Correlation of Telomere Length and Gene Expression Levels in Single Cell Resolution

Negar Pourjamal

Master Thesis
Genetics
Department of Biosciences
University of Helsinki
April, 2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Biological and Environmental Sciences	Koulutusohjelma – Utbildingsprogram – Degree Programme Genetics	
Tekijä – Författare – Author Negar Pourjamal		
Työn nimi – Arbetets titel – Title Correlation of Telomere Length and Gene Expression Levels in Single Cell Resolution		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetics		
Työn laji – Arbetets art – Level Master Thesis	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages 77

Tiivistelmä – Referat – Abstract

Telomeres are cap shaped structures at the very end part of each chromosome that protect DNA from degradation or unwanted chromosome-chromosome attachments. Telomere lengths show considerable heterogeneity in different cells of the same cell population. Reasons for heterogeneitiy and mechanisms inside cells causing them are not fully understood.

In this study, we explored the correlation between telomere length and different gene expressions. First, using FACS technique we sorted each single cell into each well of 96-well plate. Second, we used SYBR green based qPCR for telomere length measurement. Third, we used Illumina-seq for sequencing extracted mRNAs. [6]

We found a set of genes that were in strong correlation with telomere length, giving opportunity to explore the biological pathways. We compared pathways between different samples and found strong connections between genes involved in viral cycles and immune system with extracted genes that were in high correlation with telomere. We found heterogeneity of telomere lengths and transcriptomes in different cell lines. Telomere related proteins, specifically those involved in shelterin complex, are expressed highly in cancer cell lines and LPS-stimulated monocytes compared to the non-stimulated monocytes. In our study, *SLC38A2, PURB, UBR3, SSR1, NCAPH2, AIMP2, PHF21A* genes were highly correlated with telomere in mutual way and can therefore be considered as new biomarkers/novel candidates for telomere-related studies. The importance of these genes has been reported in aging/mortality. Concurrent with our findings, a recent report also suggested that NCAPH2 plays role in regulating telomere stability and maintenance through its interaction with TERF. [65]

We found new genes in correlation with telomere regulation, and our findings are therefore of high importance in research of cancer, neurodegenerative diseases and aging. Further studies are, however, required as our data is limited by small number of samples and inability to properly validate our technique.

Avainsanat – Nyckelord – Keywords
Single cell, telomere length, RNA-sequencing, cancer cell line, normal cell line

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Päivi Saavalainen, Heli Pessa, Katarzyna Leskinen, Eija Nissilä, Juha Grönholm

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Table of Contents

1. Introduction and Background

1.1 Telomere 8
1.1.1 DNA end replication problem 10
1.2 Telomerase 11
1.3 Telomerase in cancer treatment 14
1.4 Shelterin Complex 15
1.5 Techniques for telomere length measurement 18
1.5.1 Telomere Length Measurement by qPCR 19

2. Aim 20

3. Materials and methods

3.1 Cell lines 21
3.2 Cell culture 21
3.3 FACS preparation 22
3.3.1 Cell counting 22
3.3.2 Single cell sorting 22
3.4 Dynabeads and mRNA capture 23
3.5 cDNA library preparation 25
3.5.1 cDNA measurement 25
3.5.2 Library preparation and sequencing 26
3.6 Data Analyzes 28
3.6.1 Analyzes steps 28
3.7 Genomic DNA purification and pre-amplification 30
3.8 Single cell telomere length measurement by qPCR 30

4. Results

4.1 Cell count related to LPS-stimulated cell lines 33
4.2 Single cell isolation 34
4.3 cDNA library evaluation 34
4.4 Tagmentation and cDNAs prepared for sequencing 37
4.5 Illumina sequencing and seurat clustering 39
4.6 Well-known telomere related genes 41
4.7 q-PCR based telomere length measurement 42
4.8 The correlation between telomere length and DGE in single cell level 44
4.9 Pathways found for telomere highly-correlated genes 49

5. Discussion 53

5.1 Cell stimulation and cell viability 54
5.2 Shelterin Complex-Related Genes 54
5.3 DGE heterogeneity among different samples 55
5.4 Pathways Related to Extracted Genes 55
5.5 Highly correlated genes with telomere 58

5.5.1 NCAPH2 58
5.5.2 UBR3 59

6. Conclusion 62

7. Acknowledgments 64

8. References 65

9. Supplementary material 69

Abbreviations

AIMP2	Aminoacyl TRNA Synthetase Complex
ALT	Alternative Lengthening of Telomeres
BAM	Binary SAM format
CAB box	Cajal Body Box
Cq	Quantitation cycle in qPCR
DGE	Digital Gene Expression
DMEM	Dulbecco´s modified eagle medium
DTT	Dithiothreitol
EXO1	Exonuclease1
FACS	Fluorescence-activated cell sorting
FASTQ	Text based format for showing either DNA or protein
FBS	Fetal bovine serum
GAR1	GAR1 ribonucleoprotein
gDNA	Genomic DNA
G-quardruplex	Guanine-quarduruplex
HAT1	Histone Acetyltransferase 1
hg38	Human Genome 38
hTR	Human Telomere Repeat
hTERT	Human Telomerase Reverse Transcriptase
LPS	Lipopolysaccharide
MHC	Major Histocompatibility Complex
MQ	Macrophages
MSCs	Mesenchymal stem cells
NAP57 (DKC1)	Dyskerin pseudouridine synthase 1
NCAPH2	Non-SMA Condensing II Complex Subunit H2
Nhp2	Non-Histone Protein 2
NHP2	Nucleolar Protein Family A, Member 2 (H/ACA Samll Nucleolar RNPs Ribonucleoproteins)
NOP10	Nucleolar Protein 10
NSC	Neural Stem Cell

PARP2	Poly (ADP-ribose) Polymerase 2
PBS	Phosphate –buffered saline
PHF21A	PHD finger protein 21A
PMA	Phorbol Myristate Acetate
POT1	Protection of Telomeres 1
PRINS	Primed Insitu
PURB	Purin rich element binding protein B
QC	Quality control
Q-FISH	Quantitative Fluorescence In Situ Hybridization
qPCR	Quantitative Polymerase Chain Reaction
R-loops	RNA moiety loop
RIC8A	RIC8 Guanine Nucleotide Exchange Factor A
RNPs	Ribo-nucleo-protein
RT buffer	Reverse Transcriptase
RAP1B	Ras-related protein Rap-1b
SSC	Saline Sodium Citrate
SSR1	Signal Sequence Receptor Subunit 1
SLC38A2	Solute Carrier Family 38 Member 2
STELA	Single Telomere Length Analysis
THP1	Acute Monocytic Leukemia Derived Monocytes
TSO	Template Switch Oligo
TE-SDS	Tris EDTA SDS
TE-TW	Tris EDTA tween-20
TRF1	Telomeric Repeat Binding Factor 1
TRF2	Telomeric Repeat Binding Factor 2
TIN2	TERF1 Interacting nuclear Factor 2
TPP1	Tripeptidyl peptidase 1
TERF2IP (RAP1)	TERF2 interacting Protein
T-loops	Telomere loops
TRF	Terminal Restriction Fragmentation,
TERC	Telomerase RNA Component
TRF1	Telomeric Repeat Factor
TINF1	TRF1-interacting nuclear factor
TRFH	Telomere Repeat Factors Homolog

TRAP	Translocon-associated protein complex
UBR3	Ubiquitin Protein Ligase E3 Component

1. Introduction and Background

1.1 Telomeres

Telomeres are cap-shaped structures with hundreds or even thousands of repetitive hexamers $(5'\text{-}TTAGGG\text{-}3')_n$ which are protected by protein complexes called shelterin. Shelterin complexes also known as telosomes are located at the very end part of each chromosome arms, preserving DNA from instability and degradation through their interaction with telomere. [1] In each normal cell division, chromosomes encounter frequent shortening at their telomere part as a result of DNA end replication problem. DNA gradual deterioration in somatic cells has been considered as the main reason for cell senescence and aging. Therefore, telomeres could be considered as body clock for measuring each cell life span. [2] (Figure 1)

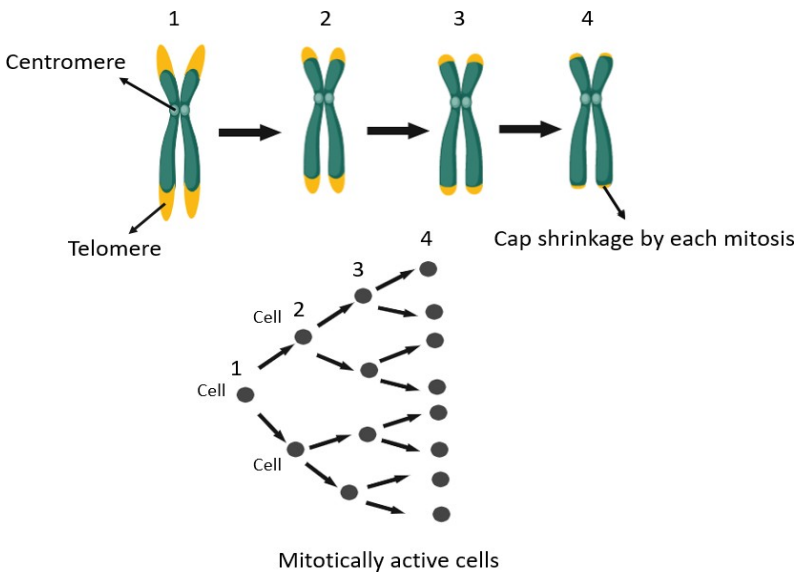


Figure 1. Telomere lengths (yellow cap) in each chromosome with the number of cell mitosis. In somatic cells telomeres go through frequent shortening by each cell division.

The length of telomere tandem repeat sequences ranges from 3 kilobase (kb) to 18kb in human somatic cells and ends to 100-280 nt G-rich single stranded DNA overhang (3´ overhang). Since each end of chromosome arms includes single-stranded G-rich overhangs (G-strand), they could be detected as damaged parts of DNA by DNA repair system and thus be exposed to degradation. However, in humans, G-strand fleeing away from DNA repair system, invades into the neighbor double-stranded complementary repeats, in order to form ultimate protective T-loops (Telomere loops). T-loop is an important factor in genome protection and DNA stability (Figure 5). [3]

On the other hand, guanines located on G-strand of the telomere overhang can form planner G-quartet by Hoogsteen hydrogen binding. 3-4 of these G-quartets attached on top of each other to form G-quardruplex sheets, which are determining factors in the formation of telomere structure and regulation of DNA replication fork. The G-quadruplexes have their highest abundance at telomere part which are considered to protect DNA end part from nucleases. Therefore, it can be concluded that the guanine-rich (G-rich) character present in telomeres are important in stabilizing telomere and specifically T-loop. [4]

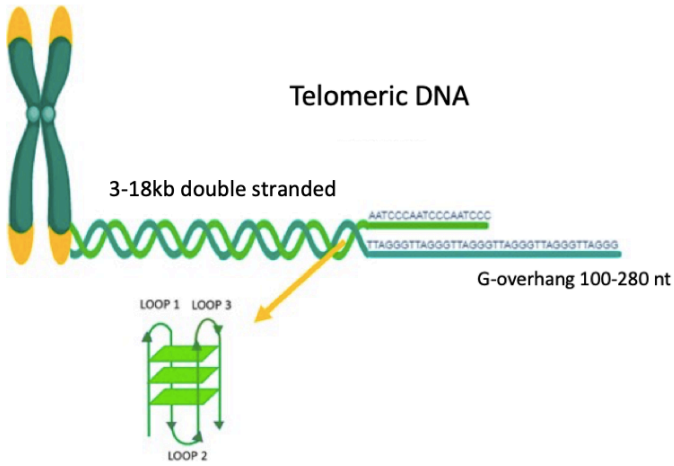


Figure 2. Telomere part of chromosome is made of 3-18 kb of double stranded tandemly repeated hexamers $(TTAGGG)_n$, which ends to a 100-280 nucleotide single stranded DNA overhang. This overhang includes large number of G nucleotides which are important in G-quadruplexes formation along single and double stranded telomere.

As it has been shown in Figure 3, telomere lengths varies between different chromosomes and different cell populations. However, their lengths stay considerably short but stable in cancer cells as a result of high expression level of telomerase enzyme and alternative lengthening of telomeres (ALT). One of the characteristics of cancer cells is their potential to divide fast and unlimited and still keep their short telomere safe from DNA damage response. [5] In stem cells telomeres have longer length and higher level of telomerase enzyme compared to the normal cells. [6] (Figure 3)

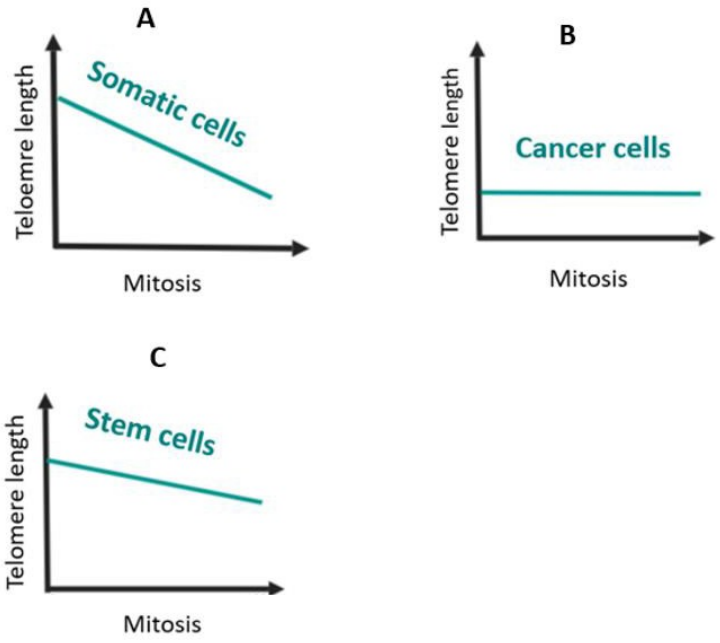


Figure 3. Comparison of telomere lengths between stem cells, normal cells and cancer cells with mitosis. (A) Somatic cells encounter telomere degradation after each mitosis because somatic cells normally do not express telomerase enzyme or express it in significantly low amount which is not detectable. (B) Cancer cells have short but stable telomere as a result of telomerase enzyme over expression. (C) Stem cells could express high levels of telomerase enzyme compared to somatic cells and therefore they have long and stable telomere length.

1.1.1 DNA End Replication Problem

During each cell division DNA goes through telomere shrinkage as a result of DNA end replication problem mechanism. DNA in eukaryotes consists of two linear strands: the leading and the lagging strands which are duplicated in semiconservative way.

During each DNA replication, the leading strand is copied ceaselessly, however, the lagging strand is copied by the help of small RNA primers (replication patterns) in different segments called Okazaki fragments. DNA polymerase I which is fragments to extend the 5´end of the single stranded DNA. Accordingly, DNA polymerase I removes RNA primer and adds DNA subunits to the free 3´ OH end of the primer in Centromere-proximal direction. [7, 8] When the last Okazaki fragment leaves the replication site, there is no free 3´ OH available for attachment of deoxyribonucleotides to the end part of replicated strand. Therefore, with continuous rounds of replications, DNA encounters with telomere shrinkage. [7]

During DNA replication procedure, DNA replication forks operate in different speeds while moving along DNA. These replication forks are fast when they start moving from centromeres and become slower when they accost telomere part. Finally, they stop moving before reaching to the very end part of telomere. These replication forks go through disintegration after their complete stop from moving along DNA, causing genetic material deterioration. Replication in telomere part are affected by several factors such as DNA end replication problem, T-loop formation and G-quadruplexes concentration in telomere part. In order to have successful telomere replication, the formation of G- quadruplexes should be prevented and T-loop structure and existed G-quadruplexes should be opened for DNA helicase easy movement along telomere region. [9]

1.2 Telomerase

Telomere shortening is because of DNA end replication problem mechanism, in which DNA is not able to be replicated completely by DNA polymerase enzyme and becomes lost during each cell division. In some cells, telomere lost part can be compensated by the help of one reverse transcriptase enzyme, called telomerase (terminal transferase). This enzyme is highly expressed in human germ line cells, hair follicular cells, stem cells and cancerous cells. Since somatic cells lack telomerase enzyme or express it in a very low amount (difficult to be detected), somatic cells go through telomere shrinkage by each cell division. Discovery of telomerase by Greider and Blackburn was one of the most important discoveries in the history of molecular biology. [10]

Telomerase enzyme includes two essential subunits: TERT (hTERT: 5p15.33) also called telomerase reverse transcriptase enzyme and RNA subunit (hTR) or hTERC (hTERC: 3q26). [11, 12] TERT is able to synthesize tandem repeats of DNA of telomere by using TER (3- CCCAAUCCC-5') as its template (Denovo synthesis of DNA). [12] Other recent elements detected in the telomerase structure includes HSP90 and p23 chaperones and AAA+ ATPase pontin and reptin, which are considered essential in assembly of macromolecules and DNA damage repair, respectively. However, the clear mechanisms in which all these elements come together and interact with telomerase structure and function in vivo environment is not clear enough. [11,13]

Telomerase RNA (hTER) includes 445 nucleotides from which only 11 nucleotides participate in hexanucleotide $(TTAGGG)_n$ elongation by providing the template for telomere end part. [6] Accordingly, the single stranded hexamers are added to the 3' end part of the chromosomes and polymerase enzyme makes it double stranded. Telomerase is normally expressed during developmental stage of the S phase of the cell cycle and becomes silent in most adult tissues. [14]

90% of cancer cell lines express considerable amount of telomerase enzyme despite of having shorter telomere length compared to normal somatic cells. Usually, the activation of p53 and p16-RB pathways causes the inhibition of cell growth in cells going through aging. Cells can go through some genetic changes such as p53 loss and therefore scape from senescence and continue with mitosis until they have very short telomeres. Short telomeres subsequently trigger some cell crisis such as telomere end to end attachment and cell apoptosis. Small percentage (less than one percent) of the cells have the opportunity to escape from these crisis and live immortal life through either increasing the expression level of telomerase enzyme or ALT pathway. [11]

The subunits of the telomerase enzyme assemble inside nucleus to help telomere length maintenance.

Below are the steps for telomerase activation:

1. hTERT protein is needed to be produced in plasma and transferred into the nucleus.

2. The hTERT is added into the accessory components (dyskerin, NHP2,NOP10 and GAR1) and hTR RNA strand.
3. The assembled complex will travel into telomere part, where they are needed for telomere elongation.

According to in vitro studies, hTR and hTERT are two main subunits of telomerase enzyme that make the enzyme to fully function along telomere. However, other 4 accessory proteins have been reported which are in association with small nuclear RNAs and are found to be very important during in vivo functioning of telomerase enzyme. [11, 15]

Pseudouridylation happens through two mechanisms of RNA dependent and RNA independent. The RNA independent pseudouridylation process does not need RNA for recognition of target site, and only protein itself (Pseudouridine synthases enzyme) is enough to make pseudouridine from uridine. However, the RNA based pseudouridylation includes a group of complex enzymes named as H/ACA Ribo-nucleo-protein (RNPs) box. This complex is made of one RNA (box H/ACA RNA) and four main proteins named as NAP57/Dyskerin, GAR1, NOP10, and NHP2. [16, 17] The ACA box with GAR1 protein companionship are used as binding sites for different proteins. GAR1 protein attaches directly to the dyskerin catalytic part without the help of NOP10 and NHP2. [17] However, other 3 main proteins Dyskerin, NHP2 and NOP10 are also very crucial proteins from the point of function. Deleting of any of these proteins causes H/ACA RNAs box destruction. NOP10 protein is used for the stabilization of Dyskerin and as a place for NHP2 attachment. NHP2 functions more independent than other main proteins in which, NHP2 deletion causes the destruction of other 3 proteins while none of the other three proteins deletion affects NHP2 survivability. [18, 19, 20]

According to the latest computer modeling, it has been reported that the hairpin-hing-hairpin- tail secondary structure is common between all H/ACA complexes. In this secondary structure, the hinge site includes an evolutionary conserved sequence (ANANNA) named as H box. [17, 16, 21] (Figure 4) Moreover, the TCAB1 protein (Telomerase and Cajal body Protein1), which is made of many WD repeats, attaches to the CAB box posited in hTER structure. This TCAB1 and CAB box assembly helps

telomerase to attach coiled bodies with 0.3-1 μm in diameter located in nucleus membrane Coiled bodies are places in the nucleus membrane in which ribo-nucleo-protein assembly and RNA modification occur. [11, 22]

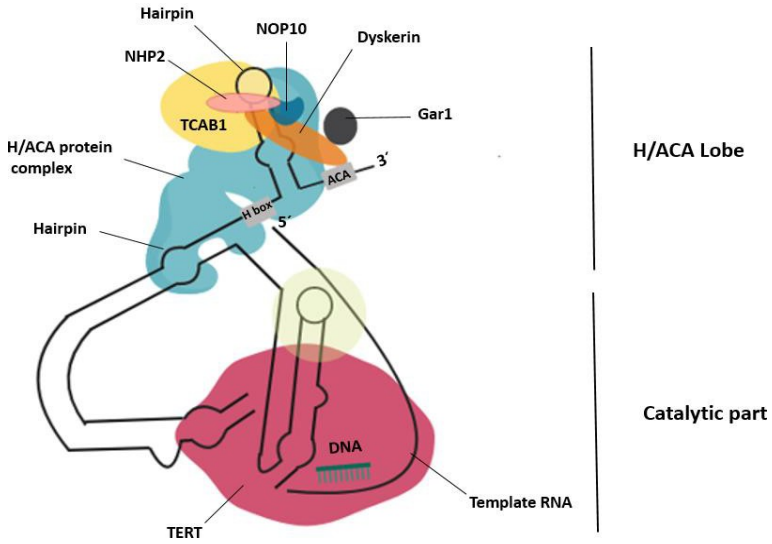


Figure 4. Structure of the telomerase enzyme and its two main lobes, H/ACA lobe and Catalytic part with their involved proteins.

