HUMBOLDT-UNIVERSITÄT ZU BERLIN

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT I
INSTITUT FÜR BIOLOGIE

MASTERARBEIT

ZUM ERWERB DES AKADEMISCHEN GRADES
MASTER OF SCIENCE

„Transkriptionsanalyse einer einzelnen Zelle mittels `next generation sequencing´"

„Single cell transcriptome analysis using next generation sequencing"

Vorgelegt von

Mirjam Blattner
geb. am 26.08.1984 in Ravensburg

angefertigt am Max Planck Institut für molekulare Genetik Berlin

Berlin, Dezember 2010

1. Gutachter: Prof. Dr. Hans Lehrach
2. Gutachter: Prof. Dr. Christian Schmitz-Linneweber

Acknowledgments

First of all, I would like to thank Professor Hans Lehrach for giving me the opportunity to realize my master's thesis work in his department.

Thanks to Dr. Andreas Dahl who supervised me throughout the project and for putting so much trust in me. It was great experience learning, working and discussing with him in a very nice atmosphere.

I am grateful to Professor Christian Schmitz-Linneweber for being my university supervisor and reviewer of my thesis.

I also want to thank my colleagues and friends, Florian Mertes, Robert Querfurth, Hannah Müller, Karin Habermann and Axel Fischer for all the help and support they gave me.

Special thanks to Sean O´Keeffe correcting my English and being there for me.

Wir haben keine Probleme, das sind nur neue Herausforderungen!

Zusammenfassung in deutscher Sprache

Die Heterogenität von Geweben insbesondere in der Tumorforschung ist ein zentrales Problem bei der Transkriptomanalyse. Daher fokussiert sich die Wissenschaft in den vergangenen Jahren immer mehr auf die Entwicklung von Methoden zur Untersuchung einzelner Zellen. Durch Einzelzellanalysen wird versucht, neue Einsichten in biologische Vorgänge von gesunden und kranken Zellen zu erhalten.

Dabei stellt man sich in der Transkriptomanalyse der Herausforderung, geringstes Startmaterial zu preparieren, amplifizieren und sukzessive zu untersuchen.

In der vorliegenden Arbeit wurden zwei grundsätzlich verschiedene Amplifikations-Ansätze mit schrittweiser Analyse durch next-generation sequencing verglichen: I. Die exponentielle Amplifikation mittels Polymerase-Kettenreaktion (PCR) und II. die lineare Amplifikation.

Arbeitsabläufe für Einzelzellgewinnung, Zellaufschluss, cDNA Synthese, cDNA Amplifikation und Präparation von next-generation sequencing libraries wurden für die jeweiligen Ansätze entwickelt. Es konnte erfolgreich gezeigt werden, dass eine transkriptionelle Analyse geringer Zellanzahlen sowohl mittels linearer als auch exponentieller Amplifikation erfolgreich durchführbar ist. Höchste Amplifikationsraten von bis zu 10^6 konnten durch exponentielle Amplifikation erreicht werden. Die lineare Amplifikation hat sich als die reproduzierbarere Methode gezeigt. Die Analyse der next-generation sequencing Daten zeigte in Einzelzell-Proben mindestens eine nachweisbare Expression von 16.000 Genen. Die gefundene Varianz zwischen den Proben weist jedoch auf die Notwendigkeit des Arbeitens mit vielen biologischen Replikaten hin. Zusammenfassend kann gesagt werden, dass Transkriptom-Einzelzellstudien mittels next-generation sequencing durchführbar sind jedoch weitere Verbesserungen der beiden verglichenen Protokolle hin zu einem größeren Anteil an sequenzierten Transkripten anstehen. In naher Zukunft können beispielsweise durch den Vergleich einzelner Krebszellen mit gesunden Zellen neue Grundlagen für eine Verbesserung von Prognose und Diagnose bei geschaffen werden.

Abstract

The heterogeneity of tissues, especially in cancer research, is a central issue in transcriptome analysis. In recent years, research has primarily focused on the development of methods for single cell analysis. Single cell analysis aims at gaining (novel) insights into biological processes of healthy and diseased cells.

Some of the challenges in transcriptome analysis concern low abundance of sample starting material, necessary sample amplification steps and subsequent analysis.

In this study, two fundamentally different approaches to amplification were compared using next-generation sequencing analysis: I. exponential amplification using polymerase-chain-reaction (PCR) and II. linear amplification.

For both approaches, protocols for single cell extraction, cell lysis, cDNA synthesis, cDNA amplification and preparation of next-generation sequencing libraries were developed. We could successfully show that transcriptome analysis of low numbers of cells is feasible with both exponential and linear amplification. Using exponential amplification, the highest amplification rates up to 10^6 were possible. The reproducibility of results is a strength of the linear amplification method. The analysis of next generation sequencing data in single cell samples showed detectable expression in at least 16.000 genes. The variance between samples results in a need to work with a greater amount of biological replicates. In summary it can be said that single cell transcriptome analysis with next generation sequencing is possible but improvements leading to a higher yield of transcriptome reads is required. In the near future by comparing single cancer cells with healthy ones for example, a basis for improved prognosis and diagnosis can be realised.

Table of contents

List of Figures

List of Tables

Scientific abbreviations

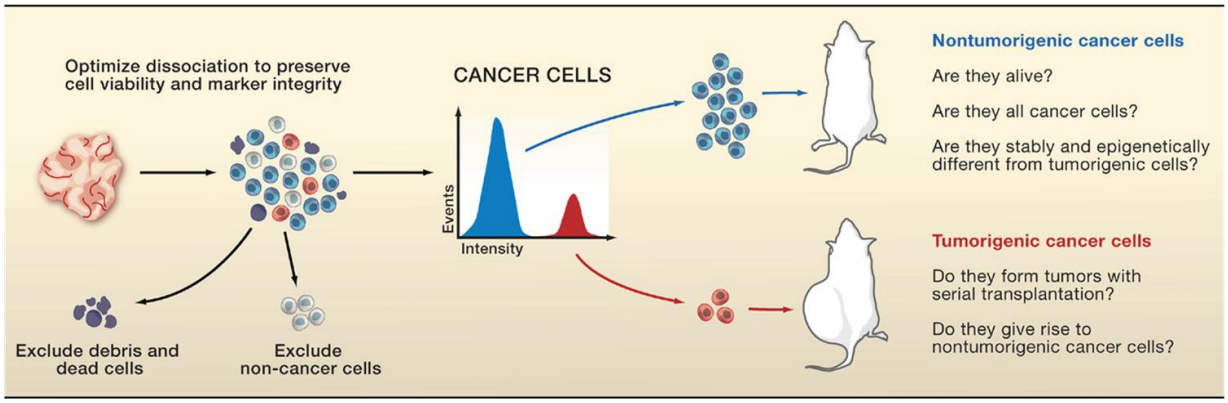
Genetik	
DNA	Deoxyribonucleic acid
cDNA	complementary DNA
3'	three prime end DNA
5'	five prime end DNA
bp	basepairs
kb	Kilo basepairs
Mb	Mega basepairs
RNA	Ribonucleic acid
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA
snoRNA	small nucleolar RNA
hnRNA	heterogeneous nuclear RNA
SINE	Short interspersed nuclear element
LINE	Long Interspersed nuclear element
Measurements	
mg	milligram
µg	microgram
ng	nanogram
pg	picogram
mM	milimolar
µM	micromolar
ml	mililiter
µl	microliter
Cp	Crossing point
RPKM	Reads per kilobase(of exon) per million(mapped reads)
Biotechnology	
WT	Whole transcriptome
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RT	Reverse transcription PCR
RNAse	Ribonuclease
DNAse	Deoxyribonuclease
dNTP	Deoxynucleotide triphosphate
dT	Deoxythymine
UP	Universal primer
FACS	Fluorescence activated cell sorting
rpm	revolution per minute
Materials	
BSA	Bovine serum albumin
ATP	Adenosine triphosphate
PBS	Phosphate buffered saline
DMEM	Dulbecco's minimum

1. Introduction

1.1 Single Cell transcriptome analysis

The transcriptome contains a set of between 13.000 and 16.000 different mRNA species with an average length of 2.2kb [1, 2]. Due to variable frequencies per transcript, in total 400.000 mRNAs per cell [2] are involved in characterising a cell in such action as its proliferation rate, developmental state and feedback to external and internal stimuli. Information written down in genes at DNA level gets transcribed into the transporter molecule mRNA and if required translated into protein [3]. That this dogma is not a one way system was demonstrated by various sets of scientists over the last number of years [4]. Recently, when talking about cell regulation, non-coding RNAs such as microRNA, snoRNA or hnRNA have increasingly become a focus of research. Controlling the expression of genes or transcripts enables a cell to adjust quickly to a certain physiological, environmental or disease related condition. Initially a tissue was believed to be a homogeneous compartment of cells all with the same function. In the meantime we have learned that tissues are highly organized compartments with many types of specialized cells in respect of their function. Therefore, gaining an insight into gene transcription and post-transcriptional modifications such as alternative splicing of healthy cell types as well as disease characteristic ones is of major interest to scientists. It will help to obtain a better understanding of processes in the human body, cell-to-cell interactions and how their misregulation leads to abnormality and disease. However, this heterogeneity in tissues is one of the major limitations for currently performed transcriptomic analyses, where the samples are normally thousands of cells coming from one and sometimes even several tissues. An extreme example of this heterogeneity are blood cells where inter-individual differences and disease-specific changes lead to high variability in composition [5]. Nevertheless, blood is the most widely sampled cellular material, which is frequently analysed for diagnostic or prognostic purposes [6]. Therefore it is critical to receive predictable transcriptome profiles for one specific cell type from a heterogeneous pool of cells. This is also the case earlier in embryonic development where differences between the first cells are of major interest. Along with this naturally occurring tissue heterogeneity come the cancer cells which display a unique heterogeneity. In tumor tissues indeed tumorigenic cancer cells are found together with normal cells, which dilute the somatic cancer cell information.

Tumor stromal cells can contain both genomic and epigenomic alterations, often distinct from those in the epithelial neoplasia [7]. In 2006 Clark *et al* defined a cancer stem cell as 'a cell within a tumor that possesses the capacity to self-renewal and to cause the heterogeneous lineages of cancer cells that comprise the tumor' [8]. In cancer tissue, cells of different function have been described but so far the so called tumour stem cell which has been postulated as a sub fraction of the cells in the tumour tissue with tumorigenic potential, has not been found (Figure 1)[9].



(Shackleton *et al*., Heterogeneity in Cancer: Cancer Stem Cells versus Clonal Evolution, Cell 2009)

Figure 1 **Cancer stem cell theory**
Just a few single cells within a heterogeneous cancer tissue have the capacity to self-renew and to start growing a new tumor.

Scientists are focused on investigating new methods to identify, capture and analyse these stem cells in the hope that by having complete knowledge of a cancer, new target medication can be developed and the relapse rate will drastically decrease. The methodological spectrum of capturing specific cells to get a homogeneous cell population or even just a single cell has massively expanded with new techniques such as laser capture microdissection (LCM) or fluorescence activated cell sorting (FACS) [5, 10]. A single cell contains roughly about 10 pg genomic DNA and 100 pg RNA from which just 1-3 % are protein coding transcripts, mRNA, and the rest is non-coding RNA, mainly ribosomal RNA [11]. The maximal content of messenger RNA is up to three picograms depending on the cell size and especially on the cell cycle stage [12]. It also seems that it can be subdivided into three groups of transcripts. The highly abundant transcripts, the medium expressed transcripts and the lowly expressed transcripts. Initial studies lead to the assumption that mainly the highly abundant transcripts lend the cell its specific characteristics [2]. To analyse a small homogeneous subset of cells or a small amount of starting material such as that from

needle biopsies or circulating tumor cells from blood, stable and easy-to-handle mRNA amplification methods are needed. This seems to have become feasible with the contributions of a small group of scientists that have been worked on the topic for the last two decades. Milestones have surely been developments such as a linear mRNA amplification procedure generating anti-sense RNA presented by Van Geldner et al. [13] in 1992. This was then used by Eberwine for single cell analysis of a neuron carried out for the first time in the same year [14]. Later methods for quantification and detection of gene expression such as real time PCR and microarray technology were added to enable genome wide gene expression profiling. However, independent of the gene expression quantification approach used, single cell experiments have to cope with minute amounts of starting RNA material. The high chance of methodically defective amplification and the strong influence of cell cycle stage on RNA content have to be considered as limiting factors, not to mention the cost of carrying out these procedures. Therefore, the approach of second generation sequencing which combines high throughput capability with the generation of massive amounts of data with a single run lead to a new level of analysis. In July 2008 Cloonan *et al.* published one of the first single cell papers using transcriptome sequencing technology [15].

1.2 Methodological Spectrum of cDNA from minute amounts of RNA

1.2.1 Reverse Transcription

The initial step in the workflow from RNA to cDNA to preserve the RNA of a cell, which is prone to hydrolytic degradation, is the reverse transcription. The reverse transcription reaction is not very well understood, and it is known in the community to be the most uncertain step in gene expression analysis. The Processivity of an enzyme , its preference for specific sequences and general faultiness are difficult to predict or measure. Therefore it is important to run experiments as similar as possible so that the condition for every template is almost the same and thus cross comparison of analyses results can be achieved. There are basically two different ways to transcribe the whole transcriptome of a eukaryotic cell into cDNA.

Oligo(dT) Priming

In mammalian cells nearly all protein coding transcripts carry a 3´ poly(A) tail [16, 17]. Newly synthesized poly(A) tails are approximately between 100 and 250 nucleotides long [2, 18]. Their function varies from promoting export from the nucleus to being part of the translation initiation complex, to protecting the transcript from degradation and (by length alteration) influencing translation efficiency [19, 20]. This tail can be used as a starting site for transcription allowing the use of total RNA as starting material [21]. After oligo(dT) primer annealing the reverse transcriptase starts to extend the newly synthesised cDNA strand. Depending on processivity, reaction time, buffer condition and secondary structure every mRNA transcript gets more or less fully transcribed into cDNA. This, however, leads to a potential 3´ bias of the transcript [22]. Giving the reaction enough time, using single strand binding proteins and working under best reaction conditions can minimize this problem.

Random Priming

A way to overcome the issue of 3´bias and to improve the outcome of full length cDNA is the use of random oligonucleotides. Standard reverse transcription protocols utilise hexamers [23, 24] or nonamers [23, 25]. A higher cDNA yield could be reached with pentadecamers [26]. There are two major drawbacks of this method: 1) a 5´bias of the transcripts because of unspecific annealing, 2) this method will lead to transcription of ribosomal RNA as well, which is about 98 % of the total RNA. It depends on the method used for quantification if ribosomal RNA has to be depleted from the starting material prior to reverse transcription or not. This can be done either by techniques such as RIBO Minus, where the ribosomal RNA is depleted by hybridization using probes targeting the ribosomal RNA, or by enrichment of polyadenylated RNA using oligo-(dT)-probes which are most often bound to beads.

However, Stahlber *et al.* 2004 [27] published results indicating that in fact reverse transcription is dependent on priming strategy but it varies for every single transcript. This means that the choice of the priming method is based more on the experimental design. The major issue working with single cells, along with RNA digestion, is the loss of transcript during experimental handling. Therefore, purification steps and switching of reaction vessels until amplification is fully complete have to be avoided. To select full length cDNA transcripts for later analysis a technique known as a switching mechanism at the 5´end of RNA template

(SMART) can be used [28]. It is in the nature of a reverse transcriptase to extend the 3´end of the cDNA with a few cytosines after transcribing the RNA strand. This overhang can be used as an anchor to anneal a specific sequence tagged with guanines and the cDNA then gets extended after re-annealing of the enzyme [29]. This new sequence is the potential starting sequence for an isothermal linear amplification or the universal priming site for an exponential amplification.

1.2.2 Amplification approaches

Using techniques such as microarrays or next generation sequencing for transcript quantification means that at least 100 ng up to several micrograms of starting material are needed. An average mRNA content of 0.1 pg per cell requires 10^5 up to 10^6 fold amplification. Amplification in this range has high potential for bias. Bias in this case means the shift in the quantitative ratio of transcripts to each other. This occurs mainly in a multi-template reaction because of preferred amplification of certain sequences (and its resulting secondary structure), template lengths or template amounts [30]. The high diversity of a complex transcriptome and the massive amplification needed requires a method of very limited bias. Two ways of amplifying mRNA have been described so far.

Exponential

Polymerase chain reaction (PCR) described in 1983 by Mullis *et al*. is the common method for an exponential amplification [31]. A maximal amplification factor of 3 X 10^{11} has been reported [32]. Nevertheless, reaching a high concentration (>10^{11} molecules per µl) of newly synthesized templates inhibits their own duplication by quickly re-annealing and therefore the PCR level move to a plateau phase [33]. During the last few cycles of the reaction higher rates of PCR bias and artefact formation occur [30]. For primer annealing, the temperature of the PCR reaction decreases and three different kinds of duplexes can be formed. Starting with a high primer concentration duplexes between primer and template normally occurs. The rate of template re-annealing, termed a homoduplex, increases proportional to the concentration and therefore templates starting with lower concentration will catch up which induces a strong bias. As a result of forming a heteroduplex between two different templates the repaired sequence is a mixture of the two parent heterologous sequences [30,

34]. An incompletely extended primer can act as a primer in the following PCR cycle. Having some homologousous sequences between different templates, extended primers from one strand can anneal to the other and a chimeric sequence results [35]. Heteroduplexes and chimera formation can lead to false detection of new gene rearrangements or SNPs. PCR bias due to differences in primer binding energy in a multi-template reaction can be reduced if all templates are ligated to a universal primer. However, guanine-plus-cytosine content of the templates will influence its amplification efficiency [36]. One way to overcome the above mentioned issues is to stop the PCR reaction as early as possible. Real time monitoring of the reaction (real-time PCR) helps to determine the best time. An amplification of at least 100.000 fold for single cell material exceeds this point and therefore the use of exponential amplification as a method for single cells has to be questioned.

Linear

The most commonly used mechanism for linear isothermal RNA amplification is based on T7 RNA polymerase-mediated *in vitro* transcription (IVT) and was first described by Eberweine and colleagues [13]. Starting at the T7 promoter sequence the polymerase transcribes cDNA into complementary RNA (cRNA) which can be then retranscribed into cDNA. In so doing, a starting material can be increased up to 1.000 fold in one round. Second and third rounds of amplification are possible [37].. It is widely believed that linear amplification is much less influenced by sequence content and therefore bias is almost negligible. The T7 IVT amplified samples, aRNA, have been shown to supply the same results as the same non-amplified material using microarray technology [38]. However, this method with three rounds of amplification is a time consuming procedure and degradation of the newly transcribed RNA is a big concern. Another way to amplify small amounts of RNA in an alternative linear approach, termed Ribo-SPIA^{TM}. It was recently described by Dafforn *et al.* [39]. Initially, a tailed DNA-RNA chimeric primer hybridizes to the 3´poly (A) tail and first strand cDNA is generated via reverse transcription. Synthesizing the second strand a short sequence of RNA-DNA results and by RNAseH treatment a new priming site is generated. Due to the strong strand displacement function of the used polymerase highly efficient linear amplification of one strand takes place (detailed information see Material & Methods). From a time and efficiency perspective linear amplification cannot keep up with exponential methods but the ability to maintain the polarity of the transcript without laborious

modification is a major advantage. This in turn gives the opportunity to identify which strand gets transcribed. Strandedness provides important information for novel genes about their possible function, both at the RNA (structure and hybridization to other nucleic acid molecules) and protein level. Antisense transcription is characteristic of eukaryotic genes and is thought to play an important regulatory role in RNA interference mainly with the sense strand. Estimates of the fraction of genes associated with antisense transcripts in mammalian cells vary from less than 2% to more than 70% of the total gene number [40-42]. Knowledge of a transcripts orientation helps to resolve colliding transcripts and to correctly determine gene expression levels in the presence of antisense transcript [43].

1.3 cDNA Quantification

After obtaining the cDNA the final step in gene expression analysis is its quantification. A number of different technologies to detect transcripts have been developed. The three main ways to quantify the gene expression are: real time PCR (qPCR), hybridization-based and sequencing-based approaches. Real time PCR is a highly specific tool with the largest dynamic range. It requires in principle just a single strand of the template. However existing knowledge about the sequences are a prerequisite and the number of transcripts which can be analysed at the same time is limited.

