checking the per base sequence content, the first 9 bases were excluded from the analysis, avoiding warnings related to a technical bias caused by library production. The sequence length distribution is excluded from the analysis, as well as k-mer and adapter content.

Besides the quality control criteria included in the fastQC analysis, three alternative criteria were included: the number of reads mapped, the percentage of reads mapped and the number of genes expressed. Passing levels were set respectively to 100000, 7% and 5000.

6.1.3. Mapping and normalization

The transcripts were mapped to the human reference transcriptome, assembly version GRCh38, using salmon beta version 0.6.1 in quasi mapping mode. First, the transcripts were indexed, setting k-mer length to 11. Secondly, the paired-end reads were quantified against the transcriptome with library type matching and unstranded (MU) using ensembl transcript identifiers, version 83. Genes that did not have at least two samples with 5 counts are removed. The counts were normalized by calculating the amount of transcripts per million (TPM). Subsequently, the

expression levels were log transformed and the ensembl transcript identifiers were mapped to ensembl gene identifiers.

6.1.4. Filter expression matrix

The mapping of the transcripts resulted in an expression matrix with 875 samples, including single cell, population and cell line samples, and 54435 transcripts, mapped to their corresponding gene. Only the single cell samples that had passed 8 of the evaluated quality control criteria were accepted for further analysis, retaining a total of 629 samples.

The genes that did not have a transcript and gene biotype of *protein coding* were excluded from the analysis. Genes that were expressed in at least 10% of the samples were ranked according to their standard deviation. The top 8000 genes with the highest standard deviation were used for network inference.

6.2. Melanoma data

6.2.1. Download data

The GEO series GSE72056 contains 4645 single cells from 19 melanoma tumors from patients with a range of clinical backgrounds. Malignant as well as nonmalignant cells were sequenced. The processed expression matrix (Tirosh et al. 2016, supplementary materials) was downloaded directly from the GEO platform. The metadata were accessed through the bioconductor GEOquery package.

6.2.2. Filter expression matrix

To create the expression matrix for network inference, only the malignant samples were included. The genes that did not have *protein coding* as gene biotype, were excluded from analysis. The remaining genes that were expressed in at least 10% of the samples were ranked according to their standard deviation among all samples. Due to the size of the dataset, only the top 6000 genes with the highest standard deviation were used for network inference.

6.3. Control data

6.3.1. Glioblastoma

A microarray analysis in the context of an experiment looking for candidate genes involved in glioblastoma relapse (GEO series GSE67089) included 5 astrocyte control samples (Mao et al. 2013). These samples were hybridized onto Affymetrix human genome U219 arrays. Log transformed GC-Robust Multiarray Averaging (GC-RMA) normalized expression values were available through GEO. The probe identifiers

were mapped to gene symbols. When several probes mapped to one gene, the average expression levels were calculated. Vice versa, no gene mapped to more than one probe.

For network inference, only genes that were in the top 8000 genes with the highest standard deviation in the glioblastoma single cell RNA-seq dataset, were retained.

6.3.2. Melanoma

The GEO series GSE3189 is a microarray analysis of 7 normal skin, 18 nevi and 45 melanoma samples (Talantov et al. 2005). These samples were hybridized onto Affymetrix human genome U133A arrays. The raw expression data and series metadata were accessed by using the bioconductor GEOquery package. Only the normal skin samples were regarded for further analysis and the probe identifiers were mapped to gene symbols. When several probes mapped to one symbol, the average expression levels were calculated. Vice versa, no symbol mapped to more than one probe.

For network inference, only genes that were in the top 6000 genes with the highest standard deviation in the melanoma single cell RNA-seq dataset, were retained.

6.4. Network inference

6.4.1. Regulators

A list of probable and possible human transcription factors was published by Vaquerizas and colleagues (Vaquerizas et al. 2009). Proteins containing selected DNA binding domains were mapped to the human genome (Ensembl version 51), which resulted in 1960 loci. These loci were classified as 'a', 'b', 'c' or 'x', according to the authors' confidence in their transcription factor functionality. 27 probable transcription factors were added from other databases and classified as 'other' (Vaquerizas et al. 2009). To construct the list of regulators for network inference, transcription factors with class 'a', 'b' and 'other' were retained. Their HCNC symbols were mapped to ensembl identifiers, version 83. This resulted in a list of 1333 probable transcription factors.