It is worth noting that the presence of telomerase alone is not enough for appropriate telomere regulation. The large expression of telomerase inside cells does not necessarily mean that those cells have long telomeres. For instance, human stimulated T-cells and 95% of cancer cells express telomerase enzyme in a high percentage, but they have short telomeres compared to other cells. [23]

1.3 Telomerase function in cancer treatment

Telomere length is significantly higher in MSCs (Mesenchymal Stem Cell) and somatic cells (7.2-9.3 kb) compared to cancer cell lines (3.9-6 kb). However, telomerase expression level is considerably higher in cancer cells. [24] Scientists have discovered cancer-related drugs which specifically target telomerase enzyme, and therefore let them to be degraded. These degraded components are transferred into the cell surface, to Major Histocompatibility Complex (MHC), and function as flags for calling various immune cells including cytotoxic T cells. Subsequently, the induced immune system attacks to alarming cells with degraded telomerase subunits

on surface and lyse them. There are many therapeutic ways under study relying on the function of telomere and telomerase and scientists are optimistic to discover more efficient approaches with less side effects for cancer treatment in near future by studying cell heterogeneity. [11]

1.4 Shelterin complex

Shelterin complex or telosome is the main protein complex known in telomere protection and telomere length control, performs by attaching to the single or double stranded DNA. Shelterin complex is made of six different proteins called; TRF1 (TERF1), TRF2 (TERF2) and POT1 which are important in recognizing TTAGGG sequences, and TIN2 (TINF2), TPP1 (ACD) and RAP1 (TERF2IP) which are crucial in DNA surveillance against DNA repair system. [25] All these proteins are attached to each other using two TPP1 and TIN2 proteins, working as bridges between different molecules. [2, 12, 26] Each protein of this complex plays a specific role in telomere has basic terminus. Two subunits along with POT1 interact directly with telomere DNA through their carboxyl-terminal Myb domains. Other proteins such as TIN2, TPP1 (TINT1) normally use TRF1, TRF2 and POT1 in order to attach indirectly into telomere DNA. [27]

TERF1 is claimed to adjust the telomere length through its interaction with different shelterin complex-related proteins, and through PINX1 telomerase inhibitor protein. [28] The TERF1 and PINX interaction gives PINX1 this opportunity to be called into telomere part and therefore, inhibit telomerase function and telomere elongation. [27] POT1-TPP1 heterodimer which is considered as the only single strand DNA binding (SSB) heterodimer unscrews telomere secondary structures such as G-quadruplexes in 5´ to 3´ direction or prevents their formation. [12, 26, 27, 29] This heterodimer is attached to the single stranded G-rich overhang through POT1. Telomeres with or without single stranded overhangs (G-overhang) are prone to be recognized as DNA damage part and therefore trigger DNA damage response. [29] Deletion of POT1 related gene correlated with telomere degradation and chromosome end fusion which starts by the displacement of T-loop. [30, 31, 32] POT1 protein attaches to the 3´ G-rich overhang and saves the telomere from being degraded. [27, 33] POT1-like proteins have also been detected in other eukaryotic species such as plants, fungi and nematodes with different telomere length adjustment mechanisms. [30, 34] For long

time it has been thought that POT1 protein has the responsibility in telomere maintenance, as the deletion of POT1 from the fission yeast caused the telomere instability and therefore its degradation. However, deletion of POT1 gene from human genome led to the telomere elongation which was opposite of what was reported in fission yeast. This result shows that POT1 protein play crucial role in telomere length regulation than telomere length protection. [35, 30, 27]

TIN2 (also known as TINF2) is the main cofactor of TRF1 which is important in gathering different telomere related DNA binding proteins. TIN2 and TRF1 correlation happens through the conserved TRFH (Telomeric Repeat Factors Homology) domain and TIN2 carboxyl domain. According to the latest studies, TRF1 is not able to control telomere length all by itself and TIN2 is required for TRF1 performance. [36] Thus, TRF1 and TRF2 are two important factors in telomere signaling network which are in collaborations with different proteins and complexes of shelterin complex, in a direct or indirect way, regulate the telomere length. [37]

The other important protein in the structure of Shelterin complex is TPP1 also called PTOP is encoded by the ACD gene and is located on chromosome 16. TPP1 plays an important role in telomere maintenance and telomerase recruitment to telomere region through its TEL patch domain. After TPP1 and telomerase assembly, TPP1 will function as a stimulator for telomerase enzyme (the number of TTAGGG repeats synthesis before telomerase separation from telomere part). [38, 27] TPP1 is considered as a connection protein between TIN2 and POT1 proteins and regulates the interaction of double stranded DNA binding TERF1 and single stranded DNA binding. [27, 39]

Telomeric repeat containing RNA (TERRA) is the other new component discovered in Shelterin complex. TERRA RNA is a non-coding RNA sequence which can attach to their genomic DNA co- or post- transcriptionally. [9] This RNA sequence has an inverse correlation with telomerase enzyme, the higher telomerase enzyme inside the cell the lower TERRA RNA exists. TERRA RNA is one of the recent important factors in telomere conservation and chromatin formation because of making DNA-RNA hybrids and R-loop structure, and due to the interacting with TERF2 protein. [10] According to the recent studies, R-loops are biological obstacles for DNA replication fork movements and create balance against chromosome instability.

Other probable proteins detected at the end part of chromosomes are not necessarily belong to Shelterin complex. Shelterin and non-shelterin proteins are discerned from each other based on different factors. For instance, shelterin proteins function only at the very end part of chromosomes and not found in other parts of the cell. [2, 10]

Shelterin complexes are crucial elements for telomerase to sit properly on telomere and have therefore significant influence on telomerase functionality. This complex and all its components are thus considered crucial for development of cancer, however, the mechanisms of how these components are regulated in each level of cancer development remain uncertain and require further research. [11]

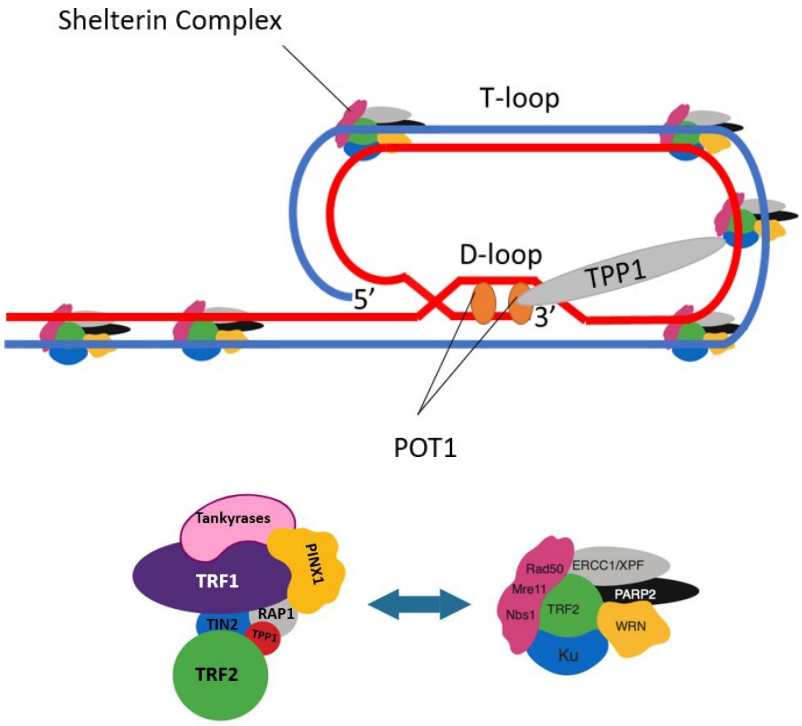


Figure 5. Shelterin complex proteins and their interaction with telomere part of the chromosome which are important in telomere homeostasis. In this complex, TIN2 is the main TRF1-associated protein, and TPP1 is the main POT1-associated protein. The interaction of POT1, TPP1, TIN-2 and TRF1 makes a new way for signals to go through telomere.

1.5 Techniques for Telomere Length Measurement

Since telomere studies are very important in aging process, age-related diseases and cancer studies, different techniques have been implemented for telomere length measurement. These techniques include the TRF (Terminal Restriction Fragmentation), Q-FISH (Quantitative Fluorescence in Situ Hybridization), STELA (Single Telomere Length Analysis), PRINS (Primed Insitu) and qPCR (Quantitative PCR).

Although TRF remained the most pioneer and accurate technique in telomere studies and is the gold standard for absolute telomere length measurement, it needs large amount of DNA (0.5-5µg). Other techniques such as STELA, PRINS and Q-FISH like TRF not only need large amount of DNA, but also some of them such as Q-FISH are not able to measure telomere lengths in mitotically active cells.

Compared to all mentioned methods, qPCR is a relatively easy and fast technique which provides some privileges over other mentioned techniques. First, lower amount of DNA is required (35-50ng), second, telomere lengths are measurable in all stages of mitotically active and not active cells, and third, it gives the opportunity to study telomere lengths in both bulk and single cell levels. Measurement of telomere length in single cell level are useful in studying cell heterogeneity in the same cell population and is more specific in our researches and treatments. [40, 41]

Measurement of telomere length (T) to a reference single-copy or multi-copy genes (S), allows the relative telomere length (T-S) measurements. [42] This calculation method (the correlation of telomere length with reference gene) is used to determine relative telomere length and it is not able to measure absolute values of telomere. (Figure 7) However, the only way to consider them as absolute values is to compare them with reference cell lines in which other methods are used for their telomere length measurement. Different laboratories may get different results for relative telomere length of the same cell line as they may use different single copy genes, or the amplified control genes may not be specific in the genome, influencing the relative telomere length measurement. [43, 44] Worth to mention that despite all these advantages, Q-PCR is not able to measure very short telomeres. [44]

1.5.1 Telomere Length Measurement by quantitative PCR

This technique is based on a fluorescent marker which detects the amplified products of target gene and attaches to them. After attachments into the small gap of double stranded DNA, these markers emit light. During PCR, after each cycle of DNA amplification, the number of double stranded DNAs increases, and as a result the amount of SYBR Green markers attached into the double stranded DNA enhances. Thus, with more PCR cycles, the detector receives strong emission of fluorescent lights.

There are two qPCR methods:

1. SYBR Green-based qPCR, in which non-specific fluorescent markers are used to detect the target products amplified with sequence specific primers. SYBR Green is a fluorescent marker which is attached to all double-stranded DNA during qPCR.
2. Probe-based qPCR, in which sequence-specific fluorescent probes are used as a marker to detect target gene. In TaqMan chemistry these probes have a fluorescent label in one end and a quencher in the other end. The target sequence amplifies with another set of surrounding unlabeled primers, and the polymerase degrades the probe including the quencher. This results in increased emission of fluorescence signal relatively to the amplification level. [42, 44]

2. Aim

The aim of this experiment is to implement new and recently published technique [6] to study telomere length of single cell, parallel with its transcriptomes, and to discover highly correlated genes with telomere length.

Earlier research on telomere and telomerase were mostly conducted in bulk or tissue level. Research on bulk or tissue level does not, however, provide strong or accurate results as they do not consider the heterogeneity of different cell populations among sub-clones of the same cell lines. Our aim is to measure the telomere lengths and link them to RNAseq profiles in single cell level. In this study, the heterogeneity of telomere length and telomere-highly correlated genes were studied in different cancer cell lines, LPS-stimulated and LPS-non-stimulated monocytes and macrophages. In addition, studying both telomeres and transcriptomes in single cell level will explore mechanisms that have not been studied in bulk level.

In more detail, this thesis aims to study the correlation of different telomere related genes' expression levels and overall transcriptomic cell phenotypes parallel with telomere length and further study these identified correlations in different cell lines – this has rarely been conducted. Related pathways will be investigated for the list of genes extracted from the data in order to compare these pathways between different samples. For this purpose, in order to sort the cells inside each well of 96 well plate, we will use FACS. SYBR-green-based qPCR will be used for studying telomere length and, illumina sequencing will be used for studying the gene expression levels in each single cell.

This study – by exploring telomere highly-correlated genes - has the potential to discover new biomarkers and efficient treatments for age-related diseases including neurodegenerative diseases, cardiovascular diseases, cancer and aging processes itself. Furthermore, studying on single cell level, rather than on bulk or tissue level, will provide more credible evidence on heterogeneity of sample. Future perspective of this thesis is to to find important factors in telomere length regulation and its correlation in cancer and againg.

3. Materials and Methods

3.1 Cell lines

The following cell lines were used in this study: A549 (human lung epithelial cancer cells), HELA (human cervix cancer) (ATCC cell bank), LPS-stimulated (LPS+) and LPS-non-stimulated (LPS-) THP-1 (Acute monocytic leukemia) and LPS-stimulated and LPS-non-stimulated monocytes from a healthy donor. Anti-CD14 antibody was used to stain the monocytes so that we could sort them by FACS.

3.2 Cell culturing

The human monocytes leukemia cell line (THP-1) was obtained from ATCC cell bank and cultured further. These cell lines are similar to primary monocytes and macrophages from the point of morphology and differentiation features. THP-1 Cells were cultured in RPMI-1640 (BioWhittaker, Walkersville, MD-USA) supplemented with 10% (v/v) heat-inactivated Fetal Bovine Serum (FBS), 0.1 mg/ml penicillin-streptomycin (Gibco, Thermo Fischer Scientific), 1% (v/v) L-glutamine 25 mM, pre-heated into 37 degree in water bath. Some THP-1 monocytes were washed with PBS carefully and later treated overnight with 1.6×10^{-7} M of PMA in RPMI cell culture medium with 10% FBS in order to let them to differentiate into macrophages. A part of these macrophages was stimulated with LPS (Lipo-poly-sacharide) for 3 hours which were named LPS+MQ and those remained non-stimulated were named LPS-MQ. On the other hand, LPS-stimulated THP-1 cell lines were named THP1-LPS+ and LPS-non-stimulated cell lines were named THP1-LPS-. Primary monocytes which treated with LPS named as Mon-LPS+ and LPS-non-stimulated ones we named as Mon-LPS.

Human A549 and HELA (ATCC) cell lines were cultured in the same approach, using Basal Media DMEM (Dulbecco's Modified Eagle's Medium) including Glucose (BioWhittaker Verviers, Belgium) and supplemented with 10% (v/v) heat-inactivated FBS (Fetal Bovine Serum) and Pen strep (Penicillin Streptomycin). All cell lines were cultured under 37 Degree incubation in the presence of 5% of CO2. Cells quality and growth rate were checked under microscope and medium change and cell splitting were done, accordingly.

3.3 FACS Preparation

14 skirted 96-well plates (Hard shell PCR 96- well plates, thin wall) were loaded with approximately 4 µl of lysis buffer containing 0.3 µl of 2% triton x-100, 0.34 µl of 0.1 mM DTT (Thermo Scientific DL-Dithiothreitol; Clelands reagent), 0.01 µl of RNAase inhibitor (40 U/µl thermos scientific) and 1 µl of nuclease free water. After plates were loaded with lysis buffer, stored in -70°C overnight. The next day, plates and samples with the cell concentration of at least 1 million/µl (Measured with TC20 ™ Biorad cell counter) and the volume of minimum 200 µl were taken into FACS for single cell sorting. Single cells were sorted by flow cytometry inside each well of 96 well plate with 4 µl of lysis buffer. Plates were stored in -70°C, immediately. Cells were re-suspended to PBS before sorting.

3.3.1 Cell counting

Cell counts were done by TC20™ automated cell counter (BioRad). Cells were mixed with Trypan blue to stain dead cells prior to counting. Cell-count machine provided the total cell numbers and cell viability percentage.

3.3.2 Single cell sorting

In order to sort monocytes, the first step was to measure the cell concentration and count the cells with cell counter. Samples were diluted with PBS in to required concentration of 1 x 10^6 cells /ml and volume of 200 µl. and if needed to be diluted and re-suspended with PBS into required concentration (1 x 10^6 cells /ml in PBS) and volume (200 µl per tube). CD14 were used as co-receptor for LPS-stimulated monocytes because recognition of LPS by Toll-like Receptor 4 (TLR4) is facilitated by CD14.

In next step, SYTOX red dead cell dye (5nM) (Invitrogen) was added to the cells before the sorting in order to label dead cells. This marker penetrates into cell nucleus from permeable membrane of dead cells and attaches to the nucleic acids and stains them. However, in live cells since the membrane is intact, the markers are not able to attach into DNA and no staining happens. Using this method live cells and dead cells are recognizable from each other. After staining process, cells were

sorted into 96-well, thin well plates (HSp9601) by using BD Influx cell sorter (BioMedicum FACS core). Plates were sealed and covered from environment heat and light on ice until they stored in -70.

The normal cell concentration for FACS is considered between $1x10^5$ to $5x10^7$ cells/ml (we used 1 x 10^6 cells /ml in PBS) and cells re-suspended to PBS before sorting. Flow cytometry machine contains 633 nm laser which cause SYTOX red to be excited. Samples were prepared according to the SYTOX dead cell stain sampler kit. Just some move the paragraph to right.

3.4 Dynabeads and mRNA Capture

Parallel with FACS, dynabeads were coated with 12 different barcode oligoes (Dynabeads^{TM} M-270 Streptavidin invitrogen by thermos fisher scientific, Catalog numero. 65305, 65306) and loaded in wells from columns 1 to 12 (barcode-1 to barcode-12) and from raw A to H of 96 well plate. However, after many trials, cDNA detection from 12 single cells pooled together with 12 different barcodes was not possible so no detectable levels of cDNA were observed by Labchip (HT DNA High Sensitivity Assay Quick Guide LabChip^R GX/GXII Touch). Therefore, further 36 new barcode oligos were ordered enabling of pooling total number of 48 together and increasing the cDNA amount from single cells.

We took the plate for each sample from -70 and vortexed whole plate for 30 seconds in order to let lysis buffer to degrade cell membrane and let gDNA and mRNA to be released inside each well. Later, wells were loaded with 10 µl of dynabeads with unique barcode oligoes (barcode1-barcode 48) (Appendixes). The beads and released mRNAs were incubated at room temperature for 5-10 min in order to let the poly-A tail of mRNAs attaches properly into poly-T tails of oligobeads. Wells were mixed regularly in order to prevent beads settlement. Beads and attached mRNAs were gathered to the wells' side by magnetic plate and separated mRNAs from the Rest of supernatant, manually. Since mRNAs from single cell are very degradable and have very short life span, after extraction they were transferred on ice while doing the experiment.

The supernatant containing single cell g-DNA and other cell-related components were transferred into new 96 well plate for later g-DNA purification (stored in +4) and qPCR.

Beads attached to mRNAs inside each well were washed two times with 30μl of washing solution, 6xSSC (Saline Sodium Citrate) and one time with 20 μl of 1xRT (reverse transcriptase) buffer in order to clean beads from other probable redundant and prepare them for RT PCR.