Hybridization-based methods usually involve the hybridization of cDNA on a microarray combined with fluorescence labeling. The possibility of custom-made microarrays including probes targeting splice variants while keeping the strand information (strandedness) make microarrays an advanced tool for transcriptome analysis. Arrays allow the mapping from several base pairs to ~100 bp [44-46]. But there are several limitations to this technique: 1) Existing knowledge about genome sequences is needed, 2)high background levels owing to cross-hybridization as well as a limited dynamic range of detection based on background and saturation [47]. Moreover, to compare results from different experiments is difficult with both technologies mentioned and they require complicated normalization methods.

In comparison sequencing-based approaches directly determine the cDNA sequence. This is termed RNA-Sequencing (RNA-Seq). Even though library preparation involves a complex workflow, the chance of producing high quantity *de novo* data in a short time will revolutionize the way in which eukaryotic transcriptomes are analysed. Unlike microarrays, sequencing costs are constantly decreasing. All this will make the sequencing based

approaches the preferably method used in genome (transcriptome) wide gene expression analysis. However, the accurate and reproducible quantitative analysis of a single cell necessitates overcoming certain obstacles. Starting with the obvious like RNA digestion, loss of material or simply labour handling issues because of small volumes right up to biological issues such as different RNA content, different cell cycle or influencing the transcriptome by abnormal treatment during selection. To workaround these issues an affordable high throughput analysis method enabling replications had to be provided.

1.4 Next generation Sequencing

1st generation

Since the early 1990s DNA sequencing has almost exclusively been done utilizing capillary-based Sanger biochemistry [48, 49]. The basic principle of this method is 'cycle sequencing' which contains the template denaturation step, primer annealing and primer extension. Known sequences flank the region of interest which provides a universal primer annealing and start point. Each round of primer extension is terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs) in a stochastic way. The incorporated ddNTPs lead to a halting of the reaction and the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. High-resolution electrophoresis leads to an identification of the discrete length of every single-stranded, end-labeled extension product such that in the end every nucleotide should at least once have a terminal position labeled with its specific fluorescence colour. Finally, software translates these traces into DNA [50, 51]. Drawbacks of this method are the limited level of parallelization. Currently a maximum of 384 independent capillaries is possible. The maximal read length is about 1.000 bp. Sanger sequencing costs in the order of $0.50 per kilobase. Initially, the Sanger method was used to sequence cDNA but this approach is relatively low throughput, expensive and non-quantitative. Tag-based methods including serial analysis of gene expression (SAGE) [52], cap analysis of gene expression (CAGE) [53] and massively parallel signature sequencing (MPSS) [54] were developed to overcome these limitations. But even though these are an improvement the technique is still based on expensive Sanger sequencing technology and a significant portion of the short tags cannot be uniquely mapped to the reference genome [55].

2nd generation

The concept of second generation sequencing in the sense of cyclical-array sequencing can be summarized as the ´sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection´ [56]. This has recently been realized in many different commercial products: 454 Genome Sequencer (Roche Applied Science, Basel); Solexa technology, Genome Analyzer (Illumina, San Diego); SOLiD platform (Applied Biosystem, Foster City); the Polonator (Dover/Harvard). Even though these platforms are quite different in their biochemistry and their arrays, the work flow is conceptually similar. The first step in library preparation is random fragmentation of DNA followed by ligation of adapter sequences. An extension of this protocol would be to generate libraries of mate-paired tags with controllable distance distributions [57]. The generation of clonal amplicons can be achieved in different ways, including emulsion PCR on the surface of microbeads (454, SOLiD, Polonator)[58] or bridge PCR on a single location in a planar substrate (Solexa) [59]. In the end every single library molecule ends up in a clustered colony of amplicons bound to solid support. The sequencing itself relies on synthesis which means alternated cycles of enzyme-driven biochemistry. Serial extension of primed template can either be done by a polymerase (Solexa,454)[60] or ligase (SOLiD) [57]. Finally data is generated by imaging of the full array at each cycle and successive image analysis.

The main advantages compared to first generation sequencing are *in vitro* construction of a sequencing library over *in vivo* cloning methods. Array-based sequencing enables a much higher degree of parallelisation than conventional capillary-based sequencing and finally only microliter-scale reagent volumes are used. Altogether as the effective size of sequencing can be on the order of 1 µm, hundreds of millions of sequencing reads can potentially be obtained in parallel and this result dramatically lowers the cost of DNA sequencing production. The most prominent disadvantage is the read-length. For all the new platforms, read-length is currently much shorter than conventional sequencing [51] though this limitation is rapidly improved with new longer read technologies.

Table 1 **Summary of commercially available second generation sequencing platforms**

Platform	Cost per instrument ($)	Cost per Mb ($)	Gb per run	Read length (bp)	Pros	Cons
454/ Roche	500,000	60	0,45	250	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homopolymer repeats
Solexa / Illumina	430,000	2	18	36	Currently the most widely used platform in the field	Low multiplexing capability of samples
SoliD/ ABI	591,000	2	30	35	Two-base encoding provides inherent error correction	Long run times
Polonator	155,000	1	12	13	Least Expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain sequencer; shortest NGS read lengths

Based on three puplications: Shendure & Ji, Nature biotechnology, 2008; Elaine Mardins, Annual reviews, 2008; Wang et al., Nature Reviews, 2009

1.5 Challenges for RNA-Sequencing

1.5.1 Library construction

The ideal way to sequence the transcriptome would be to identify and quantify all RNAs directly, small or large. Although there are only a few steps in RNA-Seq (cDNA fragmentation, library preparation, sequencing) it does involve several manipulation steps. For example, large RNA molecules must be fragmented into smaller pieces (100-500bp) to be compatible with most deep-sequencing technologies. Common fragmentation methods are RNA fragmentation (RNA hydrolysis) and cDNA fragmentation (DNase I treatment, UDG treatment or sonication). In the case of minimal input material the fragmentation step has to be done after amplification of cDNA. Each of these methods creates a different bias. Short reads that are identical to each other can be obtained from cDNA libraries that have been amplified. These could be a reflection of abundant RNA species, or they could be PCR artefacts. This has to be determined before analyses can be done [55].

1.5.2 Bioinformatics

A quite obvious issue is how to store, retrieve and process large amounts of data, which has to be overcome to reduce errors in image analysis and base calling and remove low-quality

reads. A first step after generating reads is to map them to the reference genome. However, short transcriptomic reads also contain reads that span exon junctions or that contain poly(A) ends – these cannot be analysed in the same way normal reads can. One solution is to compile a library containing the entire known and predicted junction sequences but still the challenge remains to identify novel splicing events. Large transcriptome alignment is also complicated by the fact that a significant portion of sequence reads match multiple locations [55].

1.6 Aims and focus of the project

In this project I first started to learn how to handle small volumes in general and especially small amount of RNAs. I had to learn how to work completely free of any RNAses and to find out what the requirements of a successful protocol are. The first comparative studies were carried out with a simple cell dilution series of a commonly used cell line (HEK293T). Initial studies to determine best transcription efficiency compared to various RT enzymes as well as different primer concentration had to be done. Also different ways to amplify small amounts of RNA were tested; T7 *in vitro* transcription, PCR based methods and linear amplification based on Ribo-SPIA technology. The main focus of my work was to compare linear and exponential amplification approaches regarding their bias and protocol handling. Cooperation with two different companies promoting these technologies was initiated: Miltenyi and its µMACS SuperAmp Kit as the method of choice for exponential amplification and NuGEN technologies and its WT Ovation One-Direct RNA Amplification System representing linear amplification. As a personal comment it was important to get trust and good communication with both companies so that specific changes in the respective chemistries could be made. Finally the best way to amplify the transcriptome of a single cell could be realised with all the benefits as well as limitations. Next generation sequencing implementing the ABI SOLiD platform was used to analyse the obtained spectrum of amplified cDNA as well as performing first studies on gene expression.

2. Material and Methods

2.1. Material

2.1.1. Biological material

- HEK293T, Human Embryonic Kidney 293T cells
- HeLa cells, Cervical cancer cells
- SW480, Colon cancer cell line

2.1.2. Chemicals, Buffers, Media

- 100 % Ethanol, Merck
- ATP, NEB
- BSA, NEB
- Cell lysis buffer, Ambion
- DMEM, Biochrom
- DNA away, Molecular BioProducts
- dNTP´s, dUTP, Fermentas
- Elution buffer (10mM Tris), Qiagen
- Ethidium bromide, Sigma-Aldrich
- Fetal calf serum, Bioc
- Magnesium chloride, Sigma-Aldrich
- Nocodazol, Sigma
- PBS, Biochrom
- Penicilin/Streptomycin, Biochrom
- RNase free water, Sigma
- RnaseZap, Ambion
- SybrGreen PCR Mix, Roche
- T4 Gene 32 Protein, NEB
- TAE Buffer, Inhouse
- TaqMan PCR Mix, Roche
- TrisHCL, Sigma-Aldrich
- Trypsin, Biochrom
- Ultra Pure Agarose, Invitrogen

2.1.3. Enzymes

- DNA Polymearse, TaKaRa Ex Taq, TaKaRa (5U/µl)
- DNA Polymerase I, Large (Klenow) Fragment, NEB (5 U/µl)
- DNA Polymerase, Phusion, FINNZYMES (2U/µl)
- DNase I, NEB (2U/µl)
- Exonuclease I, NEB (20 U/µl)
- Expand Long Template PCR System (5U/µl)
- M-MLV Reverse Transcriptase, Ambion (100//µl)
- M-MuLV Reverse Transcriptase, Enzymatics (200U/µl)
- Phi29, Enzymatics (10U/µl)
- Proteinase K, Inhouse (20µg/µl)
- RNase H, Ambion (10U/µl)
- RNase Inhibitor, ABI (20U/µl)
- SuperScript Transferase III Reverse Transcriptase, Invitrogen (100U/µl)
- T4 DNA Ligase, Enzymatics (600U/µl)
- T4 Polynucleotide Kinase (10U/µl)
- Terminal Transferase, NEB (20U/µl)
- Uracil-DNA Glycosylase, NEW (5U/µl)

2.1.4. Kits

- µMACS SuperAmp Kit, Miltenyi
- Cell-to-cDNA Kit, Ambion
- End Repair mix, Enzymatics
- PCR Purification Kit, Qiagen
- RNeasy Mini Kit, Qiagen
- SuperScript® III CellsDirect cDNA Synthesis Kit, Invitrogen
- WT Ovation, One-Direct RNA Amplification System, NuGEN

2.1.5. Oligo's (Primers and Adapters), DNA and RNA ladders

- 0.1 - 2 Kb RNA Ladder, Invitrogen (1µg/µl)
- 100 bp, 1 kb DNA ladder, Fermentas
- List of all used primer sequences is attached (Table 1, supplementary)

2.1.6. Devices

- Bunsen burner
- Centrifuge 5415 D, Eppendorf
- Centrifuge 5810 R, Eppendorf
- Covaris S2, Covaris
- FACS, Diva, BD Biosciences
- Gel electrophoresis, Bio Rad
- LightCycler 480, Roche
- Macs Multi Stand, Miltenyi
- Magnetic separation block
- Mastercycler gradient, Eppendorf
- Microscope
- ND 1000 Nanodrop, Thermo Scientific
- SOLiD 3+ sequencer, ABI Serie
- SPRIPlate 96R Super Magnet plate, Agencourt
- Thermocycler gradient, MJ Research PT-200, Eppendorf

2.1.7. Consumables

- 384 well plate sealing, Roche
- 384 well plate for Roche LC 480, Roche
- 5 ml, Sarstedt
- Falcon tube 15, 50 ml, Greiner bio-one
- Filtertips 10 µl, 20 µl, 200 µl, 1 ml, Biozym
- Low retention filtertips 10 µl, 20 µl, 200 µl, 1ml, Starlab
- LowBind reaction tubes 0.5 ml, 1.5 ml, 2 ml, Eppendorf
- Microcapillary tube, calibrated, 50 µl, Sigma-Aldrich
- Multiply - µStripPro, 0.2 ml, Starstedt
- PCR-Softstrips 0.2 ml, farblos, Biozym
- Reaction tubes 0.5 ml, 1.5 ml, 2 ml, Sarstedt
- T75 cell culture flask, TPP
- Tips 10 µl Gilson

2.2. Methods

2.2.1 Cell culture

All cell lines have been cultivated under standardised conditions. DMEM containing 10% FCS and 1% Penicillin/Streptomycin was used as culture medium. Culturing conditions were set as 37°C with 95% rH and 5% CO_2. In three day cycles cells were harvested by trypsinisation, washing with PBS, pelleting by centrifugation and resolving in fresh medium. Cells were diluted one to three to gain optimal growing conditions.

Equalizing transcriptome level and cell harvest

For equalizing the transcriptome level, cells were G2/M-phase arrested by Nocodazole treatment. After aspirating DMEM medium from cells, 12 ml fresh DMEM (37°C) + 1 µl Nocodazole (5 mg/ml) per T75 cell culture flask were consistently distributed and incubated for 16h at 37°C. Cells were separated by washing with 10ml PBS (37°C) and covered with 2 ml Trypsin (0.05%). Shortly afterwards cells were aspirated with 1.5 ml Trypsin and incubated for around 10min at 37°C until they were detached. Trypsin was inactivated with 10 ml DMEM and the cell lysate was dissolved and entirely separated by up and down pipetting three times.

Concentration determination

Cell number measurement was done using a counting chamber. 5 minutes slowly centrifuging (500rpm) generates a cell pellet. The pellet was resuspended in PBS to get a stock concentration of 2500 cells/µl. Cell solution was split into aliquots of 1.5 ml tubes and shock-frozen in liquid nitrogen. For further experiments these aliquots have been stored at -20°C.

2.2.2. Isolation of single cells

Fluorescence activated cell sorting (FACS)

FACSDiva Version 6.1.2. from BD Bioscience was used as described in the manual. Cells were harvested and diluted in PBS so that there would be no inhibition of further reactions. Single cells were sorted in 0.2 ml tubes containing 1.1 µl lysis buffer and 0.15 µl RNase inhibitor.

Cells were spun down immediately after sorting and continually started with the reverse transcription protocol.

Mouth pipetting

To prepare the tip of a microcapillary pipette the middle part was heated with a naked flame and by pulling gently at both ends, a really thin middle part results. After breaking at this mid-point, the resulting capillary termini were rounded off by melting the glass in a naked flame for less than a second. The resulting pipette was put on a 15-inch aspirator tube. Cells were harvested and highly diluted in PBS. Under a microscope with a 100 to 200 fold enlargement a single cell was sucked gently into the prepared pipette and blown out into 0.2 ml tubes containing 1.1 µl lysis buffer and 0.15 µl RNase inhibitor. Immediately, the reverse transcription protocol was started to avoid any degradation. A short video showing the process of collecting a single can be found in the attached CD.

2.2.3 Reverse transcription and amplification

None of the applied protocols includes a DNA digestion step due to the high risk of RNA degradation during incubation.

SuperScript III CellsDirect cDNA Synthesis System

Protocol was carried out as set up by the company. For low input material, the end volume was reduced down to 5 µl, concentrations were kept in the standard protocol.

Ambion Cell-to-cDNA

Protocol was carried out as set up by the company. For low input material, the end volume was reduced down to 5 µl, concentrations were kept in the standard protocol.

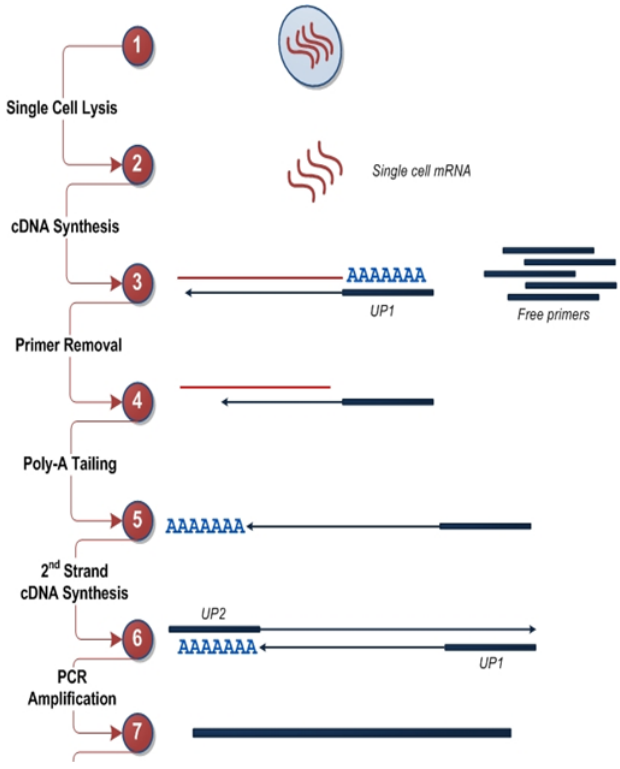
In house protocol

Regarding the Ambion Cell-to-cDNA kit, 0.5 µl of cell solution, with a maximum number of 2500 cells, were incubated and lysed in 1.25 µl Ambion lysis buffer at 75 °C for three minutes. Reverse transcription is a two step protocol. First a primer annealing step with 0.625 µl (0.5 µM) Oligo dT_{24}, 0.5 µl (2.5 mM each) dNTP mix and 0.175 T4 gene 32 protein (2 µg/µl) was performed by heating up the product reaction mix to 70°C for one minute and

then cooling down on ice. In the second step 0.56 µl RT master mix was added to each reaction containing 0.33 µl 10X RT buffer, 0.1 µl M-MuLV Reverse Transcriptase (Enzymatics) and 0.13 µl RNAse Inhibitor (ABI). Reaction was carried out in 3.61 µl end volume. Incubation temperature was 42°C for 30 min followed by heat inactivation for 10 minutes at 95°C.

Whole transcriptome preparation of a single cell, ABI

The whole transcriptome analysis preparation protocol for a single cell recommended by ABI is based on a protocol published by Kurimoto *et al.* in 2006. Little changes have been made. After lysing the cell, mRNA is extracted by using Oligo dT-UP1 labeld beads. cDNA synthesis was extended from 5 min to 30 min to get full-length first strand cDNAs. After poly(A) tailing of the newly synthesized cDNA strand this sequence is used as an anchor for the universal UP2 primer. Second strand synthesis can now be done. To avoid a bias of short primer sequences in library preparation UP1 and UP2 are modified with an amine at their 5´end. Therefore no ligation of 5´end fragments to SOLiD library adaptors can occur. Extension time in the PCR based amplification reaction was extended from three to six minutes. A schematic overview of the workflow is outlined in figure 2. This protocol, as described in the initial paper, including a few minor changes was carried out as mentioned above. Because of high primer dimerization a new UP2 sequence was designed, termed UPA.



Tang *et al.*, mRNA-Seq whole-transcriptome analysis of a single cell, Nature Methods, 2009

Figure 2 **Schematic workflow overview of whole transcriptome preparation of a single cell by ABI**

µMACS SuperAmp Kit, Miltenyi

The work was carried out as described in the protocol. No changes to the commercially available kit were made. This procedure is comparable to the protocol described by Kurimoto *et al.* in 2006. The special feature of this kit is the technique of the µMACS columns. The magnetic field of the µMacs Multi Stand gets exponentiated and focused on a specific area within the µMacs columns due to containing little pieces of iron at this part. Because of this strong magnetic field the mRNA which is connected to Oligo dT magnetic beads is held and the washing procedure can be done without any loss (Figure 3). I used one set of standard primers and one set of 5´-biotin modified primers for the PCR based amplification. Sheared amplification products generated with the biotin modified primers were primer depleted by magnetic streptavidin bead treatment, before they were subjected to the standard SOLiD library preparation.

