TELOMERE ANALYSIS BASED ON HIGH-THROUGHPUT MULTI -*OMICS* DATA

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)
im Fachgebiet
Informatik

Vorgelegt von M. Sc. Lilit Nersisyan
geboren am 19.02.1990 in Eriwan/Armenien

Die Annahme der Dissertation wurde empfohlen von:

1. PD Dr. Hans Binder (Universität Leipzig, Deutschland)
2. Prof. Dr. Peter Stadler (Universität Leipzig, Deutschland)
3. Prof. Dr. Aram Galstyan (University of Southern California, USA)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 01.09.2017 mit dem Gesamtprädikat magna cum laude.

ABSTRACT

Telomeres are repeated sequences at the ends of eukaryotic chromosomes that play prominent role in normal aging and disease development. They are dynamic structures that normally shorten over the lifespan of a cell, but can be elongated in cells with high proliferative capacity. Telomere elongation in stem cells is an advantageous mechanism that allows them to maintain the regenerative capacity of tissues, however, it also allows for survival of cancer cells, thus leading to development of malignancies.

Numerous studies have been conducted to explore the role of telomeres in health and disease. However, the majority of these studies have focused on consequences of extreme shortening of telomeres that lead to telomere dysfunction, replicative arrest or chromosomal instability. Very few studies have addressed the regulatory roles of telomeres, and the association of genomic, transcriptomic and epigenomic characteristics of a cell with telomere length dynamics. Scarcity of such studies is partially conditioned by the low-throughput nature of experimental approaches for telomere length measurement and the fact that they do not easily integrate with currently available high-throughput data.

In this thesis, we have attempted to build algorithms, *in silico* pipelines and software packages to utilize high-throughput *–omics* data for telomere biology research. First, we have developed a software package Computel, to compute telomere length from whole genome next generation sequencing data. We show that it can be used to integrate telomere length dynamics into systems biology research. Using Computel, we have studied the association of telomere length with genomic variations in a healthy human population, as well as with transcriptomic and epigenomic features of lung cancers.

Another aim of our study was to develop *in silico* models to assess the activity of telomere maintenance machanisms (TMM) based on gene expression data. There are two main TMMs: one based on the catalytic activity of ribonucleoprotein complex telomerase, and the other based on recombination events between telomeric sequences. Which type of TMM gets activated in a cancer cell determines the aggressiveness of the tumor and the outcome of the disease. Investigation into TMM mechanisms is valuable not only for basic research, but also for applied medicine, since many anticancer therapies attempt to inhibit the TMM in cancer cells to stop their growth.

Therefore, studying the activation mechanisms and regulators of TMMs is of paramount importance for understanding cancer pathomechanisms and for treatment. Many studies have addressed this topic, however many aspects of TMM activation and realization still remain elusive. Additionally, current data-mining pipelines and functional annotation approaches of phenotype-associated genes are not adapted for identification of TMMs. To overcome these limitations, we have constructed pathway networks for the two TMMs based on literature, and have developed a methodology for assessment of TMM pathway activities from gene expression data. We have described the accuracy of our TMM-based approach on a set of cancer samples with experimentally validated TMMs. We have also applied it to explore TMM activity states in lung adenocarcinoma cell lines.

In summary, recent developments of high-throughput technologies allow for production of data on multiple levels of cellular organization – from genomic and transcriptiomic to epigenomic. This has allowed for rapid development of various directions in molecular and cellular biology. In contrast, telomere research, although at the heart of stem cell and cancer studies, is still conducted with low-throughput experimental approaches. Here, we have attempted to utilize the huge amount of currently accumulated multi-*omics* data to foster telomere research and to bring it to systems biology scale.

AKNOWLEDGEMENT

My whole PhD path has been quite a creative and fun experience. I thank Arsen for providing the friendly and thought-promoting environment and for believing in me so much. You have supported me at all the levels of thesis-related activities: you were always there for scientific discussions, for helping me in brainstorming, coding (and deBOGing :)), manuscript drafting, etc. Throughout all the eight years of your supervision, you were there to support my studies, and my participation in various scientific activities. Most importantly, your strong moral support cannot be underestimated. Last, but not least, you have connected me to Hans!

Special thanks go to Hans, for his faith in Computel and PSF algorithms, and for making our life so happy during our exchange visits. Your pieces of wisdom in each email and beer-talk have had and will still have their positive influence on my mood and life-view. Special thanks to you and Henry, Lydia, Volkan, and Corrina, for happily hosting me at IZBI: I remember every single day of that visit. And finally, without Hans, this thesis wouldn't make its way to Germany. I'd also like to thank you beforehand for all the beer sessions that will follow this event :)

Thanks Lydia, for your support in all the "beaurocratic" activities related to the thesis submission. Chocolate will come!

Thanks, Anna, for your great job on "South Asian genome" study and the lung cancer cell lines. Most importantly: for your interest and patience when hearing the "telomere" stories almost every day for two years.

I thank my family for your love! Also, you've been as understanding as possible when not seeing my face for long time periods, and have freed me from everyday responsibilities when I was deeply in work.

Let all the PhD students be as happy as I've been!

CONTENTS

MTL	Mean Telomere Length
TMM	Telomere Maintenance Mechanism
TPE	Telomere Position Effect
TPE-OLD	TPE over long distances
FSHD	Facioscapulohumeral muscular dystrophy
TERRA	Telomeric Repeat Containing RNAs
ALT	Alternative Lengthening of Telomeres
LTL	Leukocyte Telomere Length
DC	Dyskeratosis Congenital
TRF	Terminal Restriction Fragment
STELA	Single Telomere Length Analysis
NGS	Next Generation Sequencing
WGS	Whole Genome Sequencing
WES	Whole Exome Sequencing
SRA	Sequence Read Archive
MRE	Mean Relative Error
SE	Standard Error (SE) of MRE
RMSE	Root Mean Squared Error
SNP	Single Nucleotide Polymorphism
GWA	Genome Wide Association
PBL	Peripheral Blood Leukocytes
AC	Adenocarcinoma
RPKM	Reads per Kilobase of transcript per Million of mapped reads
NLS	Nuclear Localization Signal
NES	Nuclear Export Signal
CB	Cajal Bodies
PML	Promyelocytic Leukaemia
APB	ALT-associated PML Bodies
SCE	Sister Chromatid Exchange
T-SCE	Telomeric Sister Chromatid Exchange
BIR	Break Induced Repair
HR	Homologous Recombination
HJ	Holiday Junction
DDR	DNA Damage Response
PSF	Pathway Signal Flow
hMSC	human Mesenchymal Stem Cell
FC	Fold Change
SVM	Support Vector Machine

ORA	Over-Representation
GSEA	Gene Set Enrichment Analysis
GO	Gene Ontology

After their discovery, telomeres have been considered as the secret to longevity and cancer treatment. In the few decades that followed, many scientific groups have devoted their time and resources to investigate telomere biology, to understand their function in health and disease, and to manipulate on them to cure cancers. However, despite the excitement and the boom, methodologies to study telomeres are still low-throughput in nature, with each study looking at only a few genes, proteins, epigenetic factors and chromosome conformations, and their link to telomeres. This contrasts to the rapid development of molecular biology fostered by the advent of high-throughput data generation techniques, such as microarrays and next-generation sequencing technologies. The exponential growth of massive data obtained from a single tissue or form single cells has prompted successful comprehension of all the processes that happen in a biological system at once. These data also contain "undisclosed" information about telomeres and telomere related regulatory processes. Therefore, we have attempted to utilize this "hidden" information to promote telomere biology research via its integration into multi-*omics* systems level studies.

The main aim of this thesis was to develop computational approaches, models and software packages to use high-throughput data derived from microarray and next generation sequencing technologies in order to analyse telomere length dynamics and telomere length maintenance processes at systems scale.

In the first chapter, we have characterized the state of the art on telomere structure and function. We talk about the known facts on structural organization of telomeres, and have highlighted the knowledge gaps in their functional roles.

In the second chapter we discuss the experimental approaches that are used for measuring telomere length and studying the link between it and gene expression. We try to identify the reasons that stand behind the relative scarcity of studies that have explored the association between telomere length dynamics and the genomic, transcriptomic and epigenomic features of the cells. We illustrate that currently accumulated large amounts of next generation sequencing data may be utilized more efficiently to also perform telomere biology research at the systems level. In the frame of this work, we have developed the software package Computel for estimation of mean telomere length from whole genome sequencing data. We have evaluated the accuracy of

Computel on synthetic data, and have compared its performance with relevant computational approaches and with experimental pipelines. We show that our approach is a valid substitute for experimental measurement of telomere length, and that it has the advantage of extracting telomere length information, aside from other genetic variations, from WGS data. In addition, availability of coupled genomic and other *-omics* data, allows for performing whole genome level studies to find associations between telomere length dynamics and transcriptomic/epigenomic features. Computel is thus aimed at utilization of next generation sequencing data to study telomeres at systems level, accounting for the global picture of processes happening in cells.

In chapter 3, we describe a couple of studies performed using Computel. One of the studies refers to finding the association between genomic variations and telomere length in a healthy population of South Asians. We have discovered polymorphisms in *ADARB2* gene, which encodes an RNA-editing enzyme, that were associated with long telomeres. This finding and the previously reported link between *ADARB2* and extreme longevity our making this gene a good candidate for future studies. In addition, we have demonstrated that the association of telomere length with age is population specific. In our second study, we have utilized WGS, transcriptomic and epigenomic datasets on lung adenocarcinoma cell lines to mine the associative relationship between telomere length, gene expression, and epigenetic changes. We have identified several genes that might be in a regulatory relationship with telomeres.

The second and third chapters of this thesis that relate to development of Computel and its applications, are mainly based on the following papers:

1. Nersisyan L, Arakelyan A: **Computel: Computation of mean telomere length from whole-genome next-generation sequencing data.** *PLoS One* 2015, **10**.
2. Nersisyan L: **Integration of telomere length dynamics into systems biology framework: A review.** *Gene Regul Syst* Bio 2016, **10**.
3. Hakobyan A, Nersisyan L, Arakelyan A: **Quantitative trait association study for mean telomere length in the South Asian genomes.** *Bioinformatics* 2016, **32(11).**
4. Nersisyan L, Hakobyan A, Arakelyan A: **Telomere-associated gene network in lung adenocarcinoma.** *Eur Respir J* 2015, **46(suppl 59)**.

Chapter 4 focuses on another important phenomenon in telomere biology: the factors that lead to activation of telomere maintenance mechanisms (TMM). There are two main TMMs known to

date: one is realized through catalytic activity of a ribonucleoprotein complex telomerase, and other occurs via homologous recombination events between telomeric sequences, called ALT. The first mechanism is mainly employed by stem cells and many cancer cells. ALT has been identified in a fewer number of cancers. The type of TMM activated in a tumor largely determines the pathomechanistic characteristics and aggressiveness of the tumor. A lot is known about the TMM mechanisms, but more aspects still remain elusive. We have attempted to gather the accumulated knowledge on drivers, activators and main players in the two TMMs, and to highlight the gaps and topics of further investigations.

The studies addressing TMMs, are performed at a low-throughput level: each study attempts to investigate the role of one or two molecular factors in these processes. A couple of studies have attempted to use standard data mining and functional annotation pipelines for investigation of TMMs. The main shortcoming of these studies is that there are currently no properly annotated functional categories and no computational model or framework that could validate computational means of TMM studies.

In this thesis we have generated an *in silico* model to study the factors leading to TMM activation and to predict the type of TMM activation in a given sample based on its gene expression data, which we have described in Chapter 5. Our approach is based on literature-based curation and generation of TMM-related molecular pathways, and pathway activity determination from gene expression data with an in-house Pathway Signal Flow algorithm. Using a set of experimentally annotated cell lines and tumor cells, we have confirmed that our pathway based approach is able to accurately predict TMM activity. Most importantly, it may serve as a convenient model to question the role of various molecular factors (proteins, RNAs, genetic mutations, etc) and molecular interactions in TMMs, based on pathway extension and accuracy estimation. Finally, we use the newly developed approach to investigate into TMM activation states of lung adenocarcinoma cell lines.

The results described in chapter 5 were presented in a poster presentation at the EMBO conference on “From Functional Genomics to Systems Biology” 2016, in Heidelberg, Germany.

In summary, we have developed a number of computational approaches to utilize high-throughput data for telomere research. We believe that these methodologies will help to utilize existing massive data in a more efficient manner for telomere-related studies and will foster telomere

research. We have also presented a few of preliminary findings made using our software and algorithms.

This work has been accomplished during the 2014-2017 years, in the Group of Bioinformatics at the Institute of Molecular Biology (National Academy of Sciences of Armenia), under the supervision of Dr. Arsen Arakelyan, and under the co-supervision of Dr. Hans Binder in the frame of cooperation with the Interdisciplinary Centre for Bioinformatics (IZBI) at the University of Leipzig, Germany. The cooperation has been realized particularly during the four months of my DAAD funded research stay at IZBI (91572203, July-November, 2015), and during the summer of 2016 (in the frames of DFG-Project for initiation of cooperation, LO 2242/1-1 (2015-16), BMBF MycSys (2016) and HNPCCSYS/CancerSys (2012-15) projects), as well as during the short term visits of Dr. Binder to Yerevan in 2015 and 2016 (the Armenian SCS MES RA 15T-1F150 (2015-17) and 13YR-1F0022 (2013-15) research grants).

CHAPTER 1
THE STRUCTURE AND FUNCTIONS OF TELOMERES

Eukaryotic chromosomes contain linear DNA, which at its beginning and the end consists of repetitions of short tandem G-rich repeats: these regions of chromosomes are called telomeres. These are not protein coding regions, but perform a variety of protective and regulatory functions in the cell. This chapter shortly outlines the state of the art on structural organization and functional aspects of telomeres, and highlights their role in age-related diseases and cancers.

1.1 INTRODUCTION AND HISTORICAL REMARKS

In the second half of the 20th century, L. Hayflick noticed that cultured somatic cells were able to divide only a limited number of times (around 60, which is also known as the Haylflick limit) [1, 2]. Later on J. Watson and A. Olovnikov talked about the "end replication problem", which stated that in each round of cell division the polymerase could not fully replicate the 5' end of the DNA strand, leading to a 3' overhang [1, 3]. The "end replication problem" was thought to be the reason for limited replicative capacity of somatic cells. Furthermore, H. Muller and B. McClintock discovered that broken chromosomes were highly unstable, in comparison to intact chromosomes with complete ends [4, 5]. Thus, they had established the protective nature of the chromosomes ends, which were named by Muller the telomeres (in Greek, "telos" - end, "meros" - part). Finally, in 1978, E. Blackburn discovered the repetitive nature of the telomeric sequence of Tetrahymena thermophila, which consisted of consecutive "TTGGGG" repeats [6]. The human telomeric sequence was later established to consist of "TTAGGG" tandem repeats [7]. Together with C. Greider and J. Szostak, they further discovered the mechanisms of elongation of telomeres by the enzyme telomerase [8] or via homologous recombination events [9]. In 2009, E. Blackburn, together with C. Greider and J. Szostak received a Nobel Prize in Physiology and Medicine *"for the discovery of how chromosomes are protected by telomeres and the enzyme telomerase"* [10].

Today, more and more experiments are conducted to discover the structural and functional roles of telomeres and their link to healthy aging and disease development [11]. It's been established that the main role of telomeres is to protect the ends of the linear chromosomes from being recognized as double strand breaks and, thus, to avoid degradation or end-to-end fusions of chromosomes. Indeed, shortening of telomeres beyond a critical threshold leads to telomere dysfunction and chromosomal instability. After this, the cells either become senescent or undergo

malignant transformations [12]. Interestingly, stem cells and cancer cells possess machinery for elongation and maintenance of telomeres to be able to divide almost "infinitely" [12]. Finally, aside from the important role of telomere length maintenance for cell cycle regulation, it has also been shown that telomere length dynamics can have its distinct role in regulation of gene expression by yet not fully established mechanisms [13–15].

Based on all of these findings, much of the attention of the scientific world has been switched towards telomere biology to understand their role in healthy development of cells, in aging, age-related diseases and cancers. In the following sections we attempted to summarize the state of the art on telomere biology.

1.2 THE STRUCTURE OF TELOMERES

Telomeres are located at the ends of eukaryotic chromosomes and are usually composed of long stretches of short consecutive repeats [16]. In humans, the repeats have the TTAGGG sequence toward the 3' end of each strand, and normally span 3-20 kb in length. Telomeres lengths vary at different chromosome ends [17], and their length range depends on the species of the organism, the age, the tissue type and the cell state [18, 19]. The sequence pattern varies among organisms: in all vertebrates, including humans, telomeric repeats have the conserved sequence TTAGGG [20]; protozoans, plants and other animals have varying sequences (the sequences for model organisms are presented in Table 1). Notably, while most of the organisms have fixed repeat patterns of length 5-8 nt (e.g. TTAGGG), a few of them, such as Sacharomyces cerevisiae, have a dynamic repeat pattern, such as G(2-3)(TG)(1-6)T (Table 1). Interestingly, some species, e.g. some mosquitos, do not possess short telomeric repeats [21].

Table 1. Telomeric repeat sequences in different species.

Group	**Organism**	**Telomeric repeat (5' to 3' toward the end)**
Vertebrates	*Homo Sapiens, Mus Musculus*	TTAGGG [20]
Ciliate protozoa	*Tetrahymena thermophile*	TTGGGG [6]
Higher plants	*Arabidopsis thaliana*	TTTAGGG [22]
Insects	*Bombyx mori*	TTAGG [21]
Roundworms	*Ascaris lumbricoides, Caenorhabditis elegans*	TTAGGC [23]
Fission yeasts	*Schizosaccharomyces pombe*	TTAC(A)(C)G(1-8)

Budding yeasts	*Saccharomyces cerevisiae*	TGTGGGTGTGGTG (from RNA template) or G(2-3)(TG)(1-6)T (consensus) [24]

The telomeres in all species normally fold in a unique way. The folding of telomeres is conditioned by the presence of a G-rich overhang: the 3' strand of telomeres, which is always longer than the C-rich strand (or C-strand). This overhang is usually 30-400 nt long in humans [25]. It folds back in the direction of centromeres to form a T-loop [26]. The T-loop is stabilized due to invasion of the 3' overhang onto the C-strands to form a displacement loop (D-loop) (Figure 1). The D-loop is preserved via formation of hydrogen bonds between the complementary strands. The overall loop is stabilized by binding of so called shelterin proteins that form a nucleoprotein complex called the telomeric cap [27]. Proper capping is essential for preserving the integrity of chromosome ends, as discussed below.

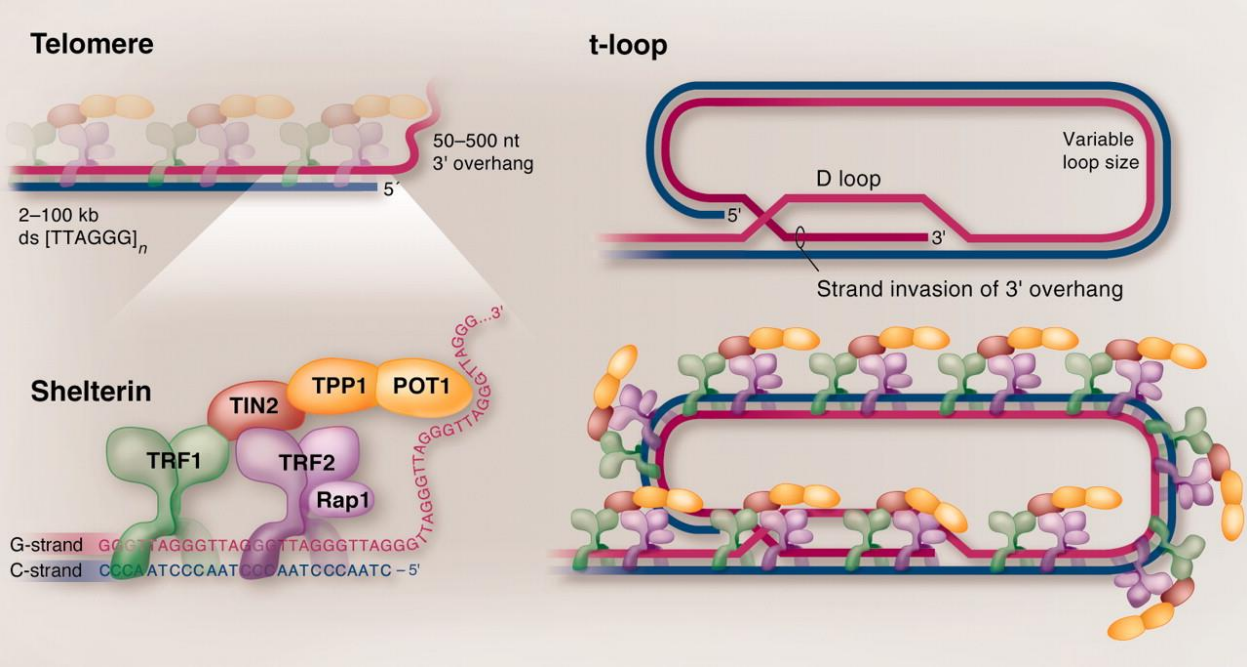


Figure 1. **The structure of telomeres (the picture source is [28]). Top-left**: the 3' end is normally 50-500 nt longer than the complementary strand. **Top-right**: the 3' overhang invasion leads to formation of the D- and T-loops. **Bottom**: the shelterin proteins bind to telomeres to cap and protect the telomeric ends.

The shelterin proteins include telomeric repeat binding factor 1 and 2 (TNF1/2), protection of telomeres protein 1 (POT1), TRF1-interacting protein 2 (TIN2), TIN2- and POT1-interacting protein (TPP1) and repressor/activator protein 1 (RAP1). POT1 binds specifically to single stranded DNA (ssDNA) at the 3' overhang (and, probably, at the D-loop) [29] and TRF1 and TRF2 bind to the double stranded DNA (dsDNA) along the telomeres [30]. TIN2 binds both to TRF1/2

and to TPP1-POT1 complex, thus bridging the shelterin proteins together [27]. RAP1 binds to TRF2 and has a role in inhibiting the processes of non-homologous end joining [27].

Mammalian telomeres are structured into nucleosomes, as is the non-telomeric chromatin, although the spacing between nucleosomes may slightly vary [31]. The compact state of telomeres largely depends on shelterin proteins. TRF1/2 and POT1 bind to dsDNA and ssDNA respectively, and are bridged together with the help of TIN2. These cross bridges lead to compacted globular-like structuring of telomeres and protect them from DNA damage response (DDR) proteins [32]. It has been hypothesized that short telomeres recruit less shelterin proteins and this may lead to an "open" telomeric state, which can attract DDR proteins and signal DNA damage response, as well as can recruit telomerase to elongate the telomeres [32].

Apart from the shelterin proteins, telomeres and the regions adjacent to them (subtelomeres) are also enriched with "silent" epigenetic marks, including trimethylation of H3K9 and H4K20, hypoacetylation of H3 and H4, high levels of HP1 protein and methylation of subtelomeric DNA [33, 34]. The H3K9 trimethylation marks are deposited by SUV39H1 and SUV39H2 histone methyltransferases. They create binding sites for HP1 proteins, which in turn recruit SUV4-20H1 and SUV4-20H1 histone methyltransferases for trimethyation of H4K20 [31, 35]. The telomeric heterochromatin plays a crucial role in regulation of telomere length and structural integrity, as well as in transcriptional activity at telomeres [31]. Proper telomeric heterochromatin formation is important for regulation of telomere length, as lack of H3K9 and H4K20 trimethylation leads to aberrant telomere lengthening [36]. On the other hand, telomere length in turn defines the epigenetic state of telomeres [31]. The telomeric heterochromatin states change during cellular differentiation and dedifferentiation [37, 38].

1.3 THE FUNCTIONS OF TELOMERES

1.3.1 Chromosome capping

The functional state of the telomeres depends on proper telomere folding and binding of shelterin proteins. In a fully functional state, telomeres perform both protective and regulatory functions. First, they protect the chromosome ends from being recognized as double strand breaks by the cell's DNA repair machinery. The DDR proteins recognize DNA double strand breaks and either attempt to fix those or trigger activation of apoptosis signaling [39]. These repair and checkpoint

proteins do not have inherent ability to distinguish between the chromosome ends and intra-chromosomal double strand breaks. However, the presence of the T-loop and the shelterin complex protects telomeric ends from DNA-damage response machinery. Additionally, the telomere capping preserves the integrity of chromosomes and does not allow for end-to-end fusions [27]. Chromosomal instability in malignant cells is commonly associated with the loss of capping due to extreme shortening of telomeres [40].

1.3.2 Regulation of gene expression

Besides the protective roles of telomeres, they also possess regulatory functions, although those have been described to a lesser extent. First, telomeres may regulate expression of nearby located genes via a phenomenon called Telomere Position Effect (TPE) [14, 41]. In TPE, genes located in subtelomeric regions near the telomeres, become reversibly silenced and this effect depends on the distance of the genes to the telomeres and telomere length [15]. The TPE silencing is explained by spreading of telomeric heterochromatin to subtelomeric regions, since telomeric heterochromatic marks are able to recruit chromatin modifiers, which in turn introduce heterochromatic marks at adjacent regions, thereby silencing nearby located genes. Therefore, the longer the telomeres, the stronger is their ability to recruit chromatin modifiers and the more distant is the spreading effect [41, 42]. Of note, epigenetic silencing of genes due to heterochromatic spreading is also observed in non-telomeric regions, at the edges of heterochromatin and euchromatin, and is generally known as position effect variegation [43]. TPE was first described in Sacharomyces *cerevisiae* and Drosophila *melanogaster*, using a series of experiments including repositioning of reporter genes from subtelomeric to distant regions and vice-versa; and examination of telomere size effect on gene silencing [41, 44, 45]. Later, TPE has been observed in a number of other species [46–48], including humans [49–52].

TPE usually spreads a few kilobases away from the telomeres (classical TPE occurring in yeast via SIR proteins), or up to 20 kb (via involvement of HAST domains in yeast) [13]. TPE-like long distance effects (known as TPE over long distances, TPE-OLD) are also observed, but those are not considered as classical TPE, since their mechanism is different and is explained by looping of telomeres towards the centromeres [13, 53].

Finally, telomeres may also regulate gene expression via specific positioning of the chromosomes relative to the nuclear lamina [54].

TPE in humans

The role of TPE in yeast and a number of other organisms varies, with some examples connected to response to nutrient deprivation or stress [55]. In these organisms telomeres are maintained at a relatively constant level due to telomerase activity [55]. In contrast, mammalian telomeres have the tendency to get shorter with age. In this context, the role of TPE in mammalian cells is not yet understood. For a long time, it's been hard to study the association of gene expression with telomere length in humans, since microarray platforms did not contain probes for subtelomeric genes. In their study, Ning *et al* [56], have constructed special microarrays involving probes against subtelomeric genes and have identified the first human gene Interferon-stimulated gene 15 (*ISG15*), whose expression was correlated with telomere length [56]. This gene is involved in immunity/stress response pathways and is located 1 Mb away from the telomeres. However, there are 8 genes located between *ISG15* and the telomeres, and their expression was not associated with telomere length. This contradicts the classical TPE mechanism, which assumes that heterochromatin spreading may not "skip" nearby located genes to affect only distantly located ones [13]. Thus, the authors have suggested that the observed association could occur via long distance effects due to telomere looping [56]. However, they have not shown whether the association of *ISG15* with telomeres is direct or indirect, i.e. occurring via other telomere-regulated (or regulating) genes.

A study conducted afterwards established the association of *DUX4* expression with telomere length, and suggested that *DUX4* might be regulated by classical TPE [52]. This gene is located near the end of the chromosome 4q and produces the DUX4-fl protein which is toxic to myoblasts, and its overexpression leads to Facioscapulohumeral muscular dystrophy (FSHD). In the genome of healthy individuals there are 100s of D4Z4 repeat elements located in the region between *DUX4* and the telomeres. FSHD patients, however, usually have only 1-10 repeats. These elements repress the expression of *DUX4*, and also act as insulators to block the spreading of telomeric heterochromatin. The authors have speculated that the loss of normal numbers of D4Z4 repeats leads to loss of repression by those elements, but also exposes DUX4 to silencing by TPE. As telomeres become shorter this silencing effect diminishes, leading to overexpression of *DUX4*,

which leads to disease progression. This may explain the delayed onset of FSHD [52]. As an additional argument for TPE, the authors have also observed that the silencing also affects neighbouring genes (*FRG2* and *FRG1*), but to a lesser extent, since these genes are located farther from the telomeres [52].