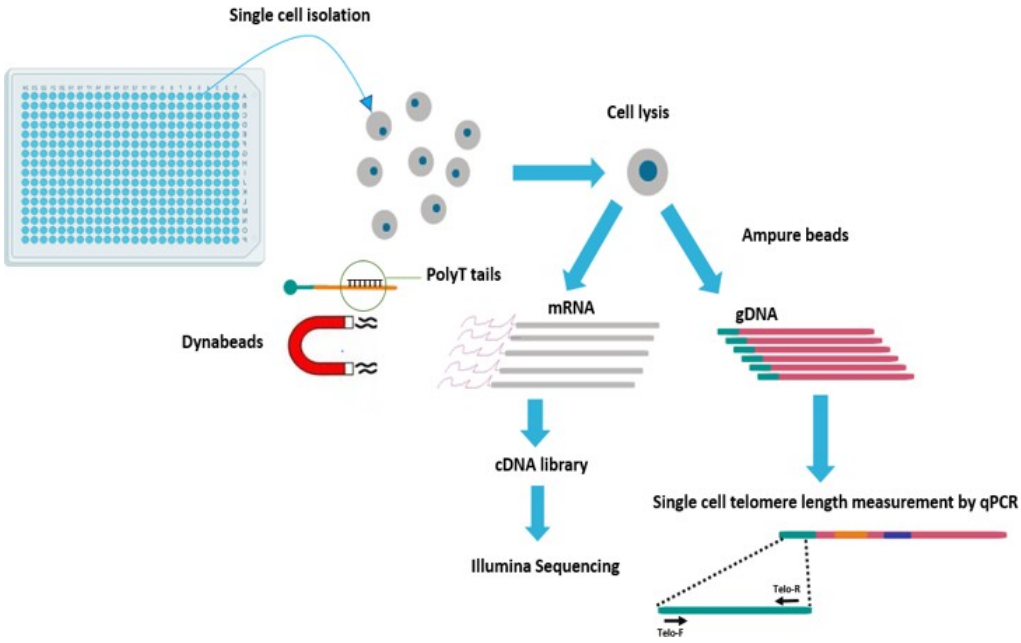


Figure 6. Schematic overview of telomere length measurement and Gene expression analysis in single cell level. [6]

3.5 cDNA Library Preparation

5 µl of RT (Reverse Transcriptase) master Mix containing Maxima 5× RT buffer (1 µl) (200 U/µl Thermo Scientific Maxima H minus Reverse Transcriptase), RNAase inhibitor (40 U/µl) (0.05 µl), 10 mM dNTPs (Thermo scientific) (0.5 µl), 50 µM template switching oligo, TSO (AAG CAG TGG TAT CAA CGC AGA GTG AAT rGrGrG) (0.25 µl), Maxima RTase (0.15 µl), 5M Betain (0.75 µl) and nuclease-free-water (2.3 µl) was added into each well with oligobeads-mRNAs. The plate then was taken into PCR machine in order to provide suitable temperature (30 min at 25 degree and 90 min at 42 degree) for reverse transcription. The cDNAs which are synthesized on the surface of beads were washed from primers and other extra elements by TE-SDS 10 mM (Tris PH 8.0 (0.5 ml) + 1 mM EDTA (0.1 ml) + 5% SDS (0.25 gram) + water (49.1 ml)) and TE- TW (10mM Tris PH 8.0 (2 ml) + 1mM EDTA (0.4 ml) + 0.01 % tween-20 (2 ml) + 195.6 ml water (195.6 ml)), respectively. Magnetic plate was used during washing steps to keep the beads separate from the supernatant.

For the removal of RNA/cDNA unbound oligos on beads exonuclease treatment was conducted. The exonuclease super mix contained 10x exo-buffer (2µl), exonuclease I (20U/µl) (1µl) and water (17µl). 20µl of exo-super-mix was added on top of beads inside each well and heated for 45 min at 37 degree using PCR thermo-cycler machine. After exonuclease treatment, beads again were washed with TE-SDS and TE-TW.

In order to do cDNA amplification, KAPA PCR was implemented, immediately by adding 10 µl of KAPA mix on beads, containing 7.5 µl of KAPA HiFi HotStart ReadyMix, 0.06 µl of SMART PCR primer (AAG CAG TGG TAT CAA CGC AGA GT) (200 µM) and 2.44 µl of nuclease-free water. (KAPA HiFi PCR Kit-KR0368- v11.17) The plate was mixed well and transferred to PCR machine according to the fallowing program: 95 for 3min and 98 for 20 sec, 65 for 45 sec, 72 degree for 3 min, 98 for 20 sec, 67 for 20 sec and 72 for 3 min (18 cycles).

3.5.1 cDNA Measurement

The single cell cDNA libraries were measured by HT DNA High Sensitivity Assay Labchip which was used for the measurement of successful cDNA libraries. The

measurement with this device includes three main steps (HT DNA High Sensitivity Assay Quick Guide Lab chip GX/GXII Touch guidelines): [45]

1. Chip preparation
2. DNA sample, buffer, and Ladder preparation
3. Chip cleaning and storage

3.5.2 Library Preparation and Sequencing

Since detecting each single cell cDNA product by Labchip machine was not possible, the amplified cDNAs were separated from wells beads by magnet and pooled together into two separate eppendorf tubes. Eppendorf tube 1 included first 48 cells pooled together (A_1-D_{12} wells) and eppendorf tube 2 included the second 48 cells pooled together (E_1-H_{12} wells) while each single cell maintaining its unique identity and barcode linked to the cDNA.

The pooled cDNAs in eppendorf tubes (volumes variable between 400-480µl) were purified by normal Ampure XP beads (Beckman Coulter, A63881) (0.6 time of the sample volume).

The purification steps were performed according to the following condition:

Beads were taken from +4 into room temperature half an hour before purification. Later, 0.6× of samples volume Ampure XP were added, vortexed shortly and incubated 5min in RT. 80% alcohol was used to wash the beads and then the beads were left on table for alcohol evaporation. Magnetic plate was used during the washing steps. Eventually, the captured cDNAs were released into 10 µl of water and cDNAs were transferred into new eppendorf tubes in order to proceed with tagmentation. The concentration for each sample was measured by Qubit 2 fluorometer for dsDNA HS in order to make product to reach the required concentration for tagmentation (0.6 ng/µl), for dilution.

Tagmentation with Illumina Nextera kit is a transposase mediated protocol through which DNA is simultaneously cut into smaller fragments and Illumina adapters added to the fragment ends. Later, SMART PCR primers (AAG CAG TGG TAT CAA CGC AGA GT) and Illumina indexing primers were used to amplify only the tagmented 3’ end fragment of the cDNAs that carry the well/cell specific barcode. [46] According to the tagmentation protocol, all samples, in first stage, were diluted with nuclease-free-

water to reach the 0.6 ng/µl concentration. Based on part A of the protocol of Nextera XT Library Preparation Kit, FC-131-1096 and the manufacturer instructions, 0.6 ng of cDNA were combined with water in total volume of 9 µl, then 10µl of Nextera TD buffer plus 1µl of amplicon tagment enzyme were added on top of 9 µl sample and mixed well, incubated at 55 degree (preheated the thermocycler machine beforehand) for 5 min. In order to stop the reaction, 5 µl of neutralization buffer was added on top of the sample and was kept in RT for 5 min.

Based on part B of the protocol all reagents were added in the fallowing order: 15 µl of Nextera PCR mix, 8 µl of water, 1 µl of New-P5-SMART PCR hybrid oligo and 1 µl of 10 µM Nextera N70X indexing oligo (16 different N70X) (Table 1). Later, HotStart PCR (×12) with 55 degree was used for enzyme activation. The cDNA library was purified twice using the AMPure beads. After first purification DNAs were eluted in 30 µl of water and after second purification DNAs were released in 10 µl of water. Labchip was used for checking successful tagmentation products. Nextera indexes (N70X) were used for Illumina sequencing clustering and barcoding during sample preparation. (Table 1) Samples were sequenced with Illumina NovaSeq device at FIMM sequencing core facility in Meilahti.

Number	sample/plate name	Concentration before Tagmentation (ng/µl)	Sample Volume	water volume needed for required dilution (µl) (C1V1=C2V2)	Indexes used for Tagmentation	Concentration after Tagmentation (ng/µl)
1	Mon´-LPS-/31st 2	10	1.2	18.8	N711	12.4
2	Mon´-LPS-/31st 1	15	0.8	19.2	N712	15.6
3	Mon´-LPS-/29th 1	18	0.7	19.3	N714	13.1
4	A549-1	56	0.2	19.8	N716	67.8
5	A549-2	25	0.5	19.5	N718	25.6
6	HELA2-1	24	0.5	19.5	N719	40.8
7	HELA2-2	12	1	19	N720	15.3
8	Mon´-LPS-/29th 2	26	0.5	19.5	N721	38.4
9	Mon´-LPS-/31st 2	28	0.5	19.5	N722	32
10	HELA-1-1	21	0.6	19.4	N723	22.8
11	HELA-1-2	44	0.3	19.7	N724	39.6
12	Mon1-LPS+	12	1	19	N726	10.9
13	Mon2-LPS-2	7	1.8	18.2	N727	8
14	Mon2-LPS-1	18	0.8	19.3	N728	18.7

Table 1. Detail information (indexes used for each sample and cDNA concentration) of different samples and different plates which proceeded with tagmentation and Illumina sequencing. All samples were diluted with nuclease-free-water until they reach to the concentration of 0.6 ng/µl.

3.6 Analyzing Data with Chipster and Read Alignment

Sequencing data were processed on the CSC's Chipster platform 60 using Drop-Seq tools (v1.13), as the library construction method is compatible with the Drop-seq protocol. [47, 48] Reads were aligned to the Human reference genome hg38 containing all human genome and plus mitochondrial DNAs.

3.6.1 Analyzing steps

1. The Quality Control (QC): Data from Illumina Novaseq sequencing were processed, and their quality was using Chipster and Single cell RNA-sequencing Program. The data was paired-end with Read 1 containing the cell barcode and the Unique Molecular Identifier (UMI) of 26 bp length and Read 2 of 91 bp length. The FastQC (Babraham Institute, Cambridge, UK) results showed that the reads were in good quality.

2. Preprocessing FASTQ Files using Drop-Seq tools: Since FASTQ files are text-based, therefore they are unable to keep the cell barcodes of reads data. To address this problem FASTQ file was changed into unaligned BAM file to: First, labeling the reads with cell and molecular (UMI) barcodes, second, filtering bad quality reads, poly-A trimming and adapter trimming and third, converting the BAM file into the format of FASTQ file again to do read alignments of genome data set to human genome version hg38 to using the STAR aligner in single end mode (STAR aligner only accepts FASTQ files as data). The aligned BAM file were successfully labeled and is shown in Chipster data part. [49]

3. Aligned and Unaligned BAM Files Integration: In this step of program, barcode and UMI sequences are retrieved from the original unaligned file.

4. Barcode error correction: barcode errors are detected and corrected if possible.

5. Digital Gene Expression Matrix (DGE) Formation: In this crucial stage, reads which covers one specific exon (Homo Sapiens GRCh38), are labeled with gene name and therefore genes were followed from each single cell. Moreover, UMIs are counted as unique transcripts, giving the gene expression value (the estimation of the number of transcripts) for each single gene as reads have already been tagged with genes names.

- Combined all same samples from the same plate by Microarray /utilities/Merge tables
- Seurat workflow

The DGE was imported into R and analyzed using Seurat package v2. [50] The data was normalized, and thereafter dimensionality reduction and cell clustering were done according to the default Seurat protocol. Genes that were expressed in at least three cells were included in the analysis. The data included barcodes with one, two and no mismatches. These mismatches were distinguished from each other by using code written by python. (Supplementary: 9.7 Code 2) We removed data with one and two mismatch barcode errors as some barcodes had only two different nucleotides and they could not be separated from sequencing. We only included data from complete matches in order to prevent probable barcode misinterpret between different samples and also thesis elongation.

3.7 Genomic DNA Purification and Pre-amplification

After mRNA extraction, the rest of the supernatant (including gDNA and cell components) was transferred into new 96-well plate and stored in +4 (to continue the qPCR experiment in less than one week), or in -70 (to continue gDNA experiment after weeks or months). Since the supernatant containing gDNAs had other impurities (from cell components) and gDNA purification was required, Ampure XP beads was used to capture gDNAs from the rest of the supernatant (the same steps used for purification during tagmentation). The amount of AMPure beads used were 0.6 times of the gDNA sample. The dilution ratio of 0.6 will give cut off of 300 nt which means gDNAs longer than 300 nt could be captured. The captured gDNAs were released into 8µl of water in the very last step of the purification.

The purified gDNA supernatant was continued with pre-amplification in which 42 µl of master mix (Thermo scientific standard protocol) was added on top of 8 µl of purified gDNA supernatant, in each well. Each reaction of Single-cell gDNA pre-amplification was set up by 5 µl of PCR buffer, 5µl of dNTPs (prepared 10 mM dNTP mix), 1 µl of each reverse and forward telomere and Alu (reference gene) primers (Supplementary: Table 9.4), 0.25 µl of Hot start dream taq DNA polymerase and water (total volume of 50 µl). PCR was performed according to the recommended thermal cycling steps (thermos scientific standard protocol), however, ×18 cycles was used instead of recommended 25-40 cycles to reduce the probability of amplification bias. The primer annealing temperature for PCR was 60 degree. Pre-amplified gDNA concentration were measured from random wells by Qubit to make sure about successful pre- amplification.

3.8 Single-cell telomere length measurement by qPCR

Single cell telomere lengths were measured using qPCR (Biorad CFX384 Real-Time System) technique. Alu was used as reference gene with millions of copies interspread overall the human genome, still maintaining same number of copies throughout the single cells. Frame star^{R} 384-well plate (clear wells, clear frame) (4ti-0384/C) were used for qPCR.

After gDNA purification and pre-amplification, the process was continued with SYBR Green based qPCR. The reaction was implemented in iQ^{TM} SYBR Green super mix in

384 well plate with total volume of 25 µl. qPCR master mix were prepared for pre-amplified gDNAs of Alu and Telomere separately and then loaded in each well. Each reaction mix contained 0.625 µl of each forward and reverse primers (Supplementary Table 9.4), 12.5 µl of 2×MM (iQ^{TM} SYBR Green Supermix) and 6.25 µl of DNAse free water. The amount of pre- amplified gDNAs used in each reaction was 5 µl. The plate was loaded half with Alu master mix and half with Telo master mix with two times repetition for each cell. Later, the plate was covered with transparent cover and centrifuged down at 1000×rcf for 2 min.

Ready extracted gDNA (stored in +4) was obtained from a volunteer as positive control. Since in each human single cell there is normally 6 to 7 pg of gDNA, representative gDNA dilutions were estimated from the ready gDNA. Therefore, first the ready gDNA concentration was measured by Qubit and according to the $C_1V_1=C_2V_2$ formula was diluted to reach the concentartion of 7ng/µl. Later serial dilution was conducted in 8-strip tube as fallowing order: 700 pg/µl →70 pg/µl →7 pg/µl →0.7 pg/µl →0.07 pg/µl.

8 µl of serially diluted samples were taken into the new 8-strip tube and continued with pre-amplification of simulated gDNAs similar as what was done for gDNAs from each single cell (Taq protocol). For negative control nuclease free water was used instead of gDNA samples. The Bio-rad qPCR was implemented with reaction conditions of 95 10 min, 95 15 sec, 60 30 sec, 72 30 sec, GoTo ×18, 60 5 sec, 95 1 min. Two primers were required to amplify telomere and two primers to amplify reference genome (reverse and forward).

After completing the qPCR, the Cq values for each well were obtained (also called Ct values in some sources). We normalized telomere length by subtracting Cq of telomere length from Cq of reference genome (Delta Cq). The higher the Cq value, the lower telomere length.

ΔCq= Cq Telomere (Target area) - Cq Alu (house-keeping-gene)

ΔCq ↑= Telomere Length↓

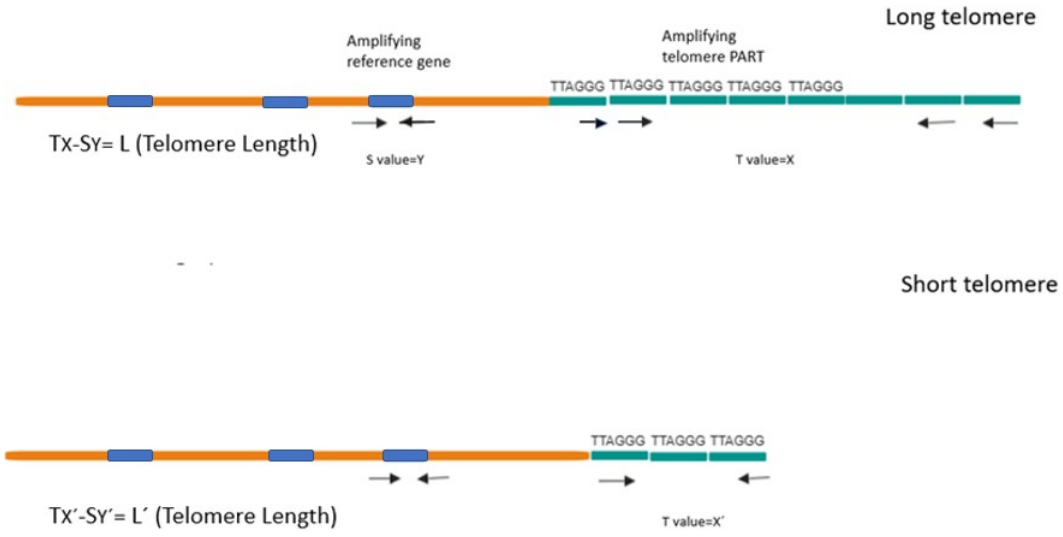


Figure 7. Telomere and reference genes amplification with primers and telomere lengths calculating. “S” represents reference genome (Alu) and “T” represents telomere.

Alu sequence used for designing the qPCR primers was taken from DFAM data base. The product was 127 bp long. Consensus sequence from DFAM>DF0000053:AluYa5 (Supplementary: Table 9.4).

4. Results

4.1 Cell Count Related to LPS Stimulated Cells

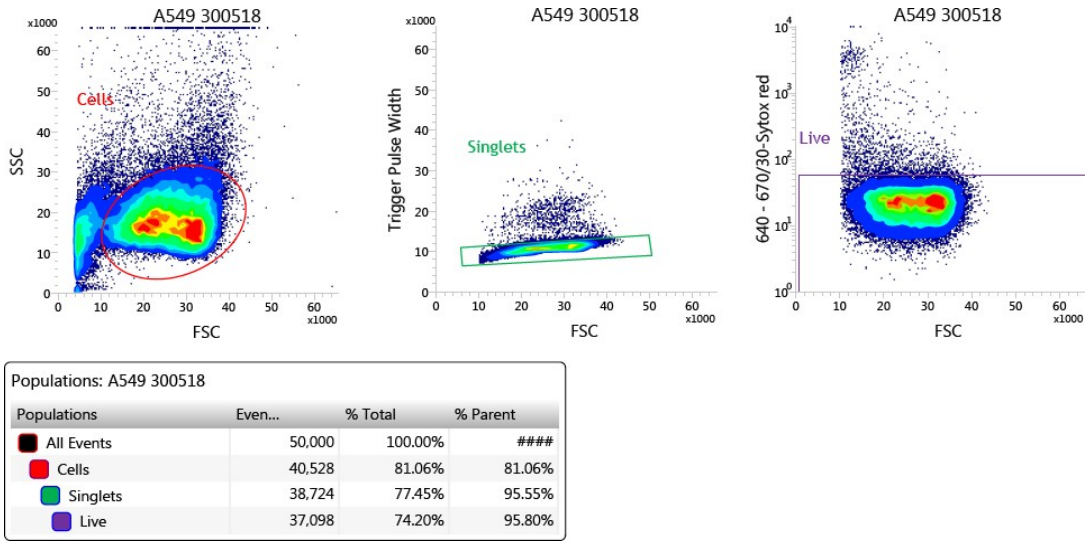
Since LPS is a highly immunogenic glycolipid found in the outer membrane of Gram-negative bacteria, we decided to study the influence of LPS (representative of gram-negative bacteria) on telomere lengths and scenesence. According to our data from the cell concentration measurements, the viability of cells was lower in LPS-stimulated monocytes and LPS-stimulated macrophages compared to LPS-non-stimulated monocytes and macrophages. (Table 2) In addition, it has been reported later in the results (page 41, Figure 11) that stimulated monocytes, also in average showed shorter telomere length than non-stimulated monocytes.

Sample	cell concentration (cells/µl)	cell viability
LPS-,Monocyte	2.2×10^6	78%
LPS+,Monocyte	1.6×10^5	89%
LPS+,Macrophage	7.0×10^5	75%
LPS-,Macrophage	1.8×10^6	91%

Table 2. LPS stimulated cell lines show smaller viability percentage compared to LPS-non-stimulated cell lines.

4.2 Single Cell Isolation

FACS (Fluorescence-Activated Cell Sorting) technique was used to isolate each single cell into each well of 96 well plates (12 plates) which were previously loaded with lysis buffer (see Materials and Methods). Detailed information about the run status and cell load of each plate was obtained as below: (Figure 8 related to FACS)



Populations: A549 300518

Populations	Even...	% Total	% Parent
All Events	50,000	100.00%	####
Cells	40,528	81.06%	81.06%
Singlets	38,724	77.45%	95.55%
Live	37,098	74.20%	95.80%

Figure 8. FACS sorting gating strategy for A549 cells (1 example of 12 plates with cells sorting information): Left: forward scatter (FSC) and side scatter (SSC) of the whole population. Middle: Singlets shown with green gate. Right: Viability of single cells population, separated by violin line from the rest of dead cells or apoptotic cells. The other samples were repeated the same way. Through gating strategy, the low qualified cells were excluded in sorting as the exclusion of cell duplicates (cells attached to each other) in middle plot or dead cells in right plot.