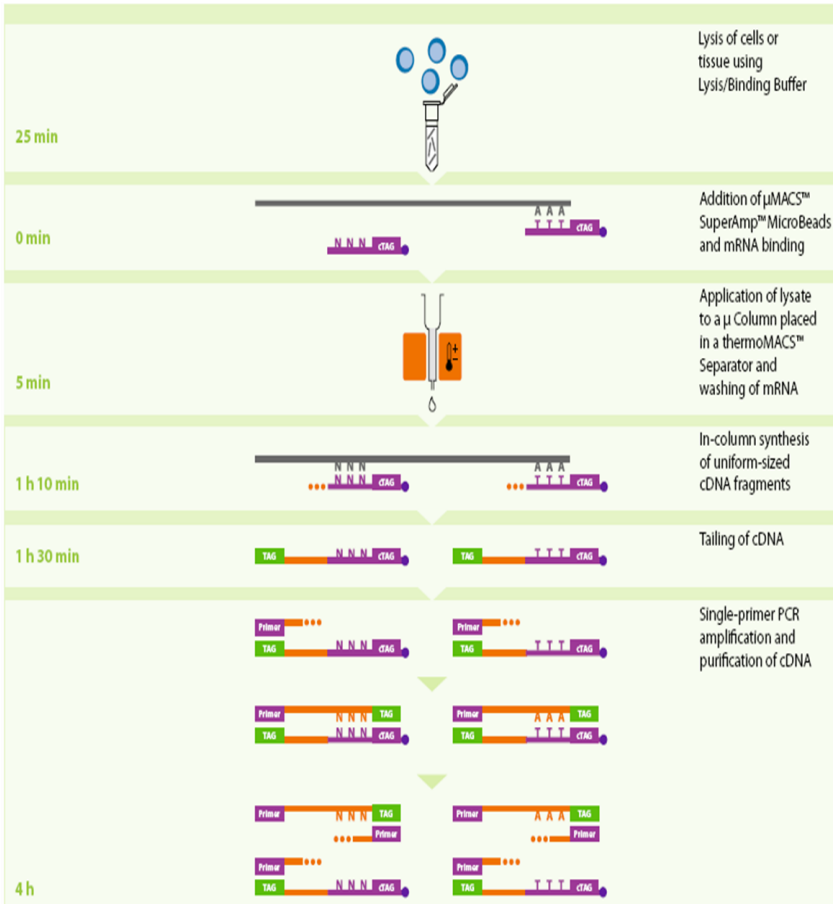


Figure 3 **Schematic overview of the principle workflow of µMACS SuperAmp**

WT Ovation, One-Direct RNA Amplification System, NuGEN

Lysis and linear amplification were carried out as described by the manufacturer. To make sure that the reverse transcriptase generates cDNA only from mRNA the NuGEN chemistry was modified from using both random priming and Oligo (dT) priming to just Oligo (dT) for cDNA synthesis. The Oligo (dT) cDNA primer is a DNA/RNA hybrid. After generating the second strand, RNAse H is able to digest the RNA of the newly synthesised DNA/RNA double strand. New primers can anneal and the used polymerase with a strong strand displacement function synthesizes copies of the initial strand (Figure 4). The limiting factor in this assay is the polymerisation length of DNA polymerase. Currently after 300 to 500 bp the enzyme stops synthesising and drops off.

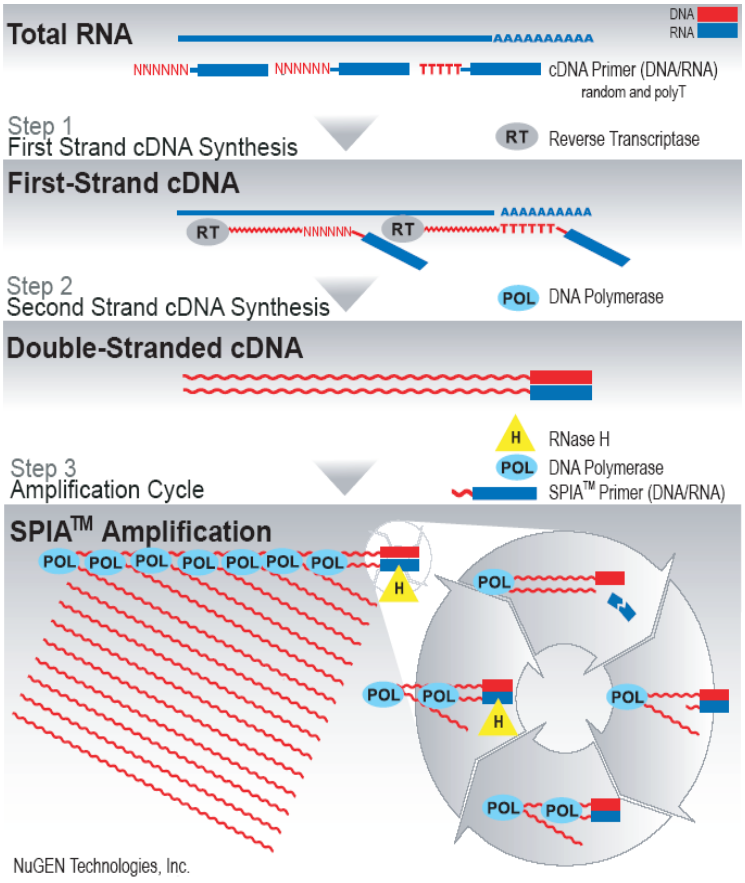


Figure 4 **Schematic overview of the linear amplification system by NuGEN**

2.2.4 SOLiD Library Preparation

Library preparation for double stranded DNA – Short fragment library, ABI

The protocol for double stranded DNA library preparation is based on the SOLiD short fragment library preparation protocol by ABI. To reduce costs, enzymes were exchanged without loss of efficiency. First of all DNA was sheared down to 100-110 bp fragments with the Covaris S2 System. Templates at a concentration ranging from 10 ng to 5 µg were mixed with water up to 100 µl. Six rounds of shearing each 60 seconds long with 200 cycles per burst, duty cycles of 10% and intensity of 5 were done. End repair of the fragments was carried out as described for the Enzymatics end repair mix. The amount of double stranded P1 and P2 adaptors needed for the reaction was calculated with the following formula:

$$X\ \frac{\text{pmol}}{\mu\text{g}}\ \text{DNA} = 1\ \mu\text{g DNA} * \frac{10^6\ \text{pg}}{1\ \mu\text{g}} * \frac{1\ \text{pmol}}{660\ \text{pg}} * \frac{1}{\text{Average insert size}}$$

$$Y\ \mu\text{l adaptor needed} = \#\ \mu\text{g DNA} * \frac{X\ \text{pmol}}{1\ \mu\text{g DNA}} * 30 * \frac{1\ \mu\text{l adaptor needed}}{50\ \text{pmol}}$$

Ligation reaction was done with 2 µl T4 Ligase (600 U/µl) from Enzymatics in a final volume of 100 µl. In contrast to the original protocol, nick- translation and PCR amplification were done after adaptor ligation and before size-selection on an agarose gel. The PCR protocol was done according to the ABI protocol except of the Phusion DNA polymerase (1 U) of FINNZYMES which was used in combination with the 5 X Phusion HF buffer in a total volume of 100 µl. The amplicon was size-selected by a 2 % agarose gel within a fragment size of 150 to 200 bp and quantified via real time PCR.

Library preparation for single stranded DNA – In house protocol

The workflow structure of our library preparation protocol from single strand DNA fragments is copied from the Small RNA Expression Kit (SREK) by ABI. The main difference is that the SREK protocol is made for library preparation straight from RNA molecules. Therefore we had to design our own adaptors and as a result changes in the components were made. The protocol contains five steps. Fragment shearing, adaptor hybridization, adaptor ligation, second strand synthesis and finally amplification. For single stranded DNA fragmentation the Covaris S2 System was used. The shearing protocol was optimized for

single stranded DNA to achieve a fragment size of 100 bp´s. Duty cycles were set to 10%, with an intensity of five and 200 cycles per burst. DNA was diluted in water to achieve the final shearing volume of 100 µl according to the procedure manual. Double stranded adaptor 1 is a hybrid of the internal linker sequence and its complementary sequence extended with a 3´overhang of six random nucleotides. Adaptor 2 is a hybrid made of the P1 adaptor sequence and its complementary sequence with a 5´overhang of six random nucleotides (Figure 5). For a successive ligation the 5´end of the template (blue dot) and 5´end of internal linker sequence (red dot) have to be phosphorylated. Adaptors were mixed to a concentration of $1.5x10^{13}$ molecules each per µl. Reaction mix contains 20µl template (70 ng/µl), 25 µl elution buffer and 1.5 µl adaptor mix and temperature was set to 65°C for 10 minutes and 16 °C for 15 minutes. Hybridization product was ligated with 5 µl of ATP (10 mM) and 2 µl T4 ligase (600 U/µl) at 37°C for 30 minutes. For the following reactions a standard PCR purification (Qiagen) was done and the product was eluted in 30 µl EB buffer. Second strand synthesis mix contains 30 µl purified template, 1 µl dNTPs(2.5 mM each), 5 µl Phi29 buffer, 1 µl Phi29 (10 U/µl), 5 µl BSA (1 mg/µl) and 8 µl H_2O. Reaction conditions were set to 30 min at 30 °C. Amplification PCR master mix contained 5 X Phusion HF buffer, 1 µl primer mix (P1 and P2-Barcode-Internal linker), 1 µl purified template, 0.5 µl Phusion (2 U/µl), 4µl dNTPs (2.5 mM each) and 33.5 µl H_20. Cycling condition were as follows: 98 °C for 30 sec. and 17 cycles of 98 °C for 10 sec, 62°C for 30 sec, 72°C for 30 sec. To determine the optimal cycling number 5 µl per template were removed after various cycles. In this way an over-amplification should be avoided. After this, large scale amplification with 5 x 50 µl PCRs was carried out with the remaining template followed by a size-selection.

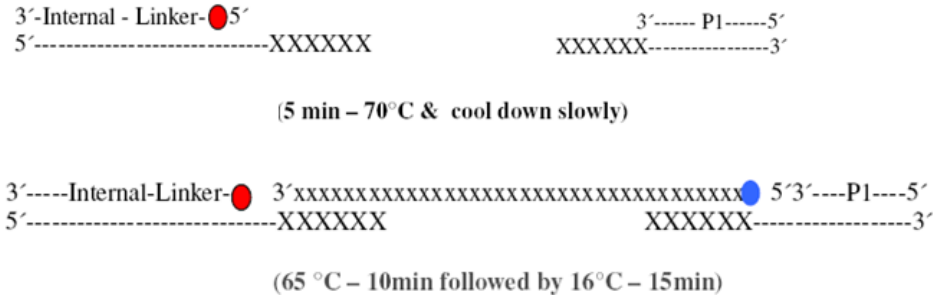


Figure 5 **Schematic overview of adaptor design and adaptor hybridization**
5´end of adaptor carrying internal linker sequence was phosphorylated (red dot) as well as the 5´end of sheared cDNA fragment (blue dot). Sense and antisense strand of adaptors were equally mixed and hybridized. The same procedure was done for hybridizing adaptors and cDNA fragment.

2.2.5 Exonuclease treatment

Digestion of remaining primer by exonuclease (E.coli,NEB) treatment was carried out with an exonuclease concentration of 1 U/µl reaction volume at 37°C for 30 min.

2.2.6 PCR

Standard PCR

The standard PCR was performed in 10 µl reactions containing the following components: 20 mM TrisHCl, 16 mM $(NH_4)_2SO_4$, 25 mM KCl, 2 mM MgCl2, 500 µM each dNTP, 1 - 1.4 U Taq Polymerase, 0.54 M betaine, 1.34 M DTT, 1.34 % DMSO, 11 µg/ml BSA and 0.3 µM forward and reverse primer and a varying template concentration. The PCR cycling program consisted of 30 cycles with 15 sec. at 95°C, 30 sec. at 60°C and 15 sec. at 72°C. PCR reactions were stopped after 2min. at 72°C for final extension.

Real time PCR

Real time PCR was performed in 10 µl reactions by either using Roche SyberGreen mix with a final primer concentration of 0.3 µM or Roche TaqMan mix and a primer-probe mix concentration of 1 µM on the Roche LightCycler 480. A two step temperature protocol was used for both assays, TaqMan and SybrGreen. Five minutes at 95°C to activate the enzyme was followed by 40 up to 45 cycles of a denaturation step of 10 sec. at 95°C and an annealing/ elongations step for 40 sec. at 60 °C. Analyses were done using the supplied Roche LightCycler software. Relative quantification analyses were made with the second derivate method and crossing points compared to each other. Human mRNA extracted out of HeLa cells and transcribed in cDNA with SuperScript III was used as positive control for all qPCR assays.

Melting curve analysis

Melting curve analysis was performed with every SybrGreen assay to check PCR products for integrity. One minute of complete denaturation at 95 °C and one minute at 60°C to form duplexes was followed by a constant temperatures increase of 0.11 °C/sec. up to 95 °C. This leads to a characteristic melt curve for every amplicon. Primer dimers and erroneously amplified products are distinguishable from the target amplicon by their melting peak.

2.2.7 Concentration determination

Nanodrop

To determine the concentration of single strand DNA or double strand DNA, 1 µl template was analysed by the ND 1000 Nanodrop (Thermo Scientific) as described in the manual.

Gel electrophoresis

Horizontal gel electrophoresis on an agarose gel leads to DNA separation based on its fragment length. By adding ethidium bromide into the gel and exposing it to UV light the fragments can be seen and an image can be taken. Agarose was used in a concentration of 1% up to 2% depending on the expected fragment length. After heating up the agarose in 1 X TAE buffer ethidium bromide was added in a concentration of 0.5 µg/ml.

2.2.8 ABI Sequencer

For performing sequencing on the ABI SOLiD 3+ sequencer quantified libraries were adjusted to 50 pg/µl for bead preparation carried out as described by the manufacturer. Bead deposition (quads, each 96 mio beads) and loading of the sequencer was carried out as described in the ABI SOLiD manual. Sequencing was performed with 35/50bp fragment sequencing kits.

2.2.9 Bioinfomatics

Sequencing analysis and maToBam conversion

Reads were processed using the whole transcriptome pipeline integrated in the Bioscope package from Applied Biosystems. Processing was split into three steps: Filtering (repetitive sequences, run chemistry, rRNA, tRNA, etc.), mapping against genome (HG19) and mapping against exon-exon junctions (generated internally using refGene.gtf version hg19). Mapping was performed in seed&extend mode seeding first 25bp of the reads allowing 2 mismatches and extending the aligned seeds with a mismatch penalty score of -2 in all three processing steps. After mapping the integrated merging pipeline was used to assign whether the reads belong to non-relevant sequences (filtering) or if they came from genomic or exon-exon-junction content. Finally color-space reads got translated into base-space.

Transcriptome analysis

Further analysis was carried out on transcriptome reads. We defined the transcriptome as all known exons in the genome as given by the EnsEMBL 56 gene annotation. Any read which overlapped by 1bp to a known exon was classified as a transcriptome read.

Homopolymer analysis

To analyse in particular the amount of adenines close to stacked sequenced repeat regions, termed pileup regions, three bioinformatic analysis steps were taken. First, all repeat regions with sequence coverage over 40 were collected in one file. In the second step sequenced regions were extended by 15 bases in the flanking regions. In the final analysis step the last twenty bases, 5 bases of repeat region and 15 bases of flankeing genome sequence, were analysed for the amount of 8-mer homopolymer stretches such as poly A/T/G/C. As a control, one million randomly generated 20mers were also analysed.

Gene expression quantification

Apart from any possible bias, it is also important to find out the robustness and reproducibility of the reverse transcription and amplification. Therefore the amount of read counts per transcript (normalized to their length) and the total amount of reads were compared between the samples of one assay. This analysis is termed RPKM and the generated score has a value for read counts per kilobase of an exon model per million mapped reads. As the median cDNA length for the linear amplification assay is between 300-500bp, the normalization for the linear amplification assay had to be set differently (2.2.3). Therefore, all genes up to 500 bp (the concatenated exon length only) were normalized using their actual length and every transcript longer than 500 bp were normalized to 500. Thus the RPKM values between the samples of the exponentially amplified assay, which gain full length cDNA, and the linear amplification can be compared. To get a better overview of the whole transcriptome, analyses were made at the gene level. We used the Ensembl 56 gene set which included an annotated gene count of 47,507. Transcription rate of a gene was subdivided into three classes, low, medium and highly abundant. The lowly transcribed genes had a minimal read count per gene of 10 and an RPKM score of 1,236. The medium level of expression had at least 50 read counts per gene and an RPKM score of 3,558. All

Genes with more than 500 read counts per gene and an RPKM score higher than 35,778 were classed as highly expressed.

Coverage length distribution

The identification of new splicing variations requires full length cDNA transcripts. Therefore, the exponential and linear amplification methods were studied for the amplified cDNA length. The complete set of reads mapping to the exome respective transcriptome was taken and the length of coverage for each gene was computed using the BEDTools software suite (BEDTools, Quinlan, AR and Hall, IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.).

3. Results

3.1 Reverse transcription

3.1.1 Processivity of reverse transcription enzymes

Processitvity of three different Enzymes, M-MuLV (Enzymatics) M-MLV (Ambion) and SuperScript III (Invitrogen) were tested. 2 µg of a polyadenylated RNA ladder, Invitrogen, was used as template (Figure 6). To digest the remaining primers exonuclease (E. coli) treatment was maintained. All reverse transcriptions assays were based on oligo(dT) priming. No differences in efficiency between the in house protocol and cell-to-cDNA kit were measured. A shift of the 1.5 kb fragment and a smear from 0.1 up to 1.0 kb in the SuperScript III assay compared to the other can be distinguished. Due to the price difference between the commercial available kits and the in house protocol, further single reverse transcription reactions without amplification were done with the in house protocol.

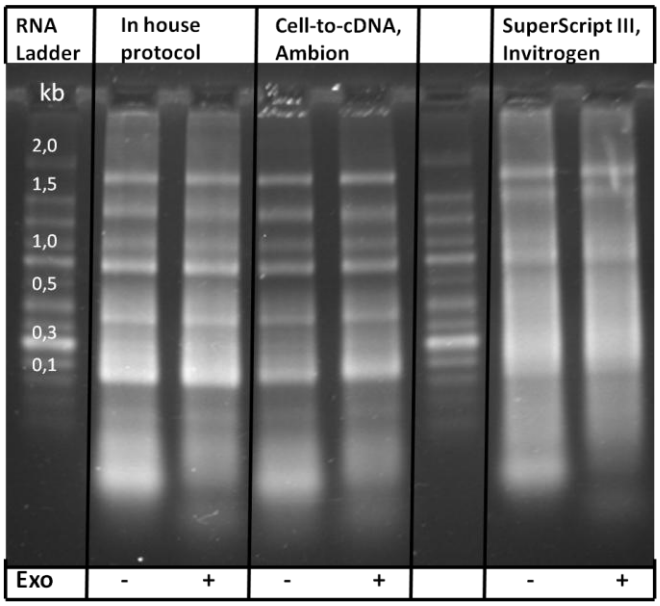
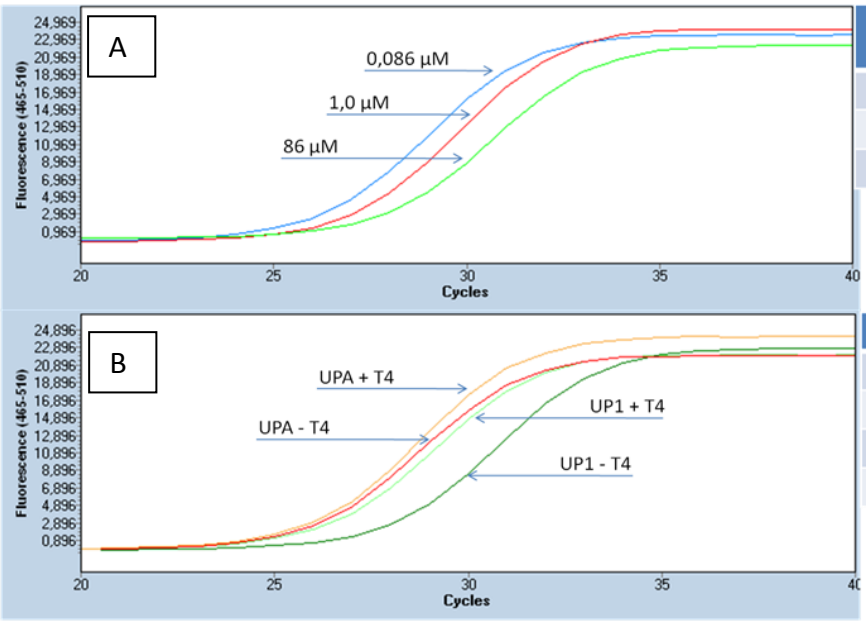


Figure 6 **Processivity of different RT – enzymes**
M-MuLV, Enzymatics, M-MLV, Ambion and SSIII, Invitrogen were compared in a UP1-Poly(T) priming based assay. Digestion of remaining primers was done with an exonucleoase treatment (+/-).

3.1.2 Primer concentration vs. Reverse transcription efficiency

To set up a reverse transcription assay with the best possible cDNA yield, primer concentration and the use of a single stranded binding protein, T4 gene 32 protein, were tested with 500 HeLa cells (Figure 7A & B). UP1-oligo (dT) Primer was used to investigate a possible dependency of concentration and cDNA yield (Figure 7A). A possible effect of the

single strand binding protein was tested in the UP1-oligo (dT) and the UPA-oligo (dT) assay (Figure 7B). cDNA yield of one µl RT reaction was measured with TaqMan ß-actin qPCR. Highest cDNA yield was reached with the lowest primer concentration. The use of ss-binding-protein has a strong positive effect on cDNA yield, ~ 3 fold higher, in the UP1 primer assay but not in the UPA assay.