A very recent study by Kim *et al* [57] has investigated the effect of TPE-OLD on expression of *TERT* in humans, which is located on the p arm of the chromosome 5, abound 1Mb away from the telomeres. The study has shown that the 5p telomere loops back and gets in touch with the *TERT* locus. Telomere shortening leads to distortion of the loop and to DNA hypomethylation at *TERT* promoter and nearby regions. While this results in expression of only the first *TERT* exon in healthy fibroblast, and it has a great impact on higher expression levels of the full length transcript during cancerogenesis [57].

1.3.3 Telomeric transcription (TERRA)

In addition to their structural and regulatory roles, telomeres also possess transcriptional activity. The C-strand of telomeres is frequently transcribed to produce telomeric repeat containing RNAs (TERRA) [58, 59]. TERRAs contain 'UUAGGG' repeats, as well as subtelomeric sequences and have length in the range 100 bp to 49 kb [60, 61]. The mechanisms regulating TERRA transcription are still not understood, however the presence of subtelomeric sequences in TERRAs suggests existence of transcriptional control elements at subtelomeres [60]. It is assumed that TERRAs perform a number of regulatory functions [62]. E.g. TERRAs usually reside in nucleus and are frequently found near telomeric regions of chromosomes. It is suggested that they may inhibit the action of telomerase by complementary binding to its RNA subunit (hTR) [61]. TERRA sequences are also found in association with inactive regions of the X chromosomes, suggesting a potential role in chromosome inactivation processes [61].

1.3.4 Summary

In summary, telomeres play a number of roles in the cell, including chromosome capping and protection from DNA damage response proteins, regulation of gene expression through heterochromatin spreading and 3D chromosome looping, as well as via expression of TERRA transcripts that have a variety of regulatory functions. While the structure of telomeres and their

role in preserving the chromosome integrity is studied extensively, there is still a lot to learn about the regulatory effects of telomeres.

1.4 REGULATION OF TELOMERE LENGTH

Telomere length maintenance depends on crosstalk of two antipodal processes: telomere attrition and elongation.

Telomere attrition leads to gradual telomere shortening during cell replication. It is assumed that there are two main mechanisms contributing to the attrition rate: the incomplete replication of the lagging strand (the "end replication problem") and nuclease driven end resection of the C-strand to generate the G-rich overhang (Figure 2) [63]. During DNA replication, the lagging strand synthesis proceeds through Okazaki fragments, which in humans are 100-200 nt long, and it's assumed that the last RNA primer is placed at a random distance of 50-100 nt from the end of the lagging strand [63]. Thus, immediately after replication, the lagging strand has a nearly mature overhang of size ~70 nt [64]. The leading strand synthesis proceeds almost to the end of the template strand, but the replication apparatus disassociates from the chromosome termini prior to adding the final 1-2 nucleotides [65]. Approximately 1-2 h after the replication, in the S/G2 phase, the newly synthesized leading C-strand gets resected by nucleases to produce a G-rich overhang of a mature size. Finally, the C-strands of both leading and lagging daughter chromatids are processed to get the CCTAAC-5' end specification [64]. All in all, the 3' overhang is generated via incomplete replication of the lagging strand and post-replicative resection of the C-strand of both telomeric ends. Even though the incomplete replication and C-stand resections are considered to be the main players in telomere shortening, the actual attrition rates are higher than if only explained by these processes. It is therefore suggested that a number of stress-induced factors, such as reactive oxygen species, damage the telomeres and contribute to faster attrition [66].

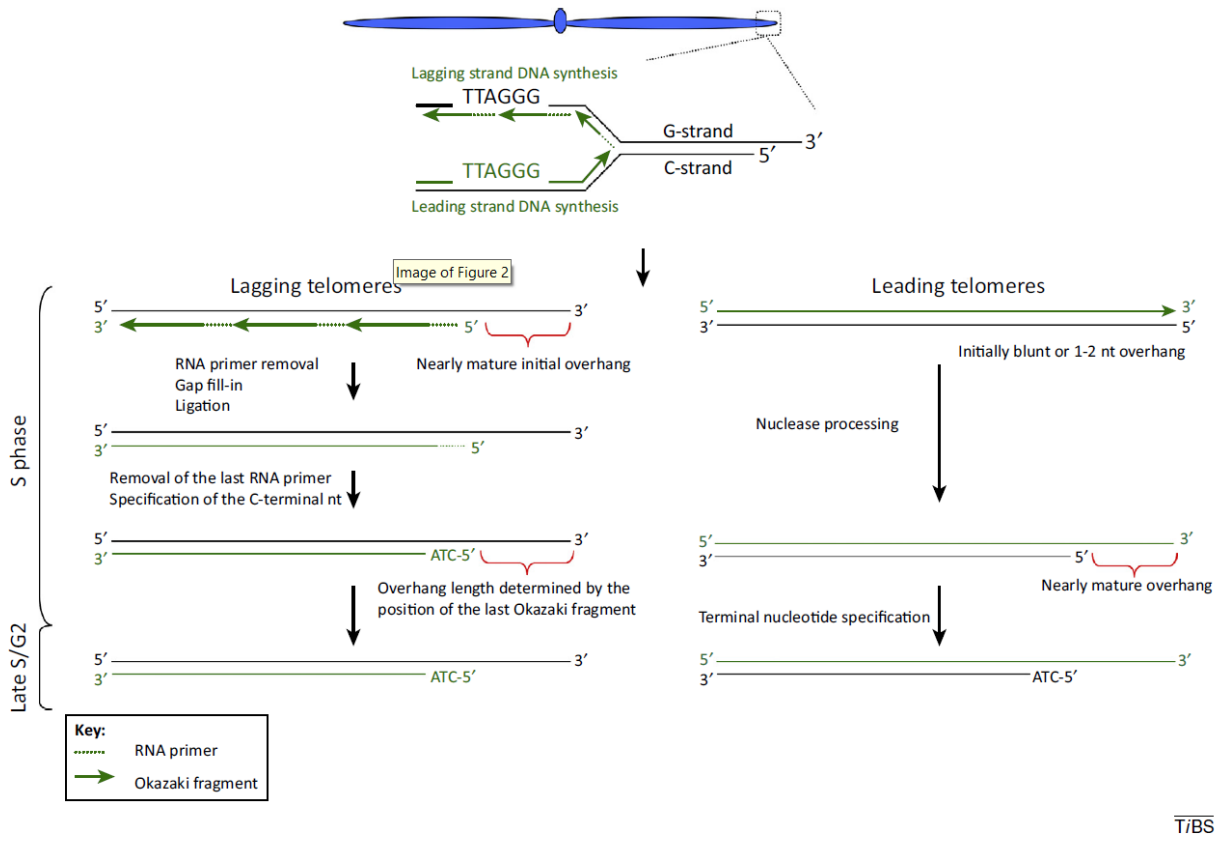


Figure 2. **The end replication problem at telomeres (the picture was taken from [63])**. **Top-center**: Canonical replication at lagging and leading strands. **Bottom-left**: the last RNA-primer at the lagging strand is positioned 70–100 nucleotides (nt) away from the ends, leading to formation of the nearly mature initial overhang. Afterwards, nuclear resection at the 5' strand produces the atc-5' end. **Bottom-right**: telomere replication at the leading strand does not produce the gaps via okazaki fragments. the polymerase synthesizes the complimentary strand up to 1-2 nt from the end. the 5' strand is then resected by nucleases to produce the 3' overhang and the atc-5' end.

Since telomeres are the limiting factor for proliferative capacity of the cells, continuously dividing cells, such as stem cells and cancer cells, employ mechanisms of telomere elongation [67]. Stem cells and the majority of cancer cells are expressing a nucleoprotein complex telomerase, which catalyzes telomere strand elongation using a special RNA template (hTR) [67]. Telomerase is mainly inactive in somatic cells, since its catalytic subunit, hTERT, is not expressed there. In some cancer cells telomere elongation may also occur in a telomerase-independent manner, via homologous recombination events. This mechanism is called the alternative lengthening of telomeres (ALT) [68]. These two mechanisms are described in detail in Chapter 4.

1.5 TELOMERES, AGING, AND DISEASES

1.5.1 Telomeres, aging and age related diseases

Telomere attrition rates depend on several internal and external factors, thus, telomere length is not constant and changes as a function of age and cell type. In contrast, telomere length at birth is a heritable characteristic. The study by Slagboom et al [69] has looked into variations in telomere lengths in the blood of between monozygotic (MZ) and dizygotic (DZ) twins and between non-related individuals, as measured by TRF. They have determined that telomere length variability was the smallest in MZ twins, larger in DZ tweens and the largest in unrelated individuals. Based on statistical regression models they have concluded that telomere length was to a large extent genetically determined. Another study has performed on blood samples from a cohort of 119 Saudi families [70]. In this study a parent-offspring MTL regression analysis was performed, with adjustment for age, sex, as well as additional phenotypic characteristics, such as body mass index and lipid profile. The study has pointed on highly heritable nature of telomere length, as well as its association with the mentioned cardiometabolic parameters. Finally, Honig *et al* [71] have looked into leukocyte telomere lengths (LTL) from long lived individuals and their offspring and relatives. They have found that the difference in LTL between long lived individuals and their offspring is much smaller, than the difference between them and relatives. And the latter was much smaller than the variability between non-related individuals. Heritability of telomere lengths may be explained by the fact that telomerase is most active in germline cells, due to which their telomeres are maintained at a relatively constant length and passed to gametes and offspring [69]. Telomerase also gets activated in other stem cells, but its activity levels are not sufficient to completely prevent telomere shortening. Thus, as the rest of the somatic cells in the body, stem cells also reach replicative arrest because of telomere shortening, though at lower rates [72], which, in turn, fosters organismal aging [73–75]. Telomere shortening is a relatively good biomarker of cellular aging, however, its association with organismal aging is not straightforward. Many studies have used leukocyte telomere length (LTL) as a surrogate marker of organismal aging [76], and have shown reverse association of LTL with age, however others have failed to reveal any significant association, leading to the conclusion that environmental and endogenous factors other than proliferative attrition may contribute to telomere length dynamics [77, 78].

1.5.2 Monogenic disorders associated with telomere dysfunction

A number of disorders are directly associated with mutations in components of telomerase that affect regenerative capacity of stem cells [79] and cause accelerated cellular aging by telomerase dysfunctions. These diseases are broadly defined as premature ageing syndromes.

One of the most studied telomerase disorders is Dyskeratosis congenital (DC) [80]. It is a rare disorder leading to abnormalities in skin pigmentation, in nails and in oral mucosa. Complications of DC lead to bone marrow failure, pulmonary fibrosis and cancer. There are several known mutations leading to DC, with three most pathogenic ones occurring in *DKC1*, *TERT* and *TERC* genes that encode the core components of telomerase. Other cases have DC-associated mutations in genes encoding proteins involved in telomerase assembly, such as *TINF2*, *NHP2* and *NOP10* [79]. All of these mutations lead to telomerase dysfunction and limited capability of stem cells to participate in tissue renewal in skin, oral mucosa, bones and liver. Notably, DC patients show disease anticipation: a phenomenon, where the disease onset happens at an earlier age in each subsequent generation. This is explained by telomerase dysfunction in germline cells, which leads to inheritance of shorter telomeres in the offspring [79]. Other severe premature aging syndromes, such as Aplastic anaemia [81] and Idiopathic pulmonary fibrosis [82], are also associated with mutations in telomerase genes.

1.5.3 Telomeres and complex diseases

The onset of many complex human diseases is associated with aging and is partly attributed to telomere shortening. On one hand, extremely short telomeres lead to telomere dysfunction and chromosome instability [12, 83]. On the other hand, regulatory effects of telomeres on gene expression, such as TPE and others (see section 1.4), may take place long before telomeres reach critically short values. These regulatory effects may be key factors leading to onset and progression of complex diseases.

As such, a number of studies have implicated LTL as a risk and/or a prognostic marker for cardiovascular diseases, ischemic strokes, type 2 diabetes, and neurodegenerative diseases, however the direct link between telomere shortening and disease development is not established for all of those cases [84–89]. Additionally, telomere length is crucial for preserving replicative capacity and clonal exhaustion of actively dividing immune cells. Age-dependent telomere shortening in B and T lymphocytes has been linked to defective immune responses and

development of many age-related diseases. In some cases, telomere shortening in T cells has been linked to development of auto-immune diseases [90]. Individuals with short telomeres or those suffering from premature aging syndromes, have susceptibility of developing infectious diseases, because of telomere shortening associated decline in immune functions [91].

In contrast to senescence related diseases, in cancers telomere length dynamics is largely affected by telomere lengthening and maintenance machinery. While in healthy cells, telomere dysfunction normally leads to cell growth arrest and apoptosis, cancer cells overcome this crisis and continue to proliferate, often leading to chromosomal instability, which is a key factor for cancer onset [12, 92]. On the other hand, further activation of telomere maintenance mechanisms is crucial for tumour cell survival and disease progression [12]. Therefore, the majority of cancer cells express telomerase [93], while others activate alternative lengthening mechanisms [94]. Activation of these mechanisms is a marker of aggressiveness and poor cancer prognosis [95]. Accordingly, therapies inhibiting these mechanisms have been shown be effective in limiting cell growth in some cancers [96]. Some studies are also implicated either long or short telomeres in cancer risk, depending on cancer type and study design [97] (see section 3.2.2).

1.5.4 Summary

All in all, the role of telomeres in healthy aging and development of age-related diseases and cancers have been undoubtedly demonstrated in numerous studies. However, even though in certain monogenic diseases the role of telomeres is clearly established, identification of their effect on the risk, onset and progression of complex diseases is a challenging task, since a wide range of diverse factors are involved in their pathomechanisms, with telomeres being just one of the contributors. Additionally, the majority of studies have concentrated on the consequences of extreme telomere shortening and telomere dysfunction, omitting the regulatory relationship between telomere length dynamics and gene expression.

CHAPTER 2
INTEGRATION OF TELOMERE LENGTH DYNAMICS INTO -*OMICS* STUDIES

Telomeres are implicated in healthy development and aging. However, the role of telomeres in regulation of transcriptome and epigenome and the influence of genomic factors in telomere length dynamics are not yet extensively analyzed. This chapter depicts the obstacles in experimental approaches that lead to scarcity of such studies and highlights the need of novel computational pipelines for systems level analysis of telomere biology.

In this thesis, we describe the development of a software package, Computel, for computation of telomere length from whole genome next generation sequencing data, which allows for integration of telomere length into high-throughput data analysis workflows.

2.1 STATE OF THE ART

2.1.1 Experimental methods for telomere length measurement

There are several experimental approaches for telomere length measurement, such as terminal restriction fragment analysis (TRF), quantitative PCR (qPCR), quantitative fluorescent in situ hybridization (qFISH), etc. [98], both for measuring mean and chromosome-specific telomere length.

The first method developed for quantitative assessment of the mean telomere length was TRF. It is based on the ability of certain restriction enzymes to cut DNA into small fragments leavening telomeric parts intact. The remaining long telomeric fragments are then separated using agarose gel electrophoresis. The isolated fragments are analysed with Southern blot using telomere probe ligation [99]. Since the fragments are of non-uniform length, they appear as dispersive smears. A crucial factor in accuracy of the average telomere length calculation is accounting for the smear length and the intensity of probe binding. TRF results obtained from different experiments depend on various factors, such as source DNA quality and quantity, choice of restriction enzymes, gel density, signal intensity calculations and length adjustment [100]. Moreover, the distance of the farthest subtelomeric restriction site to the telomeres is estimated to be 2.5-4 kb, depending on the chromosome and the restriction enzyme used. Thus, the results of different studies should be compared with caution and account for the difference in restriction enzymes. Finally, TRF requires great amount of starting DNA and is also not capable of capturing short telomeres [98].

Since TRF has been the first technique for MTL measurement, it has served as reference for further emerging methods and is thus considered the "gold standard".

The quantitative PCR approach is less elaborate compared to TRF. It is based on amplification of telomeric regions via telomere-specific primers. The basic assumption is that the longer the telomeric sequence of the source DNA, the more there are places for the primers to attach, and the more amplicons will be generated. Comparison of their relative quantity and that of the amplicons generated from single-copy gene PCR gives a T/S ratio, which is correlated with the overall telomeric content of the cell. One major disadvantage of this method is that it strictly depends on initial calibration steps, and the results derived from different experiments are difficult to compare [98]. Compared to TRF, qPCR is thought to be more prone to measurement errors, which are being addressed by further modifications and amendments [101]. A modified version of this approach, the monochrome multiplex qPCR, performs the single-gene and telomere amplifications in the same tube to reduce pipetting errors [102]. Besides measurement errors, T/S ratios obtained with these methods should be treated with caution accounting for possible copy number and chromosome number variations.

Other methods are used less frequently and serve more specific purposes. Single telomere length analysis (STELA) utilizes a 3' overhang specific linker and a subtelomeric primer to amplify telomeres at specific chromosome ends with PCR [103]. STELA uses subtelomeric primers of known lengths and known true distance from the telomeres, which makes it more accurate than TRF. Additionally, it estimates telomere lengths at specific chromosomes and is able to measure short telomeres, which is important in telomere shortening induced senescence studies and for accounting for the telomere length variability across chromosomes. However, sequence variability at most of the subtelomeric regions allows for capturing telomeres only at selected chromosome arms (Xp, Yp, 2p, 11q, 12q and 17p) [104]. Another limitation of STELA is that it is not able to capture long telomeres (more than 8 kb).

Quantitative fluorescence in situ hybridization qFISH is used for measuring telomere lengths at metaphase chromosomes via ligation of telomere-specific fluorescent probes. The signal intensity is then compared to a standard of known telomere length [105]. This method is accurate and is advantageous of providing arm-specific telomere lengths. The main drawback is that cells should

be able to divide, and, thus, it's not applicable to cell-cycle arrested cells. Also, because the method is hybridization based, it makes it difficult to quantify extremely short telomeres.

There is a wide variety of other techniques, each aimed at overcoming limitations of the existing ones [98]. There have been concerns, however, about comparability of the results obtained in different experiments, raising the need for proper calibration of the techniques based on the "gold standard" reference [106, 107].

2.1.2 The need for computational methods for telomere length measurement from NGS data: existing approaches

A major limiting factor for understanding the complex picture of telomere biology in the cell is the lack of high-throughput data coupled with experimental results on telomere length. Only a limited number of studies exist, where telomere length measurement experiments have been coupled with high-throughput genome, transcriptome or epigenome data. Importantly, in order to put the data obtained from these two experiments in the same context those should be performed on the same samples in the same time period. This poses the necessity of computational techniques, which could obtain telomere length information from whole genome sequencing data.

Next generation or massively parallel sequencing (NGS) technologies have emerged as a revolution in DNA and RNA sequencing, as they allow for simultaneous generation of a huge number of DNA and RNA short reads [108]. This, in turn, has allowed for obtaining aggregate information on whole exome, genome, transcriptome and epigenome of cell populations or even single cells [109]. Whole genome sequencing (WGS) produces short reads from the entire genome, while whole exome sequencing (WES) targets only the protein coding regions (1-2% of the genome) [110]. While WES is a relatively cheap tool, which is frequently used in variant discovery and molecular diagnostics, WGS allows to also obtain information about the non-coding parts of the genome [110]. The latter is used for various research purposes, including but not limited to phylogenetic analysis, variant discovery, analysis of regulatory sequences, tandem repeats, copy number variations, *etc*. Currently, a large amount of sequencing data are submitted to publicly available databases, such as the Sequence Read Archive (SRA) [111]. The number of entries in SRA grows exponentially [112], necessitating development of new algorithms and approaches to integrate all these data into the knowledge of global outlook of molecular processes occurring within cells and to perform systems-level association studies with telomere length dynamics.

Even though the telomeric sequences are covered by most of WGS technologies, they are masked from downstream analysis, because of their repetitive nature and the imposed difficulty of mapping telomeric reads to the reference genome. However, it's important to use the NGS capabilities to full extent and capture all possible information hidden in the generated data. We, thus, have stressed our attention to extracting telomeric content hidden in WGS data and using this information for telomere biology research.

During the last five years, a few methodologies have been developed for estimating telomere length from WGS data, all of which are based on capturing "telomeric" reads, which are presumably derived from telomeric regions of the genome. One of the pioneering works in this direction has been the study by Parker *et al* [113] on association of telomere length dynamics with pediatric cancers, where gain and loss of telomeric DNA has been estimated from WGS data by counting the number of short reads containing at least four consecutive telomeric repeats. Although the method they have used was not calibrated to return accurate results, they have shown that it's a promising approach. Another read count based software TelSeq [114] scans for reads containing more than a threshold amount of "TTAGGG" repeats and compares it to the number of genomic reads with the same GC content. This relative count is then multiplied with a GC-normalized genome length constant, which gives an estimate of absolute mean telomere length [114]. This method provides correlation of telomere length estimates with experimental results in certain settings (for 100 bp long Illumina reads). A couple of studies have already utilized TelSeq to show that major depression is associated with shorter telomeres and increased copies of mitochondrial DNA [115]; and that mutations in *ATRX* gene associated with chromatin remodelling are linked to increased telomere length in diffuse glioma [116]. However, Telseq has several limitations, which will be discussed below in detail.

To conclude, scarcity of telomere length studies arises from the lack of coupling between experimental measurement of telomere length and high-throughput data obtained from the same source. The vast availability of coupled whole genome, transcriptome and epigenome NGS data opens new opportunities for integration of telomere biology into systems level studies. The results obtained with recent methods and software for assessment of telomere length from whole genome NGS data highlight the promise of computational methodologies in fostering telomere biology research.

2.2 DEVELOPMENT OF COMPUTEL: THE TOOL FOR COMPUTATION OF TELOMERE LENGTH FROM WHOLE GENOME SEQUENCING DATA

In this section we describe our methodology for computing mean telomere length from WGS data, which is available as a software package Computal at https://github.com/lilit-nersisyan/computel and described also in [117]. At the time Computel was being developed no similar software was around. However, during finalization of our results the paper on TelSeq was published [114], and we took additional time to compare the accuracy and performance of our approach with theirs. We have shown that Computel addresses the limitations and shortcomings of previous computational approaches, including TelSeq. In this thesis and a number of preceding publications we show that Computel is also a valid substitute for experimental methods for telomere length measurement.

2.2.1 Methods and data

Algorithm description and general workflow

Computel is written in R 3.0.3 and performs command line calls to the following programs during execution: Bowtie 2-2.1.0, Samtools 0.1.19, and Picard tools 1.108. Computel can be called both from R environment and through command line, with an Rscript front-end available for Windows (versions v0.2 and lower) and Unix type systems. Detailed information about Computel installation and usage is available in its manual (see https://github.com/lilit-nersisyan/computel/blob/0.3/Computel_v0.3_User_Manual.pdf).

The general workflow of mean telomere length estimation by Computel is schematically represented in Figure 3. It consists of the following steps: (1) building a telomeric index, (2) mapping reads to the telomeric index, (3) coverage calculation at the telomeric index, (4) determination of mean coverage at reference genome (optional), (5) estimation of mean telomere length. Each of these steps is described in detail in the following subsections.

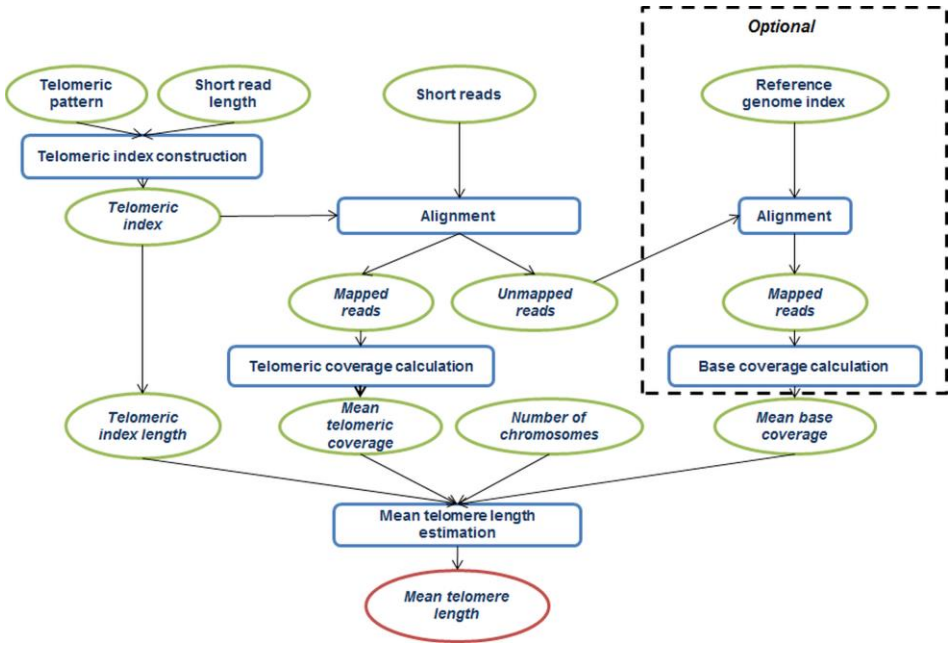


Figure 3. Schematic representation of the Computel algorithm for mean telomere length estimation. Computel takes whole-genome NGS short-reads as input; maps them to the telomeric index built based on user-defined telomeric repeat pattern and the read length; and calculates the mean telomere length based on the ratio of telomeric and reference genome coverage, the number of chromosomes, and the read length.

Building a telomeric index

The telomeric index is built using the *bowtie2-build* program. The index is designed in such a way that any read consisting of telomeric repeat patterns can map uniquely to the index. It is also important to take into consideration reads that contain telomeric repeats only partially: theoretically, we would like to also capture the reads originating from chromosome regions located at the junction of telomeric and immediate subtelomeric sequences. For this reason, the telomeric index has an additional 3'-end tail containing ambiguous nucleotide (N) bases to which any sequence can map (Figure 4). Note that the N-tail is attached only to the 3' end of the index, to minimize the number of captured reads containing interstitial telomeric repeats [118].

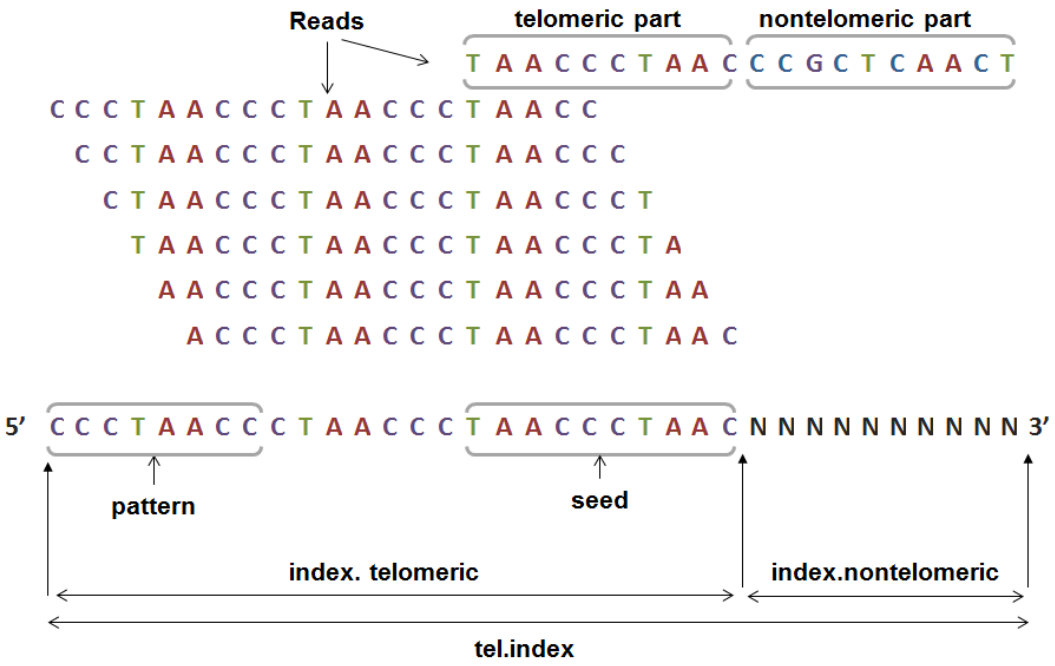


Figure 4. An example of telomeric index with reads aligned to it. The telomeric pattern is "TTAGGG" (human), which gets reverse complemented in the index ("CCCTAA"); read length = 20 nt; seed length (*min.seed* option) = 10 nt. The top read contains a non-telomeric region, which will be aligned to the non-telomeric tail of the index, the rest of the reads are six possible cyclic permutations of pure telomeric repeats.