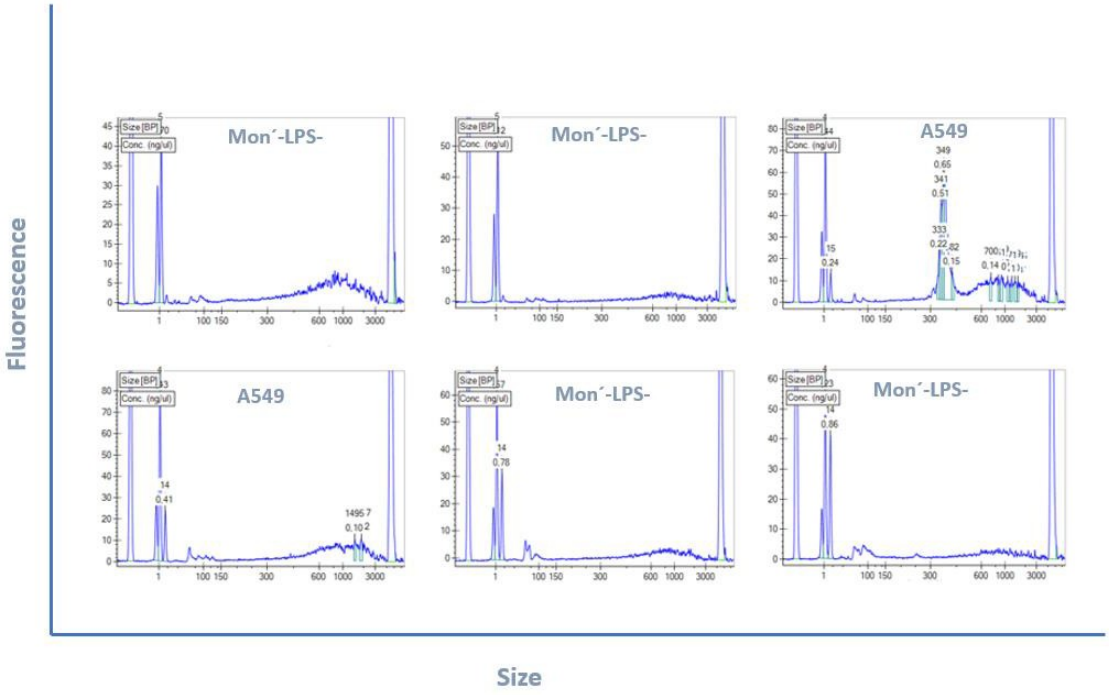
4.3 cDNA Library Evaluation

After cell lysation, released mRNAs were captured by polyT tail of barcoded dynabeads through their poly-A tails. Extracted mRNAs were converted into cDNA using Reverse Transcriptases (RT) technique.

Exonuclease treatment was implemented in order to remove unbound mRNA oligos from beads which have the potential to interfere with absolute molecule counting with UMI sequences.

KAPA PCR was thereafter conducted for cDNA amplification. The success of amplified cDNA products was checked by HT DNA High Sensitivity Assay Labchip. (Figure 9)

Since Labchip is unable to detect cDNA product from single cell, cDNAs from separate wells were pooled together into two separate Eppendorf PCR tubes; Eppendorf tube 1 included cells from the first 48 wells (A_1-D_{12} wells) and Eppendorf tube 2 included cells from the second 48 wells (E_1-H_{12} wells). Although all cells from half of each plate were pooled together, the cDNAs from each single cell saved its unique identity by having different barcodes. The pooled cDNAs were purified by 0.6 x Ampure XP beads and proceeded with tagmentation. The ratio of 0.6 for ampure beads was used for purification based on cDNA product size.



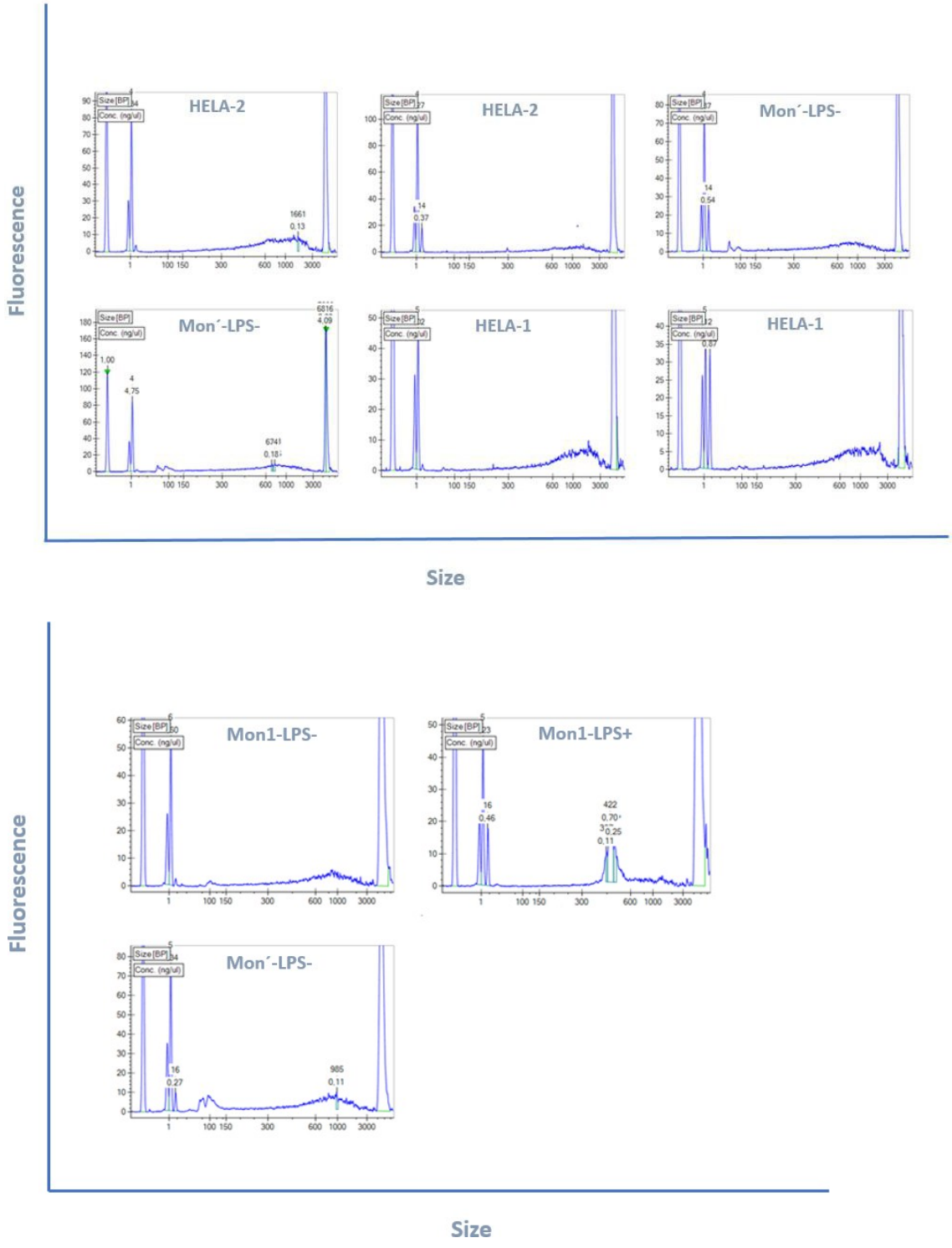
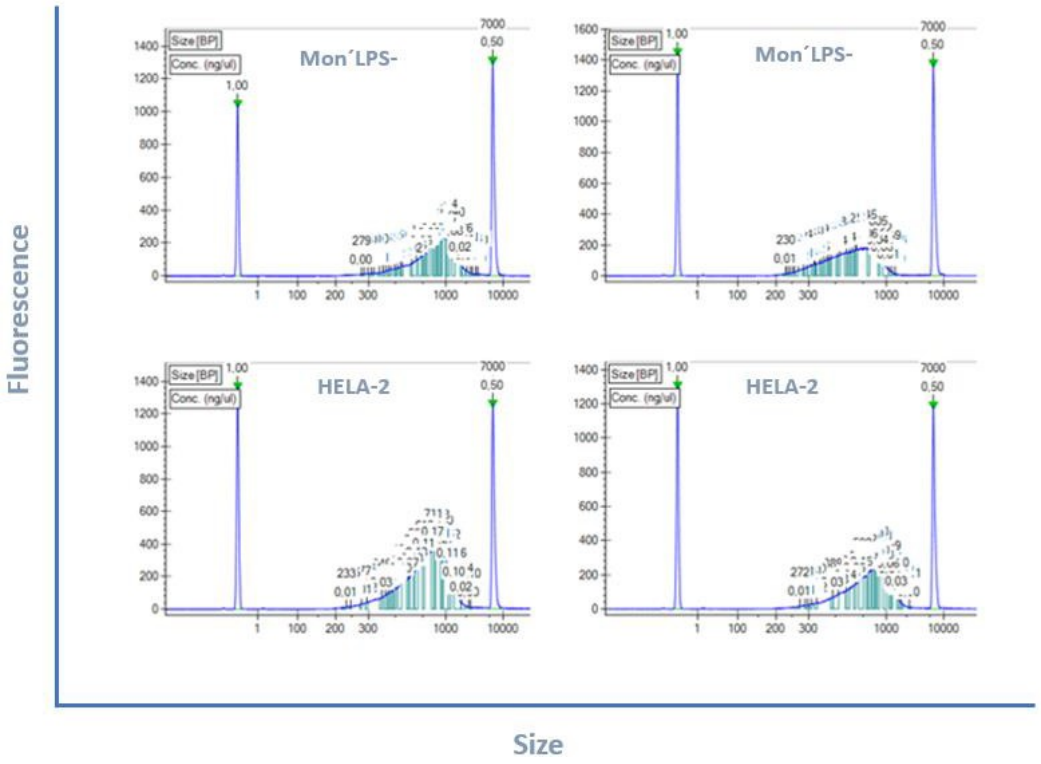


Figure 9. Successful cDNA library amplification measured by HT DNA High Sensitivity Assay Labchip after cDNAs were pooled together. x- axis shows the product (cDNA) size (ranges from 300 bp to 3000 bp) and y-axis shows the cDNA product concentration. The height of peaks shows the amount of amplified cDNA library products (ng/ul).

4.4 Tagmentation and cDNAs Preparation for Sequencing

We used tagmentation protocol of Nextera XT Library Preparation Kit, FC-131-1096 and followed the steps from manufacturer instructions (see materials and methods for more details). The cDNA libraries were purified twice using the AMPure XP beads. Labchip was used to check successful tagmentation products. (Figure 10) In this step, Nextera indexes (N70X) were used for Illumina sequencing clustering and barcoding during sample preparation (Supplementary Table 9.5).



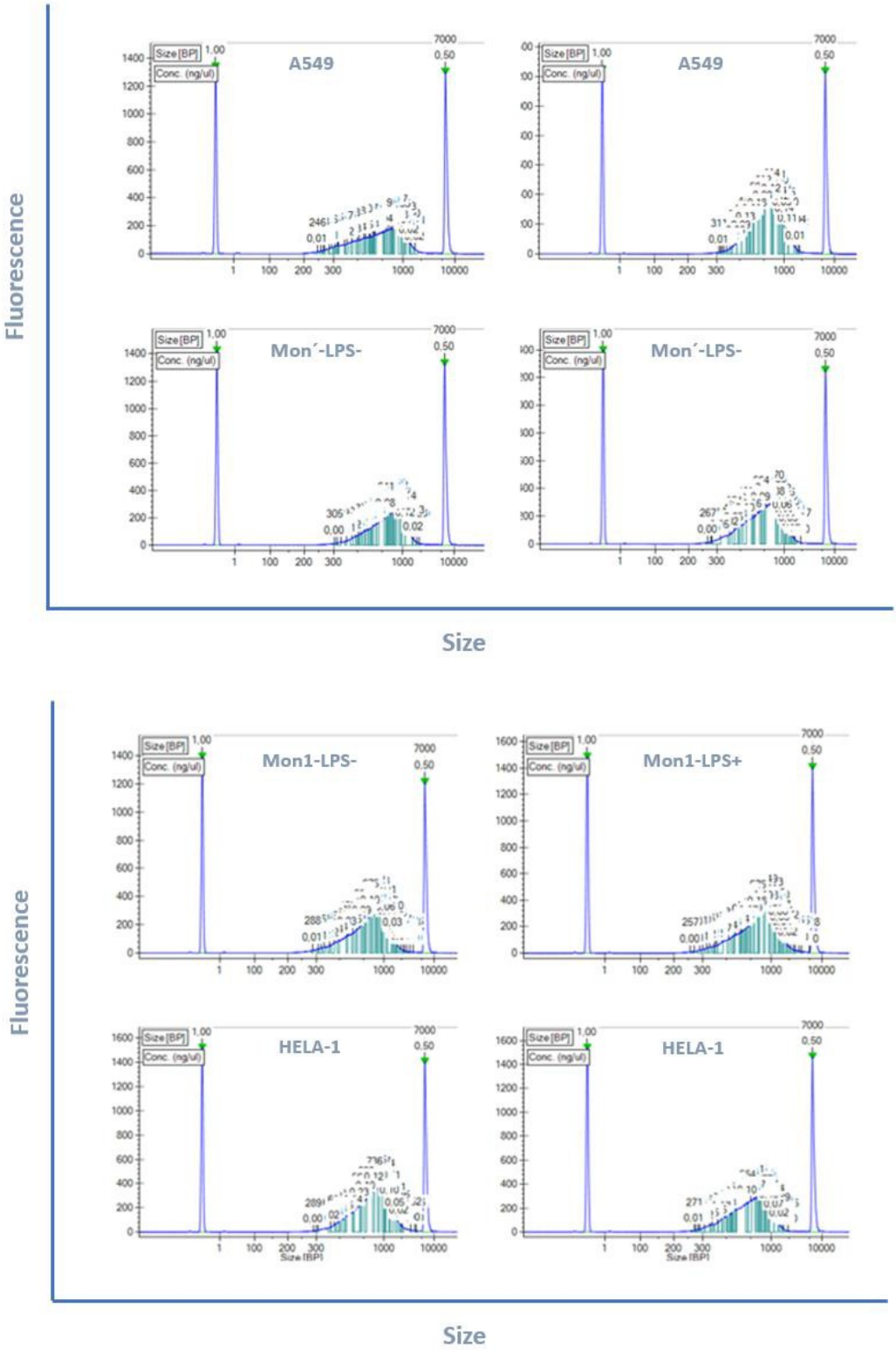


Figure 10. Electropherograms from Labchip after successful cDNA tagmentation measured by HT DNA High Sensitivity Assay Labchip.

4.5 Illumina Sequencing and Seurat Clustering

cDNAs from each single cell of all samples were clustered separately based on their gene expression using principal component analysis (PCA) and unsupervised hierarchical (HCL) to visualize the data. Accordingly, cancer cells (A549 and HELA) and LPS-stimulated monocytes (Mon-LPS+) were clustered in the same area (tSNE_1 "-5-10"), (tSNE_2 "5, -5) with high heterogeneity in gene expression level. However, LPS-non-stimulated monocytes (Mon2-LPS-, Mon'-LPS-) were clustered separately with lower heterogeneity frequency and did not show any interference with other samples. (Figure 11) Samples were selected based on their UMI barcodes which were unique for each sample.

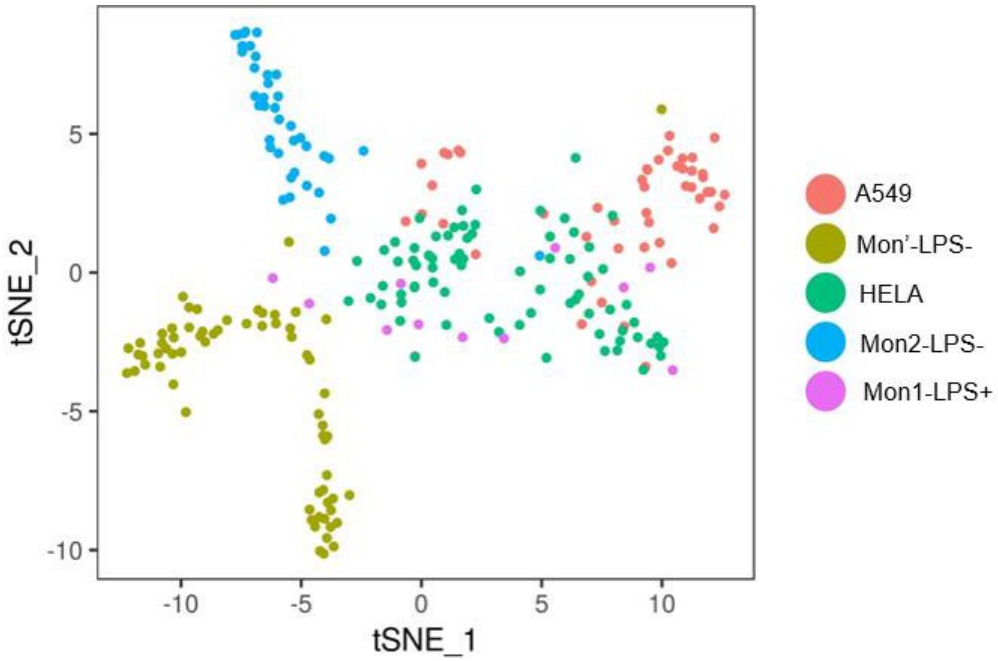


Figure 11. 5 Different cell lines and their Digital Gene Expression level. Each sample (DGE) was visualized with different colors. LPS-stimulated monocytes, A549 and HELA showed considerable heterogeneity in gene expression level and they appeared in the same area in tSNE platform. LPS-non-stimulated monocyte and primary monocytes were clustered in separate groups and showed very less heterogeneity.

In addition, DGE data (Transcriptomes) were aligned with Cq data of qPCR (Telomere length) and samples were clustered into four different groups based on their telomere lengths: Short, medium, long and NA (Cq = CqTelo – CqAlu); however, majority of cells showed medium telomere length. (Figure 12)

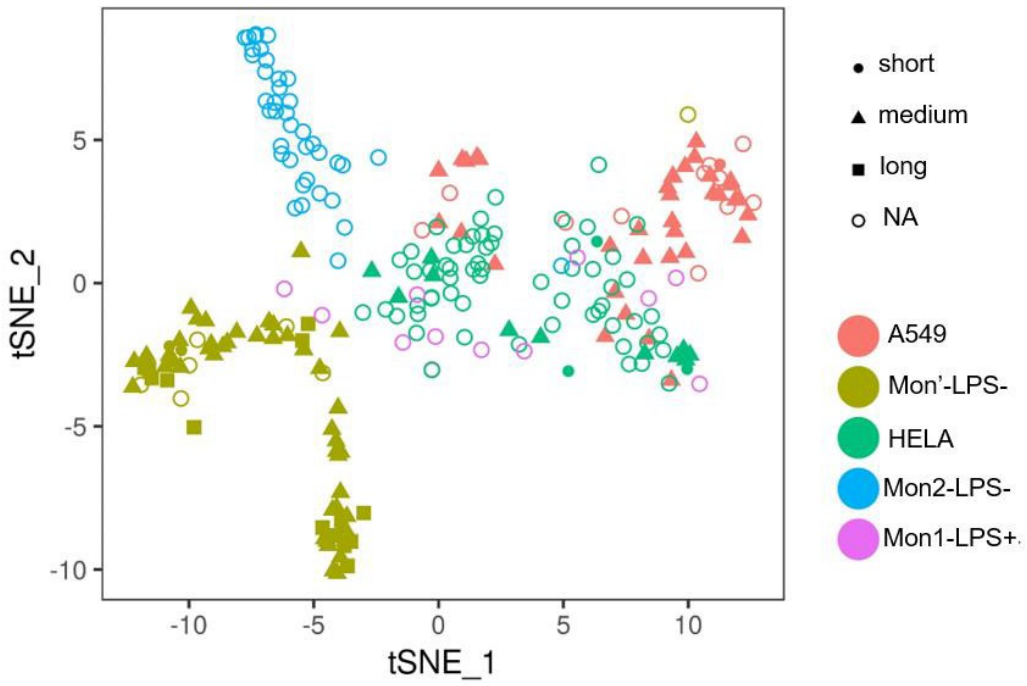


Figure 12. Cell from different cell lines with their telomere lengths in single cell level. Short telomeres have been shown in filed circles, medium size telomeres in triangles and long telomeres in rectangular shapes. NA, in this figure shows no Cq values recorded for those cells (however we have DGE data for them). In this study, although Cq values were available for Mon1-LPS+ and Mon2-LPS-, they were unavailable for Mon1-LPS+ and Mon2-LPS-, due to failure in qPCR. On contrary, DGE data was available for Mon1-LPS+ and Mon2-LPS- but not for Mon1-LPS- and Mon2-LPS+ due to failure cDNA library preparation. Mon´-LPS- was the only primary sample, which has both data from qPCR and Illumina Sequencing. Noteworthy, Mon2-LPS- includes Mon2-LPS-1 (half plate1-48 wells) and Mon2-LPS-2 (half plate 48-96 wells).

4.6 Well-Known Telomere Related Genes

Studying telomere related genes are difficult in single cell level as they are expressed in low amount such as telomerase (*TERT*). However, in this study, most of shelterin complex-related genes showed high expression levels. Shelterin-related genes such as *NOP10, DKC1, RIC8A, PARP2, HAT1, RAP1B* were highly expressed in cancer cell lines (HELA and A549) and LPS-stimulated monocytes (Mon1-LPS+) than LPS-non-stimulated monocytes (Mon1-LPS-, Mon´-LPS-). Genes such as RAP1B showed no different gene expression level based on its sample of origin. However, some genes such as CD53 was expressed highly in primary non-stimulated cell lines compared to other cancer cell lines and LPS-stimulated cell lines. (Figure 13) Although cancer cells showed high expression level of shelterin related genes compared to LPS-non-stimulated monocytes, they have short telomeres.