Primer conc. (µM)	MeanCp	STD Cp
0,068	25,67	0,11
1,02	26,56	0,21
86	27,42	0,27

Assay	MeanCp	STD Cp
UPA + T4	25,76	0,18
UPA – T4	25,60	0,12
UP1 + T4	25,92	0,11
UP1 – T4	27,45	0,43

Figure 7 **Influence of primer concentration and T4 gene 32 protein on RT efficiency**
A: Three different primer concentrations and its influence on cDNA yield were tested. B: UP1 & UPA priming assay were tested with (+) and without (-) the single strand binding protein T4 gene 32. Quantification of cDNA was done with the ß-actin TaqMan assay.

3.2 *Whole transcriptome preparation of a single cell, ABI*

3.2.1 Real time PCR analysis of the synthesised cDNA

Cells were sorted with FACSDiva and reverse transcriptase efficiency was checked by ß-actin TaqMan real time PCR. Because of high primer dimerization between UP1 and UP2 (Primer alignment and gel electrophoresis see figure 1 & 2, supplementary) a new primer was designed, UP2, and comparative tests were carried out (Table 2). None of the four one-cell samples and just two out of four two-cell samples showed a positive real time PCR result independently of the reverse transcriptase primer. Comparing the Cp after reverse transcriptase no significant difference between using UP1 or UPA as RT primer can be detected. The first round of 18 cycles of amplification lead to an increase of between 200 to 1000 fold. Figure 8A is an example of the amplification curves of one up to 100 cells. In this case reverse transcriptase was carried out with the UP1 primer.

Table 2 **ß-actin cDNA Cp values of various cells numbers**

	After reverse transcriptase				After first round of amplification			
Cell number	MeanCp		Deviation		MeanCp		Deviation	
	UP1	UPA	UP1	UPA	UP1	UPA	UP1	UPA
0	> 45	> 45	0	0	> 45	> 45	0	0
1	37,44	> 45	1,02	0	> 45	> 45	0	0
1	> 45	34,64	0	0,68	> 45	> 45	0	0
2	33,8	> 45	0,13	0	> 45	34,84	0	0,8
2	32,58	37,61	0,1	0	24,88	27,22	0,18	0,18
5	31,93	31,39	0,35	0,2	23,82	22,78	0,15	0,04
10	30,25	30,73	0,19	0,11	21,09	21,38	0,13	0,19
100	26,97	27,15	0,16	0,08				

Level of cDNA was measured after reverse transcriptase as well as after 19 cycles of PCR. Templates marked with red shaded boxes did not get a positive qPCR result up to 45 cycles.

After the second round of amplification the five and ten-cell sample amplified by UP1 & UP2 yielded a Cp of 22.75 with a deviation of 0.03 whereas the same amount of cells amplified with UPA & UP2 reached a Cp of 18.7 with a deviation of 0.03 (Figure 8B). This meant a 16 fold higher amplification with the primer pair forming much less dimerization. After the second round of amplification the difference in the amount of cDNA between the five and the ten-cell sample is not detectable anymore.

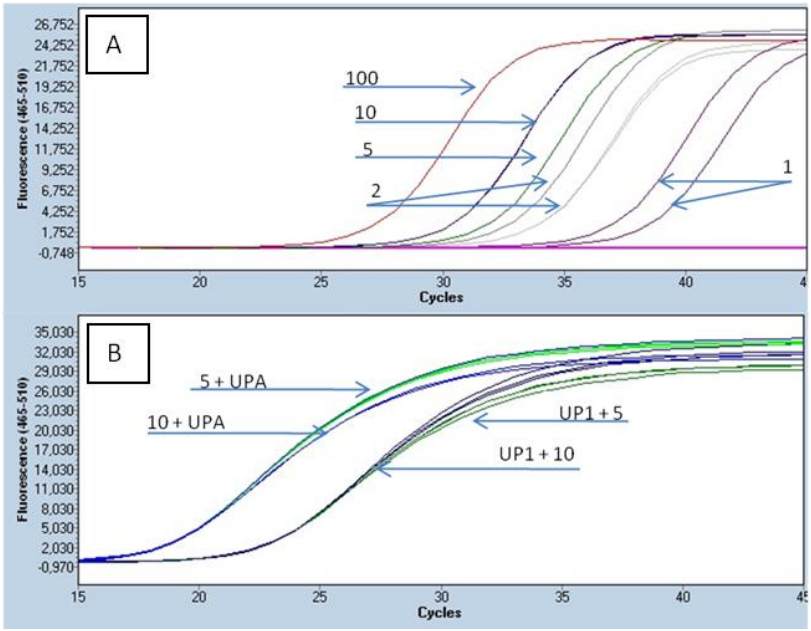


Figure 8 **Amplification curves of cDNA from none up to 100 cells**
A: 45 cycles of TaqMan ß-actin assay were performed. Numbers represent the amount of cells used as input material. B: 45 Cycles + 18 cycles of TaqMan ß-actin were performed and amount of cDNA measured. Templates were diluted 1:100. Cp values are listed in table 2.

3.3 µMACS SuperAmp Kit, Miltenyi

Three samples were generated by mouth pipetting, two one-cell samples (1-1 & 1-2) and one twenty-cell sample. Reverse transcription was maintained as described in the manual. One of the two one-cell samples (1-2 + Bio) and one half of the split twenty-cell sample (1/2 20 + Bio) were amplified with modified primer. A 5´-biotin should inhibit further ligation steps in the library preparation. After shearing fragments containing primer sequencing will not become ligated to the library adaptor sequence and therefore neither amplified nor sequenced.

3.3.1 Gel electrophoresis

In a pilot test with one zero-cell sample, two twenty-cell samples and 500 pg of polyadenylated RNA ladder, the enzyme efficiency was tested. Based on the ladder, reverse transcription and amplification was successful up to 1.0 kb. Almost no differences between the zero-cell sample and the twenty-cell sample can be detected by gel electrophoresis. Different test assays had to be set up.

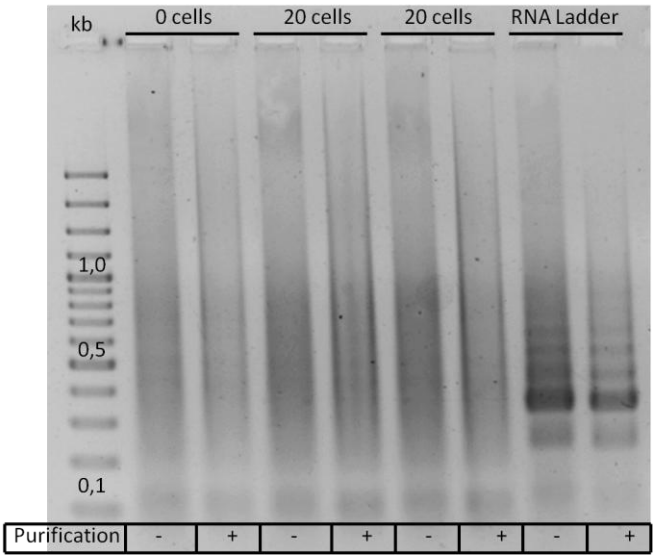


Figure 9 **Gel electrophoresis amplified cDNA** 5 µl of amplified product of zero-cell, two twenty-cell and 500 pg RNA ladder were separated on a 1,5 % agarose gel. Each product was separated before (-) and after (+) purification.

3.3.2 Real time PCR & Spectrometric measurements

Reverse transcriptase and amplification were controlled by real time quantification and spectrometer measurements. Crossing points of the real time and yield detected by the Nanodrop are listed in table 3. As seen in 3.2.1 single cell cDNA cannot be detected without amplification with this assay. Therefore, reverse transcription and especially amplification was successful, even though template 1-1 yielded a low cDNA output (Table 3 / Figure 10A).

Table 3 **Cp values and Spectrometric quantification**

Cell number	PCR primer		Mean Cp	STD Cp	Nanodrop	
	biotinylated	non - biotinylated			ng / µl	50 µl (µg)
1 (1-1)	-	+	32,71	0,09	149	7,5
1 (1-2)	+	-	18,27	0,14	240	12,0
1/2 20	-	+	19,54	0,06	154	7,7
1/2 20	+	-	17,76	0,02	248	12,4

For real time analysis templates were 1:100 diluted. Hprt SybreGreen assay was used for relative quantification. Absolute quantification was done with Nanodrop after.

Samples were 1:100 diluted for quantification via real time PCR. Biotin modification has no inhibiting influence on PCR efficiency. In this experiment it seems like that this primer modification has even improved the reaction. Amplification curves are shown in figure 10A. Sample 1-2 and both twenty-cell samples reach a Cp value of approximately 18 after a 1:100 dilution. Therefore amplification results in a ΔCp of at least 22 which means a minimal rate of 4 X 10^6 fold. The 1-1 sample amplicon has the same melting curve as the others (Figure 10B). Therefore, it is not an artefact of genomic DNA or even RNA as shown in paragraph 3.4.1. Real time PCR was carried out mainly with one assay, SybrGreen Hprt. Primers are designed to be exon spanning and therefore no genomic products are amplified (Figure 3, supplementary).

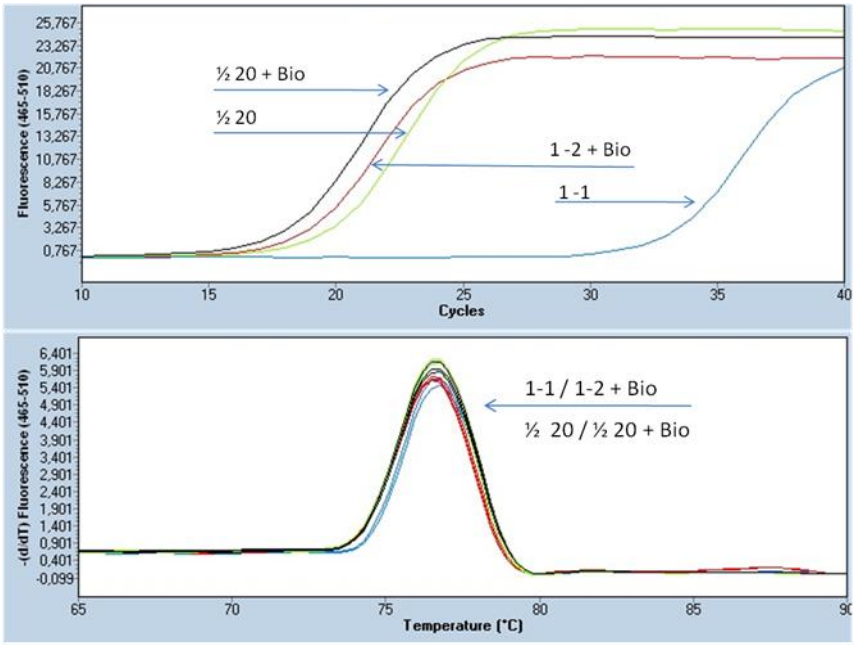


Figure 10 **Amplification curves from cDNA and melting analyses of amplicons**
A: Hprt SybrGreen assay was used to check RT and amplification efficiency of two one cell samples and one split twenty cell sample. B: Melting curve of the amplicon from sample 1-1/1-2 and both 20 cell templates.

3.3.2 Sequencing results

Biotin / non- Biotin

Table 4 **Number of read counts for amplification primer**

Sample	Normalised primer counts per 1 milion reads
1-1	574.554
1-2 + Biotin	118.671
1/2 20	24.411
1/2 20 + Biotin	32.739

To amplify the cDNA in an exponential manner a universal primer sequence is required on both sides of the fragment. With a forward and a reverse primer a PCR can be obtained with these sequences. This also means that every amplified fragment starts and ends with primer sequences which then get sequenced. To avoid this loss of sequencing capacity primers were modified with biotin. A standard fragment library was made of sample 1-1, 1-2 with biotinylated primer, ½ 20, and ½ 20 with biotinylated primer and sequenced on the SOLiD platform. After amplification and shearing of all samples to 100 bp, the in the biotinylated primer samples fragments carrying primer sequence were depleted by magnetic streptavidin bead treatment. Successively the normal procedure of library prep was carried out with the resulting supernatant and the complete non biotin templates. After normalization the read counts for the primer sequences samples with modified primer and depletion were compared to the sample with unmodified primer. Sample 1-1 without biotin labeld primer and without depletion has 4.7 times more counts for primer sequence. The twenty cell sample with biotin and depletion has 1.3 times more counts for primer sequence.

Analysis of sequencing reads

Initial analysis of the sequencing reads was made using the whole transcriptome pipeline, integrated in the Bioscope package, from Applied Biosystems. Processing was split into three steps: Filtering (repetitive sequences, run chemistry, rRNA, tRNA, etc.), mapping against the reference genome (HG19) and mapping against exon-exon junctions (generated internally using refGene.gtf version hg19). Out of these results two files were generated. One containing all filtered reads and the other containing all reads mappable to the human genome and possible exon junctions. The file containing the human mappable reads were subsequently re-mapped to the reference genome with Bowtie. In the first analysis the number of unique genome reads, meaning reads mapping just to one position, were

considered. To find out the number of transcriptome reads all genome mappable reads were aligned against all known exons in the genome as given by the EnsEMBL 56 gene annotation. EnsEMBL 56 contains about 47.507 genes, both known and predicted ones. Any read which overlapped by 1bp to a known exon was classified as a transcriptome read. Transcriptome reads were clustered into either unique mapping reads or reads mapping to one or more position within the exome, non-/unique reads. Table 5 shows counts for the different mapping results for the two one-cell samples, 1-1, 1-2 + Biotin and the two split 20 cell sample, ½ 20, ½ 20 + Biotin. Exome read counts are shaded red. Comparing the percentage of genome mappable reads and sequencing run quality leads to the result that samples 1-2, which had a poor run quality, shows the smallest mappable rate of 13%. Sample ½ 20 which had the best run quality contains 80% of mappable genome reads (Figure 5, Supplementary).

Table 5 **Absolute read counts for two one cell sample and a split 20 cell sample**

	Total reads	Filter	Unknown	Genome		Transcriptome	
Template				non-unique	unique	non-/unique	unique
1-1	25.758.637	10.267.654	-1.069.562	7.761.197	8.799.348	348.845	298.820
1-2 + Biotin	37.031.088	17.902.395	14.321.020	1.639.927	3.167.746	1.617.919	1.179.913
1/2 20	17.093.714	3.446.734	5.774.839	5.731.702	8.024.423	8.682.992	5.847.749
1/2 20 + Biotin	23.059.104	3.528.140	1.678.971	7.323.334	10.528.659	10.864.621	7.367.542

Diagram of sequencing reads

The four samples taken together generated around 103 million reads. The one-cell sample 1-1 yielded close to 26 million reads and the one-cell sample 1-2 reached 37 million reads. The split twenty-cell samples, ½ 20 and ½ 20 + Biotin lead to 17 and 23 million reads respectively. The most important part, the amount of transcriptome reads differs considerably between the samples. Sample 1-1 has 1% transcriptome reads (~ 349.000), sample 1-2 + Biotin has about 4% (~ 1,6 million) reads, sample ½ 20 reached 51% (8.7 million) reads and sample ½ 20 + Biotin contains 47% transcriptome reads (10.7 million). This number of transcriptome reads are based on the non-unique transcriptome reads. Further analyses were done with the non-/unique mapped reads on the transcriptome due to the fact that the non-unique count is based on gene duplication but still has to be considered as transcriptome. A noticeable difference between the one-cell samples and the twenty-cell samples is in the amount of filtered reads. Whereas for the one-cell samples around 40 to almost 50% were filtered out, the 20 cell sample had between 10% and 20% (Figure 11 / 12). The biotin

treated samples, 1-2 contains a high number of non-mappable reads, 39 %. Sample ½ 20 + Biotin has about 5 % unknown sequences and sample 1-1 & ½ 20 had few non-mappable reads.

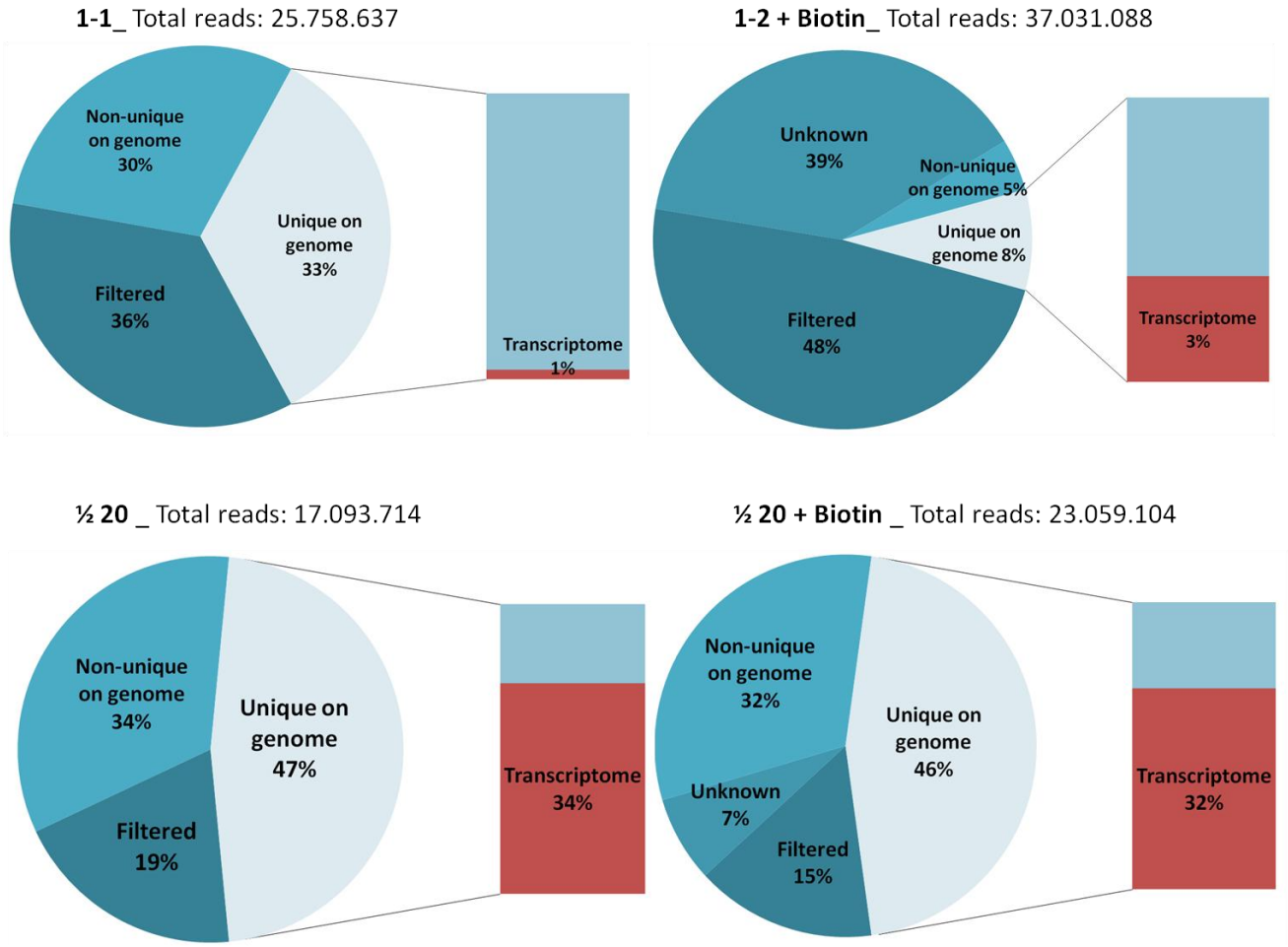


Figure 11 **Diagram of sequencing reads**
Four cDNA samples were sequenced on the ABI SOLiD platform: Two one cell samples, 1-1 and 1-2 + Biotin and one 20-cell sample which was split up after cDNA synthesis, ½ 20 and ½ 20 + Biotin. Shown is the distribution of reads mapped against the filter and the human genome. Transcriptome read counts analyzed by mapping against the exom of EnsEMBL 56 gene annotation via Bowtie.