A more recent list of predicted human transcription factors was downloaded from the AnimalTFDB version 2.0 (Zhang et al. 2015) and from Semantic catalogue of Samples, Transcription initiation And Regulators (SSTAR).

The final list of transcription factors consisted of a union of these three lists and contained 1798 transcription factors, of which 1198 were present in all three lists.

6.4.2. Gold standard

To evaluate the interactions found in the glioblastoma and melanoma dataset, a gold standard of important interactions in these types of cancer was composed.

The TCGA pilot paper on glioblastoma (McLendon et al. 2008) reported three main pathways involved in glioblastoma: the p53 and retinoblastoma tumor suppressor pathway, dysregulation of cell growth via mutations in RTK genes and activation of the PI3K pathway. The human pathways *PI3K-Akt signaling pathway* (hsa04151), *cell cycle* (hsa04110) and *p53 signaling pathway* (hsa04115) were downloaded from the KEGG database and the interactions with type protein-protein interaction (PPrel) and gene expression interaction (GErel) were written to a file. From the KEGG disease database, interactions predicted to be involved in glioma (disease H00042) were also added to the list of predicted interactions. Finally, stringDB was searched for proteins interacting with Mig-6 (Ying et al. 2010), Bcl2-L12 (Stegh et al. 2010) and PTEN and interactions between genes predicted to be involved in glioblastoma by PathCards (Belinky et al. 2015). This resulted in a total of 383 interactions predicted to be important in glioblastoma.

Further, four clinically relevant subtypes of glioblastoma were identified using TCGA data (Verhaak et al. 2010) . A list of 210 signature genes for each of these types is available, but only 679 of these genes could be mapped to an ensembl identifier. Finally, the KEGG pathway microRNAs in cancer (hsa05206) was consulted. Certain microRNAs are upregulated in glioblastoma, inhibiting tumor suppressor gene activity and the downregulation of other microRNAs may cause oncogene activation. 8 genes that are possibly upregulated and 7 genes that are possibly downregulated were added to the list of relevant genes.

An analysis of 331 melanoma patients by TCGA (TCGA. 2015) reported four melanoma subtypes: mutant BRAF, mutant RAS, mutant NF1 and Triple-WT. In the BRAF subtype, amplifications of BRAF and MITF were seen. The RAS subtype showed MAPK activation and AKT3 overexpression. In the NF1 subtype, a loss of NF1 function was described. From MGDB, a list of 527 genes reported in literature to be involved in melanoma, was downloaded. These genes had been manually curated by the authors from 682 PubMed abstracts. A distinction was made between 422 coding and 105 non-coding genes. For further analysis, only the coding genes were obtained. Secondly, the file containing interaction information for these genes downloaded. This information had been obtained by the authors through the Protein Interaction Network Analysis (PINA) (Cowley et al. 2012) platform.

6.4.3. ARACNE and CLR

The Bioconductor package minet implements several algorithms for mutual information network inference, including ARACNE and CLR. In the first step, a mutual information matrix (mim) is build based on the expression matrix, computing mutual information between all pairs of genes. The default entropy estimator is Spearman's

correlation. Secondly, this mim is used by the algorithms as input. A weighted adjacency matrix is returned. By default, the ARACNE algorithm removes the least significant edge in a triplet of nodes (Equation 2). The threshold for removal (eps) is set to 2, meaning that the weakest edge is only removed if its weight is 2 below the minimum of the other two edges.

Equation 2: Removal of the weakest edge in ARACNE – by default, eps is 0

$$Nodes(i,j,k): weakest\ edge\ (ij)\ is\ removed\ if\ (ij) < \min\{(ik),(jk)\} - eps$$

The weights of interactions between regulators and the set of all genes were sorted and the top 100000 were retained for further analysis.