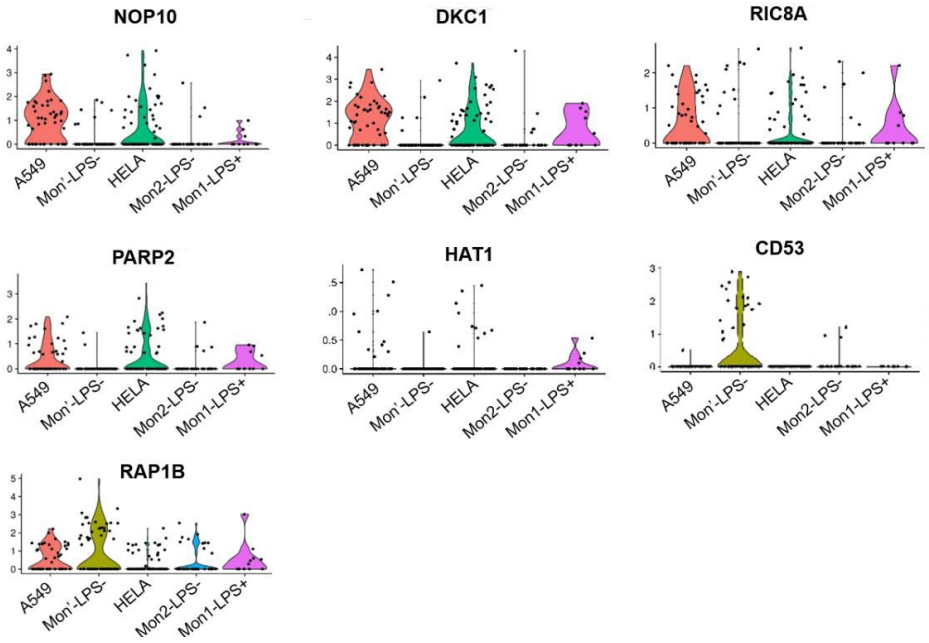


Figure 13. Violin plots for different shelterin related genes. In most cases shelterin related proteins tended to show higher expression level in cancer cell lines compared to the normal cell lines. LPS-stimulated monocytes also showed mostly higher expression for shelterin related genes in a similar way as cancer cell lines.

4.7 qPCR-Based Telomere Length Measurement

According to DGE data from Illumina sequencing and Cq values from qPCR, cancer cells not only show heterogeneity in their transcriptomes but also in their telomere length. (Figures 11 and 14)

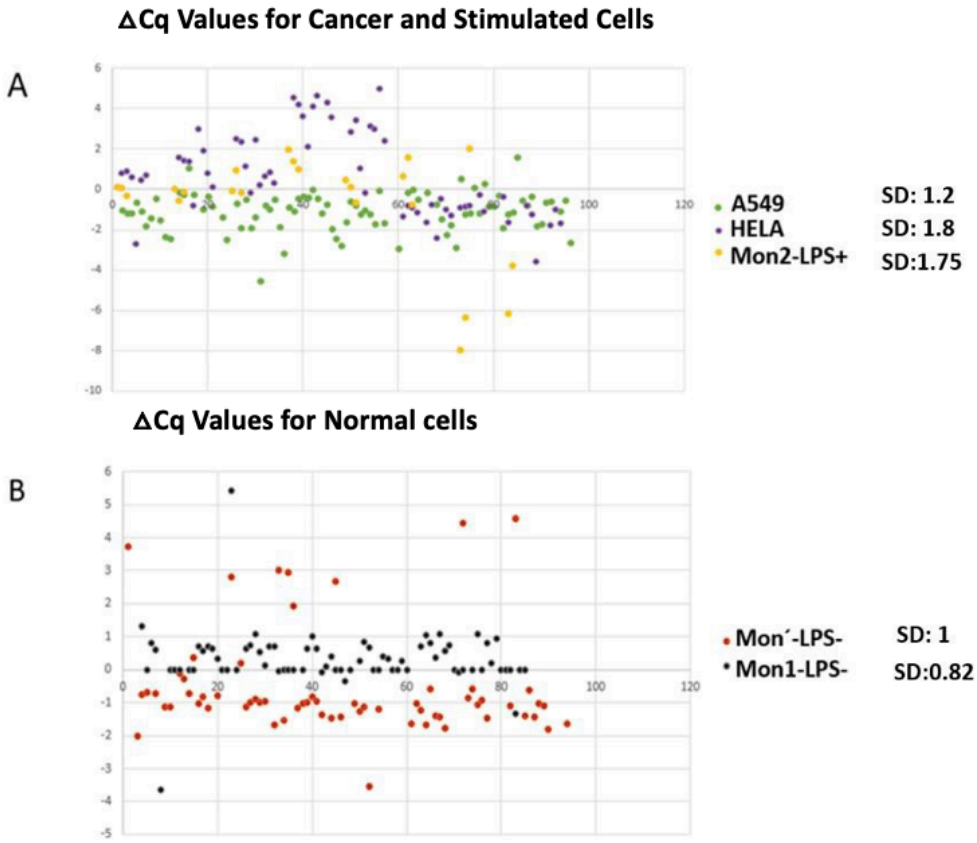


Figure 14. Scatter plot of ΔCq values (Cq telo-Cq Alu) for each sample in single cell level. The normal cell lines show less heterogeneity compared to the cancer cell lines and LPS-stimulated monocytes. Noteworthy, Mon'-LPS- and Mon1-LPS- are both the same type of sample (normal monocytes without LPS-stimulation), however from different plate. SD referes to standard deviation.

Cq refers to the cycle in which fluorescence can be detected by sensor and is the elementary result of quantitative PCR. Lower Cq values mean higher copy number of the target gene. The negative ΔCq (CqTelo-CqAlu) means the Cq value for telomere was lower than that for Alu. Thus, negative ΔCq represents longer telomere length. Alu was used as a normalization factor to correct for differences in DNA extraction and PCR efficiency between cells.

Our results revealed that cancer cell lines and LPS-stimulated monocytes (Mon1-LPS+) have shorter telomere lengths compared to the LPS-non-stimulated monocytes (Mon1-LPS-, Mon´-LPS-). (Figures 14 and 15) The probable reason is the unstoppable and fast division of cancer cells lets telomere lengths to shrink with each cell division. Although telomere lengths in cancer cells stay short, but they can escape from DNA damage response (DDR), senescence and apoptosis with frequent cell division. The mechanisms why cancer cells survive with short telomeres and which genes are important in their telomere protection still is under study. Scientists believe that cancer cells keep their short telomere stable as a result of higher level of telomerase enzyme or mechanism named as alternative lengthening of telomeres (ALT). [36] However, why LPS-stimulated cells show short telomere length in average compared to the LPS-non-stimulated cells, requires more investigation.

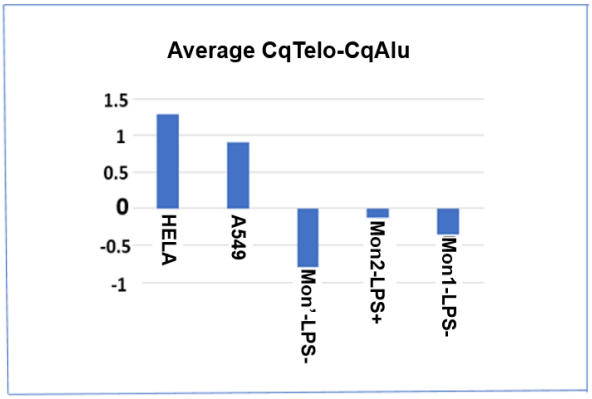


Figure 15. Average ΔCq values (not average telomere length) of single cells for each plate. Although Cq values are available for Mon1-LPS- and Mon2-LPS+, they are unavailable for Mon1-LPS+ and Mon2-LPS- due to failure in qPCR.

LPS-non-stimulated monocytes (Mon´-LPS- and Mon1-LPS-), LPS-stimulated monocytes (Mon2-LPS+), A549 and HELA classified from long to short based on their telomere length. In this category primary monocytes showed the highest average telomere lengths and HELA showed the lowest average telomere length compared to other samples.

4.8 The Correlation between Telomere Length and DGE in Each Single Cell Level

We wrote a program with python to align Cq values with DGE data of each single cell level. (Supplementary: 9.6 Code 1) Correlation was calculated between ΔCq (CqTelomere - CqAlu) and transcriptome (DGE) of each single cell. Later, genes were sorted in the order of highest to lowest in excel sheet based on their correlation value. Using this method genes which were highly correlated with telomere were sorted from high to low. Since ΔCq has reverse correlation with telomere length (ΔCq = 1/telomere length), low correlation value between transcriptome and ΔCq means high correlation with telomere length. In Figure 16 the correlation heatmap were ploted (based on delta Cq and transcriptome correlations) for all target genes (known telomere related genes) plus newly found genes (genes extracted from our data). We used heatmapper web server for data visualization. [51] As a result, *SLC38A2, PURB, UBR3, SSR1, NCAPH2, AIMP2, PHF21A* genes were found to be mutual between HELA, A549 and LPS-non-stimulated monocytes. To our knowledge, some of these 7 genes have not been considered important candidates for telomere or telomerase regulation-related pathways in a direct or indirect way. (Figure 16)

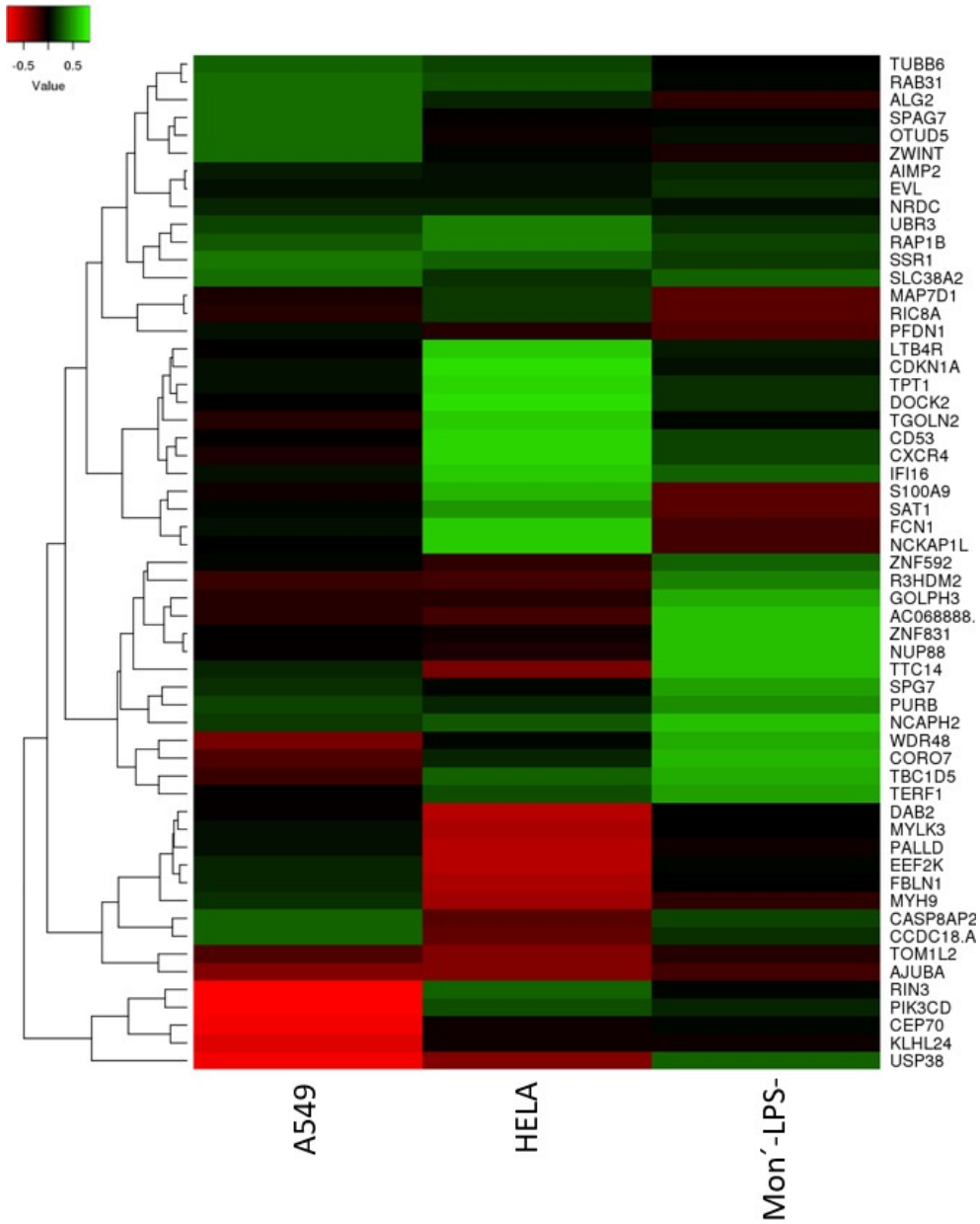


Figure 16. Heatmap based on correlation of telomere length with different telomere-related genes. This heatmap includes I) 10 genes from each sample which were the most highly correlated with telomere length and their equivalent in other samples, II) 5 genes from each sample which were the least correlated with telomere length and their equivalent in other samples, and III) 15 genes which are well-known [6,27,39,25] for their importance in telomere maintenance and stability or telomere regulation. The clustering is based on Pearson correlation coefficients.

Simultaneous measurement of telomere length and transcriptome from the same single cell highlighted that some new candidate genes are potentially playing a direct or indirect role in telomere length regulation. DGE heatmap was made based on DGE data from genes of interest (telomere highly correlated genes based on our own findings) together with well-known telomere-related genes. This heatmap visualized the differences between the expression levels of selected genes in different samples. (Figure 17)

Based on this heatmap, LPS+ monocytes expressed telomere-related genes in very high amount compared to other samples (HELA, A549, Mon2-LPS-2, Mon2-LPS-1 and Mon'-LPS-). In addition, cancer cells (HELA and A549) showed higher expression level for telomere-related genes compared to normal cell lines (Mon2-LPS-2, Mon2-LPS-1 and Mon'-LPS-). (Figure 17)

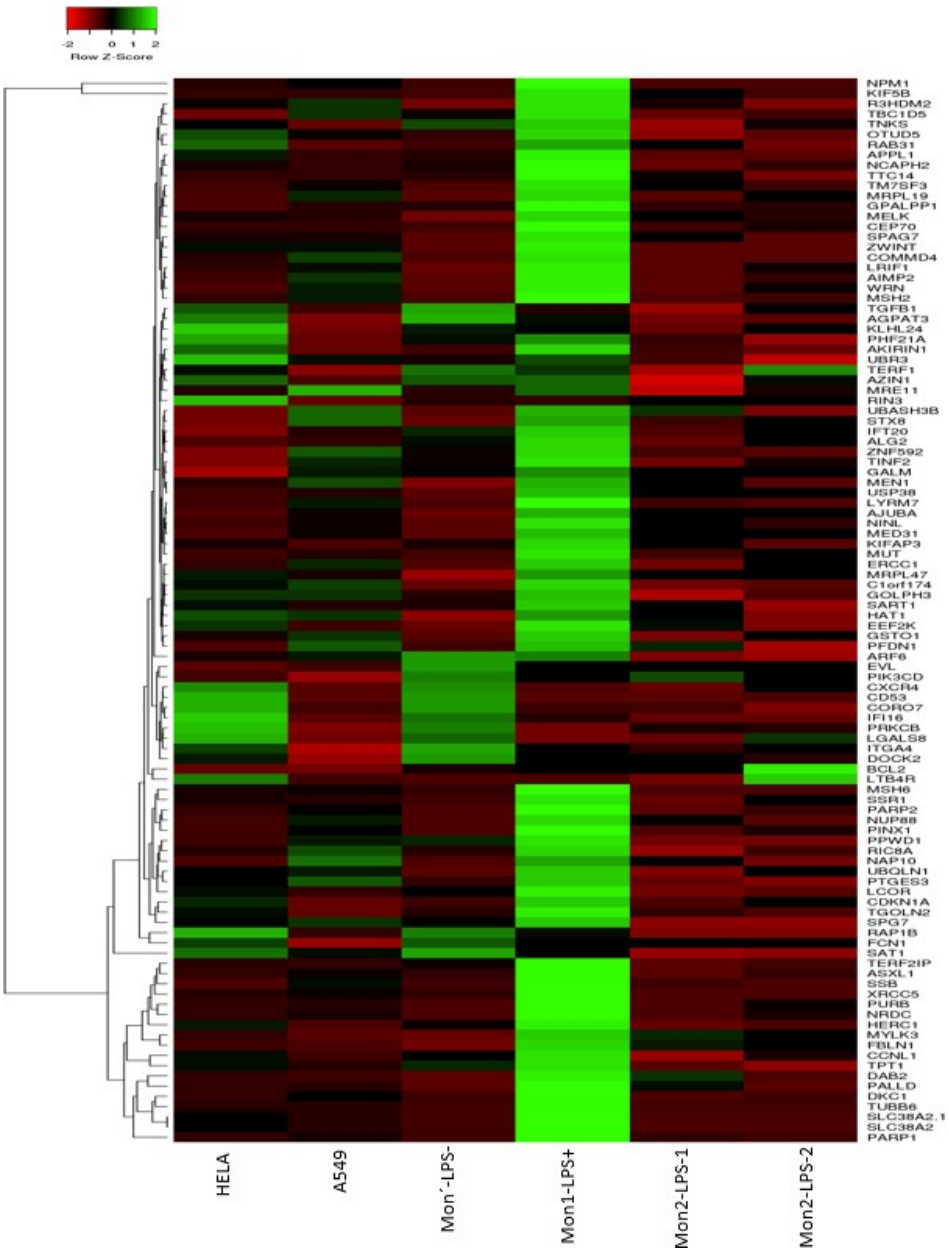


Figure 17. Heatmap shows different telomere related gene expression levels, these include the genes we extracted as well as well-known telomere related genes. Noteworthy that Mon2-LPS-2, Mon2-LPS-1 and Mon'-LPS- are normal monocytes. Green color shows high gene expression level and red shows low gene expression level (Z score).