Filtered reads

Before mapping the sequencing results to the human genome, reads were processed in a filtering step. In this step reads containing adaptor sequences (used in the sequencing chemistry), Sine, Line, Poly-N stretches, t & rRNA and E. coli were filtered out. Almost the half the reads of both one-cell samples were filtered out. Over 80% of the filtered reads mapped to E. coli, mainly to 16S and 23S RNA (Figure 12). Also the filtered reads of the

twenty-cell samples contain over 80% E. coli sequences, however, the total amount of filtered reads is just about the half compared to the single cells.

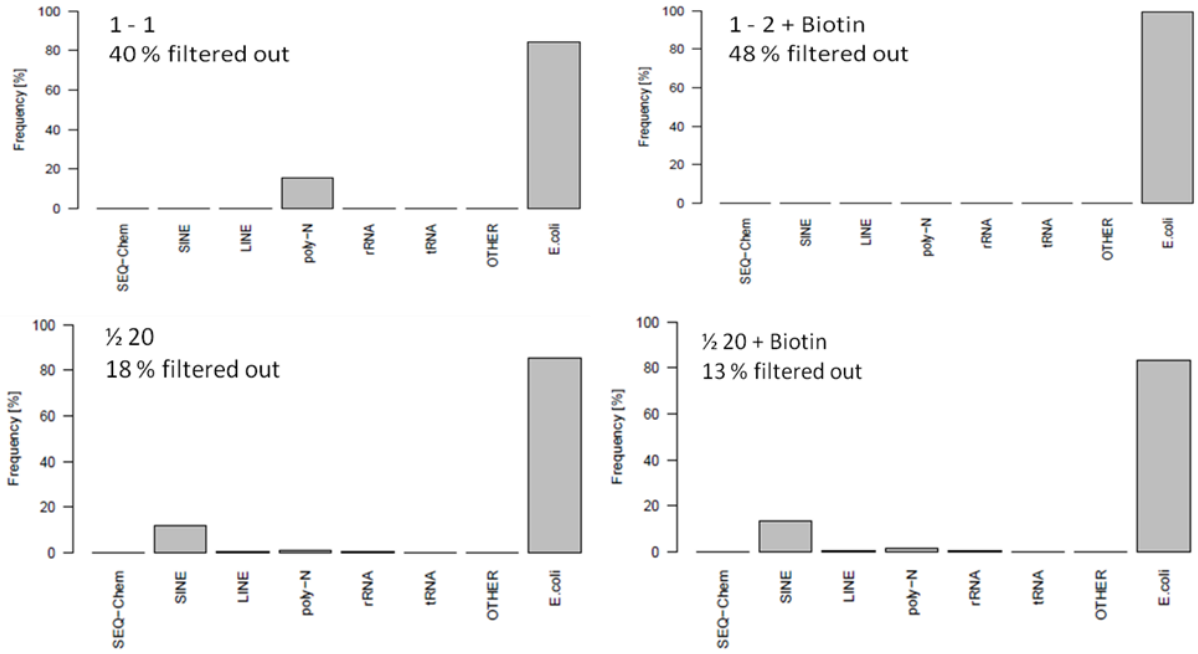


Figure 12 **Filtered reads of the four samples amplified with the µMACS SuperAmp Kit**
As a first analysis step all sequences are mapped to various filtering sequences like sequencing adaptors, rRNA or E. coli. The filtered reads are shown for samples, 1-1, 1-2 + Biotin, ½ 20 and ½ 20 + Biotin. X scale shows the various filters and the y scale the % part per filter to the total filtered reads.

3.4 WT Ovation, One-Direct RNA Amplification System, NuGEN

3.4.1 Gel electrophoresis

To test the efficiency of the reverse transcriptase used in the kit, an advanced test with zero-cells, twenty-cells and 500 pg RNA ladder was made. By simply looking at the 1.5% agarose gel it is difficult to identify differences between the patterns of the zero-cell sample, the 20 cells or the RNA ladder sample (Figure 13). In the lines of the RNA ladder no distinguishable band of the fragments can be detected. Purification of the amplified product does not lead to a big change in the smear.

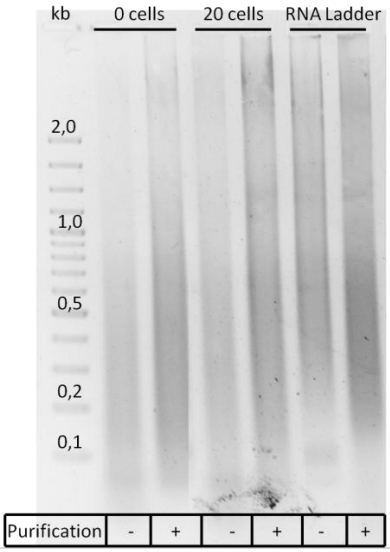


Figure 13 **Gel electrophoresis of amplified cDNA with the One-Direct RNA Amplification System**
Three reactions zero-cell, twenty-cells and 500 pg of RNA ladder were used to test the RT efficiency. Samples were analysed before (-) and after (+) purification.

3.4.2 Real time PCR & Spectrometric measurements

Samples without amplification and transcribed with the in-house protocol were compared to samples which got transcribed and linearly amplified with the One-Direct RNA Amplification System from NuGEN. With this experiment reverse transcription efficiency and especially amplification rate was measured. In particular two one-cell samples and one 50 cell sample were transcribed into cDNA using the in-house RT protocol without any further amplification. One 50 cell sample was lysed but neither transcribed nor amplified. Furthermore, three one-cell samples and one further 50 cell sample were transcribed and amplified with the One-Direct RNA Amplification System from NuGEN. The resulting non-amplified and amplified cDNA samples were quantified by real time PCR with the HPRT SybrGreen assay (Table 5). None of the samples transcribed into cDNA with the in-house protocol and no amplification showed a Cp under 40. This was also the case for the one-cell sample, 1-2, and the non-cell sample treated with the One-Direct RNA Amplification System from NuGEN. Samples 1-1, 1-3 and the 50 cell sample (NuGEN) showed a Cp of between 28 to 31 (Table 6 / Figure 14A). This meant an amplification factor of between 600 and 4000. Amplified samples were diluted 1:100 to avoid any inhibition of the real time PCR. The 50 cell sample without reverse transcription and amplification had a Cp of 32.16 but a different melting peak than the transcribed and amplified cDNAs (Figure 14B).

Table 6 **Cp values of amplified product and spectromatic quantification**

Cell number	Assay		Dilution	MeanCp	STD Cp	Nanodrop	
	RT	Ampl.				ng / µl	32 µl (µg)
1	+	-	-	> 40	0	-	-
1	+	-	-	> 40	0	-	-
50	+	-	-	> 40	0	-	-
0	+	-	-	> 40	0	-	-
1 (1-1)	+	+	1 : 100	30,91	0,1	307	9,8
1 (1-2)	+	+	1 : 100	> 40	0	193	6,2
1 (1-3)	+	+	1 : 100	29,07	0,07	291	9,3
50	+	+	1 : 100	27,84	0,33	354	11,34
0	+	+	1 : 100	> 40	0	70	2,24
50	-	-	-	32,16	1,54	-	-

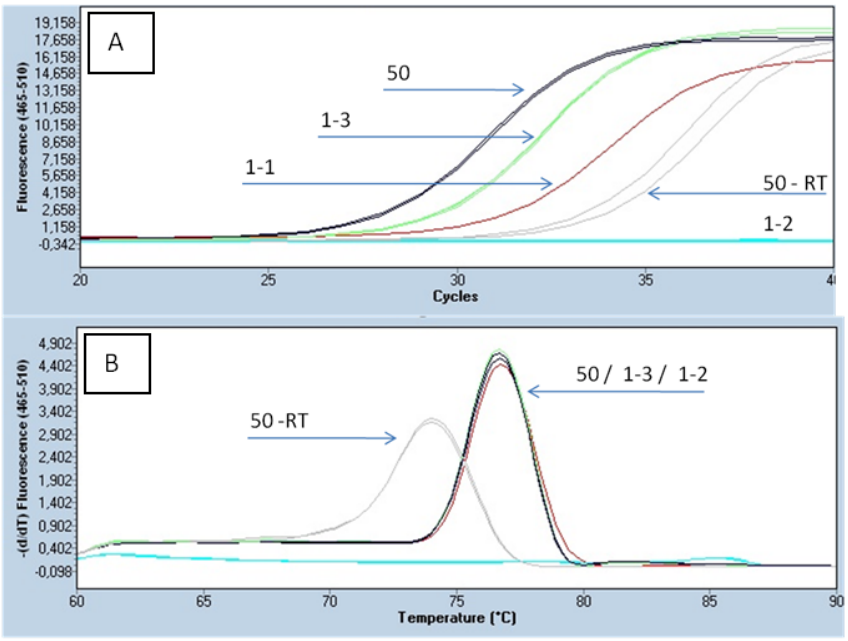
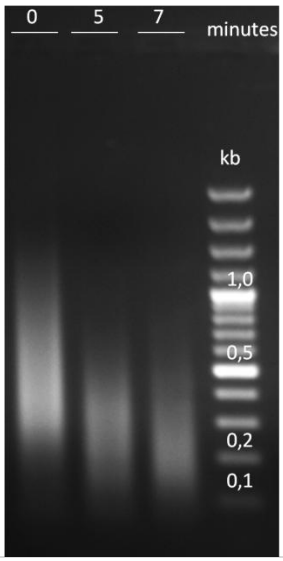


Figure 14 **Amplification curves and melting peaks**
A: Amplification curves of three one cell samples and one 50 cell sample amplified with the One-Direct RNA Amplification System from NuGEN together with one 50 cell sample without RT and amplification. qPCR was done with the Hprt SybrGreen assay. B: Melting peak of the 1-1,1-2,1-3, 50 cell sample and the 50 cell sample without RT and amplification.

3.4.3 Library Preparation

Shearing

All samples should be fragmented between 100 to maximal 200 bp for successful sequencing. This was done by the Covaris S2 system. Figure 15 shows sample 1-1 in the first lane without any shearing, in the second line with 5 minutes of shearing and in the third lane with 7 minutes. The unsheared product has as expected an average size of 300 to 500 bp.



After 5 minutes of ultrasonic treatment of the single stranded linearly amplified product the main fragments have a length of approximately 300-400 bp. After two more minutes most fragments have a size of around 200 bp. For the final library preparation sample 1-1, 1-2, 1-3 and 50 were sheared for 8 minutes.

Figure 15 **Different shearing length**
A one cell sample was sheared for 0, 5 and 7 minutes.

Trial PCR

In a pre-PCR step, termed trial PCR, the maximal number of PCR cycles is determined. This prevents the samples from getting over amplified. Trial PCR was carried out in 17 cycles and after 11, 13, 15 and 17 cycles, samples were taken and analysed by performing gel electrophoresis. In sample 1-1, 1-2 and 1-3 amplified product can be detected after cycle 15. Sample 50 is not detectable before cycle 17 (Figure 16).In the first line per sample starting material was loaded onto the gel to compare the amplicon with it. Large scale PCR was finally carried out with 17 cycles.

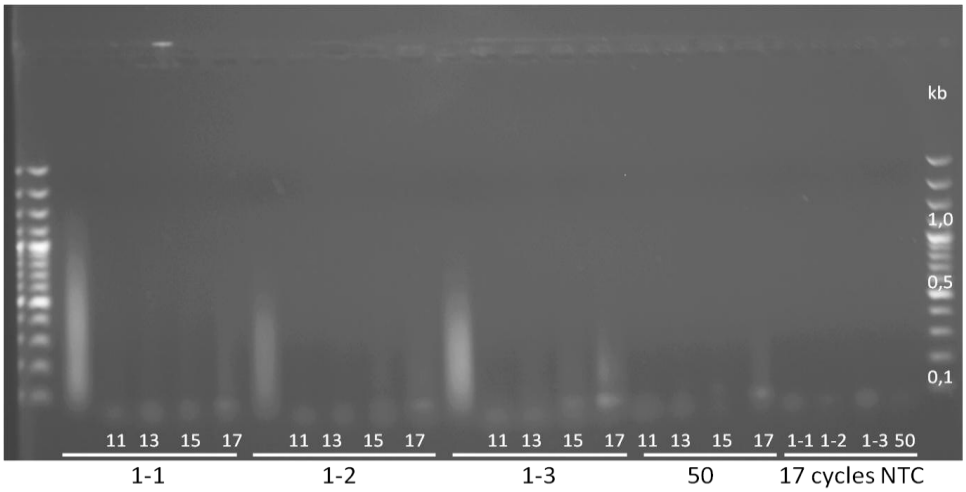


Figure 16 **Trial PCR of the sequencing library**
PCR was done with a library of samples 1-1, 1-2, 1-3 and 50. After 11, 13, 15 and 17 cycles, samples of the reaction were taken and analysed by gel electrophoresis. In the first line per sample the starting material was loaded.

3.4.4 Sequencing results

Sequencing on the ABI platform was carried out after barcoded library preparation for the single stranded product of sample 1-1, 1-2, 1-3 and 50. A full sequencing slide was used for the four barcoded samples.

Analysis of sequencing reads

Mapping and initial analysis of the sequencing reads was as described under 3.2.2.

Table 7 shows the counts for the different mapping results for the three one-cell samples 1-1, 1-2, 1-3 and the 50 cell sample 50. Exome read counts are shaded red.

Table 7 Absolute read counts for three one cell samples and one 50 cell sample

	Total reads	Filter	Unknown	Genome		Transcriptome	
Template				non-unique	unique	non-/unique	unique
1-1	118.461.094	40.082.685	24.653.460	20.871.847	32853102	16112806	13952455
1-2	105.370.964	15527262	0	36.842.111	53001591	27893618	21681537
1-3	68.729.113	25061091	0	14.932.538	28735484	11064810	9282532
50	65.572.209	27995709	0	12.840.052	24736448	1577729	1396987

Diagram of sequencing reads

The total amount of reads range between close to 70 million for the 50 cell and the 1-3 samples, 105 million for sample 1-2 to 118 million reads for sample 1-1. Each of the four samples contained between ~30 to 40% filtered reads (Figure 17).

Almost 100% of the filtered reads from sample 1-1 are library adaptors and for sample 1-2, 1-3 and 50 between 50 to 60%. The remaining filtered reads are mainly SINE and Line reads. Bar plots of filtered sequences and their frequency are shown in figure 4, supplementary. The amount of transcriptome reads ranges from 26.5% for sample 1-2 to 16% for sample 1-3 and 13.5% for sample 1-1. The sequencing results of the 50 cell sample contained just 2.5% of transcriptome reads. The number of transcriptome reads given includes also the non-unique transcriptome reads. Further analyses were done with the non-unique mapped reads on the transcriptome due to the fact that the non-unique count is based on gene duplication but still has to be considered as transcriptome. Total numbers for all reads categories are listed in table 7.

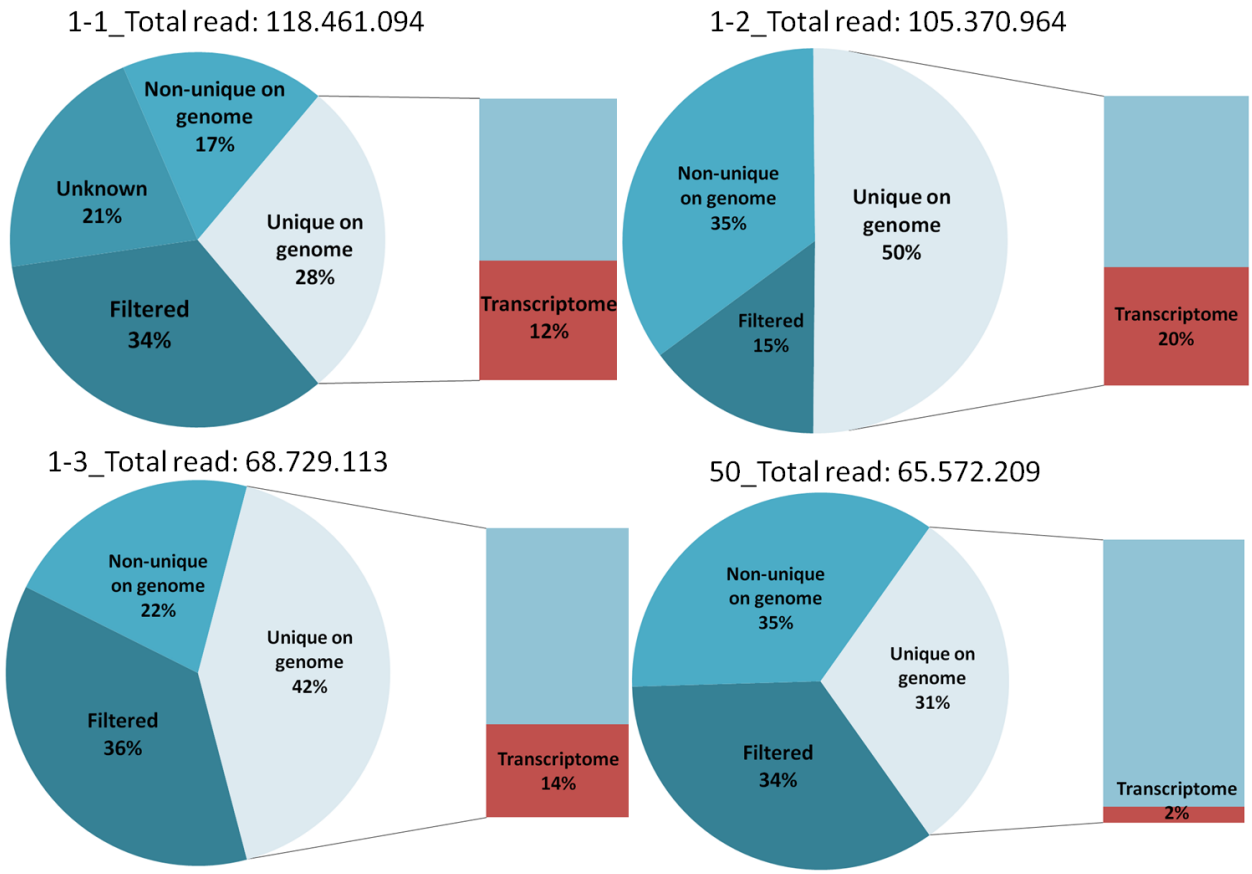


Figure 17 **Diagram of sequencing reads**
Four linearly amplified cDNA samples were sequenced on the ABI SOLiD platform: Three one cell samples, 1-1, 1-2, 1-3 and one 50 cell sample. Shown is the distribution of reads mapped against the filter and the human genome. Transcriptome red counts were analyzed by mapping against the exome of EnsEMBL 56 gene annotation via Bowtie.

Repeat region reads & content of adenines

To find the reason for the high number of random reads in genome (Figure 17), we mapped reads to repetitive sequences. The linearly amplified samples, NuGEN assay, have a significantly higher amount of repeat region reads than the exponentially amplified ones. This amount ranges from a minimal 32% for sample 1-1, 46% and 48% for sample 1-3 and 50 up to 61% for sample 1-2, whereas with the Miltenyi assay the average percentage of repetitive reads is maximal 20 (Table 8). No correlation between the amount of transcriptome and repeat reads was found.

Table 8 **Number and percentage of repeat region reads**

Assay	Sample	Repeat reads	Repeat reads [%] of total reads	Unique on genome excl. Transcriptome [%] of total reads
NuGEN	1-1	38.132.321	32	14
NuGEN	1-2	64.516.592	61	24
NuGEN	1-3	33.325.954	48	26
NuGEN	50	30.491.014	46	35
Miltenyi	1-1	14125431	55	33
Miltenyi	1-2 + Biotin	1973191	5	4
Miltenyi	1/2 20	0	0	0
Miltenyi	1/2 20 + Biotin	3681147	16	0

The stacking of repetitive sequence reads on top of each other, termed pile ups, requires high amplification & sequencing rate. Therefore, the Oligo (dT)-SPIA primer of NuGEN must somehow anneal close to these regions. To analyse this phenomenon the content of 8mers in the flanking regions of repetitive sequences were counted. A minimal coverage depth of 40 reads was required to include a repeat region in the analyses. The chance of having a stretch of eight adenosines, Poly A (8), in the flanked regions of 133.358 analysed repetitive sequences is 67%. Under same conditions the chance to have a Poly T(8) is about 16%, 9% for Poly G(8) and 8% for Poly C(8) (Figure 18B). As a reference 1.000.000 20mers randomly located on the genome were checked for any content of Poly N(8) stretches. The percentage rate of Poly A(8) and Poly C(8) is 26% and of Poly T(8) and Poly G(8) 24% (Figure 18A).