6.4.4. GENIE3

A python2 implementation of GENIE3 was downloaded and adapted for use in python3. The default parameters of this implementation were used. An array with conditions in the rows and genes in the columns was passed to the algorithm, together with a list of gene names and a list of regulators. The default tree method and number of trees is respectively Random Forest and 1000 trees. K, the number of selected attributes at each node of one tree is the square root of the number of regulators. The output is the sorted ranking of links that are directed from the candidate regulators. The top 100000 links were retained for further analysis.

6.4.5. Absolute value of Pearson's correlation

To calculate correlation between all samples, the python module pandas was used. The tsv file containing the expression matrix was read into a pandas DataFrame. Next, the function corr was used to calculated pairwise correlation between all genes. The default method is Pearson correlation. To exclude self-to-self correlation, all values on the diagonal of the correlation matrix were set to zero. Finally, the absolute values of the correlation coefficients were ranked and the top 100000 edges were retained for further analysis.

6.4.6. Hubs

For the melanoma dataset, an analysis of the hubs was executed. The change in degree with descending edge weight is calculated and plotted for each regulator, and the area under the curve is calculated. Secondly, a gene set test is used to verify whether the set of regulators is ranked more highly as opposed to a randomly chosen set of genes. Significance is verified using a Wilcoxon signed rank test. The p-values are corrected for multiple testing using the method of Benjamini-Hochberg and the top 10 genes with the most significant p-value are compared between the different methods.

6.5. Pooled single-cell vs. population RNA-seq

6.5.1. Pooling

After the conversion of the sra files to the fastq format, each run was split in two separate files, one containing the forward reads and another containing the reverse reads. To combine all the raw sequencing information per patient, all forward respectively reverse reads of cells from that patient were written to a single file. Then, counts were generated and normalized for these pooled samples using the same method as previously described (*supra* 5.1.3 Mapping and normalization).

6.5.2. Correlation

The correlation between the pooled and population samples was evaluated by plotting the expression levels for each gene in pooled versus population samples on respectively the x- and y-axis. Spearman and Pearson correlation coefficients, as well as the coefficient of determination (R^2) were evaluated.

6.5.3. Comparison of the sequencing depth

To compare the sequencing depth, the following equation was used (Equation 3):

Equation 3: Sequencing depth

$$\frac{Number\ of\ sequences * read\ length}{Genome\ size}$$

With read length 25 and genome size 3234.83 Mb. For the pooled samples, the number of sequences was calculated as the sum of the number of sequences of the individual single-cells.

6.5.4. Dimensionality reduction

Before reducing the dimensionality of the data, the top 50% most variant genes were selected. First, the different expression matrices - single-cells, population and pooled – were combined to a single expression matrix containing only genes present in all datasets. Secondly, the variances were calculated for all genes and sorted from high to low, finally selecting the top 50% genes. The correlation distances were calculated for each pair of genes. Dimensions in the gene direction were reduced to 3 using classical – Torgerson - multidimensional scaling (MDS).

Dimensionality reduction on the full dataset showed a group of outliers. Differential gene expression analysis was used to identify differentially expressed genes. Genes were defined as differentially expressed if their log fold change was larger than 4 and the p-value smaller than 0.01. The p-values were corrected for multiple testing

usingthe method of Benjamini-Hochberg. Difference in quality between these outliers and the other samples was assessed using the Mann-Whitney U test ($\alpha < 0.05$).

References

Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature medicine [Internet]. Available from: http://www.nature.com/nm/journal/v22/n1/abs/nm.3984.html.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Babu M, Luscombe N, Aravind, Gerstein M, Teichmann S. 2004. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology.283–291.

Barabási A-L, Oltvai Z. 2004. Network biology: understanding the cell's functional organization. Nature reviews Genetics.101–13.

Barabási A-L. 2009. Scale-free networks: a decade and beyond. Science (New York, NY).412–3.

Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, Lancet D. 2015. PathCards: multi-source consolidation of human biological pathways. Database (Oxford). 2015.

Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE. 2014. Comparing reference-based RNA-Seq mapping methods for non-human primate data. BMC Genomics. 15:570.