Moreover, GeneMania (online database) [52] was used in order to find further connections between 7 new extracted genes; *SLC38A2, PURB, UBR3, SSR1, NCAPH2, AIMP2, PHF21A*. GeneMania is a pliable web used for creation of hypotheses about gene function, analyzing gene lists extracted and prioritizing genes for functional assays. [52] According to GeneMania data, these genes were found to be co-expressed in 90.18 percent and interacted in 9.82 percent with each other. (Figure 18)

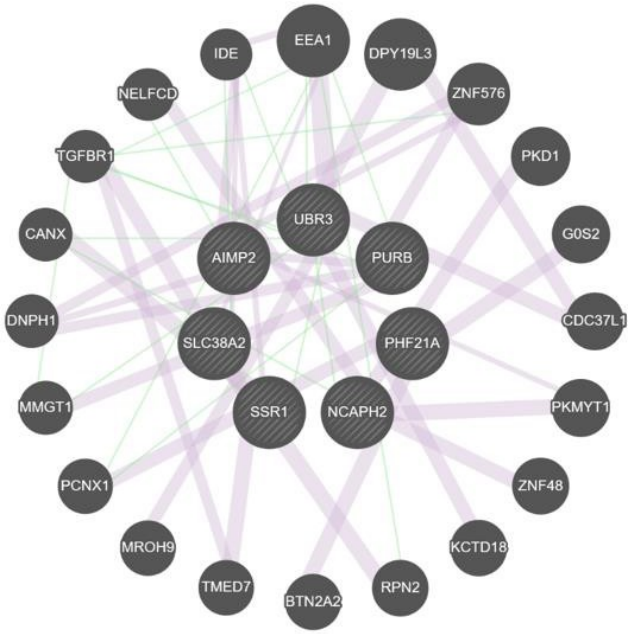


Figure 18. The genes in middle are extracted genes that we used as input in GeeMania (Online database). These genes have been studied to be co-expressed in 90.18 percent of the occasions (violet lines). [53, 54] and interact in 9.82% of the cases (Green lines). [52, 55, 56]

4.9 Pathways Found for Telomere Highly-Correlated Genes

We selected 150 highly (positively or negatively) correlated genes from each sample and run nearly 450 genes using GeneCard analytic tool [57] in order to find highly-correlated pathways for extracted genes. Keeping in mind that cancer cell lines have not correct number of reference genes and therefore the data could not be as reliable as the data from normal cell lines, we did pathway analysis of each sample separately and also all samples together. (Table 3)

Samples were run in three different forms:

1. All samples together including primary cell lines and cancer cell lines (Negatively and positively correlated genes)
2. Only Primary monocytes (Negatively and positively correlated genes)
3. Only HELA (Negatively and positively correlated genes)
4. Only A549 (Negatively and positively correlated genes)

Both positively and negatively correlated genes were included in the data analysis. We found interesting results from all types of cell lines after the importation of telomere highly correlated genes in GeneCard analytic program [57]: telomere length–related genes were remarkably involved in virus related pathways. This reveals the potential important collaboration between telomere length–related genes and virus-related pathways, as they seem to share significant number of genes in their pathways. (Table 3)

Sample	Genes Correlated	Superpathways	Scores (Gene functionality)	Matched Genes (total)	Sources
All Samples (HELA, A549, Primary Monocytes)	Positively Correlated Genes (500 Genes)	Innate Immune System	57.3	121 (2128)	Reactome
		NF-KappaB signaling	27	29 (327)	Cell Signalling Technology
		Interferon gamma signaling	24	21 (202)	Reactome
		Gene expression	20	80 (1839)	Reactome
		measles	20	25 (325)	KEGG
		RRNA processing in the nucleus and cytosol	19.5	19 (203)	Reactome
		Hematopoetic stemm cell gene regulation by GABP alpha/beta complex	19.1	7 (22)	BioSystems
		human cytomegalovirus infection	18.9	26 (360)	KEGG
		DNA damage-ATM/ATR regulation of G1/S checkpoints	18.7	9 (44)	GeneGo (Thomson Reuters)
		Immune response-Role of Integrins in NK cells cytotoxicity	18.7	9 (44)	GeneGo (Thomson Reuters)
		HIV life cycle	18.15	45 (861)	Reactome
		Influenza Viral RNA Transcription and Replication	18.14	16 (158)	Reactome
		RIG-I/MDA5 Mediated Induction of IFN-alpha/betaPathway	18	16 (159)	Reactome/KEGG/GneGo
		Activated TLR4 Signalling	16.6	15 (153)	Reactome
		Human T-Cell Leukemia virus 1 infection	16	18 (219)	KEGG
		Cellular Senescence	15.85	15 (160)	KEGG

		Apoptosis and survival caspase casacade	13.15	10 (90)	Reactome/GeneGo
		cellular response to heat stress	13.27	10 (89)	Reactome
		DNA damage	13.7	20 (296)	Cell Signaling Technology
		apoptosis and autophagy	11.9	14 (183)	Cell Signaling Technology
		DNA damage response (only ATM dependent)	10.39	10 (115)	BioSystems
		ubiquitin-proteasome dependent proteolysis	11.6	11 (123)	Qiagen
	Negatively Correlated Genes (500 Genes)	Metabolism (metabolism of lipids and lipoproteins, Metabolism)	15.49	49 (2577)	Reactome/KEGG
		Dopamin Metabolism	11.09	3 (14)	BioSystems
		Viral mRNA translation	10.53	15 (545)	Ractome/KEGG/ BioSystem/GeneGo
		Cytoskeletal Signaling	9.26	10 (308)	Cell Signaling Technology
		SMAD Signaling Network	9.09	6 (121)	Qiagen
		Protein Processing in endoplasmic Reticulum	9.05	7 (165)	KEGG
		Platelet Activation	8.91	6 (124)	KEGG
		Glucose metabolism	8.74	6 (127)	KEGG/Reactome/ BioSystems/GeneGo
		Innate Immune System	8.6	36 (2128)	Reactome
HELA	Positively Correlated Genes (150 Genes)	Innate Immune System	97.76	110 (2128)	Reactome
		Human cytomegalovirus infection	34.44	27 (360)	KEGG
		measles	32.76	25 (325)	KEGG
		apoptosis and autophagy	21.55	15 (183)	BioSystems
		Cellular senescence	18.81	13 (160)	KEGG
		HIV life cycle	11.82	28 (861)	Reactome
		senescence and autophagy in cancer cells	11.6	8 (106)	GeneGo
		Salmonella infection	10.4	4 (27)	Qiagen
	Negatively Correlated Genes	Metabolism Pathway	15.59	64 (2557)	Reactome/KEGG
		Protein Processing in Endoplasmic Reticulum	12.28	10 (165)	KEGG
		Glucose Metabolism	10.54	8 (127)	Reactome/KEGG/

	(150 Genes)				GeneGO / BioSystem
		DNA Damage-NHEJ Mechanism of DSB repair	7.48	3 (24)	GeneGO
		Viral mRNa translation	7.05	16 (545)	Reactome/KEGG/ GeneGO/BioSystem
		Cellular senescence	7	14 (452)	Reactome
		TP53 Regulates Metabolic Genes	6.81	5 (85)	Reactome
A549	Negatively Correlated Genes (150 Genes)	Prolactin Signaling pathway	18.54	12 (159)	KEGG/GeneGo/ BioSystem
		Delta508-CFTR Traffic/ER to Golgi in CF	18.19	11 (135)	GeneGo
		transcription-p53 signaling pathway	10.03	5 (54)	GeneGo
		Cell cycle/ mitotic	15.68	23 (622)	Reactome
Primary Monocytes	Positively Correlated Genes (150 Genes	Cell cycle, Mitotic	16.5	19 (622)	Reactome
		cellular response to heat stress	15.05	7 (89)	Reactome
		neurodegenerative disease	13.6	4 (23)	Reactome
		HIV life cycle	11.8	20 (861)	Reactome
		tRNA aminoacylation	10.25	5 (74)	Reactome/KEGG
		Regulation of telomerase	7.96	4 (66)	BioSystems
	Negatively Correlated Genes (150 Genes)	Metabolism (metabolism of lipids and lipoproteins, Metabolism)	14.42	26 (2577)	Reactome/KEGG
		Viral mRNA Translation	12.53	4 (121)	Reactome/KEGG/ BioSystems/GeneGo
		Innate immune system	9.91	20 (2128)	Reactome
		Defensis	9.72	4 (102)	Reactome
		SMAD signaling network	8.82	4 (121)	Qiagen
		Metabolism of Proteins	8.80	16 (1627)	Reactome
		Osteoclast differentioation	8.53	4 (128)	KEGG
		Tuberculosis	6.85	4 (179)	KEGG
		Platelet activation	8.7	4 (124)	KEGG
		Metformin Pathway, pharmacodynamics	6.44	2 (36)	PharmGKB
		Chaperonin-mediated protein folding	6.40	3 (105)	Reactome

Table 3. This table was made based on 450 highly (nearly 150 from each gene) correlated genes extracted from 28000 genes (Genes from all samples). We used GeneCard analytic tool for finding super-pathways enriched with these genes and made a table from top score pathways. Scores refer to gene functionality. Functionality status is correlated with the amount of research implemented on a particular gene and its products. [57]

5. Discussion

In this study, we first measured telomere lengths in single cell level, parallel with its transcriptomes using sc T&R seq technique [6] after several trials and protocol optimizations, we aimed to discover telomere length highly correlated genes. Noteworthy, proper validation of this technique was not possible in the scope of this work and can be done in the future study. The study of both telomeres and transcriptomes in single cell level gave us valuable results that can not be achieved in bulk level. For instance, we reported the heterogeneity of telomere lengths and telomere-highly correlated genes in different cancer cell lines, LPS-stimulated and LPS-non-stimulated monocytes and macrophages.

Second, we measured telomere lengths and linked to RNAseq profiles in order to find the correlation between these values. We thereafter chose the most correlated genes with telomere and sorted genes based on their correlation from topest to lowest. Finally, we selected 150 genes most positively and most negatively correlated for further analysis. Among hundreds of selected genes (most positively and most negatively correlated from different samples), *PURB, UBR3, SSR1, NCAPH2, AIMP2, SLC38A2 and PHF21A* were found to be mutual among all samples.

Third, we investigated related pathways for the list of hundreds of genes extracted from the data (most positively and most negatively correlated for different samples) and compared them between different samples. We found strong association between telomere-related extracted genes and genes involved both in virus cycles and immune system.

Fourth, we found that telomere-related genes including shelterin complex-related genes were highly expressed in LPS+ monocytes and cancer cell lines compared to normal cell lines. In addition, the expression of telomere related genes in LPS-Stimulated monocytes were significantly high compared to cancer cell lines.

Finally, we also found that LPS-stimulation of monocytes accelerated the cell senescence and cell apoptosis. However, we were not successful to get the DGE

data and Cq data for the same LPS-stimulated monocytes sample, and we therefore could not calculate correlation values or run pathway analysis.

Despite several advantages that this technique provided, it is, however, not a reliable technique for telomere length measurement in cancer cell lines. This technique is not reliable for studying telomere length of cancer cells as their reference single-copy gene (Alu used in this experiment) may have been duplicated or lost because of aneuploidy. Therefore, suitability of qPCR is restricted to samples that are normal diploid and karyotypically stable.

Moreover, single cell studies are able to ignore some genes. For instance, using this tehchique, we were not able to detect low-expressed genes, such as TERC and TERT, eventhough these genes have been shown to be crucial in telomere length elongation. [11, 14] Noteworthy, this experiment was conducted on low number of samples (Sample plates). However, more plates were required for supporting our (at least 3 plates for each sample) results.

5.1 Cell stimulation and cell viability

As mentioned in the Results, LPS-stimulated monocytes and LPS-stimulated macrophages showed lower viability percentage. (Table 1) In addition, cancer cell lines and LPS-stimulated monocytes showed shorter telomere lengths compared to the non-stimulated monocytes. (Figure 15) The reason why LPS-stimulated monocytes showed shorter telomere length (compared to normal monocytes but not cancer cell lines) and high heterogeneity (compare to normal cell lines and not cancer cell lines), requires further studies. In other words, LPS-stimulated monocytes showed behaviours similar to cancer cell lines rather than norml cell lines. (Figure 17, 14,13) We hypothesise that cells stimulation by LPS (representative of pathogens) could cause cell senescence and therefore cell apoptosis through their influence on telomere.

5.2 Shelterin Complex-Related Genes

Studying some telomere related genes are difficult in single cell level as they are expressed in low amount such as telomerase (*TERT*). [11] However, most (not all) shelterin related genes such as *NOP10, DKC1, RIC8A, PARP2, HAT1, RAP1B* were

highly expressed in cancer cell lines (HELA and A549) and LPS-stimulated monocytes (Mon1-LPS+) than LPS-non-stimulated monocytes (Mon1-LPS-, Mon´-LPS-). (Figure 13) We hypothesise that over-expression of telomere-related genes (particularly shelterin-related genes) in cancer cell lines could be another explanation for telomere length stabilization beside the high expression level of telomerase enzyme in cancer cells.

The mechanisms why cancer cells survive even though they have short telomeres still is under study, but studies show that both the telomerase high expression level or alternative length replacement (ALT) are the main reasons. [2,4] Therefore, cancer cells have this potential to divide frequently without encountering senescence or apoptosis despite their short telomere. In normal primary cells, telomeres shrink with each mitosis, triggering DNA damage response and subsequently senescence. [11]
In addition to shelterin related genes, other telomere highly correlated genes were found to have significantly higher expression levels in LPS-stimulated monocytes (Mon2-LPS+) compared to the non-stimulated-monocytes and compared to the cancer cell lines (HELA, A549). (Figure 17)

5.3 DGE Heterogeneity among Different Cell Lines

Cancer cell lines (A549, HELA) and LPS-stimulated monocytes (Mon2-LPS+) showed higher heterogeneity in both telomere length and different genes expression levels (DGEs) compared to the non-stimulated monocytes (Mon1-LPS-, Mon'-LPS-). (Figures 11 and 14) However, the probable reason for the heterogeneity of telomere length in cancer cells could be associated with Alu (Housekeeping gene used as normalization factor to correct for differences in DNA extraction and PCR efficiency between cells) irregular expression levels which may affect the reliability of the results in cancer cell lines. Therefore, we cannot assure that the heterogeneity of telomeres in cancer cell lines measured by this technique is related to heterogeneity of telomeres.

5.4 Pathways Related to Extracted Genes

We found strong connections between hundreds of telomere-length highly correlated genes (extracted 300 genes from each sample) and both genes involved in immune system pathways and various virus cycles. The latter include several virus cycles, such as measles pathway, human cytomegalovirus pathway, HIV life cycle pathway, Influenza Viral RNA Transcription and Replication Pathway, Human T-cell Leukemia Virus 1 infection Pathway and Viral mRNA translation pathway, regardless of sample type. (Table 3)

These findings suggest that viruses can stimulate the same genes that are important in telomere maintenance mechanism or they compete for the same genes' products needed for telomere protection and therefore affects telomere formation or stability. [59]

We ran extracted genes from different samples both separately and together in order to discover related pathways. Genecard analytic tool was used for exploring related pathways. [57] Some of related pathways for these genes are discussed below (please seeTable 3 as a main source for extracted pathways).

- Pathways related to innate immune response and virus cycles appeared in all cell lines regardless of sample type, indicating strong but not well-known connections between viruses and innate immune system with telomere. (Table 3) This connection has been evaluated from different aspects; some studies seek the similarities between telomere/telomerase function with viruses as they share many common features, such as self-replication and genome stability. Therefore, the very recent evolutionary-related hypothesis suggests that telomeres may have derived from viruses. [58] Since most of the extracted genes in our study have been found to participate in viral related pathways, it raises the question if telomere highly correlated genes could collaborate significantly in different virus-related pathways. Moreover, recent publications revealed the importance of telomere length in regulating sub-telomeric genes which are full of innate immune system-related genes and therefore, people with

different telomere length could react differently to infections caused by viruses. [59] Some other studies revealed the importance of pathogens including viruses in aging. For example, T lymphocytes aging can result from viral infection, oxidized lipids and physiological stresses which could precipitate this process. Understanding the telomerase related functions and its importance in telomere stability to delay the aging process of T cells is of great importance. According to statistics, infection is among top 5 mortality reasons for people above the age of 65 which shows its strong connection with aging. [60,61] Therfore, whether viruses (pathogens) could be considered one of the most important factors in telomere length regulation or not, requires significant studies.

- We found pathways related to telomerase. (Table 3)
- We found pathways related to senescence, autophagy and apoptosis. (Table 3)
- Finally, we found pathways related to metformin (pharmacodynamics) and glucose metabolism. (Table 3) Metformin has been considered as an anti-aging drug candidate which is currently being used for treatment of diabetes type II. [62]

Worth to mention that recently one of major research grants was devoted to a group at Karolinska Institute, Sweden, which is working on the influence of the malaria on accelerating aging process. They have earlier shown that malaria infection could reduce the life span by half in birds - through their interaction with telomere. Their research focus is to discover infections (Malaria, influenza and bacterial pathogens) on the process of aging by sorting different cell populations. They also aim to explore, by combining experimental, epidemiological and cellular studies, impact of infectious diseases – such as HIV, tuberculosis, influenza and bacterial infections – on aging hallmarks in different immune compartments as well as in different body tissues. This project has this potential to push the boundaries for aging and infection studies. [63]

Recent studies reported that significant proportion of CD8 T-cells in HIV-patients encountered replicative senescence with some genetic alterations. These patients had significant reduction in telomere length. Further experiments also showed that

the increased antiviral action by inducing the expression of telomerase gene leads to high proliferation without considerable telomere length fluctuation in CD8 T-cells. [64]

5.5 Highly Correlated Genes with Telomere and Mutual between All Samples

After comparison of extracted highly correlated genes from different samples, *PURB, UBR3, SSR1, NCAPH2, AIMP2, SLC38A2* and *PHF21A* were found to be mutual between all samples. This shows the potential connections between these genes and their role in some unknown telomere-related pathways. (Figure 16) Among these genes, we discussed NCAPH2 and UBR3 as follow:

5.5.1 NCAPH2 (Non-SMC Condensin II Complex Subunit H2)

Condensin II complexes include five subunits (2SMC4, H2, D3 and G2) which are important in chromatid condensation and chromosome protection during cell meiosis and mitosis. It has been reported that in Arabidopsis thaliana plant, condensin II could perform as DNA damage suppressor and therefore increasing plant resistance against possible stresses. [66]

Although condensin II complex has been known for its importance in chromosome compaction (across all chromosome) through their interphase activity, it remained unknown if they have any activities that target some part of chromosomes more specifically. Very recent studies revealed that NCAPH2 which is one of the most crucial proteins in condensin II complex, may have important role in telomere stability. This stabilization can be done through its interaction with TRF1 which is considered to prevent telomere elongation by inhibiting telomerase activity. [67] This same report - using fluorescence in situ hybridization (FISH), immunoblotting and immunoprecipitations techniques - showed that *NCAPH2* deletion increases the possibility of DNA damage response through the activation of *53BP1* and *YH2AX*. They also found that high amount of NCAPH2 is in a direct correlation with telomere TTAGGG repeat. NCAPH2 has therefore been considered as a candidate for telomere maintenance and stability. [68] Our data supports the idea that *NCAPH2* is telomere length-regulating gene through an unknown mechanism as all samples confirmed high correlation of these genes with telomere. (Figure 17)

Moreover, studies done on mice regading *NCAPH2* gene confirms the importance of this gene in Aging/Mortality through its interaction with condensin complexes (Condensin I and Condensin II). These complexes are thought to be crucial in chromosome formation and dissociation and therefore in NSC (Neural Stem Cell) divisions and mice survival.

Based on this report, *NCAPH2*-/- mice were not able to survive and they showed embryonic lethality. Moreover, condensin complexes deletions caused severe defects in chromosome formation and dissociation, triggering DNA damage and subsequently p53-related apoptosis. [69] Regardless of NCAPH2 importance in cell differentiation, system development, no molecular function related to telomeres has been determined till very recently. [50, 65]

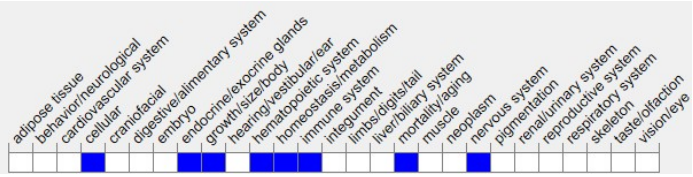


Figure 19. Phenotypic overview of mice after mutations done on *NCAPH2* gene and the correlation of *NCAPH2* with different phenotypes including Mortality/aging. [50]

5.5.2 UBR3 (Ubiquitin Protein Ligase E3 Component n-Recognin 3)

Although *UBR3* gene is considered to be *UBR2* and *UBR1* paralog because of their similarities, it does not recognize N-terminal residues which are important in triggering ubiquitination and short-lived protein degradation (through N-end rule). Mice studies revealed that *UBR2*-/- mice have normal growth and development but close to 60% of them dies right after birth and the remained 40% grow to adulthood without any problems and with normal behaviors. [70]

Studies conducted on human cells shows *UBR2* increases HR6B/UbcH2-dependent ubiquitylation of H2A and H2B, which are two main histones of nucleosome structure. This protein does not interact with H3 and H4 histones. Mechanism of ubiquitination is activated by dipeptides bearing destabilizing N-terminal residues.

Cells lacking *UBR2* show inadequate ubiquitination on meiotic chromosomes and therefore cause deficiency in double strand break repair and other meiotic processes, resulting in apoptosis. UBR2-/- somatic cells reveal parade of chromosomal irregularity such as hyper-reproduction, chromosome inconsistency, and hypersensitivity to DNA damage-inducing reagents. [71] Mice studies, on the other hand, revealed the importance of this gene in mortality/ aging. [50]

The only molecular function attributed to UBR3 is its activity as transferase. [72] The biological processes in which UBR3 is involved includes protein metabolic process and response to stimulus. This gene has been found to be highly expressed in nervous system. [70] Drosophila-related studies determined CG42593 encoding the homolog of mammalian *UBR3* plays crucial role in the regulation of apoptosis pathway through its UBR-box domain in interaction with neo-epitope of DIAP1 (Drosophila inhibitor of apoptosis protein 1) which is released as a result of caspase-mediated dispensation. This interaction lets DIAP1 protein to recruit and ubiquitinise to the underlying caspases. [73]

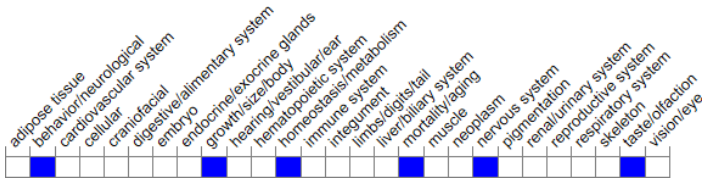


Figure 20. Phenotypic overview of mice after mutations done on *UBR3* gene. It shows the correlation of *UBR3* with different phenotypes including Mortality/aging. [50]

Ubiquitin ligase UBR3 regulates cellular levels of the essential DNA repair protein APE1 and is required for genome stability. [74] To our knowledge there is no evidence supporting UBR3 direct and specific function on telomere or telomerase regulation; however, *UBR1, UBR2* and recently *UBR4* and *UBR5* were found to bind certain destabilizing n-terminal residues.

Based on the importance of paralogs of *UBR3 (UBR1, UBR2)* in chromosome inconsistency and chromosome heterochromatinization, this gene can be another candidate for further telomere studies. Since UBR3 is among genes which highly correlated with telomere length and based on the role of its close family members

such as UBR2, UBR1, UBR4 and UBR5 in ubiquitination of chromatin histones and in nucleosome structure, it might be relevant to telomere stability in a direct or indirect way or participating in some still unknown pathways. We think that this gene probably has direct or indirect interaction with telomere or telomere related elements which could be with their interaction with other UBR members (UBR2 and UBR1).