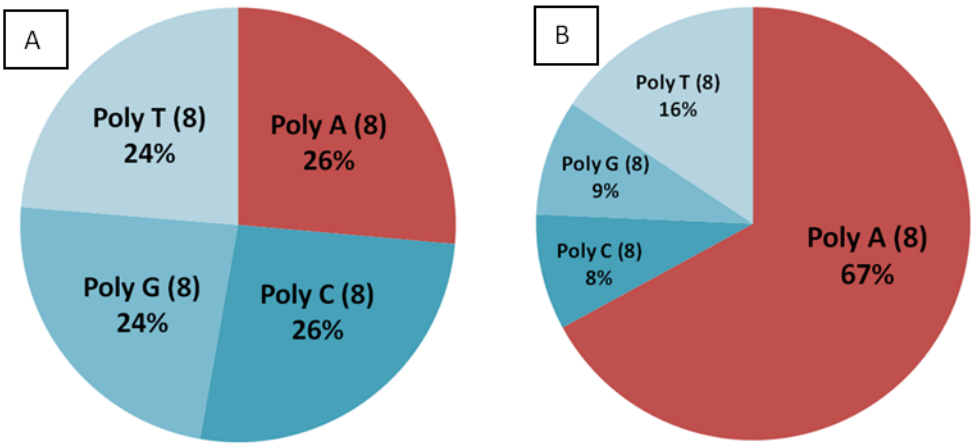


Figure 18 **Percentage distribution of Poly N(8) stretches in twentymres**
A: Randomly analysed 20mers all over the genome. B: 20mers flanking repeat regions with a higher sequencing coverage than 40.

3.5 Comparative analysis of exponential and linear amplification

3.5.1 Gene expression analysis

Sample complexity / detected genes

The number of detected genes was taken as a measure for library complexity. The Ensembl 56 gene set which included an annotated gene count of 47.507 is used as reference. Linear amplification yielded to a detected genes number of about 26.000 across all samples. Exponentially amplified samples resulted in a range of 15.000 and 23.000 detected genes. With the twenty-cell samples compared to the one-cell samples about 10 % more genes could been detected. Therefore, connection between higher input material and a more complex library can be made (Table 9). A threshold of three reads per gene had to be reached before a gene was counted as detected.

Table 9 **Number of detected genes with a coverage greater than three**

Sample	Amplification	Total reads	Transcriptome reads [%]	Detected genes	
				Total number	[%] of total Genes
1-1	Exponential	25.758.637	1	15.277	32
1-2 + Biotin	Exponential	37.031.088	4	16.932	36
1/2 20	Exponential	17.093.714	51	21.563	45
1/2 20 Biotin	Exponential	23.059.104	47	23.268	49
1-1	Linear	118.461.094	14	26.352	55
1-2	Linear	105.370.964	26	26.212	55
1-3	Linear	68.729.113	16	26.584	56
50	Linear	65.572.209	3	25.930	55

Rate of gene expression

One way to analyse sequenced results to get more information about the strengths and weaknesses of a method is to compare the transcription levels per gene. Therefore all detected genes were clustered in three levels of transcription. In the group with low expressed genes with a read count of 10 - 49 and a normalized value of 1,236-3,558 were put together. Normalization was implemented as described in paragraph 2.2.10. Medium expressed genes were those with a read count of 50 – 499 and a normalization value between 3,559 and 35,778. Highly expressed genes have a read count greater than 500 and a normalization value higher than 35,778. Results from the linearly amplified samples were taken together and the mean with its deviation was calculated. This was done for the absolute read counts and the normalized values. The same calculation was carried out with

the exponentially amplified samples and all results were illustrated in one figure (figure 19A/B).

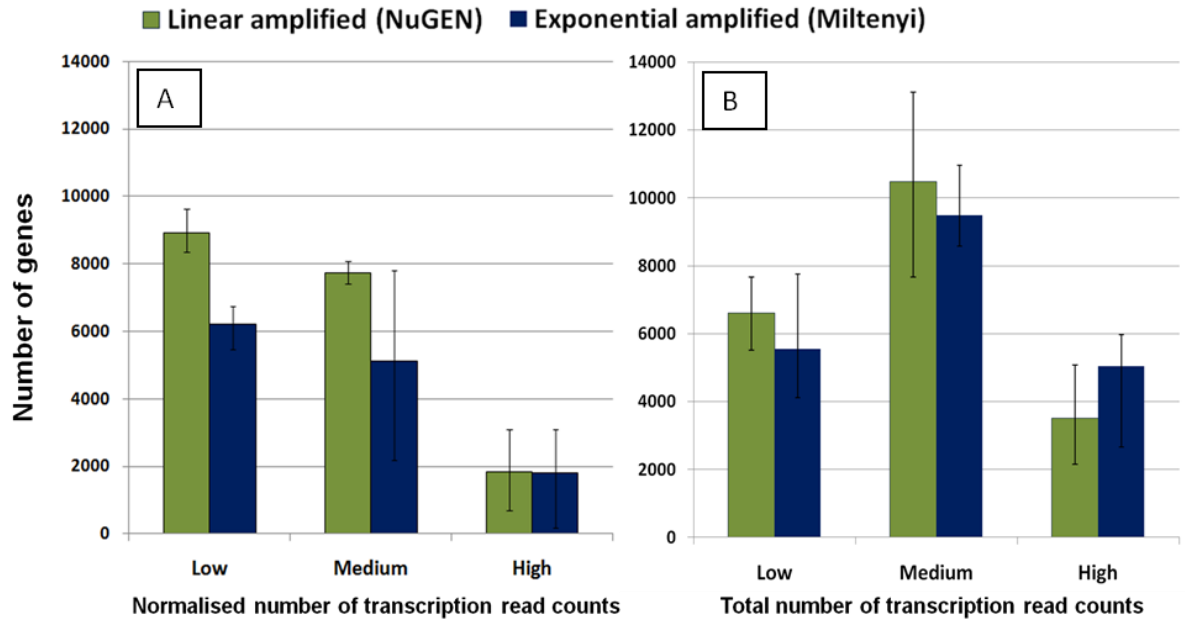


Figure 19 **Amount of detected genes clustered at low, medium and high transcription rates**
Bars show the mean gene number per transcription rate and its deviation of all linearly amplified sample respectively exponentially amplified ones. Low, medium and highly abundant means 10-49, 50-500 and more the 500 reads per gene for the total read count and 1,236, 3,557 – 35,778 and more then 35,779 for the normalized values.. A: Number of detected genes based on the total read numbers. B: Number of detected genes based on the normalized read numbers.

In this case the deviation gives a first impression about the robustness and reproducibility of the methods. Normalization has a strong influence on the number of genes per expression level cluster as well as in the observed variance. The linearly amplified samples have a good homologousy in all three groups. The exponentially amplified samples show higher variance. This cannot be concluded when considering only read counts. The gene number for the normalized set, when compared to the non-normalised read counts, is higher for the lowly transcribed genes and lower for the medium transcribed genes. In the highly transcribed gene set the number of genes decreases for both amplification methods but more drastically for the linearly amplified ones. More genes in the low and medium abundant group were detected in the linearly amplified samples and less for the strong transcribed genes compared to the exponentially amplified one. Taken together these results indicate that the normalization lead to a shift of genes into the medium transcribed level, that the linear

amplification method seems to be more reproducible and leads to a higher number of detected genes except of the highly abundant ones.

3.5.2 Strandedness of transcriptome sequencing results

One major benefit along with the predicted decrease in bias is the possibility to also get the strand information of every transcript. Figure 21 shows the high concordance between the sequences of plus and minus cDNA strand reads. Also remarkably in figure 21, sample 1-1 has fewer counts per genes which is due to the on average shorter fragment length. Linearly amplified samples in general show fewer read counts per gene and a wider range of values. But still the majority of genes express almost the same amount of plus and minus strands (22). For the NuGEN results it seems that the genome of the SW480 cells has a tendency to express more negative stranded transcripts.

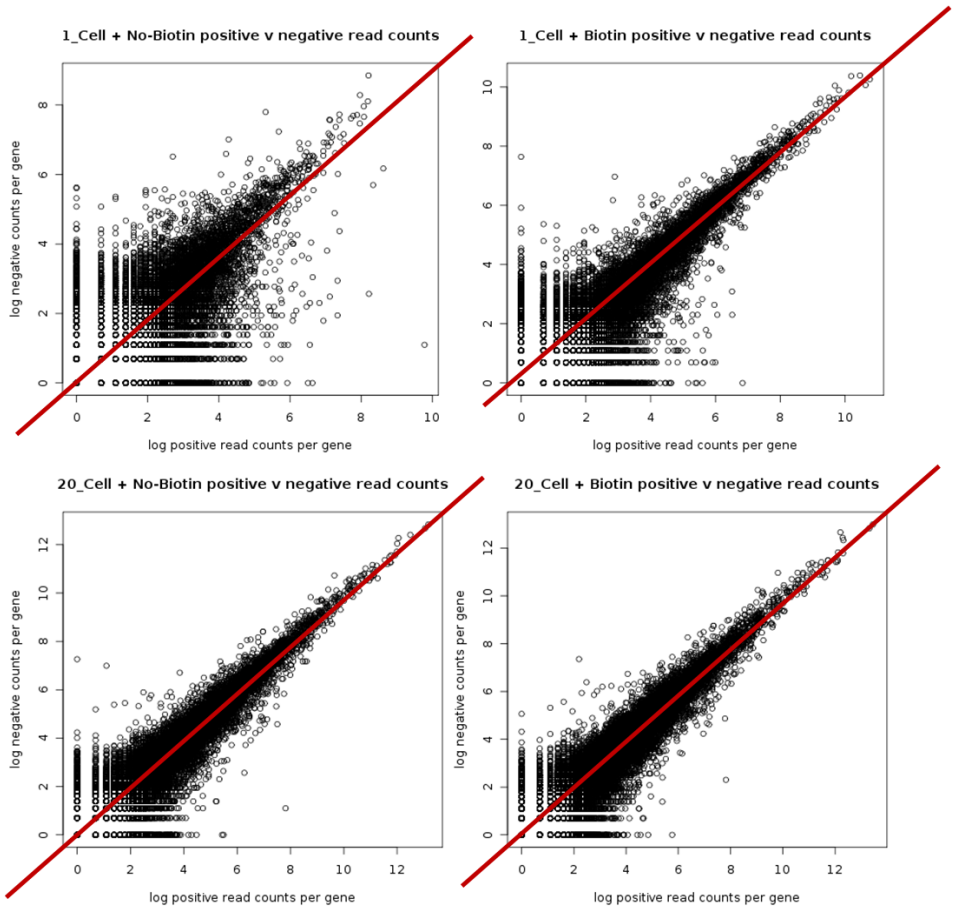


Figure 20 Correlation plots of the reads of plus and minus strands
Shown are two exponentially amplified one-cell samples and the two split twenty cell sample.

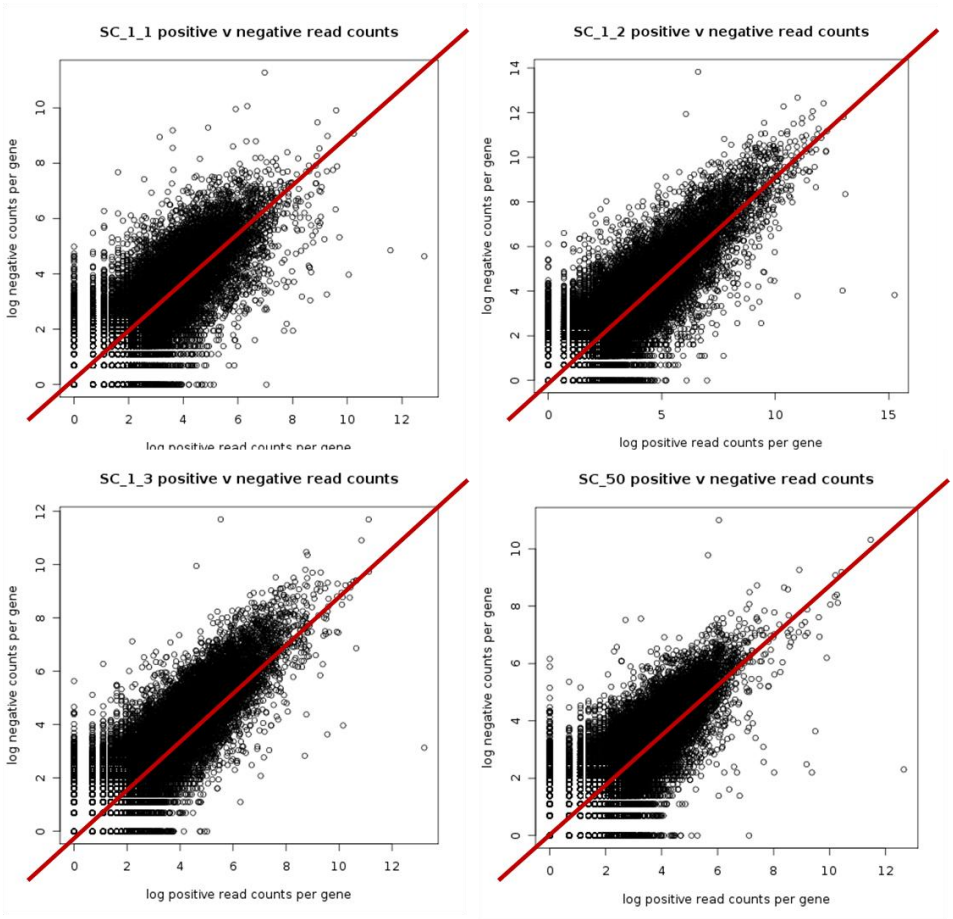
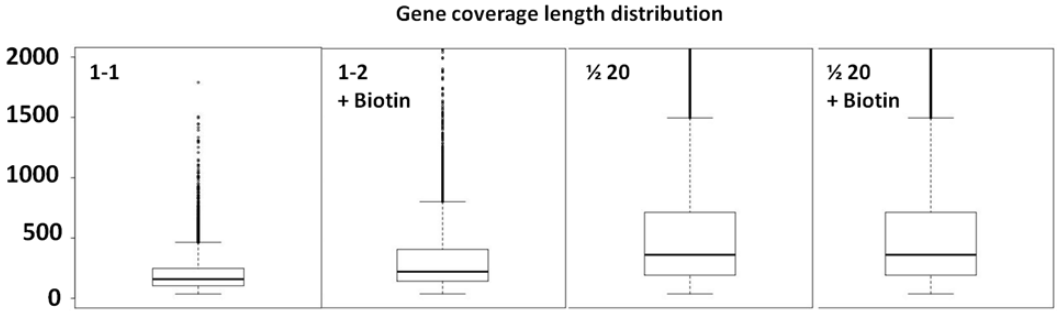


Figure 21 **Correlation plots of the reads of plus and minus strands**
Shown are three linearly amplified one-cell samples and one 50 cell sample.

3.5.3 Coverage length distribution

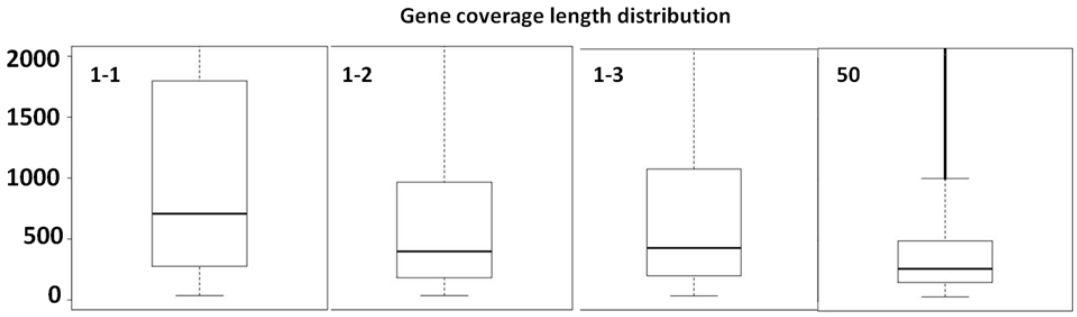
To access the length of the genes respective transcripts respectively genes that are covered by any read, the complete set of reads mapping to the transcriptome was taken and the length of coverage for each gene was computed. Due to non-strand specific sequencing the obtained data had to be corrected for bias due to overlapping genes which result in double counting for coverage calculation. The data set had been filtered as following: 1) genes covered by less than 5 reads were excluded from the analysis to reduce for background, 2) genes with a coverage length shorter than 35 bp's were excluded as the first 35 bases are high quality. For visualization of the distribution of the coverage length box-and-whisker diagrams are shown in figure 20 for the exponentially amplified samples and in figure 21 for the linearly amplified ones. Concatenating all exons per gene in the Ensembl 56 gene set the size of possible transcripts ranges from the smallest of 8 bp up to 620.154 bp´s. The median

is at around 1 kb and regarding to gel electrophoresis images the main transcripts range between 750 and 2.000 bp´s. The average covered gene length of the exponentially amplified samples is around 250 bp for the one cell sample and for the twenty cell sample between 300 and 400 bp´s. A connection between the reduced length and the amount of sequenced E. coli can be ruled out (Figure 20). The majority of sequenced transcripts have a length of up to 750 bp´s. Wherase the median length for exponetial amplified one-cell samples and twenty-cell sample were homologous the median length for the linearly amplified samples do not show any consistency (Figure 21). Sample 1-1 has a median of 700, sample 1-2 and 1-3 around 350 and sample 50 250 bp´s. Sample 1-1 shows a length distribution between 300 and 1700 bp´s. Sample 1-2 and 1-3 between 200 and around 1000 bp´s and the 50 cell sample between 200 and 450 bp´s. It has to be considered that according to the producer the average size of the amplicons is between 300 and 500 bp´s. No significant connection between the amount of repeat reads and the length of the sequenced transcripts can be drawn.



Total read	25.758.637	37.031.088	17.093.714	23.059.104
Transcriptome [%]	1	4	51	47
Repeat reads [%]	55	5	0	16
E. Coli [%]	40	48	18	13

Figure 22 **Gene coverage length distribution for the exponentially amplified samples (Miltenyi)**
Shown are box-and-whisker diagrams for the covered gene length distribution for two one-cell samples 1-1, 1-2 and a split twenty-cell sample ½ 20, ½ 20 + Biotin as well as the total reads, transcriptome, repeats and E. coli reads per sample.



Total read	118.461.094	105.370.964	68.729.113	65.572.209
Transcriptome [%]	1,5	12	5	2
Repeat reads[%]	32	61	48	46

Figure 23 **Gene coverage length distribution for the linearly amplified samples (NuGEN)**
Shown are box-and-whisker diagrams for the covered gene length distribution for three one-cell samples 1-1, 1-2,1-3 and a 50 cell sample as well as the total reads, transcriptome and repeats reads per sample.

4. Discussion

The First hurdle in working with single cells is to be able to select or separate a single cell. Therefore we started with a simple cell dilution series down to a single cell. High reproducibility and an easy handling ability were characteristics of this technique. Our aim in future is to work with single circulating tumour cells. Therefore a different way of single cell extraction has to be found. First studies working with the FACS gave promising results. The amount of input material and the high risk of contamination with unknown cells lead us to search for a more personal handling method. Therefore, I developed a technique to suck up a single cell by using a mouth pipetting strategy. After a few trials this became a simple and robust way to select single cells although not in a high throughput manner. The risks of sucking up more than just one cell accidentally or losing the cell by transferring it into the reaction tube have to be considered. Staining strategy of the target cell would simplify this method due to better discrimination of the target cell from the other cells.

The reverse transcription reaction is discussed to be not well understood and the most uncertain step in gene expression analysis [27]. Whereas oligo (dT) priming was found to lead to a 3´end bias and barely full length transcripts, random priming strategies on the other hand lead to 5´end bias. Either way the yield of cDNA especially full length cDNA for whichever strategy is used depends not on the assay but on the transcribed sequences [27]. Secondary and tertiary structure of the mRNA have a strong influence on transcriptase processivity and transcripts detected as being highly abundant may just be favoured by the reverse transcriptase because of certain sequence properties [2]. This bias should be stable and from the performed experiments a missing reproducibility of the reverse transcription was barely seen. In particular after fixation of the so called in-house protocol a stable transcription could be confirmed by real-time PCR even for low input preparations. This was the case not only within the experiments, where the same number of cells led to the same Cp´s, but also between different experiments. These advanced tests shows that in principle the reverse transcription step is not as uncertain as feared. A constant worklflow of performing the reaction, RNase free consumables, efficient protocol and a concentrated and speedy workflow are the guidelines which all need to be fulfilled. So far the reverse transcription and amplification kits used have not been combined with the in-house revere transcription protocol.