Bleeker FE, Molenaar RJ, Leenstra S. 2012. Recent advances in the molecular understanding of glioblastoma. Journal of neuro-oncology [Internet]. 108:11–27. Available from: http://link.springer.com/article/10.1007/s11060-011-0793-0

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, et al. 2013. The Somatic Genomic Landscape of Glioblastoma. Cell. 155:462–477.

Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 33:155–60.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MWW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17:13.

Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J. 2012. PINA v2.0: mining interactome modules. Nucleic Acids Res. 40:D862–5.

Cummins DL, Cummins JM, Pantle H. 2006. Cutaneous malignant melanoma. Mayo Clinic … [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S0025619611618983

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 5:e8.

Gaublomme JT, Yosef N, Lee Y, Gertner RS, Yang LV, Wu C, Pandolfi PP, Mak T, Satija R, Shalek AK, et al. 2015. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. Cell. 163:1400–12.

Genomic Classification of Cutaneous Melanoma. 2015. Cell. 161:1681–96.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. cell [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S0092867400816839

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. Cell. 144:646–74.

Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜVV. 2013. Benchmarking short sequence mapping tools. BMC Bioinformatics. 14:184.

Islam S, Zeisel A, Joost S, Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 11.

Kim J, Kolodziejczyk A, Illicic T, Teichmann S, Marioni J. 2015. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nature communications. 6:8687.

Kim K-TT, Lee HW, Lee H-OO, Kim SC, Seo YJ, Chung W, Eum HH, Nam D-HH, Kim J, Joo KM, Park W-YY. 2015. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 16:127.

Kwon T. 2015. Benchmarking Transcriptome Quantification Methods for Duplicated Genes in Xenopus laevis. Cytogenet Genome Res. 145:253–64.

Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP, Bretscher A, Ploegh H, Amon A. 2013. Molecular cell biology. 7th ed. New York: W.H. Freeman and Company.

Mao P, Joshi K, Li J, Kim S-H, Li P, Santana-Santos L, Luthra S, Chandran U, Benos P, Smith L, et al. 2013. Mesenchymal glioma stem cells are maintained by activated glycolytic metabolism involving aldehyde dehydrogenase 1A3. Proceedings of the National Academy of Sciences.8644–8649.

Marbach D, Costello J, Küffner R, Vega N, Prill R, Camacho D, Allison K, Aderhold A, Allison K, Bonneau R, et al. 2012. Wisdom of crowds for robust gene network inference. Nat Methods. 9:796–804.

Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A. 2006. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. Bmc Bioinformatics. 7:S7.

McLendon R, Friedman A, Bigner D, Meir E, Brat D, Mastrogianakis G, Olson J, Mikkelsen T, Lehman N, Aldape K, et al. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 455:1061–1068.

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, et al. 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 33:269–76.

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods.621–628.

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahil DP, Nahed BV, Curry WT, Martuza RL, et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. CANCER GENOMICS. 344.

Penna E, Orso F, Cimino D, Tenaglia E, Lembo A, Quaglino E, Poliseno L, Haimovic A, Osella-Abate S, Pittà C, et al. 2011. microRNA-214 contributes to melanoma tumour progression through suppression of TFAP2C. The EMBO Journal.1990–2007.

Phillips H, Kharbanda S, Chen R, Forrest W, Soriano R, Wu T, Misra A, Nigro J, Colman H, Soroceanu L, et al. 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 9:157–173.

Prado M, Frampton A, Stebbing J, Krell J. 2015. Single-cell sequencing in cancer research. Expert review of molecular diagnostics.

Schlitzer A, Sivakamasundari V, Chen J, Sumatoh HR, Schreuder J, Lum J, Malleret B, Zhang S, Larbi A, Zolezzi F, et al. 2015. Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. Nat Immunol. 16:718–28.

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, et al. 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 510:363–9.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics. 14:618–630.

Shirley SH, Greene VR, Duncan LM, Torres Cabala CA, Grimm EA, Kusewitt DF. 2012. Slug expression during melanoma progression. Am J Pathol. 180:2479–89.