The pseudocode for generation of the telomeric index is presented below:

Let *pattern* be the sequence of telomeric repeat pattern bases;

Let *pl* be the length of the *pattern;*

Let *rl* be the read length;

Let *min.seed* be the minimum number of telomeric read bases in the mapped reads;

Let *tel.index* be the sequence of the telomeric index;

Let *index.telomeric* be the region of the index containing telomeric repeats;

Let *index.nontelomeric* be the region of the index containing ambiguous bases {N};

The sequence of *tel.index* is computed as follows:

```
length(index.telomeric) = rl + pl - 1

count.pattern = int(length(index.telomeric)/pl)

count.substring = length(index.telomeric)/pl % pl

index.telomeric = concatenate(count.pattern * pattern,
pattern[1:count.substring])
```

```
index.nontelomeric = (rl – min.seed) * {N}

tel.index = concatenate(index.telomeric, index.nontelomeric)
```

tel.index sequence is supplied to the *bowtie2-build* program to build the index files. An example of *tel.index* sequence is given in Figure 4.

Aligning short-reads to the telomeric index

Paired- or single-end short-reads are aligned to the telomeric index with the program bowtie2-align [119]. We have chosen the most appropriate short-read alignment parameters that would combine the maximal accuracy and speed of the alignment to the telomeric index by evaluating different alignment options offered by Bowtie 2 and comparing their performance.

The alignments were performed with the Bowtie 2 preset options for --*end-to-end* alignment, i.e. --*fast* (F), --*very-fast* (VF), --*sensitive* (S) and --*very-sensitive* (VS) modes. The argument for "mismatches in a seed alignment" (–*N*), was tested for both the default value of 0 and for the value of 1, since setting -*N* to 1 increases the alignment sensitivity. Next, the -*L* option, which is the length of the seed substrings to align during multi-seed alignment, was set to one third of read length, but with minimum value of 6 and maximum value of 22 (this means that if one third of read length is less than 6, it is assigned the value of 6, and if it is more than 22, it is set to 22). Finally, the --*n-ceil* option, which is the maximum number of allowed mismatches and 'N' bases in the alignment, was set to [*rl-min.seed*], where *rl* is the short-read length and *min.seed* is the minimum number of telomeric read bases in the mapped reads. Thus, we ended up with 8 alignment modes (F-N0, VF-N0, S-N0, VS-N0, F-N1, VF-N1, S-N1, VS-N1) and compared their effect on the accuracy of telomere length estimation. Eventually, the mean relative error (MRE), standard error (SE) of MRE, root mean squared error (RMSE), and coefficient of determination (R^2) were calculated, separately for single and paired-end reads, and compared across alignment modes.

By default, the alignment is performed with the Bowtie 2 preset options for --*end-to-end* alignment, with --*very-sensitive* mode, -*N* set to 1 (to allow mismatches in a seed alignment), and -*L* ranging between 6 and 22, which is calculated automatically, depending on read length. The resulting alignment is stored in a SAM file.

Reference genome (base) coverage calculation

For reference genome coverage calculation, the generated SAM file is split into two SAM files containing mapped and unmapped reads. The SAM file containing unmapped reads is converted back to a FASTQ file using the Picard *SamToFastq* tool. Unmapped reads are then aligned to the reference genome with Bowtie 2 default options, sorted and used for base coverage (*base.cov*) calculation with the Samtools *depth* command.

Mapping short-reads to reference genome can be time-consuming. However, without significant loss in accuracy (data not shown), the base coverage can be estimated as:

$$base.cov = (\text{total number of reads}) * (rl) / (\text{total genome length}),$$

and supplied to Computel as an argument.

Telomeric coverage calculation and mean telomere length estimation

The SAM file for mapped reads is sorted and the distribution of coverage per base for the telomeric index is calculated using the Samtools *depth* command.

We have used the mean value of coverage at each base as a point estimate for coverage at telomeric index (*tel.cov*). The relative coverage at telomeric index compared to the reference genome is computed as *rel.cov* = *tel.cov* / *base.cov.* Finally, the mean telomere length (MTL) is estimated as:

$$MTL = (\text{mean}(rel.cov)) * (rl + pl - 1) \; / \; (2 * n\_chr),$$

where the number 2 in the denominator accounts for the two chromosome ends, and *n_chr* is the number of chromosomes in the haploid genome. The rest of the variables are explained above.

Computel validation with synthetic data

In order to estimate the algorithm performance, we carried out a series of telomere length calculations using synthetic data. For this purpose, we have taken a ~200 kb fragment of human reference chromosome 1 (GRCh37) from the NCBI Genome database. This fragment did not contain either pure or interstitial telomeric repeats [118], nor did it contain ambiguous bases. It was used for two purposes. First, it served as a reference genome for base coverage estimation (see below). Second, telomeric sequences consisting of human telomeric TTAGGG or CCCTAA repeats with known lengths were attached *in silico* to both ends of this chromosome fragment. The telomeric sequence lengths were randomly chosen from a normal distribution with mean 10 kb

and standard deviation 7 kb. The resulting sequences were used to generate artificial short-reads using the ART tool for Illumina sequencers [120].

The following testing scenarios have been exploited:

- short-reads of different lengths from the set {20, 36 , 51, 76, 100, 150 nt};
- short-reads with different insert sizes from the set {200, 300, 500 nt};
- short-reads with different coverage values from the set {0.1, 0.5, 1, 2.5, 5, 10, 30};
- paired-end and single-end short reads.

The 200 nt insert size was not considered for 100 nt or 150 nt length reads; and the 300 nt insert size was not considered for 150 nt length reads.

Performance comparison with TelSeq

We compared the performance of Computel to that of TelSeq [114]. Both of the software were used with their default settings, unless otherwise stated.

TelSeq computes telomeric length with the formula $l = t_k sc$, where l is mean telomere length, t_k is the abundance of telomeric reads, s is the fraction of all reads with GC composition between 48% and 52%, and c is a constant for the genome length divided by the number of telomere ends [114]. First, we performed comparisons based on the settings for short-reads generation taken from the original TelSeq paper [114]. Briefly, human chromosome 1 of the GRCh37 genome assembly was used as a reference. Terminal sequences of 30 kb length, including N-bases and telomeric repeats, were removed from each end of the chromosome and replaced with the same length of telomeric repeats. Illumina short-reads were generated with the SimSeq tool (*https://github.com/jstjohn/SimSeq*) using the following parameters: *-1 100 -2 100 --insert_size 500 --insert_stdev 200*, with coverage equal to 0.4x (498,501 reads), 2x (2,492,506 reads), or 10x (12,462,531 reads), and with duplication rate fixed at 5% for all coverages. Each setting was repeated 5 times. Mean telomere length was measured with TelSeq using exactly the same settings described in the original paper [114]. Because the genome length constant for telomeric GC content is hard-coded in the TelSeq software, we computed this parameter for chromosome 1 (21,722,000 for chromosome 1 instead of 332,720,800 for the total genome) and recompiled TelSeq with the new value. The mean telomere length estimates by TelSeq were compared to Computel's estimates.

To account for read length variation, we have repeated the settings described above, with fold coverage equal to 2x, and for read lengths equal to 36 nt (6,923,628 reads), 76 nt (3,279,613 reads), 100 nt (2,492,506 reads), and 150 nt (1,661,671 reads). TelSeq *k* threshold was kept at default value of 7 for all read lengths, except for 36 nt, for which we have tested *k* values equal to 4, 5, and 6.

Finally, we have assessed the performance of Computel and TelSeq depending on short-read generation algorithms. For this, we have generated 100 nt length short-reads with 0.4x (498,501 reads), 2x (2,492,506 reads), and 10x (12,462,531 reads) coverage using another short-read generation tool, ART Illumina read generator, with its default parameters [120].

Measures of telomere length estimation accuracy and statistical analysis

The accuracy of telomere length estimation was evaluated with the following criteria:

- Mean of relative error (MRE), and standard error (SE), where relative error of estimated (EL) over actual telomere length (L), is the ratio (EL-L)/L;
- Root mean squared error (RMSE), which represents the variance of mean telomere length estimation error;
- Coefficient of determination, R^2, which is the estimate of quality of linear fit (goodness of fit) between estimated and actual mean telomere lengths.

Paired t-tests were used for pairwise comparisons, single factor analyses were performed using ANOVA, and, finally, the effects of several factors on accuracy of estimation were assessed using multi-factor ANOVA. P values less than 0.05 were considered significant.

Algorithm validation with experimental data

In order to validate Computel performance with experimental data, we downloaded whole-genome sequencing data for paired tumor and healthy tissues of 5 neuroblastoma and 2 osteosarcoma patients (EGAD00001000135 and EGAD00001000159) [113]. For these samples, differences in telomere length between normal and tumor tissues have previously been estimated using quantitative real time PCR.

We have computed telomere lengths for these datasets with Computel and Telseq and validated the results against experimentally obtained data, described by Parker *et al* [113].

2.2.2 Results

Validation of Computel using synthetic data

We have assessed the accuracy of mean telomere length estimation by Computel in a series of computations performed on "synthetic chromosomes" with telomeres of known length attached to their ends. The lengths of attached telomeres were randomly chosen from a normal distribution with mean 10 kb and standard deviation 7 kb. The range (min-max) of generated telomeres were 194.5–21138 bp for single-end reads, and 387.5–24169.5 bp for paired end reads, respectively.

The results obtained showed very strong linear correlation between actual and estimated telomere lengths in all the experiments, with the quality of linear fit (R^2) equal to 0.95 and 0.92 for single and paired-end reads, respectively (Fig 3). The mean relative errors (MRE±SE) between estimated and actual telomere lengths were 4±0.01% for single-end reads and 7±0.01% for paired-end reads.

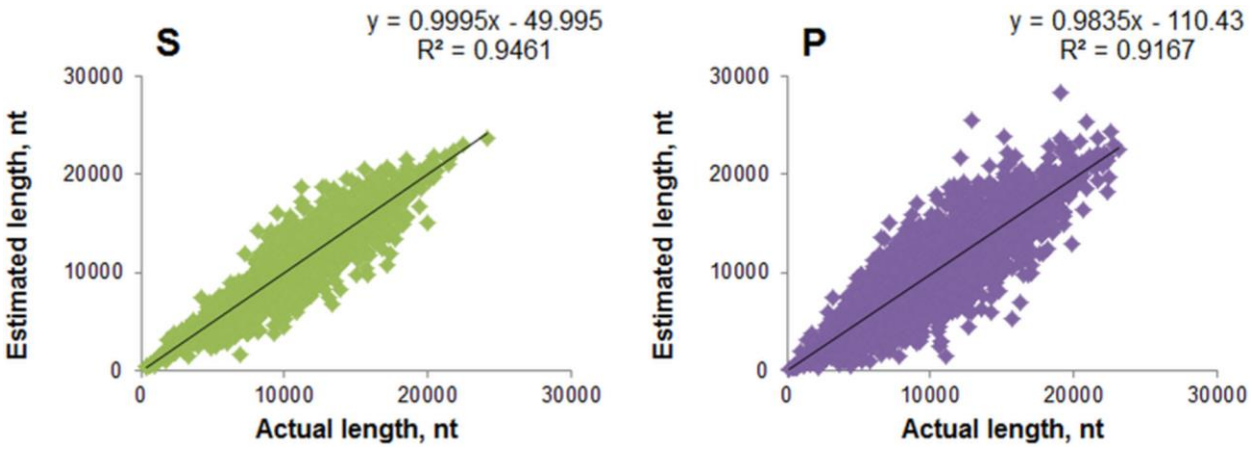


Figure 5. Correlation between actual and estimated mean telomere lengths. S—single-end reads, P—paired-end reads. Estimation of mean telomere length was performed with reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean 10 kb and SD 7 kb. The minimum-maximum range of the generated telomere lengths were: 194.5–21138 bp for single-end reads, and 387.5–24169.5 bp for paired-end reads. The read length, insert size and fold coverage ranges are described in the Methods.

Next, we compared performance of telomere length estimation based on read length, coverage and insert size (in case of paired-end reads). We have chosen the most appropriate short-read alignment parameters that would combine the maximal accuracy and speed of the alignment to the telomeric index by evaluating different alignment options offered by Bowtie 2 and comparing their performance. The alignments were performed with the Bowtie 2 preset options for *--end-to-end*

alignment, i.e. *--fast* (F), *--very-fast* (VF), *--sensitive* (S) and *--very-sensitive* (VS) modes. The argument for "mismatches in a seed alignment" (*–N*), was tested for both the default value of 0 and for the value of 1, since setting *-N* to 1 increases the alignment sensitivity. Next, the *-L* option, which is the length of the seed substrings to align during multi-seed alignment, was set to one third of read length, but with minimum value of 6 and maximum value of 22 (this means that if one third of read length is less than 6, it is assigned the value of 6, and if it is more than 22, it is set to 22). Finally, the *--n-ceil* option, which is the maximum number of allowed mismatches and 'N' bases in the alignment, was set to [*rl-min.seed*], where *rl* is the short-read length and *min.seed* is the minimum number of telomeric read bases in the mapped reads. Thus, we ended up with 8 alignment modes (F-N0, VF-N0, S-N0, VS-N0, F-N1, VF-N1, S-N1, VS-N1) and compared their effect on the accuracy of telomere length estimation. Eventually, the mean relative error (MRE), standard error (SE) of MRE, root mean squared error (RMSE), and coefficient of determination (R^2) were calculated, separately for single and paired-end reads, and compared across alignment modes.

The results showed that mean telomere length estimation was more accurate for those alignment modes, for which mismatches were allowed in the seed (F-N1, VF-N1, S-N1, VS-N1). For VS-N1, the MRE ± SE was -0.5 ± 0.17 % for single-end reads and -4.3 ± 0.01% for paired-end reads, RMSE was equal to 7% and 11%, respectively (Figure 6). Since the difference in performance among these four modes was very small and we were interested in higher sensitivity, we have chosen the VS-N1 as the default alignment option for Computel.

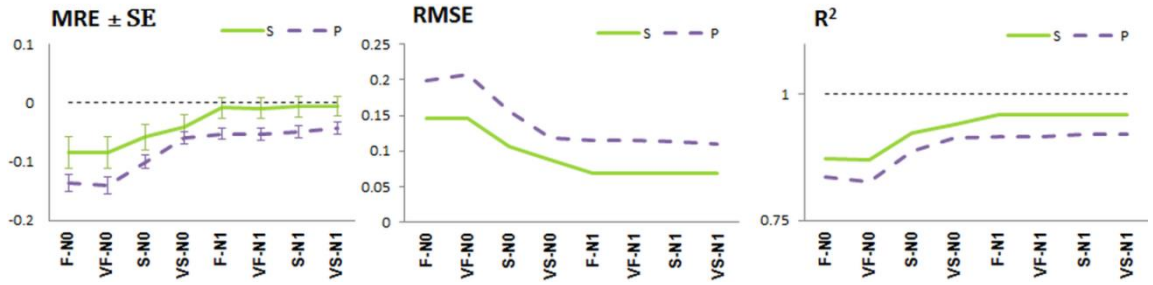


FIGURE 6. Telomere length estimation accuracy in different alignment modes. S - single-end reads, **P** - paired-end reads. Estimation of telomere length was performed with reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean and SD equal to 10 kb and 7 kb . The modes of alignment (F-N0, VF-N0, S-N0, VS-N0, F-N1, VF-N1, S-N1, VS-N1), as well as read length, insert size and fold coverage ranges are described above. MRE – mean relative error between estimated and actual mean telomere lengths, RMSE – root mean squared error

(error variance), R^2 – coefficient of determination (quality of linear fit between estimated and actual mean telomere lengths).

Comparison of performance for single-end reads

We have performed assessment of mean telomere length estimation accuracy depending on single-end read length (*rl*) and fold coverage (*fcov*), as described in Methods, Algorithm validation with synthetic data section of the manuscript. According to multi-factor ANOVA test results, there were no significant differences in mean telomere lengths depending on read length, coverage or their combinations (Table 2).

Table 2. Effects of read length and coverage on MRE of mean telomere length estimation for single-end reads.

Variable	Type III Sum of Squares	df	F	Sig.
read length (*rl*)	0.07	5	1.82	0.11
fold coverage (fcov)	0.02	6	0.48	0.83
rl-fcov interaction	0.21	30	0.95	0.54
Error	60.79	8358.00	-	-
Total	61.41	8400.00	-	-

From 42 rl-fcov combinations, in 21 cases the relative error of length estimate was not significantly different from 0, and in the rest, the calculated telomere length was underestimated by 0.2% - 0.3% (Figure 7). Meanwhile, standard errors of MRE decreased along with the increase of fold coverage. Additionally, we observed decrease of relative error variance (RMSE), as well as improvement of linear fit (R^2) with the increase in fold coverage, which is more pronounced for longer reads (Figure 7).

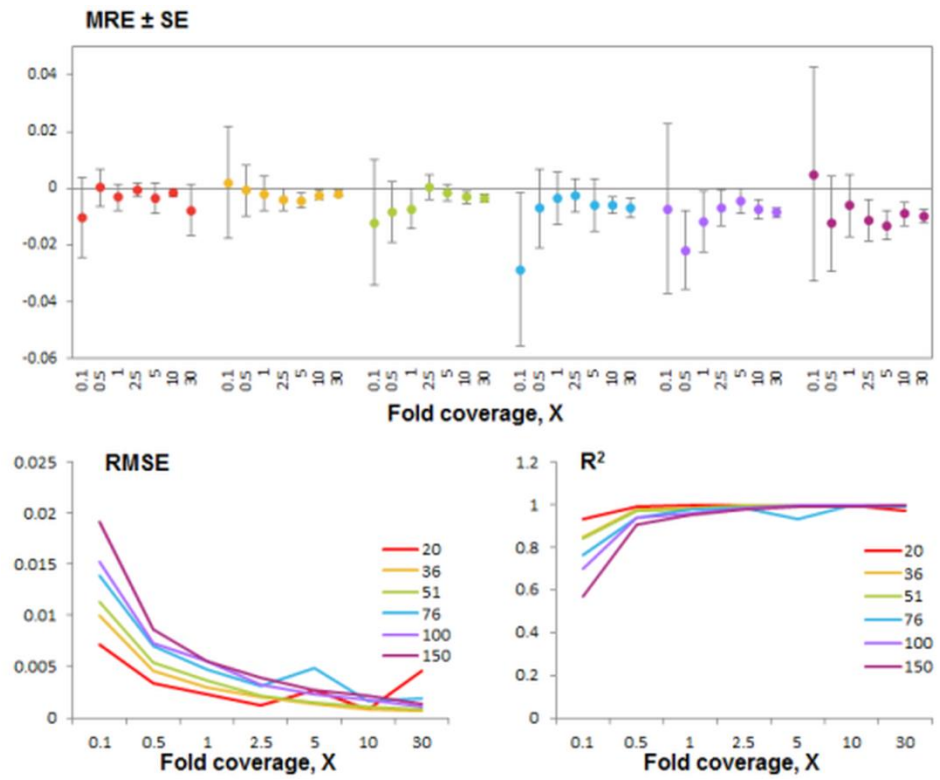


FIGURE 7. Performance metrics of mean telomere length estimation for single-end reads depending on read length and coverage. Estimation of mean telomere length was performed with single-end reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean equal to 10 kb and SD equal to 7 kb. Each read length is represented with one color. Dots on the first chart represent mean relative error (MRE) between estimated and actual mean telomere lengths; whiskers represent standard error (SE) of MRE. RMSE – root mean squared error (error variance), R2 – coefficient of determination (quality of linear fit between estimated and actual mean telomere lengths). Coverage is indicated on the x axes.

Comparison of performance for paired-end reads

We have performed similar series of analyses for paired-end short-reads, accounting also for insert size effects. Multi-factor ANOVA indicates that read length, coverage and insert size, as well as their pairwise interactions significantly influence mean relative error (MRE) of telomere length estimation in the case of paired-end reads (Table 3).

Table 3. Effects of read length and coverage on mean relative error of telomere length estimation for paired-end reads.

Variable	Type III Sum of Squares	df	F	Sig.
read length (*rl*)	1.23	5	20.13	0.00
fold coverage (*fcov*)	0.18	6	2.43	0.02
insert size (*mflen*)	1.24	2	50.65	0.00
rl-fcov interaction	0.61	30	1.67	0.01
rl-mflen interaction	0.15	7	1.75	0.09
fcov-mflen interaction	0.46	12	3.10	0.00
rl-fcov-mflen interaction	0.53	42	1.03	0.43
Error	255.88	20895.00	-	-
Total	278.60	21000.00	-	-

From the 105 combinations of read length (*rl*), fold coverage (*fcov*) and insert size (*mflen*), in 12 cases the relative error of length estimate was not significantly different from 0, the calculated telomere length was significantly overestimated in two cases (by 2% and 5.5%), and underestimated in the rest (by 0.3% - 5.8 %) (Figure 8). Similar to the case of single end reads, mean relative error (MRE) for paired-end reads decreased and quality of linear fit (R^2) improved along with increase of fold coverage (Figure 9 and 10). For paired-end reads, telomere length estimation accuracy decreased with the increase in insert size (Figure 8).

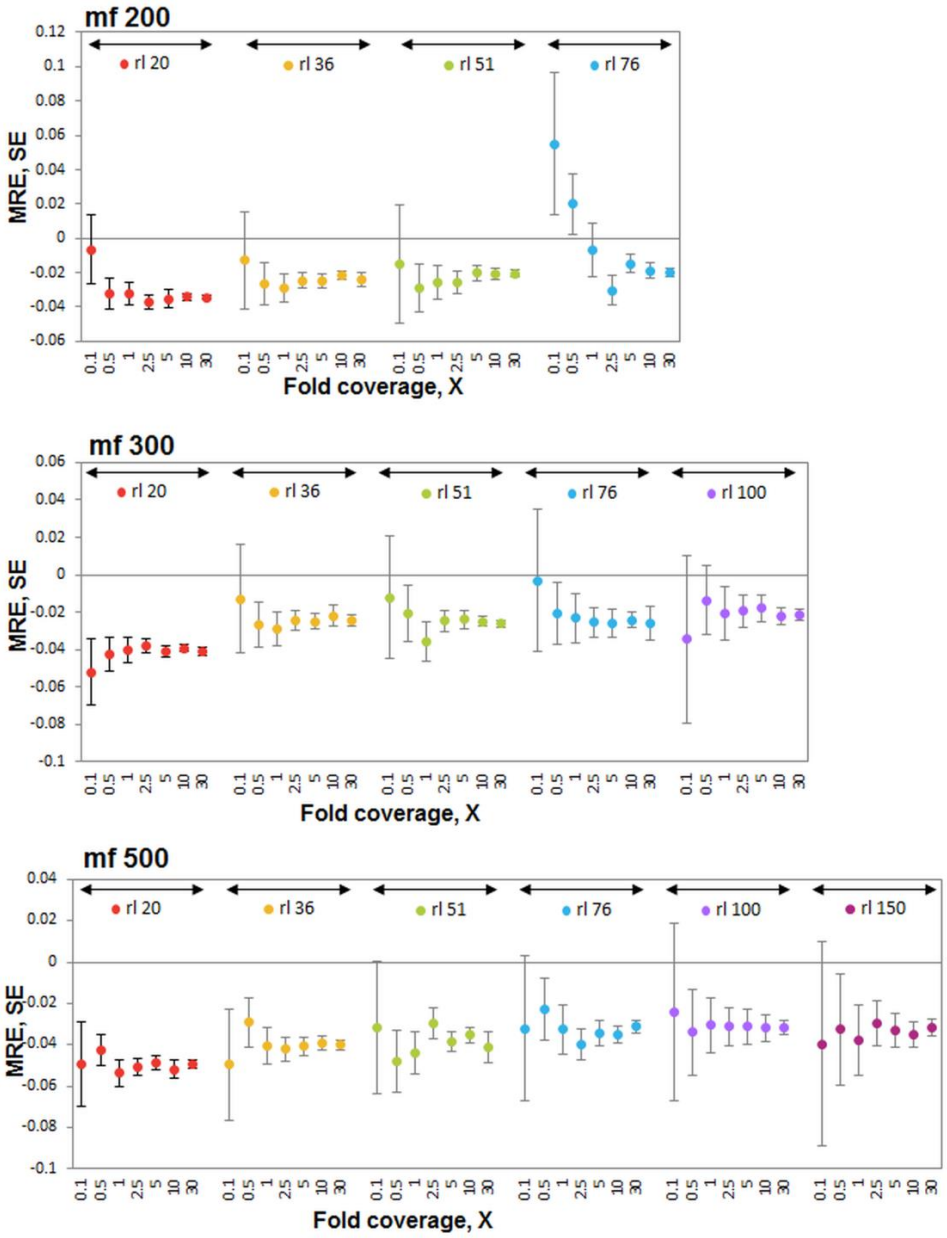


Figure 8: Performance metrics of mean telomere length estimation for paired-end reads based on read length, coverage and insert size. Estimation of telomere length was performed with paired-end reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean equal to 10 kb and SD equal to 7 kb. Insert sizes (200 nt, 300 nt and 500 nt) are indicated as chart titles. Each read length is represented with one color. Dots on the chart represent mean relative error (MRE) between estimated and actual mean telomere lengths; whiskers represent standard error (SE) of MRE. Coverage is indicated on the x axis.

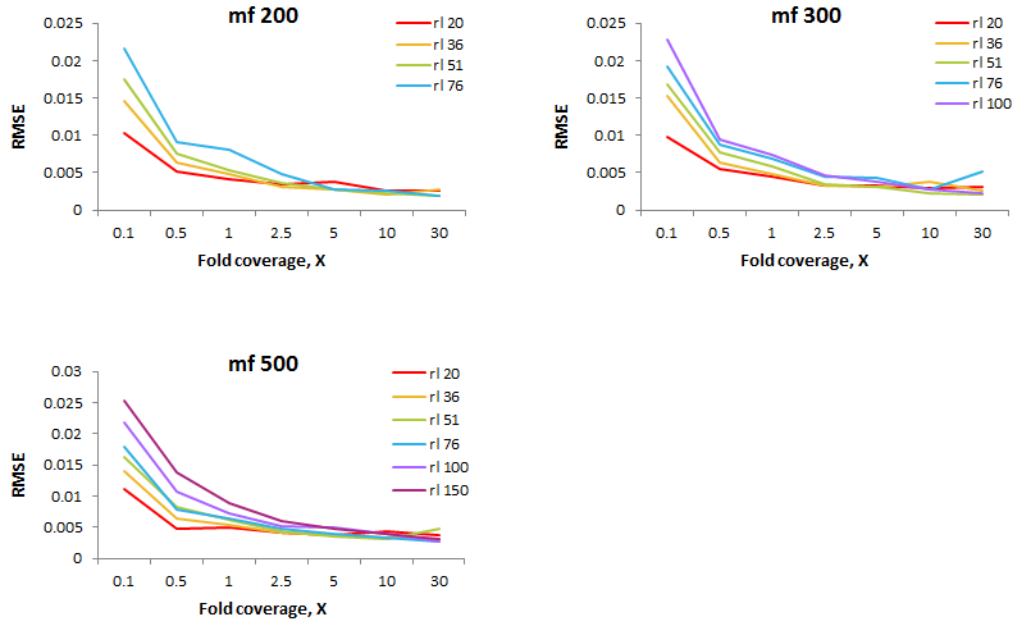


Figure 9: Root mean squared error (RMSE) of telomere length estimation for paired-end reads based on read length, coverage and insert size. Estimation of telomere length was performed with paired-end reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean equal to 10 kb and SD equal to 7 kb. Insert sizes (200 nt, 300 nt and 500 nt) are indicated as chart titles. Each read length is represented with one color. Coverage is indicated on the x axis.