Conclusion

In recent years, single cell studies have been significant. This method of study creates a path to discover differences between cells and their heterogeneity in one cell population. Studying telomere lengths parallel with different genes expression level in the same cell paves the way toward finding better treatments for cancer, age-related diseases and aging process. We measured telomere length parallel with each gene expression level of the same single cell. This technique revealed the list of genes which were in high correlation with telomere length, giving us this opportunity to get biological pathways which highly affected by telomere-correlated genes. Pathways between different samples had been compared and strong connections has been found between genes involved in viral cycles and those involved in immune system with our found highly correlated genes with telomere. There is some hypothesis that infections with viruses and bacteria can erode the immune cells and increase aging risks. The longer we live, the more we interact with environmental pathogens and we go through aging.

This experiment also provided us with the heterogeneity of telomere lengths and transcriptomes in different cell lines between each single cell. We also reported that the telomere related proteins, specifically those involved in shelterin complex are expressed highly in cancer cell lines and LPS-stimulated monocytes compared to the non-stimulated monocytes. Moreover, our results introduced 7 new potential genes which can be considered as new biomarkers and novel candidates for telomere-related studies. Further experiments are required to evaluate these genes if they involved in telomere length regulation or which pathways they may contribute in. The importance of all these genes have been reported in aging based on studies done on mice. The important correlation between some of these genes such as *NCAPH2* with telomere have been studied very recently. For instance, it is reported that NCAPH2 has an important role in regulating telomere stability and maintenance through its interaction with TERF. [58] Our results suggest that each of these genes have the potential to affect telomere regulation in a direct or indirect way. Since this technique is implemented on small number of plates and single cells, further experiments are required to support the results.

This experiment is of high importance in studying cancer, neurodegenerative diseases and aging, as it gives the opportunity to discover more genes in correlation with telomere regulation. It requires more future studies to discover the roles of these genes in aging through their interaction with telomere based on the fact that telomeres are as the aging clock for body.

7. Acknowledgments

This graduate thesis was carried out at the Faculty of Biological and Environmental Sciences, University of Helsinki.

I sincerely thank my supervisors Adjunct Professor Päivi Saavalainen, Dr. Heli Pessa, Dr. Katarzyna Leskinen, Dr. Eija Nissilä and Dr. Juha Grönholm for their insightful guidance and support. I would like to thank all co-workers. I was made very welcome and enjoyed working in the lab with you. I would also like to thank Yusef Pourjamal for his generous helps with data analysis whenever I could not move things forward. I would also like thank Adjunct Professor Kari Tikkinen with all my heart for his kind help.

I am thankful to the reviewers of this thesis Professor Juha Partanen and Adjunct Professor Kaia Achim. Your professional and thorough evaluation helped me to improve this thesis substantially.

I would like to thank Institute for Molecular Medicine Finland (FIMM) and FUGU for their excellent services in sequencing. I would like to thank Biomedicum FACS Core Facilities for providing cell sorting. I would also like to thank CSC-IT center for access to Chipster program and other computational resources.

I would like to thank my friends for always being there for me – it means a great deal for me to have you – thank you for all lovely moments! Finally, I would like to express my deep and sincere gratitude to my dear family for their kind and endless support all my life.

8. References

1. Khattar, E., & Tergaonkar, V. (2016). The role of telomeres and telomere-associated proteins as components of interactome in cell-signaling pathways. *TELOMERE*, 45.
2. Diotti, R., & Loayza, D. (2011). Shelterin complex and associated factors at human telomeres. *Nucleus*, *2*(2), 119-135.
3. Cimino-Reale, G., Pascale, E., Battiloro, E., Starace, G., Verna, R., & D'Ambrosio, E. (2001). The length of telomeric G-rich strand 3′-overhang measured by oligonucleotide ligation assay. *Nucleic acids research*, *29*(7), e35-e35.
4. Lin, C., & Yang, D. (2017). Human telomeric G-quadruplex structures and G-quadruplex-interactive compounds. In *Telomeres and Telomerase* (pp. 171-196). Humana Press, New York, NY.
5. Okamoto, K., & Seimiya, H. (2019). Revisiting telomere shortening in cancer. *Cells*, *8*(2), 107.
6. Wang, H., Zhang, K., Liu, Y., Fu, Y., Gao, S., Gong, P. & Sun, Y. (2017). Telomere heterogeneity linked to metabolism and pluripotency state revealed by simultaneous analysis of telomere length and RNA-seq in the same human embryonic stem cell. *BMC biology*, *15*(1), 114.
7. https://www.khanacademy.org/science/biology/dna-as-the-genetic- material/dna-replication/a/telomeres-telomerase
8. Gerbi, S. A. (2015). Beginning at the end: DNA replication within the telomere. *J Cell Biol*, 210(2), 177-179.
9. Maestroni, L., Matmati, S., & Coulon, S. (2017). Solving the telomere replication problem. *Genes*, 8(2), 55.
10. O'sullivan, R. J., & Karlseder, J. (2010). Telomeres: protecting chromosomes against genome instability. *Nature reviews Molecular cell biology*, 11(3), 171.
11. Jafri, M. A., Ansari, S. A., Alqahtani, M. H., & Shay, J. W. (2016). Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome medicine*, *8*(1), 69.
12. Dey, A., & Chakrabarti, K. (2018). Current perspectives of telomerase structure and function in eukaryotes with emerging views on telomerase in human parasites. *International journal of molecular sciences*, 19(2), 333.
13. Venteicher, A. S., Abreu, E. B., Meng, Z., McCann, K. E., Terns, R. M., Veenstra, T. D., ... & Artandi, S. E. (2009). A human telomerase holoenzyme protein required for Cajal body localization and telomere synthesis. *Science*, 323(5914), 644-648.
14. Tomlinson, R. L., Ziegler, T. D., Supakorndej, T., Terns, R. M., & Terns, M. P. (2006). Cell cycle-regulated trafficking of human telomerase to telomeres. *Molecular biology of the cell*, 17(2), 955-9.
15. Cohn, W. E. (1959). 5-Ribosyl uracil, a carbon-carbon ribofuranosyl nucleoside in ribonucleic acids. *Biochimica et biophysica acta*, 32, 569-571.
16. Baker, D. L., Youssef, O. A., Chastkofsky, M. I., Dy, D. A., Terns, R. M., & Terns, M. P. (2005). RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. *Genes & development*, 19(10), 1238-1248.
17. De Zoysa, M. D., & Yu, Y. T. (2018). RNA-dependent pseudouridylation catalyzed by box H/ACA RNPs. *Frontiers in Biology*, 13(1), 1-10.
18. Yu, Y. T., & Meier, U. T. (2014). RNA-guided isomerization of uridine to pseudouridine—pseudouridylation. *RNA biology*, 11(12), 1483-1494.
19. Henras, A., Dez, C., Noaillac-Depeyre, J., Henry, Y., & Caizergues-Ferrer, M. (2001). Accumulation of H/ACA snoRNPs depends on the integrity of the conserved central domain of the RNA-binding protein Nhp2p. *Nucleic acids research*, 29(13), 2733-2746.
20. Wang, C., & Meier, U. T. (2004). Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *The EMBOjournal*, 23(8), 1857-1867.
21. Ganot, P., Caizergues-Ferrer, M., & Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Development*, 11(7), 941-956.
22. Gall, J. G., Bellini, M., Wu, Z. A., & Murphy, C. (1999). Assembly of the nuclear transcription and processing machinery: Cajal bodies (coiled bodies) and transcriptosomes. *Molecular Biology of the Cell*, 10(12), 4385-4402.

23. Kim, S. H., Kaminker, P., & Campisi, J. (1999). TIN2, a new regulator of telomere length in human cells. *Nature genetics*, 23(4), 405.
24. Jeon, B. G., Kumar, B. M., Kang, E. J., Ock, S. A., Lee, S. L., Kwack, D. O., ... & Rho, G. J. (2011). Characterization and comparison of telomere length, telomerase and reverse transcriptase activity and gene expression in human mesenchymal stem cells and cancer cells of various origins. *Cell and tissue research*, 345(1), 149-161.
25. Liu, D., Safari, A., O'Connor, M. S., Chan, D. W., Laegeler, A., Qin, J., & Songyang, Z. (2004). PTOP interacts with POT1 and regulates its localization to telomeres. *Nature cell biology*, 6(7), 673.
26. Bilaud, T., Brun, C., Ancelin, K., Koering, C. E., Laroche, T., & Gilson, E. (1997). Telomeric localization of TRF2, a novel human telobox protein. *Nature genetics*, 17(2), 236.
27. Xin, H., Liu, D., & Songyang, Z. (2008). The telosome/shelterin complex and its functions. *Genome biology*, 9(9), 232.
28. Van Steensel, B., & De Lange, T. (1997). Control of telomere length by the human telomeric protein TRF1. *Nature*, 385(6618), 740.
29. Nandakumar, J., Bell, C. F., Weidenfeld, I., Zaug, A. J., Leinwand, L. A., & Cech, T. R. (2012). The TEL patch of telomere protein TPP1 mediates telomerase recruitment and processivity. *Nature*, 492(7428), 285.
30. Baumann, P., & Price, C. (2010). Pot1 and telomere maintenance. *FEBS letters*, 584(17), 3779-3784.
31. Baumann, P., & Cech, T. R. (2001). Pot1, the putative telomere end-binding protein in fission yeast and humans. *Science*, 292(5519), 1171-1175.
32. Bunch, J. T., Bae, N. S., Leonardi, J., & Baumann, P. (2005). Distinct requirements for Pot1 in limiting telomere length and maintaining chromosome stability. *Molecular and cellular biology*, 25(13), 5567-5578.
33. Hockemeyer, D., Daniels, J. P., Takai, H., & de Lange, T. (2006). Recent expansion of the telomeric complex in rodents: Two distinct POT1 proteins protect mouse telomeres. *Cell*, 126(1), 63-77.
34. Shakirov, E. V., Song, X., Joseph, J. A., & Shippen, D. E. (2009). POT1 proteins in green algae and land plants: DNA-binding properties and evidence of co-evolution with telomeric DNA. *Nucleic acids research*, 37(22), 7455-7467.
35. Hockemeyer, D., Sfeir, A. J., Shay, J. W., Wright, W. E., & De Lange, T. (2005). POT1 protects telomeres from a transient DNA damage response and determines how human chromosomes end. *The EMBO journal*, 24(14), 2667- 2678.
36. Giraud-Panis, M. J., Ye, J., & Gilson, E. (2018). TRFH domain: at the root of telomere protein evolution. *Cell Research*, 28(1), 7-8.
37. Longhese, M. P. (2008). DNA damage response at functional and dysfunctional telomeres. *Genes & Development*, 22(2), 125-140.
38. Chen, J. L., & Greider, C. W. (2003). Determinants in mammalian telomerase RNA that mediate enzyme processivity and cross-species incompatibility. *The EMBO Journal*, 22(2), 304-314.
39. Houghtaling, B. R., Cuttonaro, L., Chang, W., & Smith, S. (2004). A dynamic molecular link between the telomere length regulator TRF1 and the chromosome end protector TRF2. *Current Biology*, 14(18), 1621-1631.
40. Montpetit, A. J., Alhareeri, A. A., Montpetit, M., Starkweather, A. R., Elmore, L. W., Filler, K., & Collins, J. B. (2014). Telomere length: a review of methods for measurement. *Nursing Research*, 63(4), 289.
41. Wang, F., Pan, X., Kalmbach, K., Seth-Smith, M. L., Ye, X., Antumes, D. M., ... & Weissman, S. M. (2013). Robust measurement of telomere length in single cells. *Proceedings of the National Academy of Sciences*, 110(21), E1906-E1912.
42. Cawthon, R.M. Telomere measurement by quantitative PCR, *Nucleic Acids Research*, Volume 30, Issue 10, 15 May 2002, Page e47.
43. Aubert, G., Hills, M., & Lansdorp, P. M. (2012). Telomere length measurement—Caveats and a critical assessment of the available technologies and tools. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 730(1-2), 59-67.
44. Lai, T. P., Wright, W. E., & Shay, J. W. (2018). Comparison of telomere length measurement methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1741), 20160451.

45. Perkinelmer [website] available at http://www.perkinelmer.co.jp/Portals/0/resource/products_ls/dnarna/pdf/GX_UserGuide/DNAHiSense_UserGuide_CLS140158_Rev.E.pdf (accessed June 6th 2019)
46. Feng, Y. J., Liu, Q. F., Chen, M. Y., Liang, D., & Zhang, P. (2016). Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. *Molecular Ecology Resources*, *16*(1), 91-102.
47. Kallio, M.A. et al. (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics 12*, 507.
48. http://mccarrolllab.org/dropseq/; Macosko, E.Z. et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell 161*, 1202–1214).
49. Dobin, A. et al: STAR: ultrafast universal RNA-seq aligner (2013). *Bioinformatics* 29: 15-21.
50. Informatics [website]. Available at http://www.informatics.jax.org (accessed October 22, 2019)
51. Heatmapper [website]. Available at http://www.heatmapper.ca/expression/ (accessed October 3,2019)
52. GeneMANIA [website]. Available at https://genemania.org/ (accessed December 4, 2019).
53. Smirnov, D. A., Morley, M., Shin, E., Spielman, R. S., & Cheung, V. G. (2009). Genetic analysis of radiation-induced changes in human gene expression. *Nature*, *459*(7246), 587.
54. Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., ... & Van De Rijn, M. (2002). Gene expression patterns in human liver cancers. *Molecular Biology of the Cell*, *13*(6), 1929-1939.
55. Lin, A., Wang, R. T., Ahn, S., Park, C. C., & Smith, D. J. (2010). A genome- wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Research*, *20*(8), 1122-1132.
56. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(S1), S4.
57. GeneCards: The Human Gene Database [website]. Available at https://www.genecards.org (accessed December 4, 2019).
58. de Lange, T. (2015). A loopy view of telomere evolution. *Frontiers in Genetics*, *6*, 321.
59. Wang, Z., Deng, Z., Tutton, S., & Lieberman, P. M. (2017). The Telomeric Response to Viral Infection. *Viruses*, *9*(8), 218.
60. Kim, S. H., Kaminker, P., & Campisi, J. (1999). TIN2, a new regulator of telomere length in human cells. *Nature Genetics*, 23(4), 405.
61. Tedone, E., Huang, E., O'Hara, R., Batten, K., Ludlow, A. T., Lai, T. P., ... & Shay, J. W. (2019). Telomere length and telomerase activity in T cells are biomarkers of high-performing centenarians. *Aging Cell*, *18*(1), e12859.
62. Soukas, A. A., Hao, H., & Wu, L. (2019). Metformin as anti-aging therapy: is it for everyone. *Trends in Endocrinology & Metabolism*.
63. Asghar Muhammad's Research group [website]. Available at https://ki.se/en/meds/research-group-asghar-muhammad (accessed December 2, 2019).
64. Dagarag, M., Evazyan, T., Rao, N., & Effros, R. B. (2004). Genetic manipulation of telomerase in HIV-specific CD8+ T cells: enhanced antiviral functions accompany the increased proliferative potential and telomere length stabilization. *The Journal of Immunology*, *173*(10), 6303-6311.
65. Wallace, H. A., Rana, V., Nguyen, H. Q., & Bosco, G. (2019). Condensin II subunit NCAPH2 associates with shelterin protein TRF1 and is required for telomere stability. *Journal of Cellular Physiology*.
66. Sakamoto, T., Inui, Y. T., Uraguchi, S., Yoshizumi, T., Matsunaga, S., Mastui, M., ... & Fujiwara, T. (2011). Condensin II alleviates DNA damage and is essential for tolerance of boron overload stress in Arabidopsis. *The Plant Cell*, *23*(9), 3533-3546.
67. Garton, M., & Laughton, C. (2013). A comprehensive model for the recognition of human telomeres by TRF1. *Journal of Molecular Biology*, *425*(16), 2910- 2921.
68. Wallace, H. A., Rana, V., Nguyen, H. Q., & Bosco, G. (2019). Condensin II subunit NCAPH2 associates with shelterin protein TRF1 and is required for telomere stability. *Journal of Cellular Physiology*.
69. Nishide, K., & Hirano, T. (2014). Overlapping and non-overlapping functions of condensins I and II in neural stem cell divisions. *PLoS Genetics*, *10*(12), e1004847.

70. Tasaki, T., Sohr, R., Xia, Z., Hellweg, R., Hörtnagl, H., Varshavsky, A., & Kwon, Y. T. (2007). Biochemical and genetic studies of UBR3, a ubiquitin ligase with a function in olfactory and other sensory systems. *Journal of Biological Chemistry*, *282*(25), 18510-18520.
71. An, J. Y., Kim, E., Zakrzewska, A., Yoo, Y. D., Jang, J. M., Han, D. H., ... & De Rooij, D. G. (2012). UBR2 of the N-end rule pathway is required for chromosome stability via histone ubiquitylation in spermatocytes and somatic cells. *PloS One*, *7*(5), e37414.
72. Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, *12*(5), 449-462.
73. Huang, Q., Tang, X., Wang, G., Fan, Y., Ray, L., Bergmann, A., ... & Ye, X. (2014). Ubr3 E3 ligase regulates apoptosis by controlling the activity of DIAP1 in Drosophila. *Cell Death and Differentiation*, *21*(12), 1961.
74. Meisenberg, C., Tait, P. S., Dianova, I. I., Wright, K., Edelmann, M. J., Ternette, N., ... & Dianov, G. L. (2011). Ubiquitin ligase UBR3 regulates cellular levels of the essential DNA repair protein APE1 and is required for genome stability. *Nucleic Acids Research*, *40*(2), 701-711.

9. SUPPLEMENTARY MATERIAL

Table 9.4.

Primers	Sequences
Telomere-Forward	CGGTTTGTTTGGGTTTGGGTTTGGGTTTGGGTTTGGGTT
Telomere-Reverse	GGCTTGCCTTACCCTTACCCTTACCCTTACCCTTACCCT
Alu-Forward	GACCATCCCGGCTAAAACG
Alu-Reverse	CGGGTTCACGCCATTCTC
SMART PCR	AAG CAG TGG TAT CAA CGC AGA GT
TSO (Template Switch Oligo)	AAG CAG TGG TAT CAA CGC AGA GTG AAT rGrGrG

Table 9.5.