Stears RL, Martinsky T, Schena M. 2003. Trends in microarray analysis. Nature medicine [Internet]. Available from: http://arrayit.com/confidential/Trends_Microarray_Analysis.pdf

Stegh A, Brennan C, Mahoney J, Forloney K, Jenq H, Luciano J, Protopopov A, Chin L, DePinho R. 2010. Glioma oncoprotein Bcl2L12 inhibits the p53 tumor suppressor. Gene Dev. 24:2194–2204.

Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 16:133–45.

Talantov D, Mazumder A, Yu J, Briggs T, Jiang Y, Backus J, Atkins D, Wang Y. 2005. Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clinical cancer research : an official journal of the American Association for Cancer Research.7234–42.

Tirosh I, Izar B, Prakadan S, Wadsworth M, Treacy D, Trombetta J, Rotem A, Rodman C, Lian C, Murphy G, et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 352:189–196.

Tomczak K, ska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology [Internet]. 19:A68. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322527/

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 509:371–5.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 10:252–63.

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller RC, Ding L, Golub T, Mesirov JP. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell [Internet]. 17:98–110. Available from: http://www.sciencedirect.com/science/article/pii/S1535610809004322

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences [Internet]. Available from: http://link.springer.com/article/10.1007/s12064-012-0162-3

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57–63.

Watson IR, Wu CJ, Zou L, Gershenwald JE, Chin L. 2015. Genomic classification of cutaneous melanoma. Cancer Research [Internet]. Available from: http://cancerres.aacrjournals.org/content/75/15_Supplement/2972.short

Ying H, Zheng H, Scott K, Wiedemeyer R, Yan H, Lim C, Huang J, Dhakal S, Ivanova E, Xiao Y, et al. 2010. Mig-6 controls EGFR trafficking and suppresses gliomagenesis. Proc Natl Acad Sci USA. 107:6912–7.

Zhang D, Zhu R, Zhang H, Zheng C-HH, Xia J. 2015. MGDB: a comprehensive database of genes involved in melanoma. Database (Oxford). 2015.

Zhang H-MM, Liu T, Liu C-JJ, Song S, Zhang X, Liu W, Jia H, Xue Y, Guo A-YY. 2015. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. 43:D76–81.

Attachments

I. Supplementary tables

Patient	Mapping rate (%)
Single cell mRNA-seq_MGH26	16,73
Single cell mRNA-seq_MGH28	19,71
Single cell mRNA-seq_MGH29	24,09
Single cell mRNA-seq_MGH30	15,52
Single cell mRNA-seq_MGH31	11,86

Supplementary table 1: Average mapping rates of all single-cell samples for each patient – Patient MGH31 has a lower mapping rate than the other patients, suggesting the sequencing quality is lower for these cells

	number of QC passes	**Number of reads mapped**	**Total sequences**	**Percentage mapped**	**Genes expressed**	**Non-zero expression**
Samples subset	7,52	91124,07	2070665,22	0,05	8438,96	3882,07
Other single-cell samples	10,26	1138425,69	6160122,19	0,18	27127,79	9881,20
p-value	2,71E-30	2,708E-30	1,423E-16	2,22E-22	2,03E-24	1,3E-24

Supplementary table 2: Mean values of several quality control features for the samples of the subset versus the other single-cell samples – Each of the QC aspects are compared in the two groups using the Mann-Whitney U test for non-normally distributed data

	Pooled single-cell		**Population**	
	total sequences	sequencing depth	total sequences	sequencing depth
MGH26	9,96E+08	7,70E-01	1,54E+07	1,19E-02
MGH28	6,95E+08	5,37E-01	2,28E+07	1,76E-02
MGH29	4,90E+08	3,79E-01	8,43E+06	6,51E-03
MGH30	6,39E+08	4,94E-01	2,21E+07	1,71E-02
MGH31	6,22E+08	4,81E-01	7,70E+07	5,95E-02

Supplementary table 3: Total amount of sequences and the sequencing depth for each patient compared in pooled samples and population samples – The sequencing depth of the pooled samples is higher than in the bulk samples, which could explain the difference in expression levels (*supra* 3.2.2 Correlation).

II. Code

The code used to obtain the results presented in this thesis, can be consulted at:

https://github.ugent.be/cdvogela/masterthesis