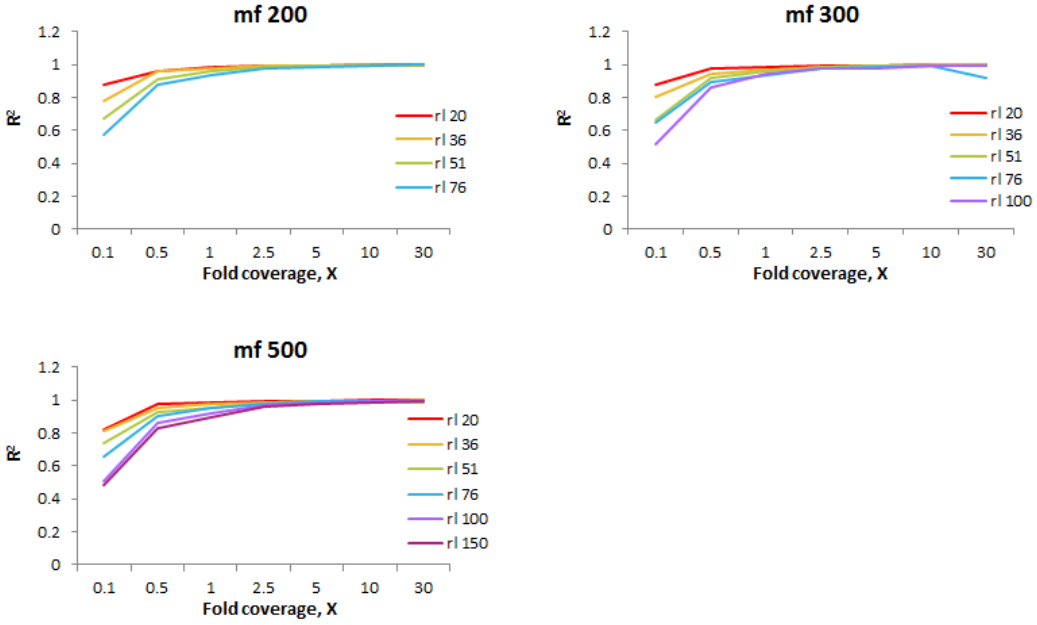


Figure 10: Quality of linear fit (R^2) of telomere length estimation for paired-end reads based on read length, coverage and insert size. Estimation of telomere length was performed with paired-end reads generated from 200 kb length region of human chromosome 1, with telomeres attached to both its ends with lengths sampled from a normal distribution with mean equal to 10 kb and SD equal to 7 kb. Insert sizes (200 nt, 300 nt and 500 nt) are indicated as chart titles. Each read length is represented with one color. Coverage is indicated on the x axis.

Performance comparison with TelSeq

We have compared the accuracy of telomere length estimation by Computel and TelSeq with short reads generated from 30 kb telomeric sequences attached to human chromosome 1 (see *Methods and data* for details) at 0.2x, 2x and 10x coverage. Computel was used with its default settings, while TelSeq source code was modified to make the computations valid for chromosome 1 (see *Methods and data*).

The results obtained indicate that Computel outperforms TelSeq in all the cases. Moreover, TelSeq fails when short-read length significantly deviates from its default value (Table 4), while the accuracy of Computel is not changed significantly in the read length ranges examined. Finally, to compare Computel and Telseq performance with an alternative short-read generation algorithm, we also used short-reads generated by the ART Illumina tool. Comparison was performed using the the same settings as described above, with 100 nt read lengths and coverage values in the range 0.2x, 2x and 10x. In this case, TelSeq significantly underestimated the actual telomere lengths, in contrast to Computel (Table 4).

Table 4. Comparison of performance of Computel and TelSeq in mean telomere length estimation from synthetic data.

Read length	Synthetic short-read generation tool	Computel mean telomere length estimate[a] mean ± SE, kb	TelSeq mean telomere length estimate[a] mean ± SE, kb
100 nt [b]	SimSeq [e]	29.2 ±0.5	28.8 ± 0.5
36 nt (k = 4) [c]	SimSeq	29.6 ± 0.4	47.2 ± 0.5
36 nt (k = 5) [c]	SimSeq	29.6 ± 0.4	47.1 ± 0.5
36 nt (k = 6) [c]	SimSeq	29.6 ± 0.4	7.8 ± 0.1
76 nt	SimSeq	29.8 ± 0.8	31.8 ± 0.8
150 nt	SimSeq	28.9 ± 0.7	NA [d]
100 nt	ART Illumina [25]	31.1 ± 0.6	24.6 ± 0.6

[a] - The actual telomere length was 30 kb attached to the Chromosome 1. [b] – The default TelSeq read length. [c] - For 36 nt read lengths, estimation of telomere length byTelSeq was performed with k (threshold of telomeric repeats in short-reads) equal to 4, 5 or 6; For all other read lengths, the default value of k = 7 was used. [d] - TelSeq fails to output results for 150 nt read length. [e] – Simseq is available at *https://github.com/jstjohn/SimSeq*.

Validation of computel using experimental data

We have computed telomere lengths with whole-genome sequences of tumor (D) - normal tissue (N) pairs from five neuroblastoma patients, using both Computel and TelSeq, with their default settings. These samples have been previously analyzed by qPCR and the log fold changes of

telomere lengths in tumor over paired healthy tissues were published [113]. In order to compare those results with the estimates obtained by Computel and TelSeq, we computed absolute values of mean telomere lengths and converted them to log2 fold change values (Table 5). The changes in telomere length predicted by Computel and TelSeq were consistent with length changes observed by qPCR (Table 6).

Table 5. Absolute values of MTL estimates for neuroblastoma samples by computel and telseq.

	Computel			TelSeq		
	Cancer tissue (D), MTL, kb	Healthy tissue (N), MTL, kb	$\log_2$(D/N)	Cancer tissue (D), MTL, kb	Healthy tissue (N), MTL, kb	$\log_2$(D/N)
SJNBL001	24.0	4.4	2.45	28.9	6.0	2.27
SJNBL002	13.8	4.5	1.61	18.7	6.1	1.61
SJNBL009	3.2	7.0	-1.12	4.4	9.9	-1.16
SJNBL030	2.8	5.4	-0.95	4.0	7.9	-1.00
SJNBL031	7.6	3.6	1.08	10.8	4.9	1.12

Table 6. Log ratios of telomere length estimates by computel and telseq compared to qPCR for five neuroblastoma (D) and matched normal tissue (N) samples.

Sample	qPCR [$\log_2$(D/N)]	Computel* [$\log_2$(D/N)]	TelSeq* [$\log_2$(D/N)]
SJNBL001	GAIN [2.89]	GAIN [2.45]	GAIN [2.27]
SJNBL002	GAIN [3.92]	GAIN [1.61]	GAIN [1.61]
SJNBL009	LOSS [-1.92]	LOSS [-1.12]	LOSS [-1.16]
SJNBL030	LOSS [-3.81]	LOSS [-0.95]	LOSS [-0.99]
SJNBL031	GAIN [5.35]	GAIN [1.22]	GAIN [1.20]

* - the absolute values of the mean telomer lengths for tumor and paired healthy tissue computed by Computel and TelSeq are presented in Table 5.

Next we used Computel to estimate telomere lengths for two osteosarcoma samples (SJOS002 and SJOS004) and compared the estimates with absolute qPCR and mTRF values [113]. Computel length estimates were partially consistent with TelSeq estimates and qPCR results, with some differences for each technique (Figure 11). In two out of the four cases, Computel estimates were closer to qPCR values than TelSeq estimates, with TelSeq estimates being closer in the other two cases.

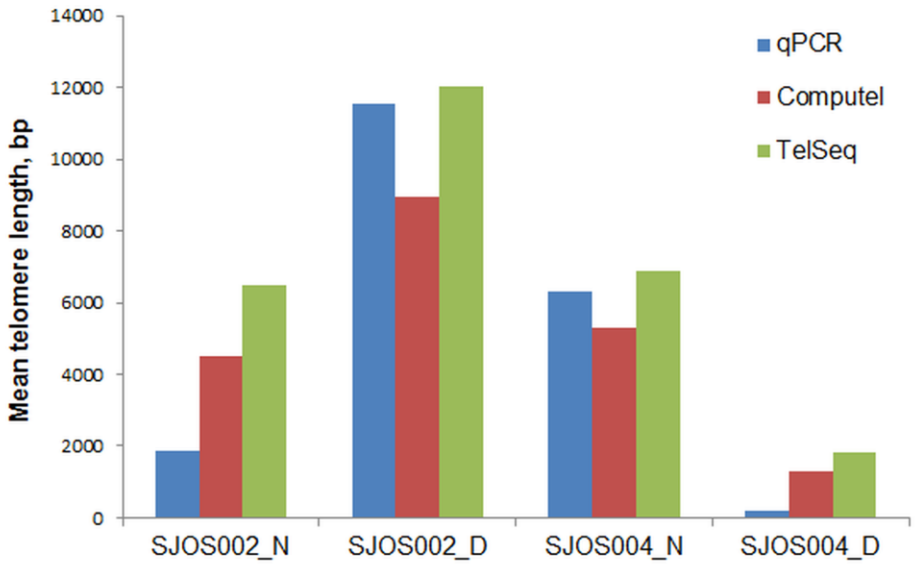


Figure 11. Mean telomere length estimates for osteosarcoma and matched normal tissues by qPCR, Computel and TelSeq. SJOS002_D, SJOS004_D - osteosarcoma tissue samples; SJOS002_N, SJOS004_N – paired healthy tissue samples.

For all the cases of telomere length estimation with experimental data, TelSeq telomere length estimates were by 2-5 kb larger than Computel length estimates. We have hypothesized that this may be the result of TelSeq capturing more reads from interstitial telomeric regions, than Computel. In order to check this, we have retrieved reads from one of the neuroblastoma sample runs (SJNBL001_D-2876158223) that Computel failed to map to the telomeric index, but that contained more than 7 telomeric repeats and were successfully captured by TelSeq. BLAST results showed that some of these reads were similar to available sequences of interstitial regions in human reference genome. The rest of the reads, however were not aligned to any known sequence, but presumably did not originate from telomeric regions, as they do not have canonical telomeric repeat patterns.

Additionally, we used SimSeq to generate short-reads (5x fold coverage) from subtelomeric 500 kb sequences of human chromosomes [121], available at http://www.wistar.org/lab/harold-c-riethman-phd/page/subtelomere-assemblies. From these short-reads, Computel mapped a total of 65 reads to the telomeric index, while TelSeq counted 327 reads. This is consistent with the hypothesis that overestimation of telomere lengths in experimental data by TelSeq compared to Computel can be partially attributed to interstitial telomeric repeats contained in the subtelomeric and other regions of chromosomes.

2.2.3 Discussion

Currently, large amounts of high-throughput NGS data for individual organisms are available [27]. Often, they contain not only WGS data, but also data from RNA-Seq, microarrays, or ChIP-Seq, which should make them valuable for associating telomere lengths with gene regulation. It is, however, difficult to calculate telomere lengths from WGS data, because a typical reference genome partially or completely lacks telomeric sequences, with chromosomal termini sometimes being denoted by runs of "N" residues. Moreover, since telomeric regions are very repetitive, traditional methods of alignment of short-reads to genomic sequence are typically confounded in this context by multiple mapping positions of the reads [23]. To overcome these limitations, we have developed the open-source software Computel, which functions by aligning short-reads to a special index, designed in such a way that only telomeric reads map to it in unique positions. Analyses have shown that Computel estimates mean telomere length with high accuracy, and its performance does not significantly depend on read length, short-read type, fold coverage and insert size.

Recently, alternative approaches have been developed for telomere length estimation from WGS data, based on count of short-reads containing certain number of telomeric repeats [113, 114, 122]. In the cases of [113] and [122], this number was fixed at 4; in case of Telseq it can vary based on read length [114]. Although TelSeq is a valuable tool, it still has some limitations that we attempted to address with Computel. First, TelSeq sets a threshold for telomeric repeat count, which makes the results of the output dependent on both the threshold and the short-read length; resulting in very poor performance, if read length considerably deviates from 100 nt (e.g., 36 nt or 150 nt). Computel, on the contrary, performs similarly well for all the short-read lengths analyzed (from 20 nt to 150 nt). Secondly, TelSeq performs relatively well on short-read data generated without reading errors in sequences; however, when sequencing errors were introduced with the ART Illumina tool, the accuracy of TelSeq results fell considerably compared to Computel. This is explained by the fact that any single error in a nucleotide sequence distorts the telomeric patterns and affects the count of telomeric reads, while the alignment approach is less sensitive to this type of errors.

An important issue concerned with NGS based telomere length estimation is the fact that there are interstitial telomeric repeats in other regions of chromosomes [118], and it is difficult to distinguish

between reads originating from these regions from true telomeric (or immediate subtelomeric) reads. The alignment-based approach utilized in Computel has the ability to reduce the number of such misclassified reads compared to TelSeq, as demonstrated with experimental data (see Results, Validation with experimental data). Notably, TelSeq underestimates telomere length in in silico experiments, where only "pure" telomeric sequences were present at chromosome ends; whereas with experimental data, where reads from interstitial telomeric sequences are presumably present, TelSeq estimates of telomere length were greater than Computel's estimates. In addition, when computing telomere lengths, Computel accounts only for the parts of the reads that have been aligned to the telomeric part of the telomeric index (index.telomeric), thus reducing the bias introduced by subtelomeric repeat-rich regions.

Finally, the hard-coded implementation of several important constants in TelSeq, such as GC-normalized genome length, and the number of chromosomes, makes this software difficult to use for analysis of telomere length of other genomes for researchers with basic programming skills, whereas all parameters of Computel can be easily set in a single configuration.

Performance assessment of Computel with experimental data have shown that telomere length estimates correlate with mean telomere length estimated with qPCR and TRF, but deviate to some extent (2-3 kb) in absolute values. In case of TRF, this difference can be attributed to the fact, that TRF also captures subtelomeric regions of chromosomes, thus overestimating telomere length by 2.5-4 kb [98]. On the other hand, estimates of absolute telomere length by qPCR, are very prone to preliminary calibration steps, therefore results obtained in different experimental settings should be compared qualitatively, rather than quantitatively [98]. It is important to note, that existing experimental methods for mean telomere length assessment all have their drawbacks and limitations [123], and, thus, cannot serve as validation methods for computational approaches, such as Computel or TelSeq. In fact, the only way to assess the accuracy of any telomere length assessment method should be based on measurements performed on a set of "artificial chromosomes" synthesized with telomeres, subtelomeric regions and interstitial telomeric repeats of known length. To our best knowledge, no experimental or computational method, including "gold standard" TRF has passed such validation. That is why validity of Computel should be recognized in terms of correlations with other measures, and not the absolute values of mean telomere length estimates.

One of the important challenges in assessment of telomere integrity is determination of telomere lengths at individual chromosomes. Computel does not allow for that, since the telomeric pattern is not chromosome specific, and it is virtually impossible to identify the chromosome source of telomeric short-reads. Currently, there are no computational methods for individual telomere length assessment (including TelSeq), nor is it measured with TRF or qPCR experiments. There are few experimental techniques (qFISH [98], chemistry based methods [124]) that allow for obtaining telomere lengths from individual chromosomes. While the data derived from these experiments is important for genome stability assessment, mean telomere length has been proven to be informative as well, and associated with various biological phenomena, such as telomere position effect [13, 15], and disease association [125–128].

Even though Computel allows for overcoming the issues described above and has relatively high accuracy, it also has a number of limitations. Its most important limitation is an inability to handle variable telomeric patterns such as those characteristic to S. cerevisiae C1-3A/TG1-3 [24]. We intend to address this in future versions of Computel. A second limitation is that alignment, the most time-consuming step in the algorithm, is performed only with Bowtie 2. In the future, we will also consider implementation with other short-read alignment programs, such as BWA [129] and SOAP [130, 131].

In summary, we have developed Computel, an open-source software package for estimating mean telomere length based on whole-genome NGS data. The overall results of performance assessment demonstrate that this methodology is valid for mean telomere length association studies based on high-throughput data. We have applied Computel to NGS data analysis to foster telomere biology research in healthy states and disease, some of which are described in this thesis (see Chapter 3 and 5).

CHAPTER 3
APPLICATIONS OF COMPUTEL: CASE STUDIES

In this chapter, we describe the findings in telomere biology obtained by applications of Computel. We first describe the association of telomere length with genetic variations, age and lifestyle of a South Asian healthy population. Next, we investigate into the regulatory relationship between telomere length and genomic, transcriptomic and epigenomic characteristics of lung cancer cell lines. These studies demonstrate the use of our methodology in bringing telomere length studies into -omics research scale.

3.1 QUANTITATIVE TRAIT ASSOCIATION STUDY FOR MEAN TELOMERE LENGTH IN THE SOUTH ASIAN GENOMES

3.1.1 Introduction

Identification of genetic and environmental factors impacting telomere length has been repeatedly addressed. From multiple loci implicated in association with telomere length, those including genes associated with telomerase were among the most validated ones. Codd *et al*. [132] conducted a meta-analysis of 37 684 individuals and reported seven loci associated with leukocyte mean telomere length (MTL), including genes coding for telomerase RNA component (*TERC*), telomerase reverse transcriptase oligonucleotide/oligosaccharide-binding fold containing 1 protein (*OBFC1*), nuclear assembly factor 1 ribonucleoprotein (*NAF1*), regulator of telomere elongation helicase (*RTEL1*), which are involved in telomere biology, and two other loci including *ACYP2* and *ZNF208* genes. These genes, apart from *ZNF208*, were validated or had supportive evidence in a study within the COGS project [133].

Furthermore, Pooley *et al*. [133] found novel telomere length association at 3p14.4 (close to *PXK*), at 6p22.1 (*ZNF311*) and at 20q11.2 (*BCL2L1*) loci. Another study of families with exceptional longevity analyzed 4289 individuals, and reported two loci (17q23.2 and 10q11.21) containing novel candidate genes, as well as validated *TERC*, *MYNN* and *OBFC1* [134]. Association of telomere length with SNPs is partially population specific, and until now there is no well-accepted genomic factor determining telomere length and telomere attrition rate. This may be partially attributed to population-specific genomic variations on one side, and lack of tools for measuring telomere length from whole genome sequencing (WGS) data, on the other.

In this study, we have performed a genome wide quantitative trait association study for MTL in South Asian population, using WGS data both for genotyping and MTL calculation.

3.1.2 DATA AND METHODS

The dataset

The study was conducted on datasets produced by the South Asian Genome project [135], deposited in European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number PRJEB5476. We used a dataset containing 2 x 101 bp WGS reads from Illumina GAIIx at 4x coverage, as well as corresponding genotypes from whole blood samples of 168 individuals. The individuals were sampled from different age groups, religious backgrounds and language groups, predominantly from India (166 India, one Kenya, one East Africa). Information on age, sex, religion and language was available for all the study subjects.

MTL calculation

MTL was calculated from 4x coverage WGS reads with Computel, using its default parameters. MTL data are presented as mean MTL 6 SD throughout the text. MTL association with age, sex and religion was evaluated using linear regression. P values <0.05 were considered significant. Multiple correction was performed with false discovery rate estimation (FDR) with Benjamini–Hochberg (BH) method. Statistical calculations were performed in R environment.

Population stratification

We have used an R package *GenABEL* [136] to analyze the population structure. For dimensionality reduction, we have performed multidimensional scaling (MDS) with 10 components using the EIGENSTRAT algorithm [137].

Association analysis

The quantitative trait association analysis for MTL was performed with *Plink* toolset [138]. SNPs with MAF < 0.1 and HWE significance threshold < 0.1 were excluded, which left us with 4, 106, 441 SNPs with genotyping rates among samples >99.93%. We used a multiple linear regression model to assess additive effect of minor alleles (0, 1 or 2 copies) on MTL for each SNP, with adjustment for age, sex, religion and top four principal components derived from population stratification analysis.

3.1.3 RESULTS

Population structure

Principal component analysis revealed that the first component was explaining 1.68% of the variability in the data, and top four components were amounting to only 5%. However, data projection over the first and the second principal components demonstrated two distinct clusters along the first component. The majority of samples were descending from India, but the exact region of birth was unknown, thus, we tried to explore the clustering based on available subject details, namely language and religion (Figure 12, Table 7). Cluster 1 contained mainly Christian, Hindu, Muslim subjects of different languages, while cluster 2 presented predominantly Punjabi speaking individuals whose religion was Hindu or Sikh, implying that these individuals were apparently originating from Punjab region. This assumption is in agreement with the sample description provided by the South Asian Genome project [135].

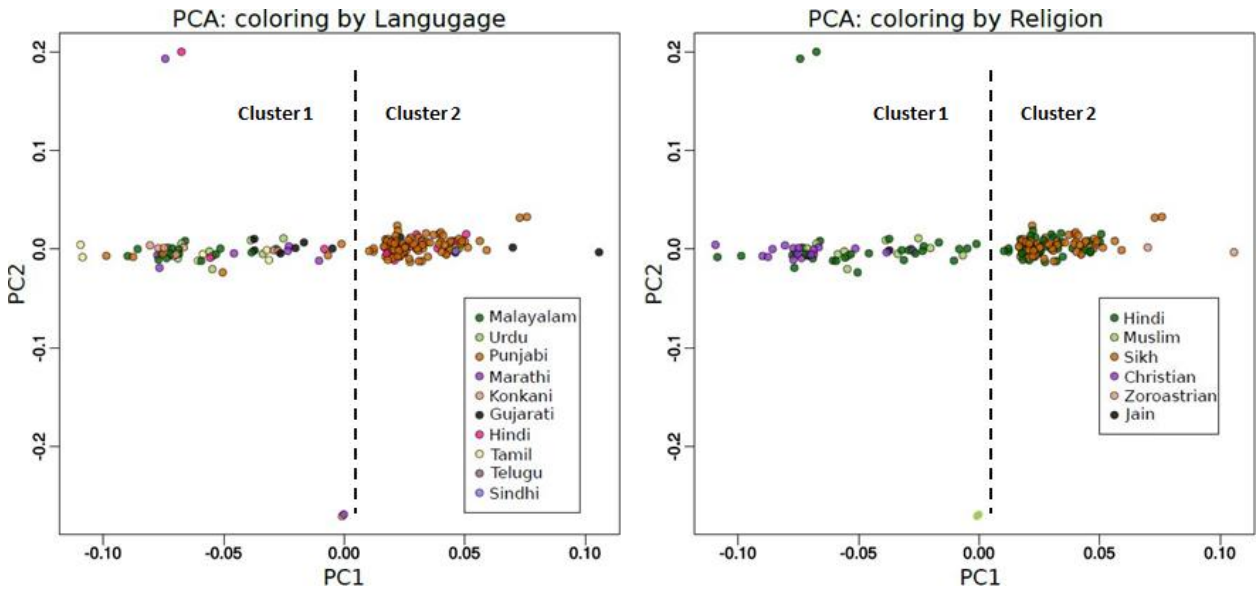


Figure 12. MDS analysis of population stratification. Projection of samples over the first (PC1) and the second (PC2) principal components (four outliers excluded). **(Left)** Distribution of samples colored by language. **(Right)** Distribution of samples colored by the region of origin. Clusters 1 and 2 depict the separation by PC1 component.

Table 7. Distribution of 168 samples from the South Asian genome by religion and language across the most discriminative principal component (see Figure 12).

Cluster 1						
Language\Religion	Christian	Hindu	Jain	Muslim	Sikh	Zoroastrian
Gujarati	0	4	1	1	0	0
Hindi	0	3	0	0	0	0
Konkani	7	1	0	1	0	0
Malayalam	8	6	0	1	0	0
Marathi	0	7	0	1	0	0
Punjabi	2	3	0	1	0	0
Sindhi	0	0	0	0	0	0
Tamil	1	5	0	1	0	0
Telugu	0	1	0	0	0	0
Urdu	0	1	0	7	0	0
Cluster 2						
Language\Religion	Christian	Hindu	Jain	Muslim	Sikh	Zoroastrian
Gujarati	0	2	0	0	0	2
Hindi	0	13	0	0	0	0
Konkani	0	0	0	0	0	0
Malayalam	0	0	0	0	0	0
Marathi	0	0	0	0	0	0
Punjabi	0	33	0	1	51	0
Sindhi	0	1	0	0	0	0
Tamil	0	0	0	0	0	0
Telugu	0	0	0	0	0	0
Urdu	0	0	0	2	0	0

MTL association with gender, age, language and religion

The MTL was calculated for all 168 samples. The values were ranging from 3234 to 8738 bp with the *mean* MTL±SD equal to 5401±1153 bp (Figure 13). Further analyses were performed on 166 individuals born exclusively in India. Linear regression of MTL with adjustment of age, sex and religion revealed no significant association with age ($P = 0.23$) and sex ($P = 0.29$), while Sikh religion was the only factor significantly affecting MTL ($P = 0.00451$). Moreover, Sikhs had significantly longer MTLs compared with the rest of the samples (mean difference 816.5 bp, 95% CI = 337.3-1295.7 bp). We then tested whether any differences in age distribution in the populations could account for deviations in MTL and no significant difference was found (two-sample *t*-test of age for Sikh versus Hindu $P = 0.3711$, Sikh versus the rest of samples $P = 0.08165$). In order to account for distinct genetic background of Sikhs as compared with other groups, we included principal components of population stratification analysis into regression

model. The results showed no significant association of MTL with Sikh religion, suggesting the religion as an indicator for genetic diversity in studied samples (Table 8). Moreover, multinomial regression model showed significant association of religion with the most discriminative PC1 component (Table 8). However, comparison of MTL in Cluster 2 revealed that Sikhs have significantly longer MTL (*mean* MTL±SD in Sikhs: 5891±1343 bp, in other samples: 5115±978 bp, *t*-test $P = 0.001$), which suggests more complex nature of influencing factors beyond genetic background (Figure 13).

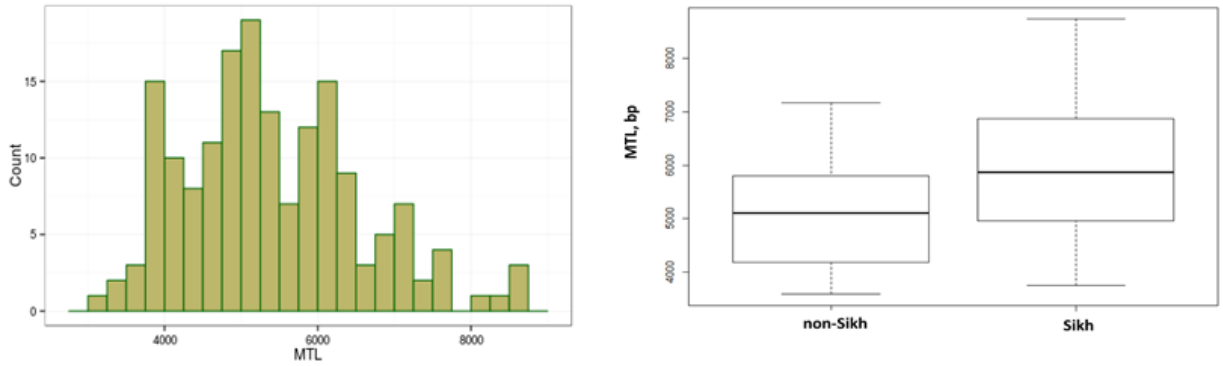


Figure 13. Mean telomere length (MTL) distribution in the South Asians. Left: MTL histogram of the 168 individuals. Right: Boxplots of MTLs in different populations of the 168 South Asian individuals.

Table 8. Multinomial regression model of association of religion with the principal components of population stratification analysis in the individuals of the South Asian Genome.

	Hindu	Jain	Muslim	Sikh	Zoroastrian
(Intercept)	0.0001	0.2710	0.0443	0.1742	0.6860
PC1	**2.31E-03**	**5.80E-01**	**1.88E-02**	**3.98E-07**	**4.77E-01**
PC2	8.00E-01	2.46E-01	5.60E-01	7.61E-01	9.96E-01
PC3	3.73E-01	1.66E-01	9.08E-01	2.27E-01	5.09E-01
PC4	7.33E-01	1.95E-01	6.27E-01	8.68E-01	7.84E-01
PC5	7.21E-02	1.63E-01	2.09E-01	**2.75E-02**	8.20E-01
PC6	6.87E-02	3.30E-01	2.38E-01	**1.87E-02**	9.84E-01
PC7	1.93E-01	4.71E-01	7.01E-01	**3.98E-02**	7.30E-01
PC8	5.15E-01	2.65E-01	9.88E-01	1.76E-01	9.54E-01
PC9	6.63E-01	4.02E-01	4.58E-01	4.25E-01	8.86E-01
PC10	2.89E-01	7.78E-01	5.61E-01	3.44E-01	9.75E-01

In order to evaluate the bias introduced by Sikh samples in MTL-age association, we excluded those and rerun the regression analysis. The resulting model (MTL= -14.46×Age + 5676.76,

R^2=0.025, *P* =0.09) became consistent with previously reported association found in South Asians, where MTL was measured with real-time PCR [139].

MTL associated loci

We analyzed 4,106,442 filtered SNPs for MTL association, with adjustment for age, sex, religion and top four principal components (λ=1.002, see Figure 14 for the QQ-plot of association values). The Manhattan plot of filtered SNPs with $P<1\times10^{-2}$ is presented in Figure 15. We used three thresholds for suggestive SNPs implicated in MTL (BH corrected $P < 0.01$, $P < 0.05$, $P < 0.2$), corresponding to 35, 42 and 51 SNPs, respectively (Table 9). The top 35 SNPs were residing in the third intron of *ADARB2* gene (10p15.3) with unadjusted $P=4.49\times10^{-8}$. Alleles in this set were in strong linkage, with allele incidence correlation $R^2 = 0.87$ for all SNP pairs. The strongest MTL association was observed in the SNP *rs1500964* ($P=1.26\times10^{-9}$). This SNP has two documented alleles (*C, T*) with *C* variant being associated with longer MTL. Heterozygous individuals (*CT*) had 812 bp longer MTL's than homozygous individuals with *T* allele ($P=7.8\times10^{-6}$). Inclusion of second *C* allele was implying an MTL increase by 572 bp, though the *P* value did not reach significance level due to small number of individuals with CC genotype ($P = 0.079$). Other significant SNPs of *ADARB2* gene had similar effects.