Although using a poly-dT priming approach for cDNA synthesis a focus of the study was to achieve a complete transcript coverage if possible. As shown in figure 6, with three tested RT enzymes similar results were achieved. Up to 2 kb fragments could be detected in all three assays. This disproves the assumption that oligo (dT) priming only yields short fragment cDNA´s. Tests using a commercially available poly-A RNA ladder between 100bp and 2 kb fragments showed that short transcripts down to 100 bp did not show a noticeably stronger transcription rate, which also questions the predicted strong 3´end mRNA bias. This could be further proven by setting up comparative real time assays located in an increasing distance starting close to the 5´end of the cDNA. Unfortunately the fragment sequences of the ladder were not publicly accessible. However, the sequencing data indicates that internal Poly A stretches might be induced if poly-dT oligos are used, which results in truncated fragments due to random priming. This also adds 'sequence noise' to the data and a strong influence of the oligo (dT) concentration on the reverse transcription reaction was shown (Figure 7A). At reduced oligo dT primer concentration a higher yield of cDNA was quantified via real time PCR. Increasing the primer concentration may lead to an annealing inhibition of primer and enzymes as well as the chance of annealing to internal sequences which would lead to truncated cDNA templates. Another explanation given by *Stahlberg et al* for the increased cDNA outcome could be the observation that less primer concentration results in a higher number of truncated fragments, generated by internal poly A priming compared to the 3´end fragments [27]. The HPRT Syber green assay used has a distance of almost 1 kb to the mRNA 3´end. No poly A stretches greater than 8 can be found in between. Therefore the increasing cDNA product is not based on more internal priming.

Random priming strategies are always considered to generate more full length coverage of transcripts than poly A based priming strategies. This would mean that in our study the assay based on exponential amplification, Miltenyi, would yield more full length covered transcripts than the linearly amplified assay which is predicted to generate between 300 to 500 bp´s long amplified cDNAs. Due to the lack of visualization methods for low input material reverse transcription could not be evaluated in completely separation from amplification. Therefore the covered read length distribution of the sequenced genes was taken as a way to analyze the initially transcribed cDNA gene covering length. The maximum length of the major fragments of the exponentially amplified cDNA samples is around 750 bp. A coverae length of up to 2 kp or more indicates that before the amplification initial

reverse transcription must have led to relatively long fragments of cDNA. Differences in the pattern in particular the high degree of short coverage length for some samples might have been caused more likely by amplification than by reverse transcription. It seems that the concentration of detected E. coli transcripts influences the length of the sequenced transcipts. More E. coli reads means less transcription length. Again the question is whether E. coli RNA inhibits the maintenance of full length human cDNA or if the E. coli cDNA transcripts and the short human cDNA transcripts lead to a loss of long products by preferred amplification.

DNA polymerase used in the linear amplification kit produces fragments between 300 to 500 bp´s even though the original fragment may have been longer. Longer fragments than expected can be seen in the covered length distribution of the sequencing reads which have the maximum length of the majority reads at around 1 kb. One one-cell sample even got closer to 2 kb. Only the 50 cell sample shows the expected length distribution.

Addressing reproducibility one has to be aware of naturally occurring biological cell to cell variability, which is for example caused by the cell cycle stage having direct impact on the mRNA content of the cell. Besides the methodological challenge of single cell transcriptome analysis this may be one of the limiting factors for a homogeneous transcriptome analysis of a single cell [12] and can be addressed only by working with high numbers of replicates. To reduce cell-to-cell variability comparability studies were firstly carreid out by working with nocodazole treated cells. This leads to a G2/M phase arrest of the cells because microtubules cannot polymerise and therefore they cannot enter mitosis [61]. Later test were done on biologically more relevant cells, non-nocodazole treated SW480 cells, a colon cancer cell line. It can be argued whether we took this step too early because the reason for gene expression variations cannot fully be distinguished between methodical issues and possible cell cycle stage variations.

About 95 – 98% of the RNA in a cell is ribosomal RNA. For transcriptome analysis only 1-3% mRNA has to be isolated from the rest to not waste sequencing capacity. As mentioned in the introduction there are several ways to avoid rRNA such as depletion approaches, oligo (dT) priming strategies or oligo (dT) based enrichment followed by a random priming strategy. The linear amplification assay used is based just on oligo (dT) selection via magnetic beads as well as oligo (dT) priming strategy. Random priming strategy cannot be carried out because the washing step would therefore have to have been made before reverse

transcription. High risk of RNA digestion while washing is the limiting factor for this step. The exponential amplification assay by Miltenyi includes a strong oligo (dT) selection with random priming strategy thereafter. Because of the specifically developed µMacs columns an easy handling and rapid washing step before reverse transcription can be carried out (2.2.3). With both strategies selection to messenger RNA was successful so none or negligible ribosomal RNA was sequenced.

Three amplification protocols were tested, modified in the process of testing and finally completely established in the lab. All have different advantages and disadvantages. Exponential amplification carried out with the ABI protocol as well as with the Miltenyi protocol lead to the highest observed amplification rates of 10^3 up to 10^6. The ABI PCR protocol contains two PCR steps. First an amplification of 20 cycles is performed and after purification, a portion of this cDNA was further amplified by nine cycles of PCR. In figure 8B it is noticeable that after the second round of amplification no differences in the Cp of the five cell sample and the 10 cell sample can be detected. A likely explanation would be that the 10 cell sample already reached the plateau phase and therefore the five cell sample finally kept up with the 10 cell sample. Therefore, amplification versus loss of dynamic range is a trade off for PCR based amplification approaches. The last cycles before the plateau phase are known to be the cycles inducing strong bias as described in the introduction.

The Miltenyi exponential amplification protocol includes a PCR step of 40 cycles. A really high amplification rate of at least 10^6 is the result. Unfortunately this most likely also results in the loss of long cDNA fragments due to preferred amplification of shorter fragments. This problem could be overcome by reducing cycle number and performing amplification by emulsion PCR to reduce the number of template molecules per reaction vessel and therefore limit competition between short and long fragments. An experimental PCR containing a one-cell sample cDNA and known spiked-in fragments of e.g. arabidopsis with various length, concentrations and sequences could provide more substance to this hypothesis. In contrast to this the linearly amplified assay leads to an increase of the input material in a range of 1000 to 4000 fold. The theory that by using linearly amplified methods the amplification rate is not big enough or that at least a second round of amplification would need to be performed is not the case from my results.

A further important issue in single cell analysis is the potential risk and dramatic impact of contamination. A single cell contains about 1-3 pg mRNA. At this level of input material every

single RNA/DNA molecule coming from the environment will be present at a high percentage of the whole sample material. That this is of particular importance for the samples with lowest input amounts was shown with the E coli contamination of the exponentially amplified samples. The percentage of E. coli contamination for the two single cell samples of 40% decreases down to 20% when just a twenty times higher input material is used. It can be thought of as a competition situation between the E. coli strands and the small human mRNA templates in the reverse transcription. The majority of the identified E. coli sequences are 16S and 23S RNA. This means that the contamination must occur during reverse transcription. Indeed in various RT enzymes E. coli 16S RNA could have been detected by real time PCR. Unfortunately I have no satisfactory answer as to why this issue is not seen with the linearly amplified samples and also not in the published single cell paper based on the ABI protocol [62]. Initial changes in the Miltenyi protocol to a reverse transcription with single Oligo dT priming did not lead to a drastic reduction of E. coli concentration after amplification (data not shown). This needs to be verified. Possible correlation between the high number of PCR cycles and the strong contamination has to be checked as well.

Apart of sequences coming clearly from other species and are therefore contaminants, the internal poly (A) priming became a working hypothesis in this study to explain the phenomenon of strong genomic repeat region abundance of 40 to 60% for samples linearly amplified with the NuGEN assay. This was proven by counting homopolymer stretches and in particular poly-A stretches in flanking regions of what we termed in this work 'genomic contamination regions' (figure 18). These regions are characterized by high counts of low complexity reads. This observation was also made in a recently published single cell mRNA sequencing by Cloonan *et a.l* who received about 20% repeat region reads [15]. We are not fully sure why we see such a high number of genomic repeat regions in only one out of the four samples amplified with the exponentially amplification based kit. The sufficient washing step before reverse transcription in this kit may be an explanation. It seems like internal priming sites having a competing position with the polyA tail of mRNAs because a high number or repeat reads are correlated with a low number of transcriptome reads. Nam *et al.* were able to show that a decrease of internal priming can be achieved by using anchored oligo (dT) primers [63].

Additional to the cDNA synthesis and amplification part in the workflow certain steps during NGS sequencing library preparation might have influence on final results as well. The

fragmentation of samples is known to be a step which introduces GC content dependent bias. Methods, especially those based on ultrasonic shearing have poor efficiency and are often non-isothermal. Figure 15 showing the fragmentation results after 5 and 7 minutes of ultrasonic treatment support this idea. Even after 7 minutes of shearing a wide smear is seen and sizes of fragments are too long. But further extension of the duration of ultrasonication increases the chance of losing shorter fragments. The first gel electrophoresis band per sample in Figure 16 shows the sheared non-amplified samples. Even though they were all treated in the same manner the results varied strongly and even after 8 minutes of ultrasonication the fragments are too long, which means a high loss of material in the size selection step. A possible reason for the poor efficiency especially when compared with the linearly amplified samples would be that the amplified product is single stranded leading to secondary structures avoiding efficient fragmentation. First experiments working with uracil-DNA glycosylase cleavage on every position containing a uracil were made. Uracil was incorporated during reverse transcription. Higher concentration of UTPs leads to a decrease of fragment length. Unfortunately the first experiments did not lead to expected results. Further studies have to be carried out with this method of fragmentation.

The loss of cDNA fragments during library preparation cannot be completely avoided. The hybridization step of the adaptor P1 and P2 with the templates as well as the following ligation reactions are known to be inefficient. As seen in the sequencing results of the filter step of the linearly amplified samples (figure 4, supplementary) a high percentage of adaptor sequences was detected. Analysis of these filtered reads showed that cross-hybridised products of P1 and P2 were generated and sequenced. Changing hybridization conditions and possible amine modification of adaptor have to be discussed to reduce the amount of adaptor dimerization and concatermerization.

Even though high throughput second generation sequencing is still drastically reducing the cost for whole transcriptome analysis, sample and sequencing library preparation should be proven before sequencing. Two steps in the workflow of single cell transcriptome analysis seem to be extremely important. 1) Reverse transcription and 2) Amplification. As seen in table 2 and figure 8 either no signal or really poor quality signal was achieved with real time PCR even for a highly expressed gene like HPRT. Besides this, it remains very difficult to take quality control aliquots after reverse transcription in both of the methods. Solutions to this issue have to be found. This doesn't just apply to the sample quality but also to address the

question of whether the full length cDNA ever got transcribed or if the PCR is the reason for loss, as discussed above. Discovering the sample complexity after amplification is not trivial but it might worth trying high resolution melt analysis for the whole sample[64].

Beside the sample quality, the quality of a sequencing run and the amount of resulting information are dependent on various factors. An obvious one is the influence of the quality of the sequencing workflow itself. Compared to systems such as the market leader Illumina, the handling of the SOLiD platform is not trivial. Bad bead deposition, multi template beads or insufficient emulsion PCR (leading to reduction in the amount of templates per beads) have a major influence on reproducible results of the SOLiD platform. Our two exponentially amplified one-cell samples, and the two split twenty cell samples were loaded on one sequencing slide. The four samples were not labeld with a barcode. Therefore every sample could be tracked for their respective sequencing quality. The ratio of usable beads versus non-usable beads per cycle gives the first indication of quality. Figure 5, supplementary, already shows the connection between the ratio of 'good' beads to 'bad' beads and the percentage of mappable reads. There is a slight tendency that the samples containing biotin labeld primer and treated with streptavidin beads have a greater number of non-usable beads. The four linearly amplified samples were barcoded and loaded on one full slide. Therefore the sequencing quality per sample could not be checked. A general ratio of good to bad beads on the full slide is given in figure 6, supplementary.

However, there are still differences in the mappability of every sample. Sample 1-2 yielded 85% genomic mapped reads, sample 1-3 64%, sample 1-1 45% and sample 50 just 57%. In general the library quality is the key for the quality of a sequencing experiment (pers. comm., Dr. Andreas Dahl) For one sample it is relatively clear that the low percentage of mappability is due to a suboptimal bead preparation process. However, whether the reason for low mappability is the library preparation process itself or the cDNA amplification can not clearly be explained without any repetition. Due to the newly developed adaptor design for the single stranded DNA only limited data is available for comparison. The new adaptor design was required to maintain the feature of strand specific sequencing for the linear amplification, one of the major advantages of this methodology,

The paper published in 2009 by Tang *et al.* was the first reference to get an idea about the amount of transcriptome reads which yield from a single cell experiment. In this study the same protocol i was initially working with was carried out and the sequencing also took place

on the ABI SOLiD platform. Therefore our results were roughly comparable. However, it has to be mentioned that they were working with blastomeres and oocytes which are much bigger cells than normal tissue cells. Therefore our single cell conditions were even more challenging. The seven analysed cells resulted in 26 to 45% transcriptome reads except for the splice junction reads which we did not analyse. Exponentially amplified one-cell samples which have almost the same amplification protocol lead to transcriptome output of 1-4%. Much less efficient than described in the paper. But the analysed 20-cell sample, which is more comparable to blastomeres also yielded around 50% transcriptome reads. The high amount of filtered E. coli reads which took up almost 50% of the sequencing power may be an explanation for the low amount of transcriptome reads for the single cells (Table 5). The linearly amplified samples have the same number of genomic mapped reads than described in the paper, around 30-50% but again the transcriptome counts are much less, 2,5 – 26%. The reason for this is the extremely high number of sequenced repetitive reads.

Another way to measure library complexity is to calculate the number of detected genes. Even though samples were barely transcribed, such as the exponentially amplified 1-1 sample, the amount of detected genes is still 16.000 genes. This means that even low read counts can have a high complexity. The linearly amplified samples have in general a higher number of detected genes despite the amount of transcriptome reads being less compared to the exponential amplification based method (Table 9). In my view this could be evidence that linear amplification contains much less bias and cDNA gets amplified independently of the sequence and especially the length. New analysis sorted by the amounts of reads grouped into increasing gene - exon length should lead to the same mean amount of counts whereas with the exponentially amplified reads genes with longer transcripts should be less represented. Again this could also be tested by known spiked in sequences of different length and quantities.

In a second analysis genes were clustered into three expression groups: low, medium and high. Lowly expressed genes had read counts per gene between 3 and 49, medium ones between 50-499 and highly expressed genes having at least 500 counts. Working with the normalized gene expression results in that most genes have a low transcription rate (Figure 19). These results confirm what Klein et al. published about a single micrometastatic cell [65]. Also Hastie and Bishop showed that the lowly expressed genes are the majority and the strongly transcribed genes are quite rare. For the mean amount of genes clustered into the

three abundant classes the average between all four samples per method was calculated. Therefore the deviation per group is an index for methodical reproducibility. As already seen in the comparable amount of detected genes the linearly amplified samples have remarkably fewer deviations. This indicates again a more stable amplification (Figure 19). Another benefit of linearly amplified samples is the strand information which was predicted to be maintained. Indeed compared to the exponentially amplified samples the NuGEN correlation plots display a wider range of values which indicates a larger discrepancy between plus and minus strand but only for some genes (Figure 20). Therefore it can be said that for at least these four samples limited transcription orientation was achieved. Reasons might be false positive counts due to internal Poly A priming during cDNA synthesis. Furthermore hairpin loops of the mRNA which might lead to a reverse amplification have to be discussed and new experiments are required to test these.

I stopped working with the whole transcriptome preparation of a Single Cell using the ABI protocol because of non-homogeneous reverse transcription and amplification results. The time it takes to extract a single cell and complete the whole protocol is also too long and reduces therefore the chance of highly reproducible results. Therefore we decided to focus on the µMACS SuperAmp Kit from Miltenyi as an example of exponential amplification of single cells. The protocol procedure is based on the same idea as the ABI protocol but because of the described µMacs column technology, better purification steps and better handling in general is provided. The amount of transcriptome reads for the twenty cell samples was surprisingly high and the concordance of the plus and minus strand was really high. Never the less, read out for full length transcripts has to be increased by better reverse transcription efficiency and/or improved PCR conditions. The issue with E. coli contamination should be negligible when using for example special enzyme pre-treatment. The transcriptome read outcome of WT Ovation, One-Direct RNA Amplification System from NuGEN was fairly low even for the 50 cell sample. Amplification rate was as efficient as predicted and samples were much more homologousous then exponentially amplified ones. Higher amounts of detected genes could be also an indicator for a more sensitive assay. The expected maintenance of strand information did not materialise which requires further investigation.

As it will not be possible to achieve an amplification method for single cell transcriptome analysis without any bias, scientific development has to focus on methods with hopefully

moderate but, more importantly, reproducible bias. This study has given insights into the issues related to the investigation of single cell transcriptomes. Two promising methods of single cell transcriptome amplification could be identified and investigated using next generation sequencing.

6. References

1. Lewin B: **Cells**. Sudbury, Mass.: Jones and Bartlett Publishers; 2007.
2. Hastie ND, Bishop JO: **The expression of three abundance classes of messenger RNA in mouse tissues**. *Cell* 1976, **9**(4 PT 2):761-774.
3. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**(5258):561-563.
4. Mendes Soares LM, Valcarcel J: **The expanding transcriptome: the genome as the 'Book of Sand'**. *EMBO J* 2006, **25**(5):923-931.
5. Herzenberg LA, De Rosa SC: **Monoclonal antibodies and the FACS: complementary tools for immunobiology and medicine**. *Immunol Today* 2000, **21**(8):383-390.
6. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M: **Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach**. *BMC Bioinformatics* 2010, **11**:27.
7. Fukino K, Shen L, Patocs A, Mutter GL, Eng C: **Genomic instability within tumor stroma and clinicopathological characteristics of sporadic primary invasive breast carcinoma**. *JAMA* 2007, **297**(19):2103-2111.
8. Clarke MF, Dick JE, Dirks PB, Eaves CJ, Jamieson CH, Jones DL, Visvader J, Weissman IL, Wahl GM: **Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells**. *Cancer Res* 2006, **66**(19):9339-9344.
9. Shackleton M, Quintana E, Fearon ER, Morrison SJ: **Heterogeneity in cancer: cancer stem cells versus clonal evolution**. *Cell* 2009, **138**(5):822-829.
10. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection**. *Science* 1996, **274**(5289):998-1001.
11. Alberts B: **Molecular biology of the cell**, 5th edn. New York: Garland Science; 2008.
12. Darzynkiewicz Z, Evenson DP, Staiano-Coico L, Sharpless TK, Melamed ML: **Correlation between cell cycle duration and RNA content**. *J Cell Physiol* 1979, **100**(3):425-438.
13. Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH: **Amplified RNA synthesized from limited quantities of heterogeneous cDNA**. *Proc Natl Acad Sci U S A* 1990, **87**(5):1663-1667.
14. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P: **Analysis of gene expression in single live neurons**. *Proc Natl Acad Sci U S A* 1992, **89**(7):3010-3014.
15. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G *et al*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nat Methods* 2008, **5**(7):613-619.
16. Darnell JE, Wall R, Tushinski RJ: **An adenylic acid-rich sequence in messenger RNA of HeLa cells and its possible relationship to reiterated sites in DNA**. *Proc Natl Acad Sci U S A* 1971, **68**(6):1321-1325.
17. Edmonds M, Vaughan MH, Jr., Nakazato H: **Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship**. *Proc Natl Acad Sci U S A* 1971, **68**(6):1336-1340.
18. Brawerman G: **The Role of the poly(A) sequence in mammalian messenger RNA**. *CRC Crit Rev Biochem* 1981, **10**(1):1-38.
19. Furuichi Y, LaFiandra A, Shatkin AJ: **5'-Terminal structure and mRNA stability**. *Nature* 1977, **266**(5599):235-239.