Barcodes	Well position A-D	Well position E-H	Primer Diffrences	Sequences
DS-BiotinBulkindex16	A01	E01	TAGCGAGTAATT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TAGCGAGT AATT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex18	A02	E02	GTAGCTCCTTAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GTAGCTCC TTAA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex19	A03	E03	TACTACGCCCGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TACTACGC CCGG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex20	A04	E04	AGGCTCCGGGCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AGGCTCCG GGCC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex21	A05	E05	GCAGCGTAAACC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GCAGCGTA AACC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex22	A06	E06	CTGCGCATTTCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CTGCGCAT TTCC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex23	A07	E07	GAGCGCTACCTT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GAGCGCTA CCTT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex24	A08	E08	CGCTCAGTGGAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CGCTCAGT GGAA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex26	A09	E09	GTCTTAGGAAGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GTCTTAGG AAGG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex27	A10	E10	ACTGATCGTTGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC ACTGATCG TTGG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex28	A11	E11	TAGCTGCACCAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TAGCTGCA CCAA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T

DS-BiotinBulkindex29	A12	E12	GACGTCGAGGTT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GACGTCGA GGTT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex16_B	B01	F01	AATTTAGCGAGT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AATT TAGCGAGT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex18_B	B02	F02	TTAAGTAGCTCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TTAA GTAGCTCC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex20_B	B03	F03	CCGGTACTACGC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CCGG TACTACGC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex20_B	B04	F04	GGCCAGGCTCCG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GGCC AGGCTCCG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex21_B	B05	F05	AACCGCAGCGTA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AACC GCAGCGTA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex22_B	B06	F06	TTCCCTGCGCAT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TTCC CTGCGCAT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex23_B	B07	F07	CCTTGAGCGCTA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CCTT GAGCGCTA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex24_B	B08	F08	GGAACGCTCAGT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GGAA CGCTCAGT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex26_B	B09	F09	AAGGGTCTTAGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AAGG GTCTTAGG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex27_B	B10	F10	TTGGACTGATCG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TTGG ACTGATCG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex28_B	B11	F11	CCAATAGCTGCA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CCAA TAGCTGCA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex29_B	B12	F12	GGTTGACGTCGA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GGTT GACGTCGA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex16_C	C01	G01	TAGCAATTGAGT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TAGC AATT GAGTN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex18_C	C02	G02	GTAGTTAACTCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GTAG TTAA CTCCN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex19_C	C03	G03	TACTCCGGACGC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TACT CCGG ACGCN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex20_C	C04	G04	AGGCGGCCTCCG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AGGC GGCC TCCGN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex21_C	C05	G05	GCAGAACCCGTA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GCAG AACC CGTAN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex22_C	C06	G06	CTGCTTCCGCAT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CTGC TTCC GCATN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex23_C	C07	G07	GAGCCCTTGCTA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GAGC CCTT GCTAN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex24_C	C08	G08	CGCTGGAACAGT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CGCT GGAA CAGTN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex26_C	C09	G09	GTCTAAGGTAGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GTCT AAGG TAGGN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex27_C	C10	G10	ACTGTTGGATCG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC ACTG TTGG ATCGN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T

DS-BiotinBulkindex28_C	C11	G11	TAGCCCAATGCA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TAGC CCAA TGCAN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex29_C	C12	G12	GACGGGTTTCGA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GACG GGTT TCGAN NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex16_D	D01	H01	AATAGCGAGTTT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AA TAGCGAGT TT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex18_D	D02	H02	TTGTAGCTCCAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TT GTAGCTCC AA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex19_D	D03	H03	CCTACTACGCGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CC TACTACGC GG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex20_D	D04	H04	GGAGGCTCCGCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GG AGGCTCCG CC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex21_D	D05	H05	AAGCAGCGTACC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AA GCAGCGTA CC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex22_D	D06	H06	TTCTGCGCATCC	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TT CTGCGCAT CC N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex23_D	D07	H07	CCGAGCGCTATT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CC GAGCGCTA TT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex24_D	D08	H08	GGCGCTCAGTAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GG CGCTCAGT AA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex26_D	D09	H09	AAGTCTTAGGGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC AA GTCTTAGG GG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex27_D	D10	H10	TTACTGATCGGG	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC TT ACTGATCG GG N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex28_D	D11	H11	CCTAGCTGCAAA	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC CC TAGCTGCA AA N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T
DS-BiotinBulkindex29_D	D12	H12	GGGACGTCGATT	TTT TTT TAA GCA GTG GTA TCA ACG CAG AGT AC GG GACGTCGA TT N NNN NNN NTT TTT TTT TTT TTT TTT TTT TTT TTT T

9.6 Code 1: Allignment of Cq values with DGE data:

```
global well_position
class nomeque:
    def impoort(self):
        # a=open('nom1.xlsx', 't')
        a=[

'AACGCTCAGT', 'AAGTAGCTCC', 'AATAGCTGCA', 'ACTCCGGACGC', 'ACTGATCGGG', 'AGAACCCGTA',
'AGCGTAAACC', 'AGCTCCTTAA', 'AGGCTCCGCC', 'AGTTAACTCC', 'CCAGGCTCCG', 'CCCTGCGCAT',
'CCGCAGCGTA', 'CGCTCAGTAA', 'CGGGTTTCGA', 'CGTCGAGGTT', 'CTAAGGTAGG', 'CTACGCCCGG',
'CTCAGTGGAA', 'CTGCGCATCC', 'CTGGAACAGT', 'CTTAGGAAGG', 'GACGTCGATT', 'GAGCGCTATT',
'GCAATTGAGT', 'GCAGCGTACC', 'GCCCAATGCA', 'GCCCTTGCTA', 'GCGAGTAATT', 'GCGCATTTCC',
'GCGCTACCTT', 'GCGGCCTCCG', 'GCTGCACCAA', 'GCTTCCGCAT', 'GGACTGATCG', 'GGCTCCGGGCC',
'GGGTCTTAGG', 'GGTACTACGC', 'GTAGCTCCAA', 'GTCTTAGGGG', 'TACTACGCGG', 'TAGCGAGTTT',
'TAGCTGCAAA', 'TGATCGTTGG', 'TGTTGGATCG', 'TTGACGTCGA', 'TTGAGCGCTA', 'TTTAGCGAGT',]

        b=[
'AGAACCCGTA', 'AGCGTAAACC', 'AGCTCCTTAA', 'AGTTAACTCC', 'CCAGGCTCCG', 'CCCCAGCGTA',
'CCCGGCGCAT', 'CCCTGCGCAG', 'CCCTGCGCAT', 'CCCTGCGCCT', 'CCGCAGCGTA', 'CCGCAGCGTC',
'CCGCTGCGTA', 'CCGTGCGCAT', 'CCTGGCTCCG', 'CGCTCAGTAA', 'CGCTGCGCAT', 'CGGTCAGTAA',
'CGTTAACTCC', 'CTACGCCCGG', 'CTCAGTGGAA', 'CTCCGCATCC', 'CTCCGGACGC', 'CTCTGCGCAT',
'CTCTGTGGAA', 'CTGCGCATCC', 'CTGCGGACGC', 'CTGGAACAGT', 'CTTAGGAAGG', 'CTTCGCCCGG',
'CTTTGGAAGG', 'GCAATTGAGT', 'GCAGGCTCCG', 'GCCCAATGCA', 'GCCCTATGCA', 'GCCCTTGCTA',
'GCCTGCGCAT', 'GCGAGTAATT', 'GCGCAATGCA', 'GCGCAGCGTA', 'GCGCATTTCC', 'GCGCTACCTT',
'GCGCTGCGTA', 'GCGCTTTTCC', 'GCGGCCTCCG', 'GCGTCTTAGG', 'GCGTGCGCAT', 'GCGTGTAATT',
'GCTATTGAGT', 'GCTCCGGGCC', 'GCTGGCTCCG', 'GCTTTTGAGT', 'GGACTGATCG', 'GGCGTAAACC',
'GGCTCAGTAA', 'GGCTCCTTAA', 'GGCTCTTAGG', 'GGCTGCGCAT', 'GGGCTACCTT', 'GGGGCTTAGG',
'GGGTCAGTAA', 'GGGTCTTAGG', 'GGGTGCGCTA', 'GGTACTACGC', 'GGTCTGATCG', 'GGTGGCTCCG',
'GGTTAACTCC', 'GGTTCTACGC', 'GTCAGTGGAA', 'GTCCGGACGC', 'GTCTGTGGAA', 'GTCTTAGGGG',
'GTGAGCGCTA', 'GTGCGCATCC', 'GTGCGCTATT', 'GTGCGGACGC', 'GTGGAACAGT', 'GTGTGCGCTA',
'GTTAGGAAGG', 'GTTCGCCCGG', 'GTTTGGAAGG', 'TCTGATCGGG', 'TGCGTAAACC', 'TGCTCCTTAA',
'TGGTCCTTAA', 'TGGTGCGCTA', 'TGTTAACTCC', 'TGTTCCCGTA', 'TGTTCGTTGG', 'TGTTTACTCC',
'TTCAGCGCTA', 'TTCGCTCAGT', 'TTCTGCGCTA', 'TTGACGTCGA', 'TTGAGCGCTA', 'TTGTCGTCGA',
'TTGTGCGCTA', 'TTGTGCGCTC', 'TTTTGCGAGT', 'TTTTGCTGCA',]
        genes=[
'957', '1189', '3184', '2521', '2664', '1654', '720', '805', '6502', '829', '5278', '1404',
'2918', '2535', '3662', '3857', '617', '1563', '1719', '2419', '2308', '1074', '3796', '629',
'2604', '4240', '1355', '1206', '2449', '3960', '2651', '413', '888', '4205', '2051', '1530',
'4266', '2063', '2001', '4501', '4828', '3646', '1685', '2426', '1849', '796', '1587', '2645',
'491', '1060', '2436', '771', '1753', '935', '2611', '2312', '750', '772', '910', '646', '1094',
'2807', '1417', '1976', '2649', '641', '2476', '2147', '1266', '2875', '1869', '497', '2937',
'3151', '524', '1226', '633', '4853', '1172', '2952', '1901', '1021', '1538', '3688', '1503',
'914', '3356', '1184', '2290', '1457', '1152', '1309', '1026', '1114', '5238', '1565', '6179',
'785', '582', '1000',]
        UMIs=[
'2480', '2494', '10000', '6218', '4986', '3231', '973', '1174', '32044', '1269', '25246',
'2569', '7837', '5533', '8493', '15056', '917', '3378', '3448', '7002', '6292', '1805', '10045',
'899', '7584', '15303', '2133', '2187', '5248', '32420', '5809', '512', '1152', '13814', '4292',
'2892', '12761', '4813', '4288', '17791', '19436', '12121', '3345', '5522', '10712', '1326',
'2794', '6724', '606', '1902', '4289', '1057', '2860', '1695', '7377', '5756', '1219', '1168',
'1545', '993', '2025', '8293', '2122', '3876', '4976', '802',  5985', '4300', '2503', '6122',
'4457', '934', '5849', '8553', '939', '1850', '1045', '13869', '1807', '11769', '3530', '2632',
'3368', '12261', '3108', '1261', '9610', '5861', '6015', '2641', '1750', '2269', '1469', '1673',
'16613', '2597', '21992', '1077', '993', '1815']
        well_position=[
'F08', 'F02', 'F11', 'G03', 'H10', 'G05', 'E05', 'E02', 'H04', 'G02', 'F04', 'F06', 'F05',
'H08', 'G12', 'E12', 'G09', 'E03', 'E08', 'H06', 'G08', 'E09', 'H12', 'H07', 'G01', 'H05',
'G11', 'G07', 'E01', 'E06', 'E07', 'G04', 'E11', 'G06', 'F10', 'E04', 'F09', 'F03', 'H02',
'H09', 'H03', 'H01', 'H11', 'E10', 'G10', 'F12', 'F07', 'F01']
        return (a, b, genes, UMIs, well_position)

def issimilar(a, b, genes, UMIs, well_position):
    geneComparment = open('genesCompareSame.txt' , 'w')
    s=[]
    k=0
    print('>>>>>>>>>>>>>>>   EXACTLY SIMILAR CODES   <<<<<<<<<<<<<')
```

```
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs, GENE_CODE,
well_position')
    print('>>>>>>>>>>>>>>>>   EXACTLY SIMILAR CODES   <<<<<<<<<<<<<<', file= geneComparment, end=
' \n')
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs, GENE_CODE,
well_position', file= geneComparment, end= ' \n')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]==b[j]:
                k += 1
                if k < 10:
                    #d = k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j]
                    #print(d)
                    print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')

                else:
                    # d = k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j]
                    # print(d)
                    print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')
    return s

def OneCharDiff(a, b, genes, UMIs, well_position):
    geneComparment = open('genesCompareOne.txt' , 'w')
    s=[]
    k=0
    print('\n')
    print('>>>>>>>>>>>>>>>>   ONE CODE MISSMATCH   <<<<<<<<<<<<<<')
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs, GENE_CODE,
well_position')
    print('>>>>>>>>>>>>>>>>   ONE CODE MISSMATCH   <<<<<<<<<<<<<<', file= geneComparment, end= '
\n')
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs, GENE_CODE,
well_position', file= geneComparment, end= ' \n')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]!=b[j] and len(a[i])==len(b[j]):
                mem=0
                for m in range(len(a[i])):
                    aa=a[i]
                    bb=b[j]
                    if aa[m]==bb[m]:
                        mem += 1
                if mem == len(a[i])-1:
                    k += 1
                    if k < 10:
                        #print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')
                    else:
                        #print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')
                    continue

    return s

def TwoCharDiff(a, b, genes, UMIs, well_position):
    geneComparment = open('genesCompareTwo.txt' , 'w')
    s=[]
    k=0
    print('\n')
    print('>>>>>>>>>>>>>>>>   TWO CODE MISSMATCH   <<<<<<<<<<<<<<')
    print('Row   BARCODE   GENE_CODE    Barcode_row   GENE_row   genes   UMIs, GENE_CODE,
well_position')
```

```
    print('>>>>>>>>>>>>>>>>   TWO CODE MISSMATCH   <<<<<<<<<<<<<<', file= geneComparment, end = '
\n')
    print('Row   BARCODE   GENE_CODE    Barcode_row   GENE_row   genes   UMIs, GENE_CODE,
well_position', file= geneComparment, end= ' \n')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]!=b[j] and len(a[i])==len(b[j]):
                mem=0
                for m in range(len(a[i])):
                    aa=a[i]
                    bb=b[j]
                    if aa[m]==bb[m]:
                        mem += 1
                if mem == len(a[i])-2:
                    k += 1
                    if k < 10:
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i])
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')
                    else:
                        print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i])
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j],'  ', b[j],'  ', well_position[i], file = geneComparment, end = ' \n')
                    continue

    return s

def main():
    geneComparment0 = open('genesCompareSame.txt' , 'w')
    geneComparment1 = open('genesCompareOne.txt' , 'w')
    geneComparment2 = open('genesCompareTwo.txt' , 'w')
    comp=nomeque()
    # issimilar(comp.impoort())
    f, g, genes, UMIs, well_position=comp.impoort()
    issimilar(f, g, genes, UMIs, well_position)
    OneCharDiff(f, g, genes, UMIs, well_position)
    TwoCharDiff(f, g, genes, UMIs, well_position)

main()
```

9.7 Cod 2: Extraction of complete matches, 1 mismatches and 2 mismatches.

```
class nomeque:
    def impoort(self):
        # a=open('nom1.xlsx', 't')
        a=[

'AACGCTCAGT', 'AAGTAGCTCC', 'AATAGCTGCA', 'ACTCCGGACGC', 'ACTGATCGGG', 'AGAACCCGTA',
'AGCGTAAACC', 'AGCTCCTTAA', 'AGGCTCCGCC', 'AGTTAACTCC', 'CCAGGCTCCG', 'CCCTGCGCAT',
'CCGCAGCGTA', 'CGCTCAGTAA', 'CGGGTTTCGA', 'CGTCGAGGTT', 'CTAAGGTAGG', 'CTACGCCCGG',
'CTCAGTGGAA', 'CTGCGCATCC', 'CTGGAACAGT', 'CTTAGGAAGG', 'GACGTCGATT', 'GAGCGCTATT',
'GCAATTGAGT', 'GCAGCGTACC', 'GCCCAATGCA', 'GCCCTTGCTA', 'GCGAGTAATT', 'GCGCATTTCC',
'GCGCTACCTT', 'GCGGCCTCCG',  'GCTGCACCAA', 'GCTTCCGCAT', 'GGACTGATCG', 'GGCTCCGGGCC',
'GGGTCTTAGG', 'GGTACTACGC', 'GTAGCTCCAA', 'GTCTTAGGGG', 'TACTACGCGG', 'TAGCGAGTTT',
'TAGCTGCAAA', 'TGATCGTTGG', 'TGTTGGATCG', 'TTGACGTCGA', 'TTGAGCGCTA', 'TTTAGCGAGT',]

        b=[
'AGAACCCGTA', 'AGCGTAAACC', 'AGCTCCTTAA', 'AGTTAACTCC', 'CCAGGCTCCG', 'CCCCAGCGTA',
'CCCGGCGCAT', 'CCCTGCGCAG', 'CCCTGCGCAT', 'CCCTGCGCCT', 'CCGCAGCGTA', 'CCGCAGCGTC',
'CCGCTGCGTA', 'CCGTGCGCAT', 'CCTGGCTCCG', 'CGCTCAGTAA', 'CGCTGCGCAT', 'CGGTCAGTAA',
'CGTTAACTCC', 'CTACGCCCGG', 'CTCAGTGGAA', 'CTCCGCATCC', 'CTCCGGACGC', 'CTCTGCGCAT',
'CTCTGTGGAA', 'CTGCGCATCC', 'CTGCGGACGC', 'CTGGAACAGT', 'CTTAGGAAGG', 'CTTCGCCCGG',
'CTTTGGAAGG', 'GCAATTGAGT', 'GCAGGCTCCG', 'GCCCAATGCA', 'GCCCTATGCA', 'GCCCTTGCTA',
'GCCTGCGCAT', 'GCGAGTAATT', 'GCGCAATGCA', 'GCGCAGCGTA', 'GCGCATTTCC', 'GCGCTACCTT',
'GCGCTGCGTA', 'GCGCTTTTCC', 'GCGGCCTCCG', 'GCGTCTTAGG', 'GCGTGCGCAT', 'GCGTGTAATT',
'GCTATTGAGT', 'GCTCCGGGCC', 'GCTGGCTCCG', 'GCTTTTGAGT', 'GGACTGATCG', 'GGCGTAAACC',
'GGCTCAGTAA', 'GGCTCCTTAA', 'GGCTCTTAGG', 'GGCTGCGCAT', 'GGGCTACCTT', 'GGGGCTTAGG',
'GGGTCAGTAA', 'GGGTCTTAGG', 'GGGTGCGCTA', 'GGTACTACGC', 'GGTCTGATCG', 'GGTGGCTCCG',
'GGTTAACTCC', 'GGTTCTACGC', 'GTCAGTGGAA', 'GTCCGGACGC', 'GTCTGTGGAA', 'GTCTTAGGGG',
'GTGAGCGCTA', 'GTGCGCATCC', 'GTGCGCTATT', 'GTGCGGACGC', 'GTGGAACAGT', 'GTGTGCGCTA',
'GTTAGGAAGG', 'GTTCGCCCGG', 'GTTTGGAAGG', 'TCTGATCGGG', 'TGCGTAAACC', 'TGCTCCTTAA',
'TGGTCCTTAA', 'TGGTGCGCTA', 'TGTTAACTCC', 'TGTTCCCGTA', 'TGTTCGTTGG', 'TGTTTACTCC',
'TTCAGCGCTA', 'TTCGCTCAGT', 'TTCTGCGCTA', 'TTGACGTCGA', 'TTGAGCGCTA', 'TTGTCGTCGA',
'TTGTGCGCTA', 'TTGTGCGCTC', 'TTTTGCGAGT', 'TTTTGCTGCA',]
        genes=[
'957', '1189', '3184', '2521', '2664', '1654', '720', '805', '6502', '829', '5278', '1404',
'2918', '2535', '3662', '3857', '617', '1563', '1719', '2419', '2308', '1074', '3796', '629',
'2604', '4240', '1355', '1206', '2449', '3960', '2651', '413', '888', '4205', '2051', '1530',
'4266', '2063', '2001', '4501', '4828', '3646', '1685', '2426', '1849', '796', '1587', '2645',
'491', '1060', '2436', '771', '1753', '935', '2611', '2312', '750', '772', '910', '646', '1094',
'2807', '1417', '1976', '2649', '641', '2476', '2147', '1266', '2875', '1869', '497', '2937',
'3151', '524', '1226', '633', '4853', '1172', '2952', '1901', '1021', '1538', '3688', '1503',
'914', '3356', '1184', '2290', '1457', '1152', '1309', '1026', '1114', '5238', '1565', '6179',
'785', '582', '1000',]
        UMIs=[
'2480', '2494', '10000', '6218', '4986', '3231', '973', '1174', '32044', '1269', '25246',
'2569', '7837', '5533', '8493', '15056', '917', '3378', '3448', '7002', '6292', '1805', '10045',
'899', '7584', '15303', '2133', '2187', '5248', '32420', '5809', '512', '1152', '13814', '4292',
'2892', '12761', '4813', '4288', '17791', '19436', '12121', '3345', '5522', '10712', '1326',
'2794', '6724', '606', '1902', '4289', '1057', '2860', '1695', '7377', '5756', '1219', '1168',
'1545', '993', '2025', '8293', '2122', '3876', '4976', '802', '5985', '4300', '2503', '6122',
'4457', '934', '5849', '8553', '939', '1850', '1045', '13869', '1807', '11769', '3530', '2632',
'3368', '12261', '3108', '1261', '9610', '5861', '6015', '2641', '1750', '2269', '1469', '1673',
'16613', '2597', '21992', '1077', '993', '1815',]
        return (a, b, genes, UMIs)


def issimilar(a, b, genes, UMIs):
    s=[]
    k=0
    print('>>>>>>>>>>>>>>>>   EXACTLY SIMILAR CODES   <<<<<<<<<<<<<<<')
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]==b[j]:
                k += 1
```

```
                if k < 10:
                    print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                else:
                    print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])

    return s

def OneCharDiff(a, b, genes, UMIs):
    s=[]
    k=0
    print('\n')
    print('>>>>>>>>>>>>>>>   ONE CODE MISSMATCH   <<<<<<<<<<<<<<')
    print('Row,  BARCODE, GENE_CODE,  Barcode_row, GENE_row, genes, UMIs')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]!=b[j] and len(a[i])==len(b[j]):
                mem=0
                for m in range(len(a[i])):
                    aa=a[i]
                    bb=b[j]
                    if aa[m]==bb[m]:
                        mem += 1
                if mem == len(a[i])-1:
                    k += 1
                    if k < 10:
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                    else:
                        print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                    continue

    return s



def TwoCharDiff(a, b, genes, UMIs):
    s=[]
    k=0
    print('\n')
    print('>>>>>>>>>>>>>>>   TWO CODE MISSMATCH   <<<<<<<<<<<<<<')
    print('Row   BARCODE   GENE_CODE    Barcode_row   GENE_row   genes   UMIs')
    for i in range(len(a)):
        for j in range(len(b)):
            if a[i]!=b[j] and len(a[i])==len(b[j]):
                mem=0
                for m in range(len(a[i])):
                    aa=a[i]
                    bb=b[j]
                    if aa[m]==bb[m]:
                        mem += 1
                if mem == len(a[i])-2:
                    k += 1
                    if k < 10:
                        print(k,'  ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                    else:
                        print(k,' ', a[i],'  ', b[j],'  ', i+1,'  ', j+1,'  ', genes[j],'  ',
UMIs[j])
                    continue

    return s

def main():
    comp=nomeque()
    # issimilar(comp.impoort())
    f, g, genes, UMIs=comp.impoort()
    issimilar(f, g, genes, UMIs)
```

```
    OneCharDiff(f, g, genes, UMIs)
    TwoCharDiff(f, g, genes, UMIs)

main()
```