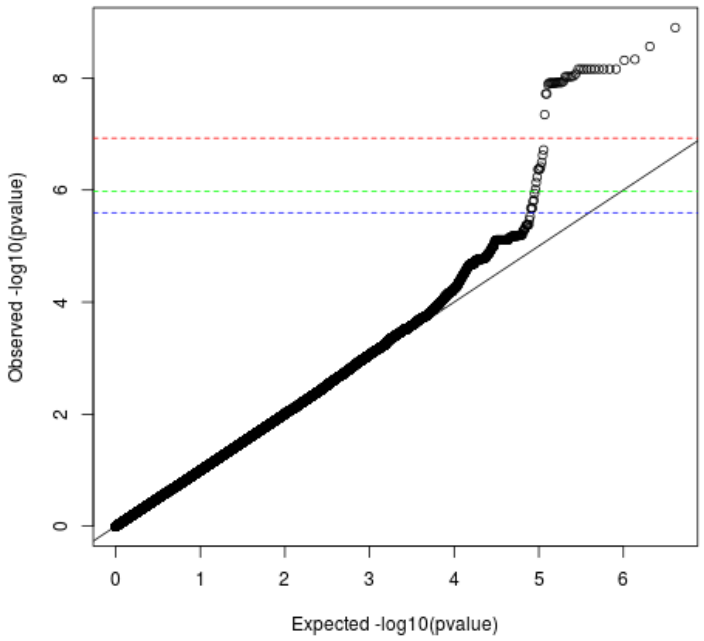


Figure 14. QQ-plot of observed versus expected p values for MTL association with SNPs in the South Asian genomes. Red, green and blue dashed lines match BH corrected 0.01, 0.05 and 0.2 thresholds.

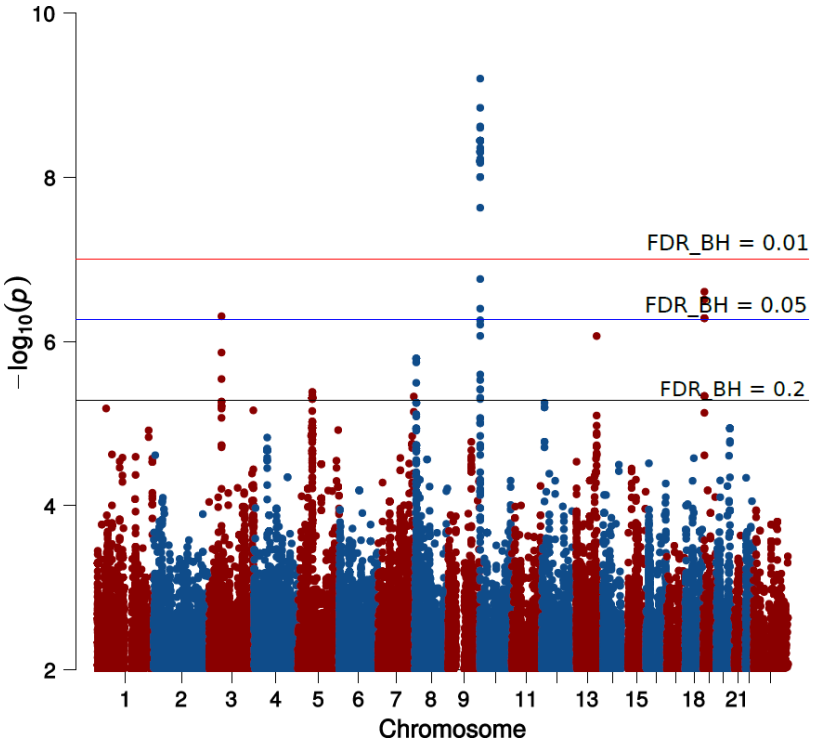


Figure 15. Manhattan plot of SNPs with MTL association in the South Asian genomes ($p < 10^{-2}$). Horizontal lines represent FDR_BH thresholds for BH corrected p values

The second significance threshold of BH adjusted P <0.05 extended the associated SNP list with another SNP from *ADARB2* gene, and six SNPs from two additional regions. Five of these SNPs were located in 19p13.3 locus, in a close proximity to tumor necrosis factor (TNF) family and TNF ligand family proteins (*TNFSF9*, *CD70*, *TNFSF14*).

The broader list of SNPs with BH adjusted P <0.2 was additionally encompassing SNPs from intergenic regions of 13q33.3, 8p23.1, loci (Table 9) and *rs7643501* ($P=4.937\times10^{-7}$) in an intronic region of CACNA2D3 gene (13p14.3), which is a putative tumor suppressor [140].

Table 9. MTL associated SNPs in the South Asian genomes. Only SNPs with BH adjusted p value < 0.2 are presented.

CHR	Position	SNP_ID	P value	Adjusted P	Chr_locus	Gene
chr10	1352078	rs1500964	1.26E-09	0.001621	p15.3	ADARB2
chr10	1342493	rs11598727	2.725E-09	0.001621	p15.3	ADARB2
chr10	1341438	rs2387646	4.637E-09	0.001621	p15.3	ADARB2
chr10	1340620	rs3849975	4.81E-09	0.001621	p15.3	ADARB2
chr10	1342318	rs61832043	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341417	rs2387645	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341262	rs2387644	6.908E-09	0.001621	p15.3	ADARB2
chr10	1343866	rs10794736	6.908E-09	0.001621	p15.3	ADARB2
chr10	1351088	rs1007147	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341493	rs2387647	6.908E-09	0.001621	p15.3	ADARB2
chr10	1342033	rs12217541	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341098	rs2387642	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341507	rs2387648	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341691	rs4471349	6.908E-09	0.001621	p15.3	ADARB2
chr10	1341183	rs2387643	8.381E-09	0.001621	p15.3	ADARB2
chr10	1340032	rs10903416	9.27E-09	0.001621	p15.3	ADARB2
chr10	1343614	rs12570479	9.438E-09	0.001621	p15.3	ADARB2
chr10	1343596	rs58991694	9.438E-09	0.001621	p15.3	ADARB2
chr10	1343591	rs56718200	9.438E-09	0.001621	p15.3	ADARB2
chr10	1343558	rs35620030	9.438E-09	0.001621	p15.3	ADARB2
chr10	1342596	rs10903417	1.146E-08	0.001621	p15.3	ADARB2
chr10	1344052	rs4880815	0.000000012	0.001621	p15.3	ADARB2
chr10	1343726	rs10903420	0.000000012	0.001621	p15.3	ADARB2
chr10	1343042	rs11593264	0.000000012	0.001621	p15.3	ADARB2
chr10	1343198	rs11594721	0.000000012	0.001621	p15.3	ADARB2
chr10	1341897	rs12217329	1.223E-08	0.001621	p15.3	ADARB2
chr10	1341842	rs4417195	1.223E-08	0.001621	p15.3	ADARB2
chr10	1341958	rs12220563	1.223E-08	0.001621	p15.3	ADARB2
chr10	1341858	rs12220555	1.223E-08	0.001621	p15.3	ADARB2

chr10	1341852	rs12220554	1.223E-08	0.001621	p15.3	ADARB2	
chr10	1341914	rs12218307	1.223E-08	0.001621	p15.3	ADARB2	
chr10	1341003	rs2387641	0.000000013	0.001673	p15.3	ADARB2	
chr10	1342819	rs10903419	1.906E-08	0.002302	p15.3	ADARB2	
chr10	1342809	rs10903418	1.906E-08	0.002302	p15.3	ADARB2	
chr10	1340503	rs3849974	4.49E-08	0.005268	p15.3	ADARB2	BH p < 0.01
chr19	6542135	rs112914484	1.913E-07	0.02182	p13.3	None	
chr19	6543496	rs8103412	2.385E-07	0.02647	p13.3	None	
chr10	1343928	rs10794737	3.211E-07	0.0347	p15.3	ADARB2	
chr13	107626059	rs2391396	3.818E-07	0.0402	q33.3	None	
chr19	6548175	rs348378	4.267E-07	0.04172	p13.3	None	
chr19	6546785	rs10407602	4.267E-07	0.04172	p13.3	None	
chr19	6547397	rs348376	4.267E-07	0.04172	p13.3	None	BH p < 0.05
chr3	54316431	rs7643501	5.855E-07	0.05591	p21.1	CACNA2D3	
chr10	1343017	rs11599315	7.291E-07	0.06805	p15.3	ADARB2	
chr10	1370899	rs11597169	0.000000972	0.0887	p15.3	ADARB2	
chr10	1339185	rs4880814	0.000001149	0.1026	p15.3	ADARB2	
chr10	1370915	rs11598750	0.000001503	0.1313	p15.3	ADARB2	
chr3	54316755	rs6799791	0.000001609	0.1376	p21.1	CACNA2D3	
chr8	8114543	rs1915986	0.000002106	0.1696	p23.1	None	
chr8	8115578	rs2945269	0.000002106	0.1696	p23.1	None	
chr8	8114141	rs2980419	0.000002106	0.1696	p23.1	None	BH p < 0.2

Table 10. MTL regression model for the South Asian genomes adjusted for principal components of population stratification analysis.

	Estimate	Std. Error	t value	P	Estimate	Std. Error	t value	P
(Intercept)	5665.73	803.72	7.05	6.24E-011	5658.42	871.53	6.49	1.39E-09
Age	-10.95	8.87	-1.23	0.219	-8.32	9.18	-0.91	0.37
Sexmale	-362.87	319.86	-1.13	0.258	-355.45	322.98	-1.10	0.27
ReligionHindu	310.01	409.76	0.76	0.451	147.15	461.40	0.32	0.75
ReligionJain	1040.24	1294.45	0.80	0.423	995.38	1333.50	0.75	0.46
ReligionMuslim	1053.63	548.44	1.92	0.057	1212.13	666.72	1.82	0.07
ReligionSikh	1128.36	453.51	2.49	0.014	832.38	535.60	1.55	0.12
ReligionZoroastrian	-364.70	1005.33	-0.36	0.717	-857.25	1210.14	-0.71	0.48
LanguageHindi	115.12	558.97	0.21	0.837	72.85	598.09	0.12	0.90
LanguageKonkani	212.25	686.08	0.31	0.757	534.39	754.86	0.71	0.48
LanguageMalayalam	466.71	599.81	0.78	0.438	930.71	651.50	1.43	0.16
LanguageMarathi	237.53	619.80	0.38	0.702	567.38	687.10	0.83	0.41
LanguagePunjabi	46.67	510.54	0.09	0.927	-88.30	553.69	-0.16	0.87
LanguageSindhi	752.21	1245.49	0.60	0.547	251.30	1269.21	0.20	0.84
LanguageTamil	-17.07	642.75	-0.03	0.979	450.71	705.29	0.64	0.52
LanguageTelugu	-394.20	1228.06	-0.32	0.749	-63.95	1291.19	-0.05	0.96
LanguageUrdu	-352.30	679.11	-0.52	0.605	-378.87	722.77	-0.52	0.60

PC1					6272.26	3969.02	1.58	0.12
PC2					2764.47	2788.04	0.99	0.32
PC3					283.32	2678.58	0.11	0.92
PC4					-1235.93	2600.94	-0.48	0.64
PC5					-360.44	3087.34	-0.12	0.91
PC6					89.56	3032.34	0.03	0.98
PC7					-231.74	3200.43	-0.07	0.94
PC8					-5297.38	3157.35	-1.68	0.10
PC9					2084.04	2993.06	0.70	0.49
PC10					7017.63	3072.14	2.28	0.02

3.1.4 Discussion

Our findings revealed that MTL in the studied population is not correlated with age, which is in agreement with previously published results on South Asian population, where quantitative PCR was used to measure relative telomere length [139]. These results deviate from other studies on different populations (samples recruited from Austria, France, China and Denmark) that showed strong correlation of MTL with age [141–144] thus suggesting that the association might be population specific [145, 146].

Surprisingly, we have observed that Sikhs have significantly longer telomeres compared with the rest of the samples. The results of regression analysis with adjustment for genetic background suggested about possibility of complex influence of genetic background and environmental factors. It is worth noting that smoking, drug taking and using tobacco are banned in Sikhism, alcohol is rarely consumed and many Sikhs are lifelong vegetarians. Unfortunately, we could not assess the effects of these factors due to lack of details regarding lifestyle of sampled individuals. However, the impact of oxidative stress, including smoking, and lifestyle on telomere length was repeatedly investigated, indicating that telomere length is negatively affected by smoking, while the impact of healthy lifestyle is positive [147] Additional investigations are needed to assess whether long telomeres are characteristic to Sikh population, and are preserved due to high heritable nature of telomere lengths [148].

The genome wide association (GWA) scan of the South Asian population revealed 51 SNPs associated with MTL. Among these, the most significant ones were residing in *ADARB2* gene. This is an RNA editing gene of double-stranded RNA adenosine deaminase family. [149] detected 10 SNPs in *ADARB2* gene, strongly associated with extreme longevity. Two of those SNPs, *rs10903420* ($P=1.199\times10^{-8}$) and *rs1007147* ($P=6.908\times10^{-9}$) were among the most implicated

SNPs in MTL, reported in our study. Additional three SNPs (*rs2805562*, *rs884949* and *rs2805535*) had small association *P* values, but did not reach the significance threshold. In total, five SNPs of *ADARB2*, associated with extreme old age, are also supposedly associated with MTL. These findings call for further investigation aimed at understanding molecular mechanisms through which *ADARB2* is involved in telomere length regulation and longevity.

At this point of discussion it is worthwhile to identify several limitations of our study. The first and the most important limitation in GWA analysis was the small sample size, which we had access to. This dramatically reduced the GWA power to detect associated SNPs, thus, many functionally relevant SNPs did not reach the significance threshold after multiple correction. Further, we had limited information about the samples: BMI, smoking habits and other lifestyle details could serve for adjustment in the association analysis phase and give some important insights over the MTL aberrations. The next limitation was the inhomogeneity of the studied population. Even though sampled individuals were descending from South Asia, the genetic footprints of different religious groups were considerably diverse.

3.1.5 Conclusion

Here, we have presented a pilot study of GWA for MTL in South Asians, where we exploited WGS data for SNP information and for MTL calculation. Concordance of certain findings with previously published results validated our approach and demonstrated that the usage of WGS data can be extended to be utilized in telomere studies. This eliminates the necessity of conducting additional experiments for telomere length measurement, greatly facilitating further research. Moreover, there is large amount of already existing WGS data focused on age-related diseases and cancers, where telomeres play an important role, and our approach can be used to exploit available datasets to enrich the results with telomere length data. Our study showed that Sikhs are distinguished with longer telomeres compared with other religious groups from South Asia. This phenomenon needs further investigation in order to assess the involvement of genetic and/or environmental factors on telomere length dynamics. Moreover, our results suggest that not only telomere length, but also its association with age can be affected by ethnicity. Finally, we have identified that *ADARB2* gene highly impacts telomere length in South Asians. Longevity-related nature of this gene is characterized in other populations, thus combination of this information with

our results calls for more investigation to understand the role of this gene in telomere regulation and ageing in general.

3.2 LUNG ADENOCARCINOMA: ANALYSIS OF TELOMERE-ASSOCIATED GENES AND PATHWAYS IN LUNG ADENOCARCINOMA CELL LINES

3.2.1 Introduction

Several studies have addressed the question of how telomere length is involved in initiation and progression of cancers. So far the evidence has been conflicting, with some studies supporting the association of short telomeres in peripheral blood leukocytes (PBL) [150], some showing no association [151], and others implicating longer telomeres in PBL with cancer risk [152]. A few studies suggest that while telomere length is not associated with cancer risk, longer telomeres may be implicated in poor cancer prognosis [153]. The study by Rode *et al* has identified that genetically determined long telomeres are implicated in cancer mortality, and have suggested that the previously reported association of short telomeres with cancer mortality is the result of confounding factors that generally lead to increase in overall mortality rates [154]. They have studied mutations in *TERT*, *TERC* and *OBFC1* genes to be the main determinants linked to the ability of cancerous cells to survive, and have established that lung cancers and melanoma are the main cancer types affected by these variations [154]. Importantly, all of these studies have measured telomere length in PBL using quantitative PCR methods.

All in all, there is a non-trivial link between telomere length, cancer initiation, progression and mortality. The results of studies addressing these associations should be treated with caution: primarily paying attention to the tissue where the telomere length has been measured (in the tumor or in the blood), and also to other factors, such as the cancer type, the study design, the method for telomere length measurement, adjustment for confounding factors, etc. In this sense, utilization of NGS data give an opportunity to measure telomere length and assess genomic, transcriptomic and epigenomic status of cells derived from the same tumor tissue, thus, highly increasing the accuracy of telomere-related studies.

In this thesis, we have addressed the association of telomere length dynamics with genetic and epigenetic factors in lung adenocoarcinoma cell lines using Computel. Lung adenocarcinoma is the most common form lung cancers, originating in the peripheral lung tissue. It is described by a

number of genetic alterations including mutations in TP53, CDKN2A, KRAS and EGFR genes, as well as global changes in epigenetic modifications and gene expression profiles [155, 156]. Long telomeres have recently been linked to higher risk of lung adenocarcinoma, as identified by a large cohort study including several other cancer types [157]. On the other hand, another study has linked short telomeres with poor prognosis of non-small cell lung cancers in general [158]. However, the exact role of telomere length dynamics in the progression and pathomechanisms of lung adenocarcinoma has not been extensively studied yet.

3.2.2 Data and methods

Data

The data used in this study has previously been collected by Suzuki et al [159]. They had generated whole genome DNA-seq, RNA-seq, ChIP-seq and Bisulfite-seq data from 26 lung adenocarcinoma cell lines and made those publicly available at DNA Data Bank of Japan (DDBJ), as well as at http://dbtss.hgc.jp/. 13 of the cell lines had Japanese origin, 9 were from Caucasians, 1 Black and two were of unknown origin.

Sequencing was performed using Illumina HiSeq platforms. WGS data was obtained under 10x coverage, with 100 bp paired-end reads (Accession numbers: DRA001859 and DRA001858). Gene expression data was available as an RPKM table, generated by the authors, from RNA-seq data (Accession numbers: DRA001846). RNA-seq data was also available for one cell line from the lung of a healthy individual (SAEC, accession number: DRA002311). The ChIP-seq data was obtained for RNA Polymerase II, H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9/14ac, and H3K27ac histone marks (Accession numbers: DRA001860 and DRA002311). DNA methylation was analyzed by Bisulfite sequencing (Accession number: DRA001841) [159].

Mean telomere length calculation

Raw whole genome sequencing data was supplied to Computel v1.01 for MTL estimation. The program was run with its default parameters, with base coverage roughly estimated from the number of reads in the Fastq files (10x).

Partial correlation analysis

For assessment of direct correlation of gene expression with MTL, we have used partial correlation analysis with the R package GeneNet https://cran.r-project.org/web/packages/GeneNet/index.html. Partial correlation estimates the direct correlation between two variables, accounting for a set of controlling variables. It does so by estimating the correlation between each of the variables and the control variables, and then estimating the correlation between the obtained residuals. GeneNet implements the methodology of Schäfer *et al*, where they have inferred large scale gene association networks by estimating partial correlation for each pair of genes, accounting for the rest of the genes in the studied set [160]. After estimation of partial correlations, GeneNet constructs a directed network between the genes. An edge direction corresponds to the direction of the smaller regression coefficient, and is chosen in a way that it points from the node with larger standardized partial variance (less explained or independent variable), to that of the lower (well explained or "dependent"). This means that the direction of the edges imposes causality [161].

We have taken the normalized log2 (0.1+RPKM) values and bounded those to the normalized log2 telomere lengths of the 26 cell lines. The MTL values and the gene expression values were treated together, as a single set of features. We have performed pairwise partial correlation analysis by GeneNet and constructed a network by filtering out edges with correlation values less than 0.1. We have then selected the node corresponding to telomere length (MTL node), along with the genes directly correlated with MTL.

Combined correlation analysis of gene expression and epigenetic modifications

The ChIP-seq sequencing reads for RNA Polymerase II, and H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9/14ac, and H3K27ac histone marks were mapped to the human reference genome GRCh38/hg38. The ChIP and Input read counts in regions 2000 bp up- and downstream of the gene transcription start sites were analyzed with R package edgeR [162]. The ChIP and Input groups were treated as different experimental conditions. We have filtered the genes for which a ChIP-seq value was available for at least half of the cell lines. We have also filtered for the H3K27Ac, H3K36me3, H3K4me1, H3K4me3 and H3K9_14Ac histone marks, in order to maximize the number of selected genes. The filtering, overall, left us with 12245 genes and five histone marks.

MTL (L) association with gene expression (E), DNA methylation (M) and ChIP-seq was assessed with multivariate linear regression approach, accounting for the Japanese and non-Japanese (that was largely European) origin of the cell lines (O). The three separate linear regression models were set as:

(1) L ~ M + O and M ~ L + O

(2) L ~ E + O and E ~ L + O

(3) L ~ H3K27Ac + H3K36me3 + H3K4me1 + H3K4me3 +H3K9_14Ac

In (1) and (2) regression models, we consider the dependence of telomere length on methylation (or gene expression), as well as the dependence of methylation (or gene expression) on telomere length, accounting for cell line origin. We have considered only the genes where regression was significant for both dependence types. Such a two-directional dependence couldn't be considered in the (3) model, since the number of histone marks would impose an unreasonably big number of possible models.

Based on these models, we have got three lists of genes, where MTL (L) was significantly associated with gene expression (E), DNA methylation (M) or histone modifications (H). We have then taken pairwise overlaps of these lists based on gene names. We have reasoned that if a gene's expression is associated with MTL and at the same time with DNA methylation at its promoter, this implies a possible mechanistic link between MTL, DNA methylation and gene expression: MTL might regulate gene expression through DNA methylation, or differential gene expression governed by DNA methylation changes might regulate telomere length. A similar reasoning applies for the L~E and L~H lists, where an overlap would identify genes, potentially associated with telomere length via histone modification changes.

3.2.3 Results

The MTL computed from WGS data on the 26 cell lines ranged from very short (1.7 kb) to abnormally long ones (33 kb) (Figure 17). Most of the cell lines had MTL in the range 2-6 kb, with median MTL being 4.4 kb. The distribution was skewed to the right, with eight cell lines possessing more than 10 kb MTL.

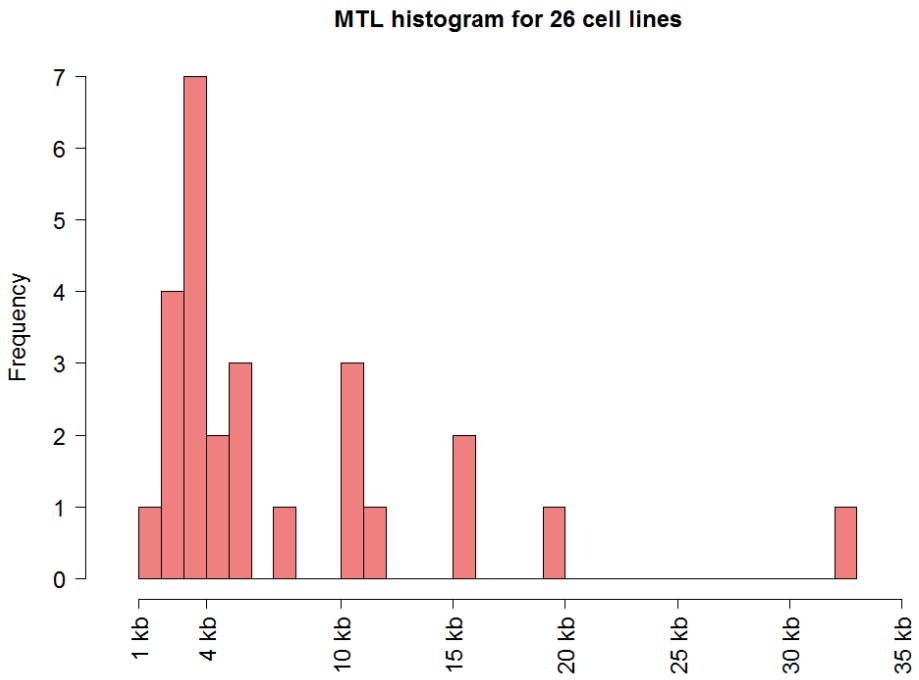
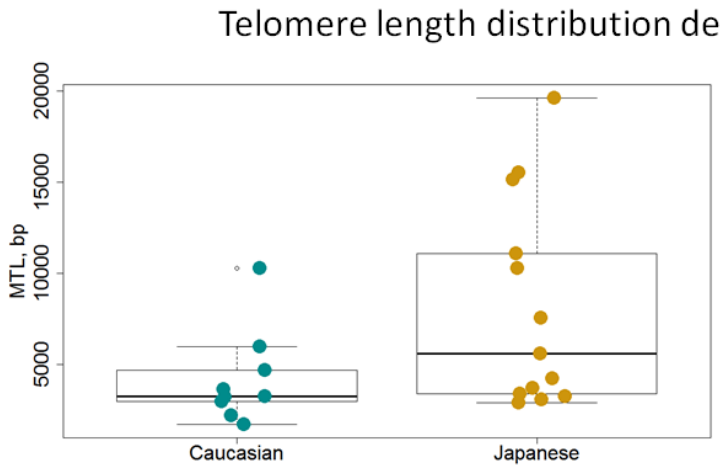


Figure 16. Distribution of mean telomere lengths (MTL) in 26 lung adenocarcinoma cell lines. MTL was measured with Computel from WGS data. The median MTL is 4.4 kb, the minimum is 1.7 kb (H2347), and the maximum is 33 kb (H1703).

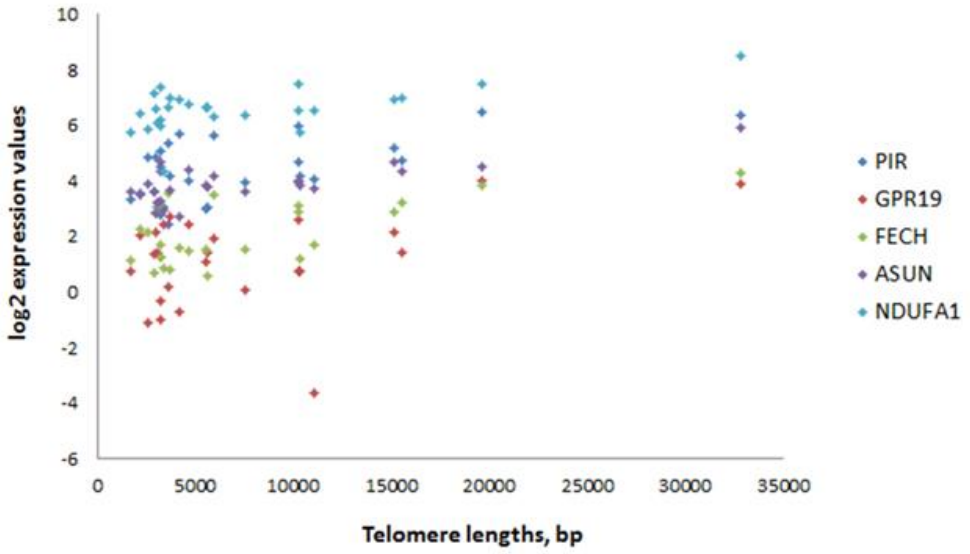
The distribution of telomere lengths varied between the Japanese and Caucasian cell lines. After the H1703 cell line with extremely long (32 kb) telomeres was removed, the mean ± sd of the MTL became 8096±5707 bp for Japanese, and 4212±2606 bp for Caucasians. The mean difference of 3884 bp between Japanese and Caucasians was significant, according to Welch t-test (p value = 0.0454). Notably, the telomeres in the Japanese cell lines were more widely distributed compared to Caucasians (Figure 17).