20. Wickens M, Stephenson P: **Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation**. *Science* 1984, **226**(4678):1045-1051.
21. Myers TW, Gelfand DH: **Reverse transcription and DNA amplification by a Thermus thermophilus DNA polymerase**. *Biochemistry* 1991, **30**(31):7661-7666.
22. Brooks EM, Sheflin LG, Spaulding SW: **Secondary structure in the 3' UTR of EGF and the choice of reverse transcriptases affect the detection of message diversity by RT-PCR**. *Biotechniques* 1995, **19**(5):806-812, 814-805.
23. Decraene C, Reguigne-Arnould I, Auffray C, Pietu G: **Reverse transcription in the presence of dideoxynucleotides to increase the sensitivity of expression monitoring with cDNA arrays**. *Biotechniques* 1999, **27**(5):962-966.
24. Luo L, Salunga RC, Guo H, Bittner A, Joy KC, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR *et al*: **Gene expression profiles of laser-captured adjacent neuronal subtypes**. *Nat Med* 1999, **5**(1):117-122.
25. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J *et al*: **Within the fold: assessing differential expression measures and reproducibility in microarray assays**. *Genome Biol* 2002, **3**(11):research0062.
26. Stangegaard M, Dufva IH, Dufva M: **Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA**. *Biotechniques* 2006, **40**(5):649-657.
27. Stahlberg A, Hakansson J, Xian X, Semb H, Kubista M: **Properties of the reverse transcription reaction in mRNA quantification**. *Clin Chem* 2004, **50**(3):509-515.
28. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A: **Amplification of cDNA ends based on template-switching effect and step-out PCR**. *Nucleic Acids Res* 1999, **27**(6):1558-1560.
29. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD: **Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction**. *Biotechniques* 2001, **30**(4):892-897.
30. Kanagawa T: **Bias and artifacts in multitemplate polymerase chain reactions (PCR)**. *J Biosci Bioeng* 2003, **96**(4):317-323.
31. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H: **Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986**. *Biotechnology* 1992, **24**:17-27.
32. Iscove NN, Barbara M, Gu M, Gibson M, Modi C, Winegarden N: **Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA**. *Nat Biotechnol* 2002, **20**(9):940-943.
33. Kainz P: **The PCR plateau phase - towards an understanding of its limitations**. *Biochim Biophys Acta* 2000, **1494**(1-2):23-27.
34. Thompson JR, Marcelino LA, Polz MF: **Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'**. *Nucleic Acids Res* 2002, **30**(9):2083-2088.
35. Shuldiner AR, Nirula A, Roth J: **Hybrid DNA artifact from PCR of closely related target sequences**. *Nucleic Acids Res* 1989, **17**(11):4409.
36. Dutton CM, Paynton C, Sommer SS: **General method for amplifying regions of very high G+C content**. *Nucleic Acids Res* 1993, **21**(12):2953-2954.
37. Zhao H, Hastie T, Whitfield ML, Borresen-Dale AL, Jeffrey SS: **Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis**. *BMC Genomics* 2002, **3**(1):31.
38. Wang E, Miller LD, Ohnmacht GA, Liu ET, Marincola FM: **High-fidelity mRNA amplification for gene profiling**. *Nat Biotechnol* 2000, **18**(4):457-459.

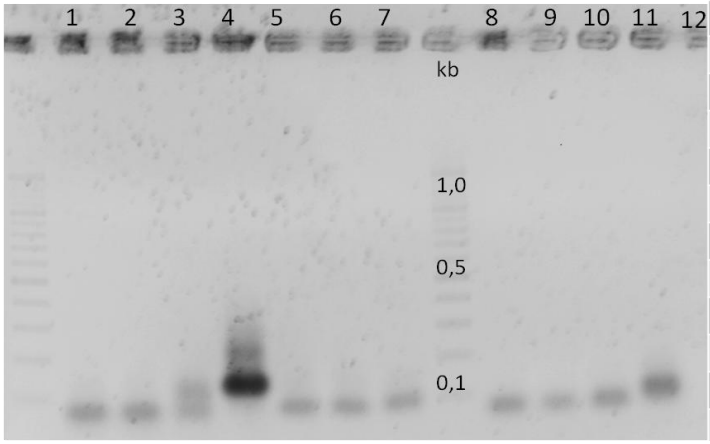
39. Dafforn A, Chen P, Deng G, Herrler M, Iglehart D, Koritala S, Lato S, Pillarisetty S, Purohit R, Wang M *et al*: **Linear mRNA amplification from as little as 5 ng total RNA for global gene expression analysis**. *Biotechniques* 2004, **37**(5):854-857.
40. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW: **The antisense transcriptomes of human cells**. *Science* 2008, **322**(5909):1855-1857.
41. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**(5830):1484-1488.
42. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**(5740):1564-1566.
43. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA**. *Nucleic Acids Res* 2009, **37**(18):e123.
44. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM: **A high-resolution map of transcription in the yeast genome**. *Proc Natl Acad Sci U S A* 2006, **103**(14):5320-5325.
45. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al*: **Global identification of human transcribed sequences with genome tiling arrays**. *Science* 2004, **306**(5705):2242-2246.
46. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G *et al*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**(5725):1149-1154.
47. Okoniewski MJ, Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations**. *BMC Bioinformatics* 2006, **7**:276.
48. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA**. *Nature* 1977, **265**(5596):687-695.
49. Swerdlow H, Gesteland R: **Capillary gel electrophoresis for rapid, high resolution DNA sequencing**. *Nucleic Acids Res* 1990, **18**(6):1415-1419.
50. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.
51. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat Biotechnol* 2008, **26**(10):1135-1145.
52. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**(5235):484-487.
53. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M *et al*: **CAGE: cap analysis of gene expression**. *Nat Methods* 2006, **3**(3):211-222.
54. Reinartz J, Bruyns E, Lin JZ, Burcham T, Brenner S, Bowen B, Kramer M, Woychik R: **Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms**. *Brief Funct Genomic Proteomic* 2002, **1**(1):95-104.
55. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
56. Mitra RD, Church GM: **In situ localized amplification and contact replication of many individual DNA molecules**. *Nucleic Acids Res* 1999, **27**(24):e34.
57. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**(5741):1728-1732.

58. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B: **Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations**. *Proc Natl Acad Sci U S A* 2003, **100**(15):8817-8822.
59. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E: **Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms**. *Nucleic Acids Res* 2000, **28**(20):E87.
60. Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM: **Fluorescent in situ sequencing on polymerase colonies**. *Anal Biochem* 2003, **320**(1):55-65.
61. Zieve GW: **Nocodazole and cytochalasin D induce tetraploidy in mammalian cells**. *Am J Physiol* 1984, **246**(1 Pt 1):C154-156.
62. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A *et al*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**(5):377-382.
63. Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, Chen J, Rowley JD, Wang SM: **Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription**. *Proc Natl Acad Sci U S A* 2002, **99**(9):6152-6156.
64. Schutze T, Arndt PF, Menger M, Wochner A, Vingron M, Erdmann VA, Lehrach H, Kaps C, Glokler J: **A calibrated diversity assay for nucleic acid libraries using DiStRO--a Diversity Standard of Random Oligonucleotides**. *Nucleic Acids Res* 2010, **38**(4):e23.
65. Klein CA, Seidl S, Petat-Dutter K, Offner S, Geigl JB, Schmidt-Kittler O, Wendler N, Passlick B, Huber RM, Schlimok G *et al*: **Combined transcriptome and genome analysis of single micrometastatic cells**. *Nat Biotechnol* 2002, **20**(4):387-392.

7. Supplementary

Table 1 **Sequences of used primers**

Primer name	Sequence 5´- 3´
Hprt forward	GGTGGAGATGATCTCTCAACTTTAA
Hprt reverse	AGGAAAGCAAAGTCTGCATTGTT
E. coli forward	CTCCTACGGGAGGCAGCAG
E. coli reverse	GWATTACCGCGGCKGCTG
16S rRNA forward	CAGACAACCTTAGCCCAAACCA
16S rRNA reverse	TTCATCTTTCCCTTGCGGTA
UP1	ATATGGATCCGGCGCGCCGTCGACTTTTTTTTTTTTTTTTTTTTTTT
UP2	ATATCTCGAGGGCGCGCCGGATCCTTTTTTTTTTTTTTTTTTTTTTT
UPA	ATACGGATCCGTCAGCGCCGCGACTTTTTTTTTTTTTTTTTTTTTTT
Adaptor 1 sense	NNNNNNATCACCGACTGCCCATAG
Adaotor 1 antisense	CTATGGGCAGTCGGTGAT
Adaptor 2 sense	TGCTGTACGGCCAAGGCGNNNNNN - Woobles -
Adaptor 2 antisense	CGCCTTGGCCGTACAGCAG
P1	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT
P2	CTGCCCCGGGTTCCTCATTCTCTNNNNNNTGCTGTACGGCCAAGGCG
Beta actin	Hs03023880_g1 (ABI)
Oligo dT24	TTTTTTTTTTTTTTTTTTTTTT



Nr.	Primer	PCR	Enzyme (DNA Pol.)
1	UP1	-	-
2	UP1	+	-
3	UP1	+	+
4	UP1 + UP2	+	+
5	UP2	-	-
6	UP2	+	-
7	UP2	+	+
8	UPA	-	-
9	UPA	+	-
10	UPA	+	+
11	UPA + UP2	+	+
12	-	+	+

Figure 1 **Primer dimerization of UP1 – UP2 / UPA – UP2**
Monitoring of different primer combination and its dimerization with (+) or without (-) amplification and with (+) or without (-) DNA polymerase.

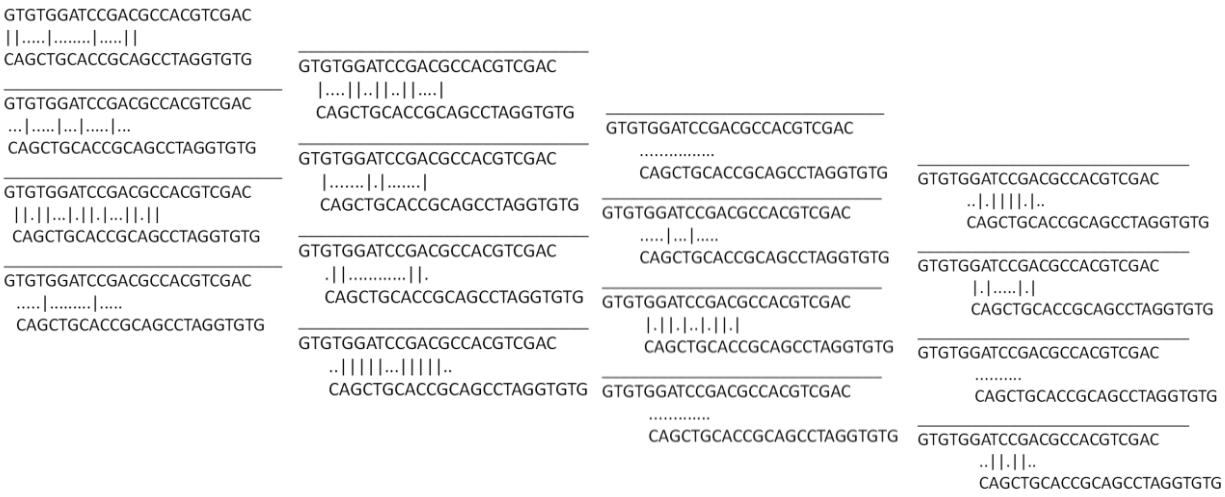
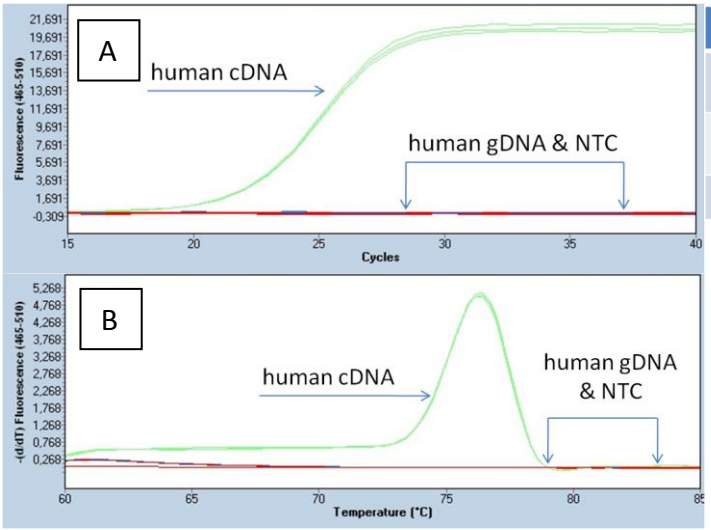


Figure 2 **Alignment of Primer UP1 with UP2**
To visualise homology and possible resulting primer dimerization between UP1 and UP2 alignments with one base shift per line were carried out computationally.



Template	MeanCp	STD Cp
Human cDNA	21,81	0,01
Human gDNA	> 40	0
NTC	> 40	0

Figure 3 **Amplification curve and melting peak for human cDNA Hprt amplicon**
A: Amplification curves of the Hprt SybrGReen with human cDNA, from HeLa cells, and genomicDNA (gDNA), HeLa, as well as without any template control (NTC). B: Melting peak of cDNA, HeLa, and genomicDNA (gDNA), HeLa, as well as of the non template control (NTC)

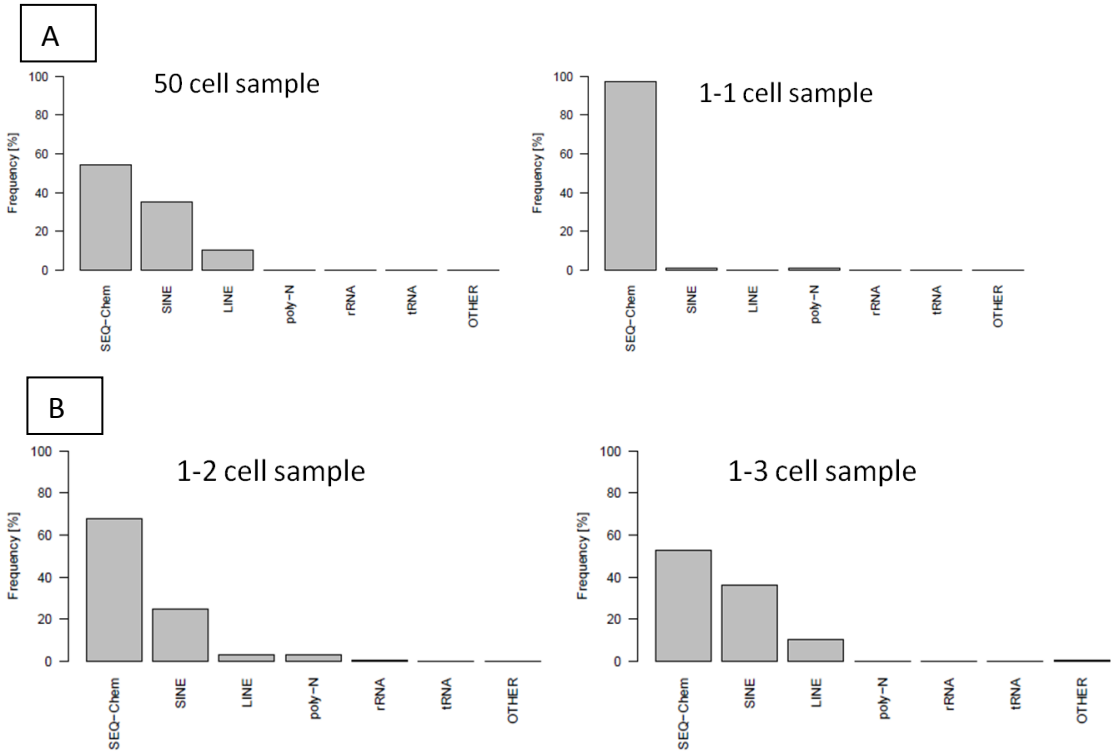


Figure 4 **Filtered reads of the four samples amplified with One-Direct RNA Amplification System, NuGEN**

As a first analysis step all sequences are mapped to various filtering sequences such as sequencing adaptors, rRNA or Poly-N. The filtered out reads are shown for samples, 1-1, 1-2 + 1-2 and 50.

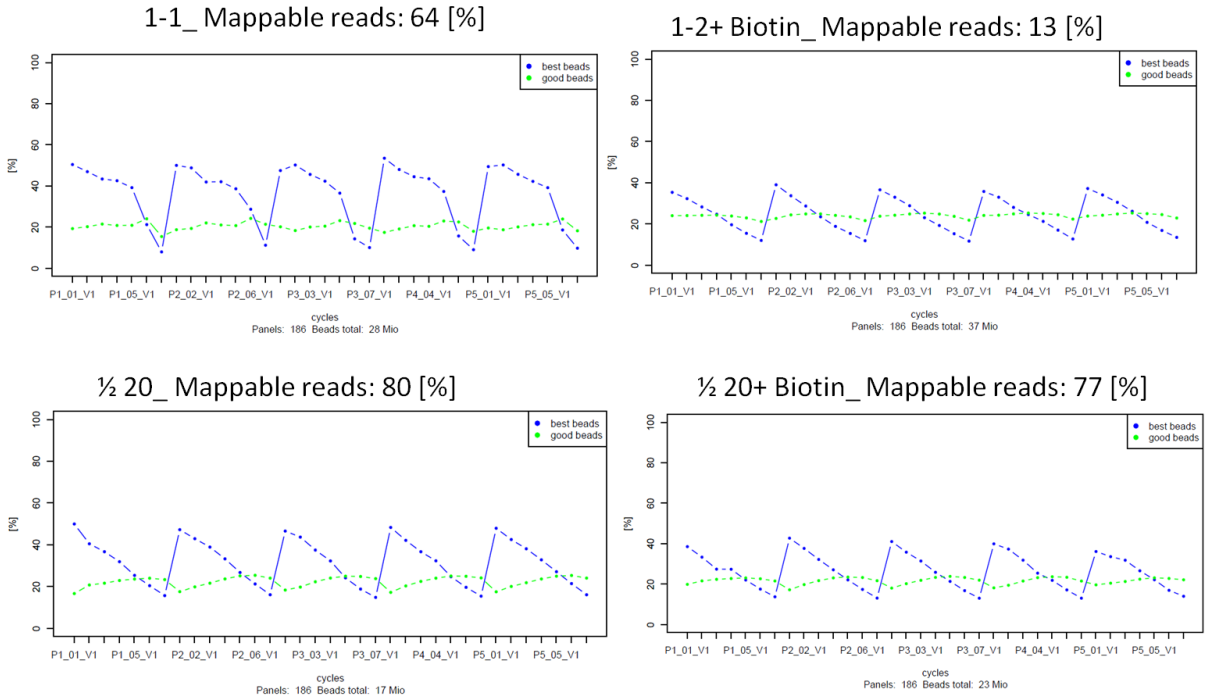


Figure 5 **Good (blue) and bad (green) beads derivation for the exponentially amplified samples**

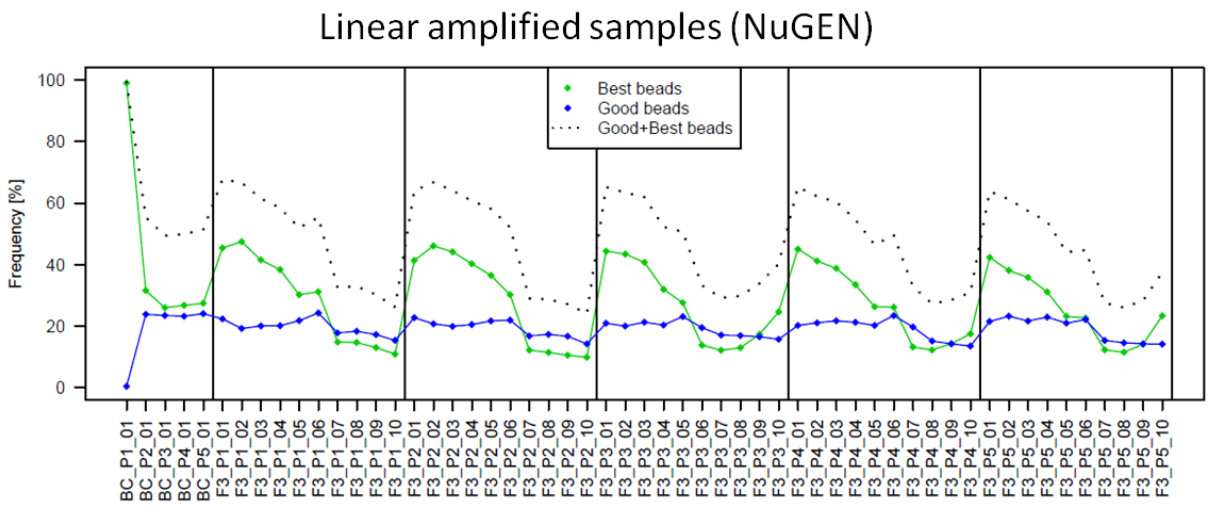


Figure 6 **Good (blue) and bad (green) beads derivation for the linearly amplified samples**

8. Eigenständigkeitserlärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit erstmalig einreiche, selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Berlin, den 08.12.2010