Cell line	MTL	Origin	Cell line	MTL	Origin
H2347	1696	Caucasian	LC2/ad	2887	Japanese
H1650	2202	Caucasian	PC-7	3065	Japanese
H1819	2976	Caucasian	RERF-LC-Ad1	3244	Japanese
A427	3218	Caucasian	RERF-LC-Ad2	3383	Japanese
H2126	3247	Caucasian	VMRC-LCD	3691	Japanese
H1437	3624	Caucasian	RERF-LC-MS	4210	Japanese
H1299	4685	Caucasian	RERF-LC-KJ	5599	Japanese
H322	5981	Caucasian	RERF-LC-OK	7531	Japanese
A549	10282	Caucasian	II-18	10265	Japanese
H1703	32844	Caucasian	ABC-1	11074	Japanese
H1648	10357	Black	PC-3	15146	Japanese
H2228	2615	Unknown	PC-14	15527	Japanese
H1975	5561	Unknown	PC-9	19636	Japanese

Figure 17. The distribution of mean telomere lengths (MTL) in cell lines of Caucasian and Japanese origin. The cell line H1703 of Caucasian origin was not accounted, since it had abnormally long telomeres (33 kb). The mean±sd for the Caucasian group was 4212±2606 bp, and 8096±5707 for the Japanese group. The difference between the means is statistically significant (Welch t-test p value = 0.0445).

Partial correlation analysis on the gene expression values and MTL has revealed five genes to be in direct correlation with telomere length (Figure 18): *PIR* (p = 0.008), *GPR19* (p = 0.013), *FECH* (p = 0.015), *ASUN* (p = 0.027), and *NDUFA1* (p = 0.035). The direction of these associations in all the five cases was from the MTL node to the genes. This means that MTL had bigger partial variance compared to the target nodes, and according to Opgen-Rhein *et al* [161] implies that telomere length "regulates" the expression of these genes.



Gene	Gene description	Adjusted p value
PIR	pirin (iron-binding nuclear protein)	0.00755
GPR19	G protein-coupled receptor 19	0.01260
FECH	ferrochelatase (protoporphyria)	0.01449
ASUN	Asunder Spermatogenesis Regulator	0.02686
NDUFA1	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa	0.03514

Figure 18. Genes directly correlated with telomere length. Top: the correlation of log2 gene expression values with MTL are shown: each gene has its color. **Bottom:** the list of genes and the p values of partial correlation of each gene with MTL.

Furthermore, we have explored the nodes adjacent to the five genes in the same partial correlation network (Figure 19). Gene set enrichment analysis with the program DAVID (https://david.ncifcrf.gov/) with the genes involved in this network revealed that 7 of those (*BOK, NDUFA1, NDUFA10, ADCY10, METAP1D, SGK1*) are associated with mitochondrion and are involved in mitochondrial oxidation, 15 have nuclear localization, of which 7 (*CITED2, EID3, ETS1, MAPK11, PIR, ZNF131, ZNF488*) are involved in regulation of transcription. However, none of these enrichment clusters had significant BH corrected p-values.

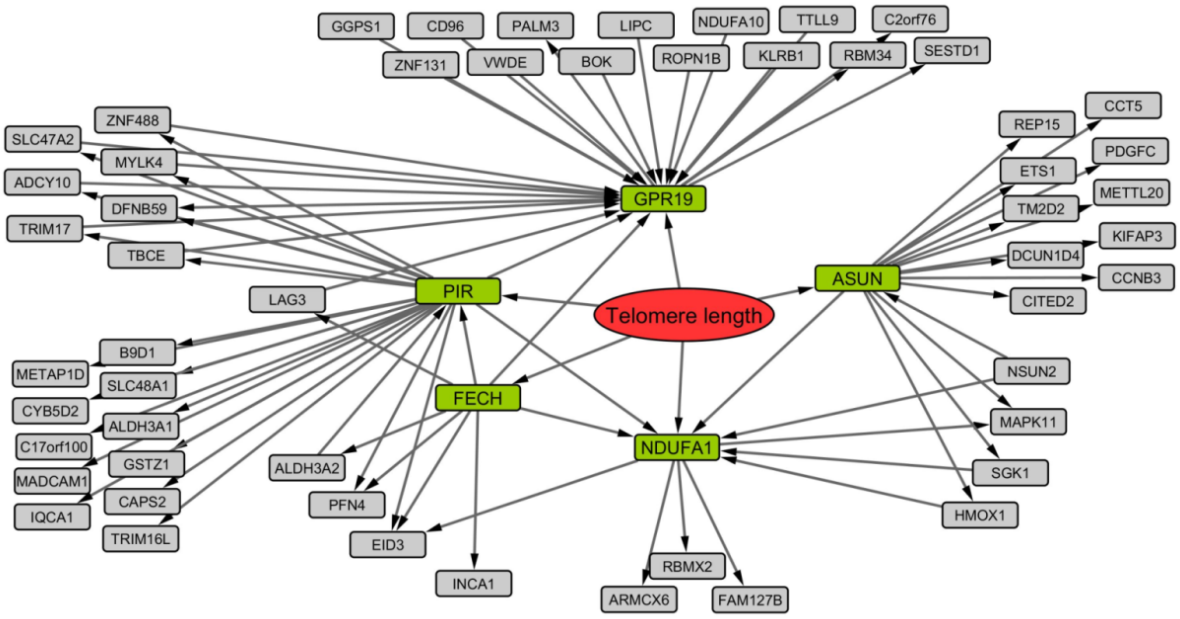


Figure 19. The directed network of partial correlations of gene expression values and MTL. The MTL node is highlighted with red (labeled: Telomere length). The nodes for the five genes directly correlated with MTL are colored in green. A adjacent nodes are colored in gray. The delta shaped targets of the edges indicate the direction of causality of the associations.

The next set of analysis was aimed at identification of genes whose epigenetic modifications and expression were concurrently correlated with telomere length. For this, we have run multivariate linear regression to identify genes, for which MTL was associated with either gene expression, or DNA methylation, or modifications of five histone marks (H3K27Ac, H3K36me3, H3K4me1, H3K4me3, H3K9_14Ac). As a result, we have obtained lists of 809, 609, and 104 genes, where MTL was separately associated with gene expression, methylation and modification of at least one histone mark, respectively. Intersections of those lists revealed 20 genes that had both expression and methylation marks, while only two genes that had both histone modification and gene expression marks associated with MTL (Table 11). In order to assess the significance of these intersections, we have performed 1000 cycles of bootstrapping randomly permuting the telomere lengths and cell line origins. The p-value for the 20 genes with simultaneous MTL correlations with methylation and expression was 0.25, and 0.6 for the case of FAM84B, where modification for only one histone mark was concordantly correlated with gene expression and MTL. The only statistically significant result was obtained for the *VPS37B* gene, where the p-value for finding overlapping association between two histone marks, gene expression and MTL was 0.001.

Table 11. Estimates of MTL association with gene expresion and histone marks (TOP) or DNA methylation (BOTTOM). The values represtent the estimates of slope of each respective linear regression model. Each of the individual associations is statistically significant ($p < 0.05$). The overlap sizes for methylation and single histone marks were not statistically significant ($p = 0.25$, $p = 0.61$, respectively), and only the probability of the overlapping association between MTL and two histone marks and gene expression for VPS37B is significant ($p = 0.001$). Positive and negative slope estimates point on positive and negative correlations, respectively.

Slope Estimates of MTL association with gene expression and histone marks

	Expression	H3K27ac	H3K36me3
VPS37B	-0.3	-0.5	-10
FAM84B	-0.95	-0.6	

Slope estimates of MTL correlation with gene expression and DNA methylation

Gene	Expression	Methylation	Gene	Expression	Methylation
RPS12	0.68	10.00	**FARSA**	0.77	-4.21
METTL16	1.28	7.75	**OR1B1**	40.04	-4.59
SLC25A29	-0.60	5.15	**PRPS1L1**	1.60	-6.27
SF3B5	0.83	5.12	**PLEKHA6**	-0.28	-6.61
SEPT6	0.56	3.34	**FEM1C**	-0.65	-6.81
DERL3	-0.53	1.94	**TM4SF18**	0.28	-7.42
HAT1	0.80	-2.08	**RWDD1**	0.77	-8.04
TXNRD1	0.49	-2.17	**GATSL3**	0.78	-9.08
PLXNA3	-0.49	-3.10	**MRPS2**	0.97	-11.22
SPATA9	-0.94	-4.10	**TPI1P2**	1.02	-16.57

3.2.4 Discussion

Lung adenocarcinoma is a complex disease with multiple changes observed at genomic, transcriptomic and epigenomic levels. We have tried to apply two different methodologies to understand which fraction of these changes might be associated with telomere length dynamics. In theory, association of gene expression/epigenetic modifications with telomere length may be caused by regulatory effects of telomeres on gene expression (TPE, TPE-OLD); or on the contrary: by genes that regulate telomere maintenance and lengthening mechanisms; or it could be caused by a confounding factor (indirect associations), as depicted in Figure 20.

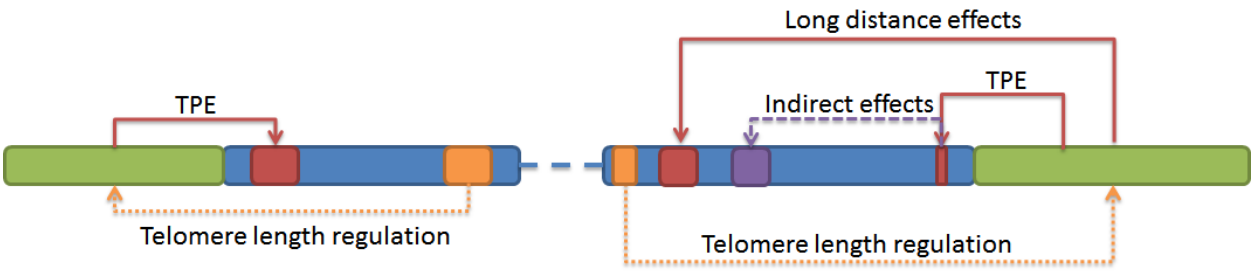


Figure 20. Causal links possibly leading to association of telomere length with gene expression changes. A single chromosome is represented, with telomeres colored in green, and the centromere with dashed line. Red, orange and violet rectangles point on genes or genetic loci. Such associations may take place within a single chromosome or between chromosomes.

Identification of directly correlated features from high dimensional data is a challenging task. The main issue with standard correlation analysis followed by multiple test corrections is the effect of confounders that conceal the features that have direct causal relationship. We have faced the same challenge in studying the association of telomere length with gene expression. Here we have applied two approaches to tackle the issue. First, we have constructed the partial correlation network of MTL associated genes (Figure 19), and identified five genes, PIR, *GPR19*, *ASUN*, *FECH* and *NDUFA*, directly associated with MTL. The network has also revealed that the direction of causality was from MTL to those genes. The pirin protein (*PIR*) is a transcription factor, and is involved in apoptosis. It is known that pirin plays a role in lung diseases, and cancers, where it regulates epithelial to mesenchymal transition in metastasis by helping the cells to overcome the senescence barrier [163]. *GPR19* is usually overexpressed in lung cancers, and it's been shown that it accelerates G2-M cell cycle transition in lung cancer cells [164]. *ASUN* is a critical regulator of cell cycle and division [165]. *FECH* and *NDUFA1* play a role in mitochondrial processes, and their function in lung cancers has not been widely investigated, however it is known that mitochondria are closely implicated in cellular ageing. Interestingly, several finding have also implicated telomere attrition with mitochondrial malfunction [166, 167].

Our next approach relies on the reasoning that if an identified association between gene expression and telomere length is conditioned by a mechanistic link, gene expression changes will most probably be accompanied by epigenetic changes at their promoters. For this we have looked at the genes whose expression and either DNA methylation or histone modifications were simultaneously correlated with telomere length. It should be noted that this approach aims at identifying genes epigenetically regulated by or regulating telomere length dynamics, and that it

will fail to reveal cases where an association is mediated by altered levels of a transcription factor, or posttranscriptional and posttranslational modifications.

We have identified that the expression of the vacuolar protein sorting 37 homolog B (S. cerevisiae) (*VPS37B*) gene and H3K27Ac and H3K36me3 modifications at its promoter were both negatively correlated with MTL. This gene has previously been reported to be differentially regulated in aging, and to encode a protein that is a part of ESCRT-I complex, which is a regulator of telomere lengthening processes in yeast [168]. This makes this gene a potential target for future investigations in mammalian cells as well.

The observed association of VPS37B was statistically significant, as the probability of observing simultaneous correlation of MTL with two histone marks and gene expression was 0.0001, according to bootstrapping results. The rest of the results, however, were not statistically significant, since overlaps of these sizes (1 for histone marks, 20 for methylations) are expected to be obtained by chance. However, it should be noted that the high-dimensionality and the small sample size of our dataset (24000 genes versus 26 samples) considerably decrease the statistical power of the test. Therefore, we find it reasonable to address the biological significance of the obtained gene lists, despite the high p-values.

An interesting association has been identified for the *PRPS1L1* gene. Its expression is positively correlated with MTL, while DNA methylation at its promoter is negatively correlated with MTL. This gene is specifically expressed in testis and its product is similar to phosphoribosylpyrophosphate synthetase enzyme that catalyzes conversion of nucleotides to mono-nucleotides [169]. A study by Giannone *et al* [170] has identified that PRPS1L1 protein interacts with TRF2, one of the shelterin proteins, which binds telomeric DNA. The effect of this interaction and the role of PRPS1L1 in general are still not known.

PLXNA3 is also linked to MTL with association of its expression and DNA methylation changes. This gene encodes Plexin-A3 protein that participates in cytoskeletal remodeling, tumor progression, apoptosis and cell differentiation. It plays an important role in axon pathfinding during development of nervous system in zebrafish [171] and mice [172]. *PLXNA3* is located only 1.5 Mb away from the end of the q arm of chromosome X. This, and the association with telomere

length, makes this gene a potential target for investigation of existence of TPE-OLD mechanisms at this locus.

FARSA encodes a tRNA synthetase that adds amino acids to tRNAs and plays an important role in mitochondrial translation processes. We have found that its expression and DNA methylation are also associated with MTL. Interestingly, this is the first tRNA synthetase known to be expressed in differentiation dependent manner. Interestingly, another tRNA synthetase, *FARS2*, has recently been identified to be in close contact with telomere end of the chromosome 6 and whose expression is associated with telomere length [13].

Chromosome position analysis has revealed that a few genes in our list are located close to the telomeric ends (*PLXNA3* – 1.5 Mb, *METTL16* – 2.5 Mb from 17p, and *MRPS2* – 2.8 Mb from 9q end). This tempts to speculate that the observed association might be caused by TPE-OLD like mechanisms.

Enrichment analysis over the whole list of genes has shown that eight of them (out of 20) are associated with acetylation, some play role in mitochondrial protein translation and oxidative reactions, and some are parts of ribosomes and play a role in protein translation in the cytoplasm and in ER. Tissue enrichment analysis has identified that nine genes from our list are normally highly expressed in testis (*DERL3, FEM1C, HAT1, FARSA, PRPS1L1, RPS12, SEPT6, SPATA9, TXNRD1*). The similarity between the processes related to spermatogenesis and tumorigenesis has already been described [173], and it's tempting to speculate that proliferative activity might concordantly be associated with expression of these genes, and changes in telomere length dynamics.

Finally, enrichment analysis on transcription factor binding sites has identified 14 genes to be associated with XBP1 and 14 with NMYC transcription factors. XBP1 is known to promote expression of genes related to immune cell differentiation and ER response to stress [174]. Most notably, it regulates proliferation of endothelial cells and leads to angiogenesis [175]. NMYC is a protooncogene associated with a variety of tumors and also regulates telomerase activity [176].

3.2.5 Conclusion

Altogether, our data have revealed genes directly linked to telomere length dynamics and some presumably associated with telomere length via epigenetic regulatory mechanisms. A large part of the genes we have identified have previously been linked to either aging, or telomere biology or cancers. The rest of the genes have not been studied previously in that context and should be further investigated. Functional studies may elucidate the causality of found associations, their role in normal development and cell differentiation on one hand, and in lung adenocarcinoma development on the other hand.

CHAPTER 4
TELOMERE LENGTH MAINTENANCE: STATE OF THE ART

The mechanisms employed by highly proliferative cells to maintain the length of their telomeres may vary, but are classified into telomerase dependent and independent mechanisms. Activation of a particular telomere maintenance mechanism (TMM) specifies the behavior of a cell. In cancer cells, the TMM activity is a predictor of tumor aggressiveness and cell survival. It also predicts the cell's response to anti-cancer therapies targeting TMMs.

Here we describe the state of the art on known triggers of those mechanisms, the known players and the order of events leading to telomere elongation. This chapter is mainly based on literature published in the last decade, with a considerable amount only within the last couple of years: some papers have even been published in the first quarter of 2017. We mention this to highlight the importance and timeliness of the topic for the scientific community.

4.1 INTRODUCTION

The main telomere maintenance mechanism (TMM) employed by most of the stem and cancer cells, depends on the catalytic activity of telomerase [27, 67, 93]. However, a part of cancer cells elongate their telomeres via an alternative mechanism (ALT), which dependents on homologous recombination events between telomeric sequences [68, 94, 177, 178]. Activation and switch between these two TMMs is regulated by various known and unknown factors [179], which are described in detail in the following sections.

4.2 TELOMERASE DEPENDENT TELOMERE MAINTENANCE MECHANISM

Telomerase is a ribonucleoprotein complex that is able to catalytically elongate the ends of chromosomes. It is composed of the core catalytic subunit hTERT (encoded by *TERT*), the dyskerin protein (encoded by *DKC1*), and the RNA-component hTR (encoded by *TERC*), which contains a telomeric template sequence [180]. While *TERC* and *DKC1* genes are ubiquitously expressed in somatic cells, *TERT* is often silenced [181]. Therefore, mere expression of *TERT* is often (but not always) enough to induce telomerase complex formation and, thus hTERT is considered as the main limiting component for telomerase assembly [181]. However, hTERT also possesses a number of extra-telomeric functions, such as regulation of gene expression, apoptosis, proliferation, cellular signaling, DNA damage response, etc., and its expression is not always

correlated with telomerase activity and telomere length maintenance [182]. In this sense, study of the whole pathway leading to assembly of the telomerase complex, and its catalytic activation is of considerable value.

The hTR RNA component of telomerase is of 451 nucleotides in length and contains a telomeric template domain, and domains for interaction with hTERT [183]. Expression of *TERC* is activated through downregulation of p53, or up-regulation of c-Myc and through the action of sex and growth hormones [181]. A large part of *TERC* transcripts undergo degradation by nuclear and cytoplasmic exosomes, and a small part is processed to maturity and carried to the locations of telomerase complex assembly [184].

Dyskerin is a telomerase subunit that binds to hTR and stabilizes the telomerase complex. Mutations in the gene *DKC1* lead to a premature aging syndrome called dyskeratosis congenital (see Chapter 1). Aside from its role in telomerase complex assembly, dyskerin possesses a number of other roles, such as rRNA processing and ribosome production [185].

It's important to note that while the presence of these three core components is obligatory for telomerase complex assembly, there may be many forms of telomerase, and other accessory proteins, such as WRAP53/TCAB1, pontin and reptin, participate in its formation and are sometimes also considered to be parts of the complex [186, 187].

4.2.1 Expression of hTERT and its nuclear import

Expression of hTERT is regulated by multiple factors, including transcriptional activators and repressors, promoter mutations, epigenetic modifications [188] and interactions with telomeric regions of chromosome 5 [57]. Among the known transcriptional activators of hTERT are c-Myc, NF-κB, STAT and Pax proteins, and the estrogen receptor [188]. There are also transcriptional repressors, such as CTCF, E2F1, and hormone nuclear receptors [188]. Hypo- or hyper-acetylation of H3 and H4 histone marks regulate silencing and activation of *TERT* transcription, respectively, and DNA methylation at the *TERT* promoter is known to play a complex regulatory role [188]. Differentiation of pluripotent stem cells is usually accompanied with epigenetic modifications at the *TERT* promoter to induce its silencing, while the opposite process may lead to *TERT* overexpression in cancers [188]. Overall, the regulation of transcriptional activity of *TERT* is quite complex, with multiple players acting at different regulatory levels.

While transcriptional activity of *TERT* is of crucial importance, many events should follow it to ensure proper assembly and activity of the telomerase complex. For example, nuclear import of hTERT, it's recruitment to the place of complex assembly, proper post-translational modifications and formation of a correct conformational state, as well as availability of the rest of the subunits are necessary for the complex to form, while certain post-translational modifications ensure enzymatic activity of hTERT in the already assembled holoenzyme [187].

At the first place, hTERT should be recruited to the nucleus after translation. hTERT is a large protein (~124 kDa), which means that it cannot pass through the nuclear membrane via passive transport. Therefore, it gets into the nucleus via active transport, through its bipartite nuclear localization signal (NLS) located at residues 220-240. The NLS sequence is recognized by importin-α, importin-β, which regulate the process of import. Akt-mediated phosphorylation of S227 of hTERT leads to efficient binding of importin-α 5 to the NLS, and promotes its nuclear import [189]. Well known upstream activators of Akt are phosphoinositide-dependent kinase-1 (PDK-1) [190], phosphoinositide 3-kinase (PI3K), as well as heat shock protein 90 (Hsp90) and protein phosphatase 2A (PP2A) [191].

It's also thought that importin 7 may mediate hTERT nuclear localization, possibly by an alternative nuclear import pathway. Molecular chaperons Hsp90 and p23 bind to hTERT and maintain its proper conformation for the nuclear import to take place. In contrast to this, binding of CHIP ubiquitin ligase and Hsp70 leads to cytoplasmic accumulation and degradation of hTERT [189]. Another kinase that activates telomerase via hTERT phosphorylation is protein kinase C (PKC). The study by Chang *et al* [192] suggests that PKC-mediated phosphorylation of hTERT promotes its binding to Hsp90, therefore leading to proper telomerase complex assembly.

Certain proteins may also perform posttranslational modifications of hTERT that can either inhibit its nuclear import, or complex assembly, or enzymatic activity. An example is c-Abl protein tyrosine kinase, which is known to inhibit the activity of telomerase by phosphorylation of hTERT. However, the exact role of this modification is not yet understood [193].

Aside from the NLS signal, hTERT also possesses a nuclear export signal (NES). Therefore, hTERT shuttles in and out of the nucleus. In *yeast*, hTERT is exported from the nucleus to the cytoplasm, where the assembly of the telomerase complex takes place. In humans, it is not yet known whether the assembly happens in the nucleus or in the cytoplasm. Studies have shown that

mutations in the NLS signal decrease telomerase activity, but the effect of NES mutations still has to be investigated [187].

4.2.2 hTR transcription, processing and degradation

In contrast to *TERT*, *TERC* is not considered as a rate-limiting factor for telomerase activity, as it is relatively constantly expressed. However, when *TERT* becomes overexpressed, hTR levels may limit the amount of assembled complexes [194]. This has been demonstrated in a series of experiments in mice, where it was shown that *TERT* overexpression in the absence of *TERC* does not lead to telomerase activity, and even inhibits tumorigenesis [195].

The rate of *TERC* transcription *per se* is not enough to ensure required amounts of hTR available for telomerase assembly: one should take into account that the amount mature hTRs depends on the competition between the processes of maturation and degradation of newly synthesized *TERC* transcripts. These processes are controlled by nuclear and cytoplasmic exosomes that degrade those nascent hTR transcripts that were marked by certain modifiers immediately after transcription; or by other modifiers that remove degradation marks and promote the export of hTRs to Cajal bodies, where they become mature [184]. Among hTR processing complexes are the CBCA complex (*NCBP1, SRRT, NCBP2*) which leads to 5' capping, and the NEXT complex (*RBM7, ZCCHC8, MR4*) that adds oligo-A tails to hTRs [184]. CBCA and NEXT associate with hTR co-transcriptionally and form the CBCN complex, which recruits the nucleolar exosome [184]. Long hTR transcripts get degraded by the exosome (largely by its EXOSC10 component), while the short ones are transported to the nucleolus and partly become oligo-adenylated by the TRAMP complex (*MTR4, ZCCHC7 and PAPD5*) [184]. Oligo-adenylation also marks these transcripts for degradation by the exosome. A part of these transcripts escape the exosomal degradation due to the action of PARN, which removes oligoA tails from the hTRs [196]. Polymerases may add polyA tails to the hTRs, to which PABPN1 will bind to promote PARN-directed hTR maturation [184]. The mature hTRs that are 451 nts in length are then properly folded and bound by dyskerin-NOP10-NHP2 complex and NAF1, which provides final protection against degradation [197]. This complex is recruited to Cajal bodies, where NAF1 gets replaced by GAR1 [197]. Besides nuclear exosomes, cytoplasmic exosomes DCP2 and XRN1 are also factors that may degrade hTR after its export, particularly when dyskerin binding to hTR in nucleus is compromised [196].

4.2.3 Telomerase assembly

The assembly of telomerase is aided by ATPases Pontin and Reptin (encoded by *RUVB1* and *RUVB2*) [186]. It has been shown that Pontin and Reptin affect hTR levels through their interaction with dyskerin, however, it has not been shown if they also interact with hTERT directly. In yeast the site of telomerase assembly is in the cytoplasm, in humans, however, the site is not yet identified [197]. It has been observed that telomerase localizes to Cajal bodies (CB) for most of the cell cycle, which is driven by interactions of WRAP53/TCAB1 with hTR, but the role of CBs in complex assembly is not yet clear [198].

4.2.4 Recruitment to telomeres and catalytic activity

The localization of the assembled nucleoprotein complex to telomeres is further guided by interactions with shelterin proteins POT1, TPP1, TRF1 and TRF2 [186, 197]. An important role in recruitment of telomerase to telomeres is played by WRAP53/TCAB1, however it is not yet established whether WRAP53 mediates interaction of telomerase with telomeres, or it induces posttranslational modifications that promote telomerase-telomere interactions [187].

The catalytic activity of the holoenzyme is also regulated by shelterin proteins [199]. For example, TRF1 inhibits telomerase activity, and the longer the telomeres, the more of TRF1 is recruited, and thus the stronger will be the inhibition. Another shelterin protein, POT1, competes with telomerase by binding to the single stranded 3' telomere overhang. Similarly, other shelterin proteins, such as TIN2, TPP1 and TNF2 also act as negative regulators. Partly due to this regulatory mechanism, and for some yet undefined factors, telomerase elongates mainly the shortest telomeres and the mean telomere length of the cells expressing telomerase usually stays at a relatively constant level [200].

4.2.5 Summary

Taken together, the regulation of telomerase dependent TMM occurs at multiple levels of cellular activity: starting from the expression of the enzyme components, their proper processing and maturation, and ending with proper complex assembly, recruitment to telomeres and catalytic

activity. While this process is well characterized in *yeast* and other model organisms, its many aspects are still unknown in humans.

4.3 ALTERNATIVE LENGTHENING OF TELOMERES

A part of cancer cells keep their proliferative potential via a telomerase-independent mechanism, which depends on homologous recombination events, and is called alternative lengthening of telomeres (ALT). ALT is observed in many common tumors, but is mainly inherent to tumors of mesenchymal origin [201]. Activation of ALT is an indication of aggressiveness of tumors and poor prognosis [202]. Additionally, tumor cells may switch from telomerase positive to ALT phenotypes during development [203], and these two TMMs may also coexist in the same cell [204]. As mentioned above, researchers are attempting to target cancers with telomerase inhibiting therapies. In these cases, tumor cells often switch from telomerase positive to ALT phenotype [205]. The mechanisms leading to ALT activation are therefore of considerable interest, however, in spite of the significance of the topic, the existing knowledge is yet scarce. Below we describe the state of the art on known ALT mechanisms and the factors leading to their activation.

4.3.1 Descriptors of ALT phenotype

ALT positive cells are described by a number of phenotypic markers. They contain extrachromosomal telomeric repeat containing DNA, such as T-circles that are generated via resolution of telomeric T-loops by recombination enzymes, and C- and G-circles, which are largely single stranded circular sequences of C-terminal or G-terminal telomeric repeats, with C-circles being more abundant. ALT cells usually have longer mean telomere lengths: about 20 kb on average, as opposed to telomerase positive cells that usually have less than 10 kb of mean telomere length [206]. The telomere lengths at individual chromosomes of ALT cells are highly heterogeneous with simultaneous presence of extremely short and extremely long telomeres, and the length of individual telomeres rapidly changes during proliferation. In ALT cells, there is a greatly elevated level of recombination events at telomeres, and, finally, a large number of PML (promyelocytic leukaemia) bodies containing telomeric chromatin, also called ALT associated PML bodies (APBs). The role of PML bodies is not well understood, but it is known that these are involved in senescence and DNA damage response (DDR). APBs are highly characteristic to ALT cells, and can be observed both attached to the chromosome ends, and to extrachromosomal

telomeric sequences. APBs contain recombination proteins and it is assumed that ALT associated recombination events happen at APB sites. Therefore, presence of APBs, as well as telomeric C-circles are used as markers for identification of the ALT phenotype [207].

Additionally, it is known that the majority of ALT cells have mutations in *ATRX* and/or *DAXX* genes; however, their role in ALT initiation is not yet identified [208]. Finally, telomerase is usually under-expressed in ALT cells, however the functional consequences of its down-regulation are also not known.

4.3.2 (Possible) mechanisms of homologous recombination in ALT

There is a growing experimental knowledge to understand the sequence of events happening in ALT and the main players involved in each event. A comprehensive review by Cesare *et al* [209] has summarized the existing knowledge on ALT regulators, and the paper by Pickett *et al* [210] has discussed involvement of these factors in particular stages of homologous recombination (HR) happening in ALT.

Two possible mechanisms of telomere elongation in ALT cells are currently considered: unequal telomeric sister chromatid exchange (T-SCE) and HR dependent DNA replication. T-SCE events are generally observed with increased frequency in ALT cells and the hypothesis assumes that due to unequal length distribution of telomeres on sister chromatids, one of the daughters receives chromosomes with long telomeres and the other one with short telomeres, and thus, the first cell will gain proliferative capacity as a result of uneven segregation of the longer telomeres. The existence of a mechanism leading to such non-random segregation of long and short telomeres into separate daughter cells is still hypothetical [209]. The second mechanism referred to as HR dependent DNA replication is thought to be employed through break induced replication (BIR). In this case, the 3' G strand hybridizes with another telomeric C strand of a template sequence. The latter may be telomeric end of another chromosome, a sister chromatid, an extra-chromosomal telomeric sequence (C-circles), as well as a T-loop. The G strand is elongated up to the end of the template C strand, and afterwards the strands are separated [209]. Of these two possible mechanisms, we will further refer to the ALT pathway in reference to the HR-dependent mechanism, since it's been studied most extensively.

The BIR mechanism can occur intra-chromosomally, inter-chromosomally, or extra-chromosomally, depending on the source of the telomeric template (Figure 21) [211]. The invasion leads to formation of a Holiday Junction (HJ), and replication mediated elongation of the invaded strand. The HJ is then dissolved, and the C strand of the short telomeres is elongated using the newly copied G-strand [210]. It is also thought that extra-chromosomal telomeric repeats, such as C-loops and T-loops may serve as simple templates for strand elongation (Figure 21, B) [212]. Moreover, the T-loop of the same chromosome may be used as a template in a rolling-based replication [212]. Finally, the other possibility for intra-chromosomal copying may occur through looping of the G-strand and its simple replication based on the C-strand (Figure 21,C) [213]. It is subject for further investigation to reveal which of these mechanisms takes place in ALT and to which extent. Here, all further discussions will be based on consideration of the classical inter-chromosomal BIR mechanism. It is worth noting that the inter-chromosomal BIR that may occur through telomere copying from a sister chromatid assumes that sister chromatids may have varying telomere lengths (Figure 21, D). There are some studies that discuss this possibility as well [213].

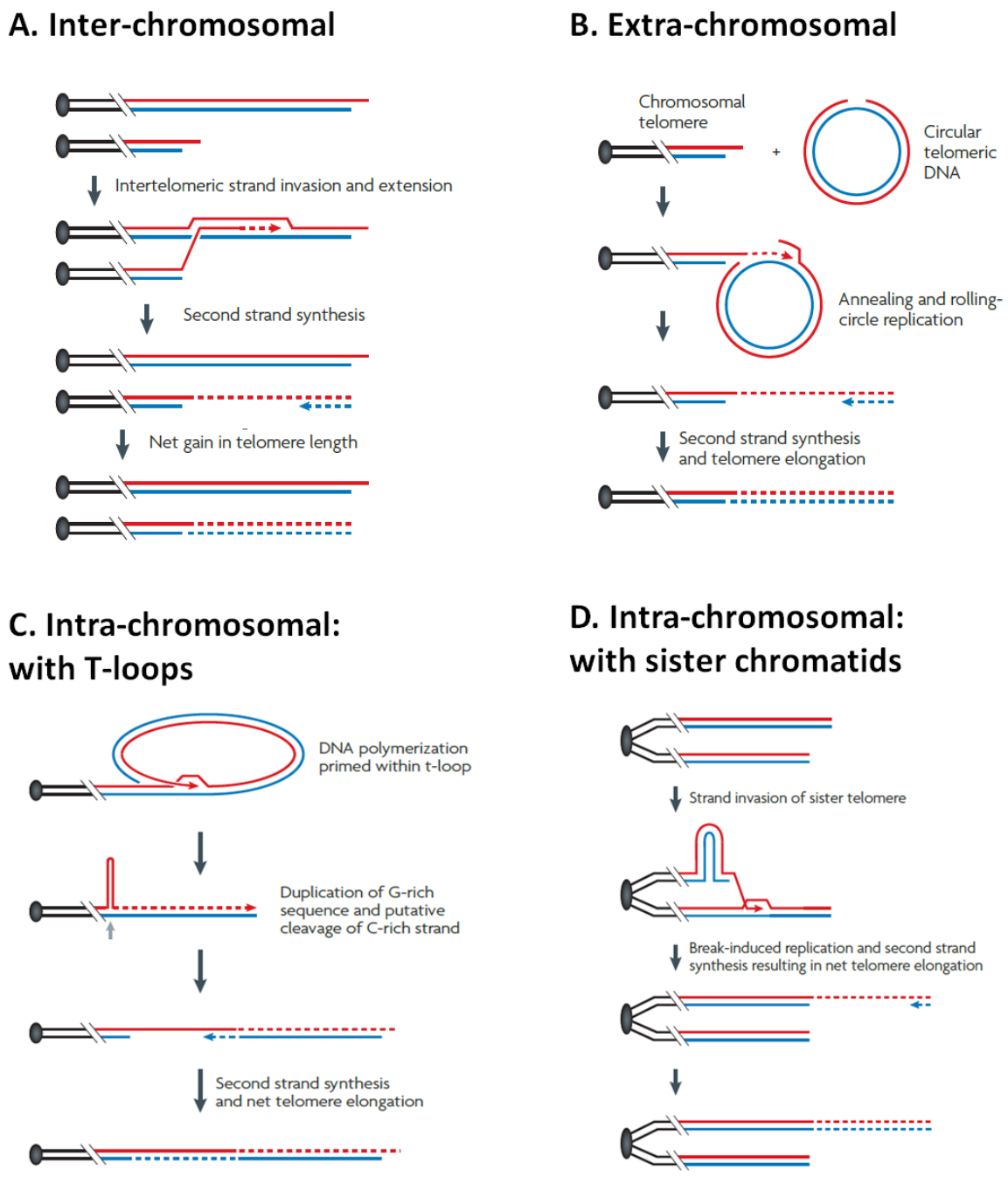


Figure 21. Possible mechanisms of homologous recombination in ALT (adapted from [209]**). A –** Inter-chromosomal recombination happens via G-strand invasion onto homologous or a distantly located chromosome. **B –** Recombination occurs using extra-chromosomal telomeric sequences (C-circles, T-circles, linear telomeric sequences) as a template. **C,D –** Intra-chromosomal recombination is based on telomeric sequences in the same chromosome, such as the T-loop **(C),** or the sister chromatid **(D)**.

HR is a complex multistep process that involves interactions between various proteins participating in strand invasion, template directed synthesis and resolution of recombination intermediates.

The first step in the overall process is the strand invasion, which lead to formation of an HR specific structure known as Holiday Junction. The long-range movement of the telomeric G-strand to a homologous or non-homologous chromosome required for inter-chromosomal copying may be conditioned by Hop2-Mdn1 heterodimer interaction with RAD51 [214]. The ATR and Chk1 lead to recruitment of Hop2 to the telomeres [215]. Noteworthy, RAD51 independent mechanisms have also been observed in yeast, and such a possibility is discussed in humans as well [215]. The telomeric DNA is protected from such invasions via shelterin proteins, particularly POT1, which binds to ssDNA at telomeres. RAD51 is a known promoter of DNA invasion in HR, and it is thought that POT1 should be replaced by RAD51, and that an intermediate step in this process is the loading with replication protein A (RPA) complex. The heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) has been shown to inhibit the replacement of POT1 with RPA. RAD52 is a positive regulator of RAD51 loading [210].

The second step in HR events is the strand directed synthesis. A recent paper by Dilley *et al* [215] reports that RFC1-mediated PCNA loading at the telomeric breaks recruits the polymerase δ through its POLD3 subunit, and RFC1-PCNA is thought to be the initial sensor of telomere damage. They have shown that POLD3 is critical for break induced telomere synthesis in the majority of ALT cells. Worth to mention, the study has also shown that the break induced telomere synthesis is not dependent on the ATR induced damage signaling and RAD51. ATR-Chk1 signaling, thus, regulates telomere integrity and ALT cell survival, while RFC1-PCNA-POLD3 independently participate in telomere synthesis [215].

After the synthesis phase the Holiday Junctions and HR intermediates are dissolved by the BTR complex composed of TOP3A, BLM, RMI1and RMI2 [210]. Several proteins, such as SLX4/1 and MUS81-EME1 complexes and GEN1, may inhibit this process via resolution of HR intermediates. It seems that dissolution happens mostly after telomere synthesis, while resolution of HR intermediates interferes with telomere synthesis thereby inhibiting ALT mediated telomere lengthening [210].

4.3.3 ALT-specific heterochromatic states

Regulation of the ALT phenotype strictly dependents on chromatin structure and the compact state of telomeres. Lack of shelterin proteins and telomeric histone modifiers leads to decompaction of telomeres and formation of an 'open' chromatin state. It is assumed that this open state recruits recombination proteins to the accessible telomeric DNA, which is the main potent trigger of ALT phenotype establishment [32, 62]. Particularly, all the experiments with impaired telomeric chromatin, have observed large numbers of APB bodies [36, 62, 216–218].

According to this notion, the proteins found to be regulating (or participating in) ALT are functionally related to one or many of the following processes: loosening of heterochromatin (histone (de)acetylases/(de)methylases, NuRD-ZNF827 complex), T-loop breakage through replacement of shelterin proteins with ALT initiators (POT1, RPA, RAD51), homologous-recombination (SLX4-SLX1, MUS81-EME1, BTR complex), and APB formation (MRN, SMC5/SMC6 complexes, BRCA1, BLM).

It has been suggested that nuclear receptors NR2C2 (TR4) and NR2F2 (COUP-TF2) bind to telomeric C-type variants and recruit a recently discovered protein ZNF827 [219]. ZNF827 in turn recruits the NuRD complex to telomeres to promote ALT via increased T-SCEs, APBs and C-circles. A number of proposed mechanisms may act to support the ALT-promoting activity of the NuRD-ZNF827 complex. First, it's thought that NuRD is able to simultaneously interact with several ZNF827 proteins. If these proteins are located at telomeres of different chromosomes, this will lead to so called "telomeric bridge" formation, which in turn may promote telomeric strand migration as an initiator of HR at telomeres. Second, the NuRD-ZNF827 complex may recruit HR proteins, such as BRIT1 and BRCA1 to telomeres [219]. Finally, the hystone deacetylases HDAC1 and HDAC2 that are part of the NuRD complex may contribute to deacetylation and reduced chromatin compaction at telomeres [219].

The NuRD complex consists of six functional subunits, and performs a number of roles in regulation of replication, transcription and genomic stability. Depending on the protein composition of the subunits, NuRD may perform specific functions, which may have opposing effects, such as tumor suppressive or promoting [220]. It is not established whether the composition of NuRD subunits affects recruitment by the ZNF827 complex. Of note, the expression of some NuRD components, as reported in [219] is not significantly different in ALT

positive and negative phenotypes, leading to the assumption that the recruitment of NuRD to telomeric sites affects ALT more (if not at all), than the general abundance of NuRD proteins.

It has also been demonstrated that depletion of chromatin modifiers SUV39H1/H2, and SUV420H1/H2 which deposit the repressive H3K9me3 and H4K20me3 marks respectively, leads to chromatin decompaction, and formation of an ALT permissive environment [31, 36].

4.3.4 APB BODIES

Promyeolocytic leukemia bodies (PML) are generally present in all somatic cells, and they grow in size and number as the cell undergoes senescence [221]. The PML bodies that are associated with telomeric sequences are frequently found in ALT cells (thus, the name: ALT associate PML bodies (APB)), and are considered to be the main site where HR events occur [222]. Owing to the increased number of APB bodies in ALT cells, those have been used as markers of ALT activity [207]. APB bodies are found near telomeres at chromosome ends, as well as extra-chromosomal telomeric repeats. In addition to the regular components of PML bodies, such as the PML, Sp100 and shelterin proteins, APBs also contain additional components, such as RAD1, RAD9, RAD51, RAD52, RPA, RAD51D, BLM, WRN, RAP, BRCA1, MRE11, RAD50, and NBS1, *etc* [223]. While some of the APB components participate in HR events, the role of others in ALT in not known. These are the SMC5/6 complex (NSMCE2 (MMS21), SMC5, SMC6) and the MRN complex (NBN, RAD50 and MRE11A). The MRN complex appears in the early stages of dsDNA repair, and functions as a regulator of cell cycle checkpoints. In ALT cells, the MRN complex colocalizes with HR proteins, and depletion of this complex leads to reduced telomere length [224]. Overexpression of SP100 sequesters the MRN complex away from APBs and inhibits ALT [223]. The SMC5/6 complex proteins sumoylate telomere binding proteins such as TRF1 and TRF2, leading to increased APB formation at telomeres. Inhibition of SMC5/6 complex suppresses HR at telomeres [225]. The RecQ helicase WRN is found at APBs, however, its role in ALT is not established, since it's shown to be required in some, but not all of the ALT cells [226].

4.3.5 The role of ATRX and DAXX

The majority of ALT cells have loss-of-function mutations in either component of the ATRX/DAXX complex. There are several hypothesis of how ATRX and DAXX affect ALT, however their exact role has not been established yet [227]. The ATRX/DAXX complex is shown to deposit the histone H3.3 variant at telomeres, which may lead to repressed telomeric transcription and reduction of TERRA transcripts [228]. The role of H3.3 variant at telomeres is not clearly established, however association of ATRX/DAXX with H3.3 at telomeres has been shown to have stabilizing effect on telomeric heterochromatin, since the ATRX/DAXX complex possesses chromatin remodeling activities [229]. Finally, in mice, ATRX has been shown to reduce HP1α and lead to chromatin decompaction [227]. Despite the frequent mutation rate of ATRX/DAXX, its worth to mention that their mutations alone are not enough to initiate ALT, as has been shown for SV-40 hTERT immortalized cell lines [208].

4.3.6 Telomere fragility and sister chromatid loss

Flap structure specific endonuclease 1 (FEN1) is a canonical Okazaki fragment processing protein. Recently it has been shown that it plays a major role in telomere stabilization during replication, both at the leading, as well as the lagging strands [230]. At the lagging strand, FEN1 participates in replication fork progression, reinitiation and telomere stability [230]. At the leading strand, FEN1 cleaves the RNA:DNA hybrids that are generated during TERRA transcription. Removal of RNA fragments from the DNA as the replication progresses is important to limit telomere fragility and promote DNA repair at the leading strand [230]. Owing to its ability to stabilize telomeres, FEN1 may be important for telomere lengthening in ALT cells. However, there is a controversy, since FEN1 inhibits TERRA RNA:DNA hybrid formation, and these hybrids are known to support recombination events in ALT [230]. Studies have revealed that depletion of FEN1 leads to sister chromatid loss in ALT telomeres. It has been shown that FEN1 stabilizes telomeres in ALT positive cells. Its depletion from ALT positive cells leads to generation of telomere dysfunction induced loci (TIF) and end-to-end chromosome fusions. In telomerase positive cells, loss of telomeres at sister chromatids caused by FEN1 depletion was compensated by the action of telomerase [231].

4.3.7 Summary

Taken together, the mechanism through which ALT takes plays may vary, depending on the source of the template telomeric sequence and the mode of replication. Numerous studies have tried to reveal the molecular factors involved in ALT events, and these pieces of puzzle rapidly come together. It is possible that many pathways exist that lead to ALT induction and realization, however which of them is targeted by this or that study is largely not known. Therefore, currently accumulated knowledge on ALT may reveal a generalized picture of all the possible mechanisms that lead to recombination driven telomere elongation.

CHAPTER 5
RECONSTRUCTION OF TMM PATHWAYS

Activation of telomere maintenance mechanisms (TMM) is one of the key processes leading to cancerogenesis. These mechanisms are being actively targeted by anti-cancer therapies in the recent years. It is thus important to understand the factors leading to activation of TMMs. While there is a considerable amount of experimental data on TMM, combination of data- and knowledge-driven approaches based on utilization of high-throughput gene expression have the potential to foster further developments in this direction.

In this thesis, we have reconstructed in silico models for TMM phenotype prediction. These models represent pathways that include the key proteins, RNAs, and processes involved in the TMMs and the functional interactions between them. We have assessed the activity of these pathways based on gene expression data and have analysed the prediction accuracy against experimentally annotated samples. The results provide evidence on the validity of our model and show that it may become a useful tool in TMM research.

5.1 DATA AND METHODS

5.1.1 Datasets

In order to construct valid TMM pathways, able to predict the TMM phenotype of a cell/tissue from gene expression data, we have obtained samples with experimentally validated TMM phenotypes, for which gene expression data was available. For this purpose, we have downloaded the datasets deposited by Lafferty-Whyte *et al* in the Gene Expression Omnibus (GEO) repository [232]. These datasets contain microarray gene expression profiling data for ten cell lines of different tissue origin, and seventeen liposarcoma tumor samples along with four human Mesenchymal Stem Cells (hMSC) isolated from the bone marrow of healthy individuals (GEO accession: GSE14533). Among the cell lines, four were ALT positive, four were telomerase positive, and two didn't have any active TMM (normal cell lines). Among the liposarcoma samples, nine were ALT positive, as assessed by the presence of APBs, and eight were telomerase positive, as tested by the telomeric-repeat amplification protocol for telomerase activity detection [233].

5.1.2 Pathway construction and extension

The pathways were constructed based on protein-protein, protein-RNA interactions that lead to activation of the ALT and Telomerase TMMs, curated from the literature. The pathway maps were built in the Cytoscape environment (http://cytoscape.org/).

Each node in the pathway represents a gene product (protein or RNA), a protein complex, or a process. The edges in the pathway describe functional associations between the nodes. Those are of two types: activation and inhibition, depending on the effect of the source node on the functionality of the target node. It should be noted that the associations in the pathway refer to functional effects of direct protein-protein, protein-RNA interactions between the nodes: indirect effects or regulatory effects on expression are not considered.

We have first constructed initial pathways based on genes with clearly established roles in either TMM, and assessed their ability to predict TMM phenotypes of the studied samples. We then extended the core pathways iteratively, by adding nodes and changing their positions in the pathways. We have evaluated the prediction power of the TMMs at each extension step to arrive at the 'final' pathways with best scores (Figure 22). Measures for prediction accuracy are described in detail below.

5.1.3 Data preprocessing

The gene expression values for technical replicates were averaged. The cell lines and tissue samples (liposarcoma tumors and hMSCs) were further processed in separate sets. For each set, the gene expression values were converted to fold change (FC) values by dividing the expression of each gene to its mean expression across the samples. The FC values were then quantile normalized. FC values higher than the 99% percentile were set to the value at that percentile. These FC values were then mapped onto pathway nodes and used for assessing pathway activity (pathway signal flow analysis, see below).

5.1.4 Pathway signal flow analysis

Assessment of pathway activity was performed using the Pathway Signal Flow (PSF) algorithm [234–236], with the PSFC app for Cytoscape [234] using its default parameters and the signal

propagation rules described below. The PSF algorithm computes the strength of the signal propagated from the pathway inputs to the outputs through pairwise interactions between nodes, based on their fold change expression values. For each source-target interaction, the FC values are multiplied for edges of type 'activation' (FC_{source} * FC_{target}) and divided for edges of type 'inhibition' ($1/FC_{source}$ * FC_{target}). The weighted sum of multiple signals from many sources is assigned as the signal at the target node (the 'proportional' option in PSFC). The signal propagation starts from input nodes, spreads through the intermediate nodes and arrives at the sink nodes. In our case, there is a single sink node for each TMM pathway (labeled "ALT" and "Telomerase" respectively). The higher the initial FC value of a node, and the more activation signals it gets from upstream nodes, the higher will be its activity value (or PSF score), and vice versa. The PSF algorithm returns a single PSF score for each node in a pathway. We are, however, interested only in the sink nodes of each TMM pathway: the 'Telomerase' and the 'ALT' nodes. The PSF scores of these nodes reflect the overall activity of the pathways.

We have upgraded PSFC to also apply functions onto protein complexes: if involvement of all the subunits is obligatory for the complex to function, then the minimum signal of all the subunits is assigned to the complex, otherwise the complex is treated with PSF default rules. Note that nodes, where gene expression values were not available were omitted in calculations and assigned a PSF value of 1.

The PSF algorithm is calibrated in a way that if all the nodes in a pathway have FC values of 1, that produces unity PSF values at sink nodes. In our case, the FC value for each gene is computed by taking as a reference the average expression across the samples. This leads to FC values higher or lower than 1, depending on the differential expression of the gene across samples with different TMM phenotypes on one hand, and on the phenotypic composition of the studied samples on the other. Thus, while high PSF values indicate on higher activity of the TMM pathway, there is no predefined PSF threshold that classifies the pathways to be either 'active' or 'inactive'.

5.1.5 ASSESSMENT OF PREDICTION ACCURACY

Given gene expression data of a set of samples with either ALT^+, $telomerase^+$, ALT^-/$telomerase^-$ or ALT^+/$telomerase^+$ phenotypes, the TMM pathways should be able to predict those based on the TMM pathway activities, or PSF values. Ideally, $telomerase^+$ samples should have high PSF values

for the Telomerase pathway and low PSF values for the ALT pathway. ALT^+ samples should score high on the ALT pathway, and low on the Telomerase pathway. The samples with ALT^-/$telomerase^-$ phenotypes should score low at both pathways.

To assess the prediction power of the ALT pathway, we have computed the mean difference in PSF scores at the "ALT" sink between ALT^+ and $telomerase^+$, and between ALT^+ and ALT^-/$Telomerase^-$ samples. For the Telomerase pathway, the difference of mean PSF scores at the "Telomerase" sink node between $telomerase^+$ and ALT^+ or between $telomerase^+$ and ALT^-/$Telomerase^-$ samples was computed. The higher the difference, the more is the prediction accuracy. The significance of these mean differences was tested with Kruskal-Wallis rank sum test.

As it was mentioned above, there is no PSF threshold to define if the pathway is 'active' or 'inactive'. Thus, we have used linear support vector machines (SVM), to divide the samples into ALT^{high} or ALT^{low}, and $Telomerase^{high}$ or $Telomerase^{low}$ groups for each TMM, respectively. We have then plotted the samples onto a 2D space, where Telomerase PSF values were on the *x* axis, and ALT PSF values were on the *y* axis. The prediction accuracy was then assessed based on the ratio of correct versus incorrect predictions. In the datasets of cell lines, correct predictions were the cases were normal cell lines had ALT^{low}/$telomerase^{low}$ values, ALT^+ cell lines had ALT^{high}/$telomerase^{low}$, and $telomerase^+$ cell lines – ALT^{low}/$telomerase^{high}$ values. For the liposarcoma and hMSCs group, correct predictions were the cases where the ALT^+ liposarcoma tumors had ALT^{high}/$telomerase^{low}$ values, $telomerase^+$ tumors had ALT^{low}/$telomerase^{high}$ values and hMSCs - ALT^{low}/$telomerase^{low}$ values. All other predictions were considered as incorrect.

One may argue that since stem cells should possess an active TMM mechanism, ALT^{low}/$telomerase^{low}$ values should not correctly predict hMSCs' phenotype. However, no evidence exists for hMSCs to have either of the TMM mechanisms activated: according to some reports they have low (or not detectable) levels of telomerase activity [237] and no signs of ALT activity [238]. It should also be noted that even though it's possible for ALT and telomerase dependent TMMs to coexist in the same cell [204], all the cases with ALT^{high}/$Telomerase^{high}$ values in our datasets were considered incorrect. This is reasoned by that fact that none of the samples in our datasets were annotated to possess both of the phenotypes.

5.2 RESULTS

5.2.1 Reconstruction of the TMM pathways

Using currently accumulated knowledge on the mechanisms of activation and realization of Telomerase and ALT TMMs (described in Chapter 4), we have constructed several versions of ALT and Telomerase pathways. We have started with an initial pathway with well documented members and interactions, and iteratively added nodes and edited their positions in the pathways to arrive at the best predictive power. Figure 22 shows the changes in prediction accuracy metrics with each pathway extension step for cell lines and for liposarcoma tumors/hMSCs. The overall prediction accuracy of both pathways is shown with the green line, while each of the TMM's power to distinguish between the phenotypes is shown by the rest of the lines (see the legend for details). It is seen from the figure that the final iterations have achieved high accuracy for almost all of the metrics together.

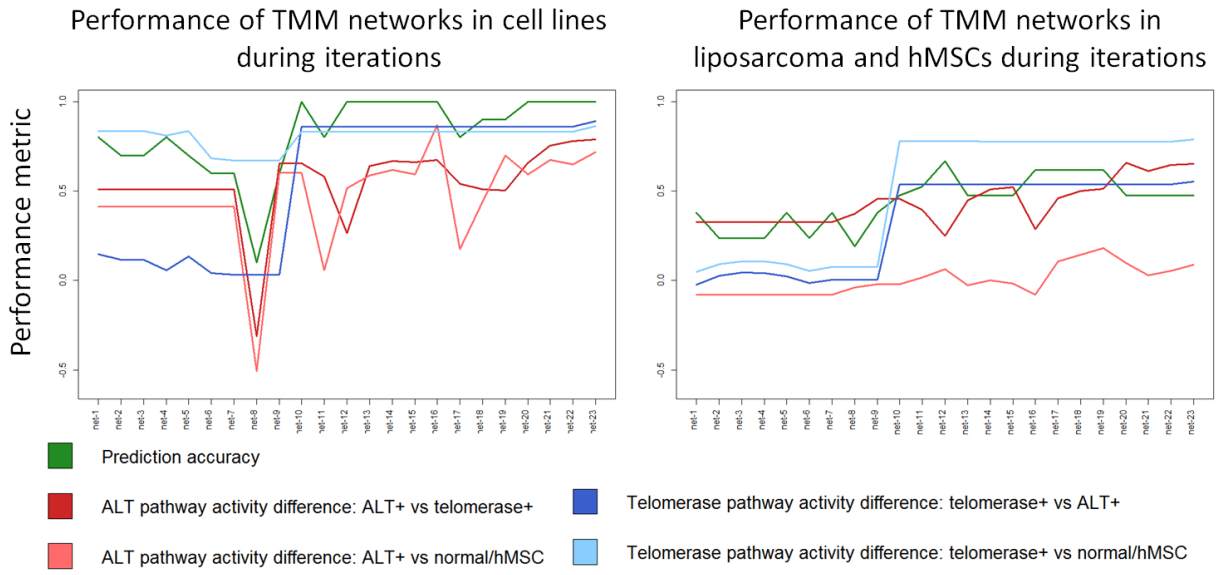


Figure 22. Prediction accuracy dynamics of TMM pathways during pathway extensions. The two plots represent pathway accuracy scores for cell lines (**left**) and liposarcoma tumors/hMSCs (**right**) respectively. The x-axis labels (net-*) correspond to the pathways at each extension step: either the Telomerase or ALT pathway is modified at each extension. Performance scores are represented on the y axis. The green lines correspond to the prediction accuracy changes, as predicted by SVMs (see methods, Pathway performance assessment). The dark red and light red lines represent the average difference of the ALT pathway activity between ALT^+ and $telomerase^+$ samples (dark red); and between ALT^+ and ALT^-/$telomerase^-$ samples (light red). The dark blue and light blue lines show the average difference of the Telomerase pathway activity between $telomerase^+$ and ALT^+ (dark blue); and between $telomerase^+$ and ALT^-/$telomerase^-$ samples (see methods, Pathway performance assessment).

The Telomerase pathway at the final iteration is depicted in Figure 23 A. The pathway consists of three main branches. The yellow branch includes the factors that lead to nuclear localization, enzymatic activation of hTERT. Of the activating factors, PI3K, PP2A and PKC are complexes with several possible subunit compositions. Chapter 4 describes their role in hTERT activation, however the exact subunit compositions that lead to such effects are not known. Therefore, the pathway is constructed in a way to account for the complex with the maximum activity. Similarly, the Hsp90 isoform with maximum abundance is considered in the network. The yellow branch also contains the *RUVBL1* and *RUVBL2* genes that code for pontin and reptin. These proteins have a crucial role for telomerase complex assembly, but have been included in the yellow branch, since they are mainly involved in hTERT activation changes to promote complex formation.

The brown branch involves the genes and complexes involved in hTR transcription and maturation, while the green one includes the factors leading to degradation of hTR transcripts. The amount of mature hTR transcripts depends on the interplay between these two processes.

Finally, the blue nodes highlight the core subunits of the telomerase complex and the sink node of the pathway ("Telomerase").

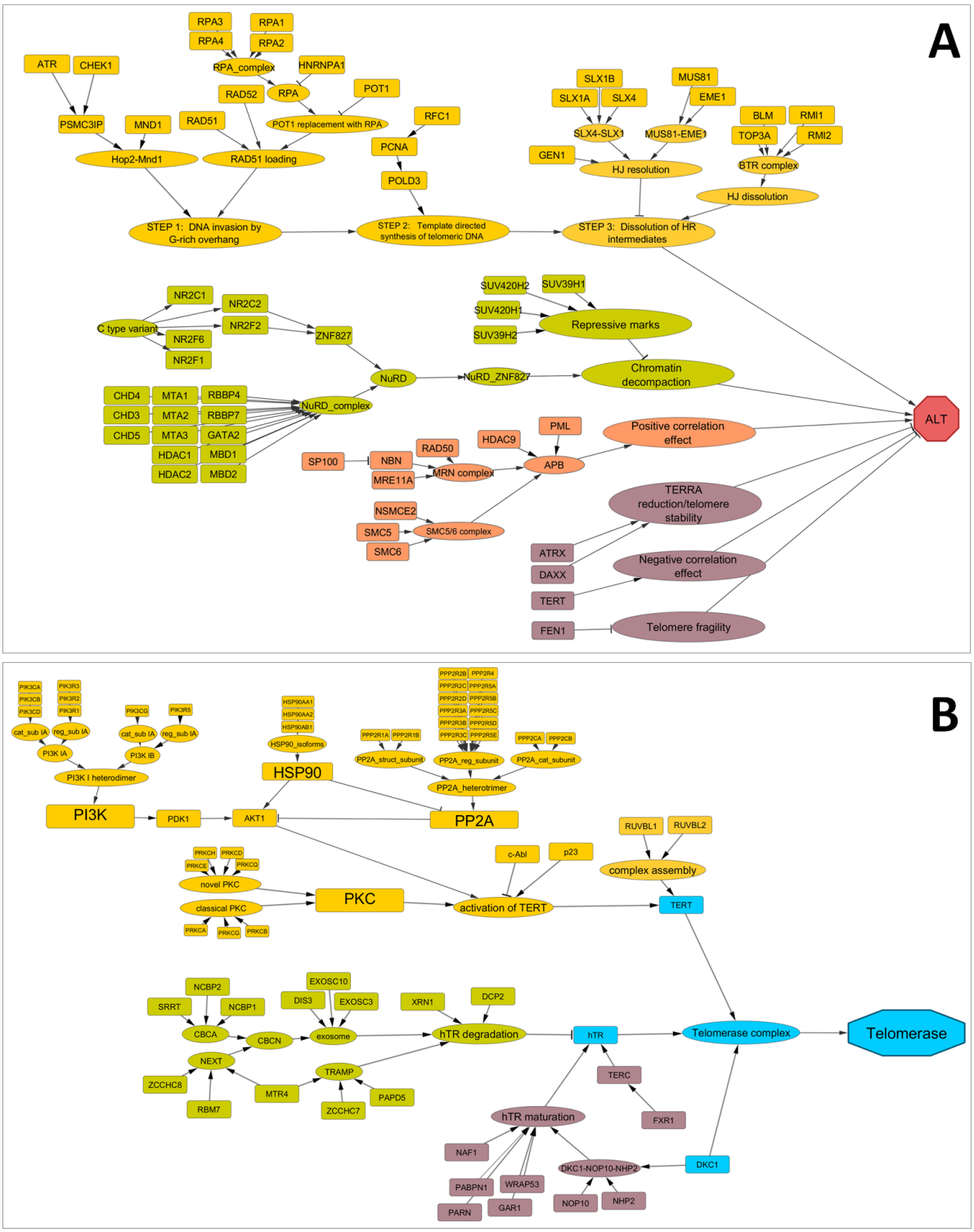


Figure 23. The Telomerase (A) and ALT (B) pathways. The "Telomerase" and "ALT" nodes are the targets of each pathway. In **(A),** the core components of the telomerase complex are colored blue. The rest of the colors in both plots are for enhancement of readability. Rectangular nodes represent genes, while oval-shaped nodes represent processes or complexes. The edges with delta and T shaped targets indicate activation and inhibition interactions, respectively.

The ALT pathway is depicted in Figure 23 B. As is described in Chapter 4, multiple mechanisms exist that allow for template directed synthesis of telomeres in ALT: the template may be from sister chromatids (intra-chromosomal), from distantly located chromosomes (inter-chromosomal) and from extra-chromosomal telomeric sequences. In construction of the ALT pathway, we have accounted for the factors involved in inter-chromosomal break-induced repair and extra-chromosomal template-driven mechanisms. It is important to note, however, that our pathway mostly represents a generalized picture, where the mechanism that is realized by involvement of this or that pathway entity is not clearly indicated.

In the ALT pathway, the yellow branch involves the main processes leading to HR in ALT. These processes are divided into three consecutive steps. The first step presented with the node "Step 1: DNA invasion by G-rich overhang" involves the factors leading to ssDNA loading with RAD51, which leads to invasion of the G-rich strand onto a telomeric template. It is worth noting that the RAD51-dependent pathway is observed in the majority, but not all of the ALT cells. The Hop2-Mnd1 heterodimer is also not an exclusive player in ALT. It plays an essential role in some ALT cells by promoting strand invasion onto distantly located chromosome ends. The second step represents the telomere strand synthesis after strand invasion, and is presented by the node "Step 2: template directed synthesis of telomeric DNA". It's been shown that this step is performed by DNA polymerase δ. It is reported that its POLD3 subunit, but not POLD4, is essential for this step. Indeed, inclusion of the POLD4 node did not improve the pathway performance and thus was removed. The final step is the dissolution of the Holiday Junctions after the telomere synthesis is finished (node "Step 3: Dissolution of HR intermediates"). An opposite process to this is the resolution of HR intermediates that occurs in the absence of telomere synthesis.

The rest of the branches involve factors that promote ALT by creation of an ALT permissive environment or those that are positively/negatively correlated with ALT by yet unknown mechanisms. The green branch represents the factors leading to decompaction of chromatin and establishment of ALT permissive environment, such as chromatin modifiers (the SUV family), and the NURD-ZNF827 complex. Notably, this complex has additional roles in ALT, other than chromatin decompaction, as mentioned in Chapter 4. The node named "C-type variants" represents telomeric TCAGGG repeats that are often found in ALT and recruit nuclear receptors. This node

did not have an initial FC value in our calculations. However, if there were whole genome sequencing data along with gene expression, one could compute the abundance of C-type variants and use it in the PSF calculations.

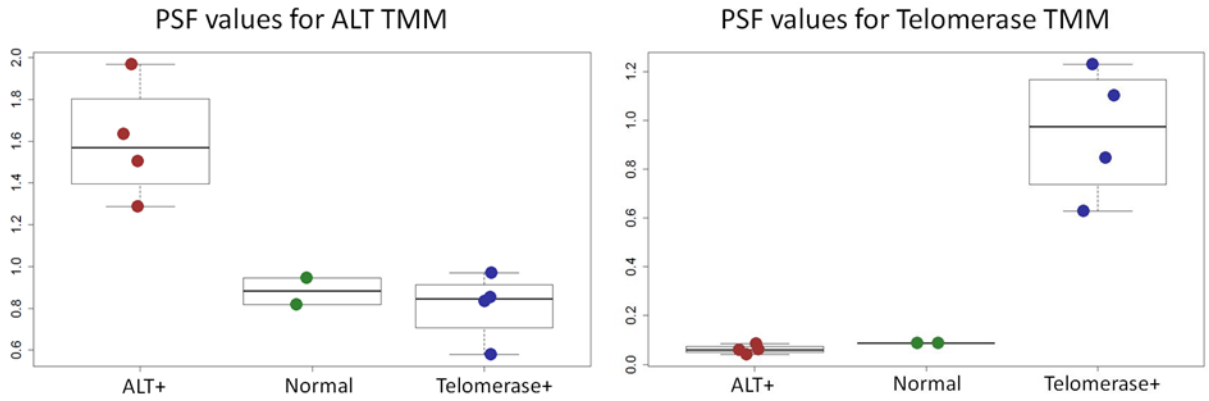
The orange branch involves proteins found in APB, and factors supporting or inhibiting APB formation.

The brown branch involves factors that are associated with ALT activation by not yet established mechanisms. The ATRX/DAXX genes are mutated in the majority of ALT cells. We did not have genome variant data for the studied samples, and thus have considered higher expressions of these genes to have inhibitory effects on ALT. Availability of information on genetic variations would probably lead to more accurate results. We have also included hTERT in the pathway. Even though it does not play a direct role in ALT, its suppression or low abundance usually provokes the cells to escape senescence via the alternative ALT mechanism. Since addition of *TERT* also considerably improved the performance, we have kept it in the pathway. Finally, the *FEN1* node represents that gene's role in stabilization of telomeres and inhibition of sister chromatid loss. As mentioned in Chapter 4, the role of FEN1 is controversial, since it also reduces the RNA:DNA hybrids that are known to promote ALT. This was also seen in our simulations, as presence of *FEN1* led to improved performance in case of cell lines, and a bit worsened it in the case of tumors. The role of *FEN1* and its inclusion in the network, thus, needs additional testing on a larger dataset.

5.2.2 Pathways' prediction accuracy for cell lines

We have computed PSF values for ALT and Telomerase TMMs using gene expression data from four ALT+ (SKLU, WI38-SV40, SUSM1, KMST6), four Telomerase+ (5637, C33A, HT1080, A2780), two normal (ALT-/Telomerase-) (IMR90, WI38) cell lines. The boxplots of PSF values for both TMM pathways are depicted in Figure 24. The ALT^+ cell lines had significantly higher (mean differences = {0.7 (vs normal), 0.8 (vs $telomerase^+$)}, $p = 0.04$) PSF values compared to the normal and $telomerase^+$ cell lines. A similar picture is observed for the Telomerase TMM, with $telomerase^+$ cell lines scoring significantly higher than ALT^+ and normal ones (mean differences = {0.9 (vs normal), 0.9 (vs $telomerase^+$)}, $p = 0.02$). Thus, in this case, the ALT and $Telomerase^+$ TMMs alone are powerful enough to distinguish between ALT^+ and $telomerase^+$ cell lines from other phenotypic groups (Figure 24).

Even though using each TMM separately allows for distinguishing between ALT$^+$ and ALT$^-$, or between telomerase$^+$ and telomerarse$^-$ phenotypes (Figure 24), the combined assessment of their PSF values with SVMs allowed us to distinguish between the three phenotypic groups simultaneously (Figure 25).



Summary statistics for PSF values

TMM	ALT			Telomerase		
Samples	ALT+	Telomerase+	Normal	ALT+	Telomerase+	Normal
Min.	1.3	0.6	0.8	0.0	0.6	0.1
1st Qu.	1.5	0.8	0.9	0.1	0.8	0.1
Median	1.6	0.8	0.9	0.1	1.0	0.1
Mean	1.6	0.8	0.9	0.1	1.0	0.1
3rd Qu.	1.7	0.9	0.9	0.1	1.1	0.1
Max.	2.0	1.0	0.9	0.1	1.2	0.1
median.diff	0.0	0.7	0.7	0.0	0.9	0.9
mean.diff	0.0	0.8	0.7	0.0	0.9	0.9

Figure 24. Boxplots of group-wise PSF values for each TMM for cell lines. The red, green and blue dots represent ALT$^+$, Normal and telomerase$^+$ cell lines respectively. ALT$^+$ cell lines have higher PSF values for the ALT TMM compared to both normal and telomerase$^+$ groups **(left)**, while the Telomerase TMM shows considerably higher PSF values for telomerase$^+$ cell lines, compared to ALT+ and normal cell lines **(right)**. These differences are statistically significant.

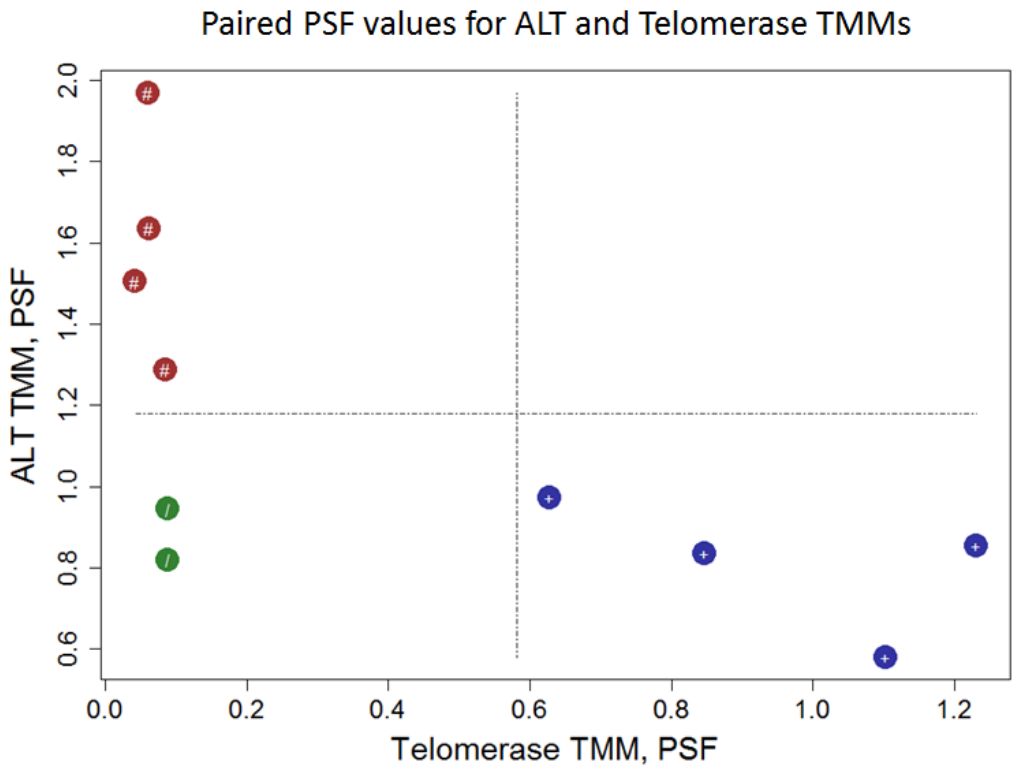
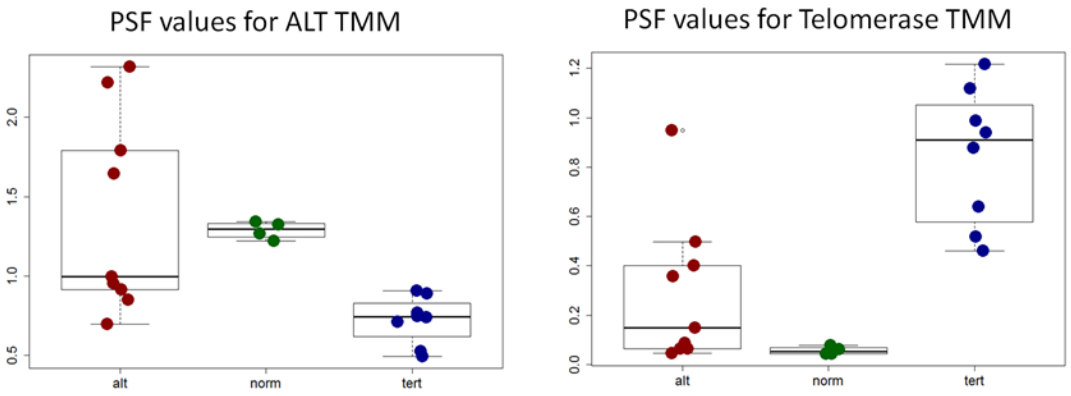


Figure 25. Paired PSF values for ALT and Telomerase TMMs in cell lines. The PSF values for the Telomerase TMM are on the x axis, and for the ALT TMM those are on the y axis. The red, green and blue dots correspond to ALT+, normal and telomerase+ cell lines, respectively. The horizontal and vertical dotted lines separate ALT^{high} from ALT^{low}, and telomerase^{high} from telomerase^{low} phenotypes, respectively. These lines are located based on with one-dimensional linear SVMs (see methods). The bottom-left quadrant is predicted to be ALT^{low}/telomerase^{low}, as indicated by the "/" white signs. The upper-left quadrant predicts ALT^{high}/telomerase^{low} phenotypes ("#" white signs), and the bottom-right predicts ALT^{low}/telomerase^{high} phenotypes ("+" white signs). The prediction accuracy, computed by the ratio of correct versus incorrect predictions, is 1 in this case, as all the predictions correspond to the actual phenotypes.

5.2.3 Pathways' prediction accuracy for the tumors/hMSC group

The tumor/hMSC group comprised liposarcoma tumor samples, where nine had ALT+ and eight had telomerase+ phenotype; and four hMSC samples were derived from the bone marrow of healthy individuals. We have found that the ALT$^+$ tumors scored considerably higher on the ALT pathway than telomerase$^+$ tumors, and vice-versa, telomerase$^+$ tumors had significantly higher activity of the Telomerase pathway compared to ALT$^+$ tumors. The activity of the Telomerase pathway was very low for hMSCs, however there was no statistical difference of ALT pathway activity between hMSCs and ALT$^+$ tumors (Figure 26).



Summary statistics for PSF values

TMM	ALT			Telomerase		
Samples	ALT+	Telomerase+	hMSC	ALT+	Telomerase+	hMSC
Min.	3.0	1.0	14.0	3.0	12.0	1.0
1st Qu.	11.0	3.5	14.8	6.0	14.8	1.8
Median	13.0	5.5	15.5	9.0	16.5	3.0
Mean	13.9	5.5	15.5	9.2	16.8	3.5
3rd Qu.	19.0	7.5	16.3	11.0	19.3	4.8
Max.	21.0	10.0	17.0	18.0	21.0	7.0
median.diff	0.0	7.5	-2.5	0.0	7.5	13.5
mean.diff	0.0	8.4	-1.6	0.0	7.5	13.3

Figure 26. Boxplots of group-wise PSF values for each TMM for liposarcoma tumors/hMSCs. The red, blue and green dots represent ALT$^+$, telomerase$^+$ tumors and hMSCs, respectively. ALT$^+$ tumors have considerably higher PSF values for the ALT TMM compared to telomerase$^+$ tumors, and were almost similar in the mean values to hMSCs **(left)**, while the Telomerase TMM shows considerably higher PSF values for telomerase$^+$ cell lines, compared to both ALT+ tumors and hMSCs **(right)**. All these difference, except for ALT$^+$ versus hMSC comparison for the ALT TMM, are statistically significant.

Combination of the ALT and Telomerase PSFs on a 2D-space revealed that four ALT$^+$ tumors clearly clustered together with ALT^{high} phenotype, and six telomerase$^+$ and one ALT$^+$ tumors also showed clear clustering at the other corner of the plot (Figure 27). hMSCs were classified as ALT^{high}/Telomerase^{low} but, having ALT acitivity levels very close to the ALT^{high} - ALT^{low} decision line, and at the same time having very low Telomerase activity values. Four ALT$^+$ and two telomerase$^+$ tumors were misclassified as having neither TMM (ALT^{low}/Telomerase^{low}). The overall prediction accuracy was 0.6, not accounting for hMSCs. We didn't include these cells in prediction accuracy estimation, as it is not clearly established which TMM is active in these cells. When including these samples and considering them as ALT$^-$/telomerase$^-$ we arrived at a lower prediction accuracy of 0.48.

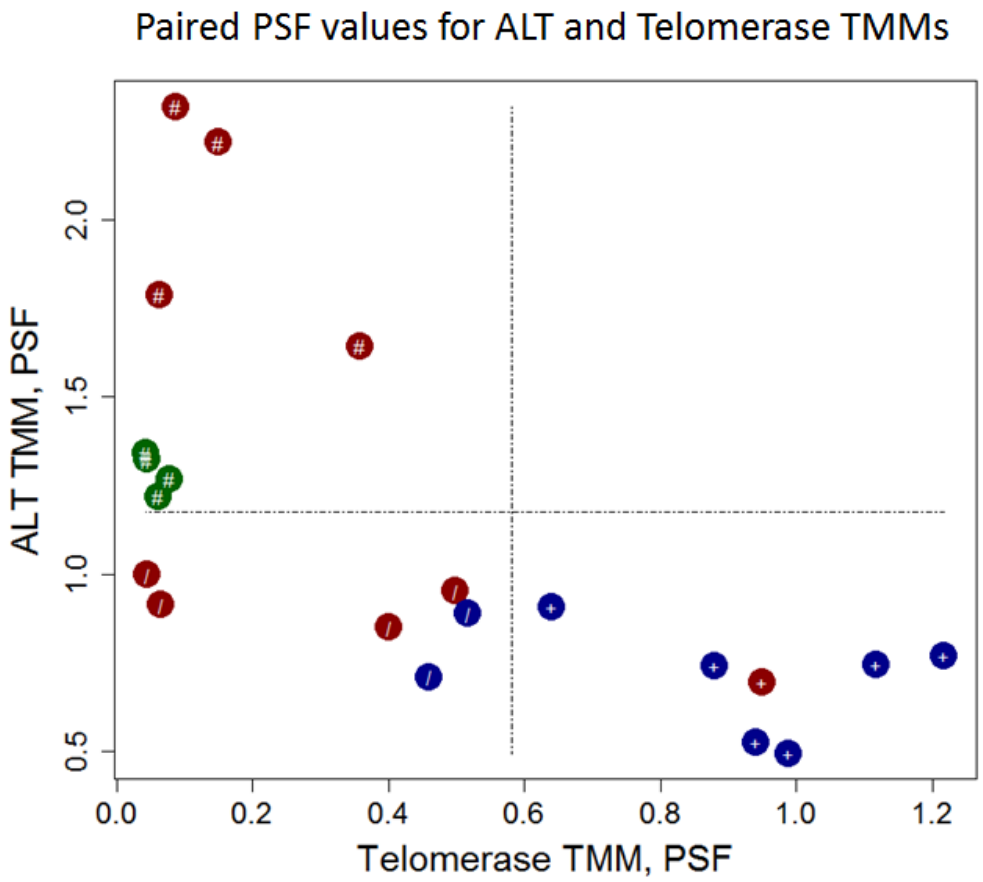


Figure 27. PSF values for ALT and Telomerase TMMs in liposarcomas and hMSCs on a 2D space. The PSF values for the Telomerase TMM are on the x axis, and for the ALT TMM those are on the y axis. The red, blue and green dots correspond to ALT^{+}, telomerase^{+} tumors and hMSCs, respectively. The horizontal and vertical dotted lines separate ALT^{high} from ALT^{low}, and telomerase^{high} from telomerase^{low} phenotypes, respectively. These lines are located based on with one-dimensional linear SVMs (see methods). The bottom-left quadrant is predicted to be ALT^{low}/telomerase^{low}, as indicated by the "/" white signs. The upper-left quadrant predicts ALT^{high}/telomerase^{low} phenotypes ("#" white signs), and the bottom-right predicts ALT^{low}/telomerase^{high} phenotypes ("+" white signs). The prediction accuracy, computed by the ratio of correct versus incorrect predictions, is 0.6 in this case (hMSCs).

5.2.4 TMM DETECTION IN LUNG ADENOCARCINOMA CELL LINES

After validation of the TMM pathways we moved on to apply those on datasets with unknown TMM phenotypes. We have measured the activity of the TMM pathways in the lung adenocarcinoma cell lines dataset described above. Using the ALT and Telomerase TMM PSF scores, we have performed hierarchical clustering with complete linkage and the Euclidean distance metric. This led to identification of four large clusters of cell lines described by different activities of the TMM pathways (Figure 28). The rightmost cluster (the red cluster) in the

dendrogram presents nine cell lines where the predominant TMM was ALT. Six cell lines with high Telomerase and low ALT pathway activities clustered together in one large cluster (the blue cluster). The SAEC cell line derived from the lung of healthy subjects had low activities at both TMM pathways. It resides in a small sub-cluster with another lung adenocarcinoma (Lung AC) cell line (RERF.LC.Ad1), and in a larger cluster with seven other Lung AC cell lines (the green cluster). Finally, two Lung AC cell lines showing high activity for both TMMs (PC.7 and PC.3) clustered together in the violet cluster.

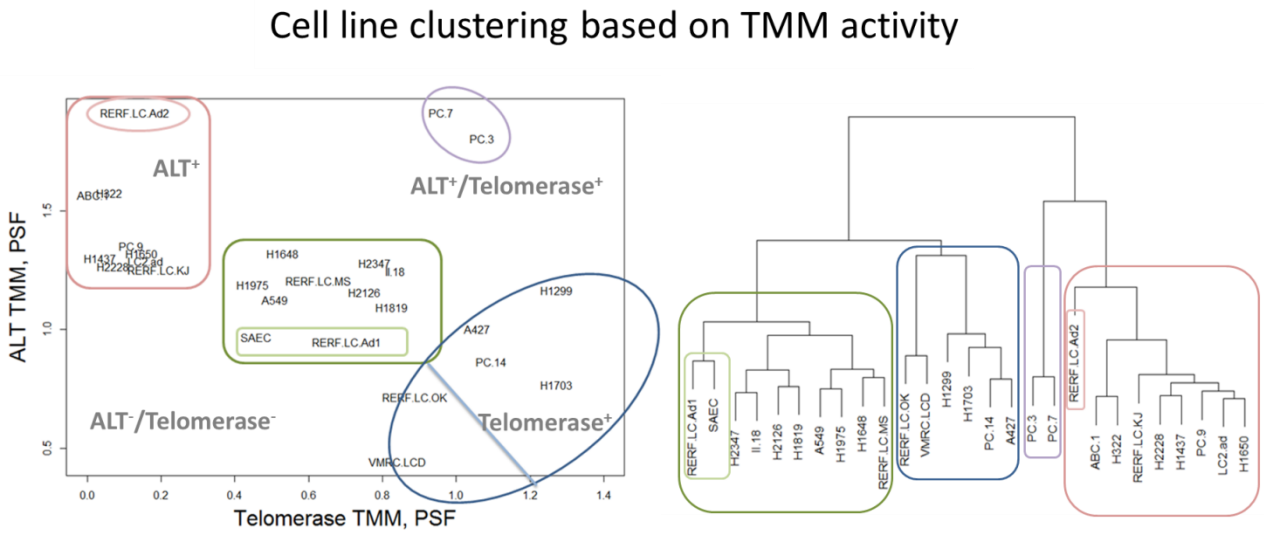


Figure 28. Clustering of lung adenocarcinoma cell lines based on ALT and Telomerase TMM activity values. (Left): The mapping of cell lines and clusters on a 2D coordinate space, where the x and y axes show the PSF values for Telomerase and ALT TMMs, respectively. (**Right**): hierarchical clustering dendrogram based on complete linkage and Euclidean distance defined by Telomerase and ALT TMM PSF values. The clusters are colored similarly in the left and right plots. The red and blue clusters correspond to the ALT$^+$ and Telomerase$^+$ predicted phenotypes, respectively. The violet cluster has PSF values corresponding to the ALT$^+$/Telomerase$^+$ phenotype. The green cluster has low PSF values for both TMMs. The healthy lung cell line (SAEC) resides in this cluster, and is in a smaller sub-cluster with one RERF.LC.Ad1 lung adenocarcinoma cell line.

As mentioned previously, we had also computed the mean telomere length (MTL) of the Lung AC cell lines using Computel (see Chapter 3 for details). We have then performed correlation analysis and group comparisons to identify whether MTL was associated with TMM activity, and found no significant association. We have also not identified a link between TMM activity and the origin of cell lines. It can, however, be noticed that cells with extremely long telomeres (> 15 kb) are found in clusters with significant activity of TMMs: PC.9 (19 kb) and ABC.1 (11kb) in the ALT$^+$ cluster,

PC.3 (15 kb) in the ALT$^+$/Telomerase$^+$ cluster, and H1703 (33 kb) and PC.14 (16 kb) in the Telomerase$^+$ cluster.

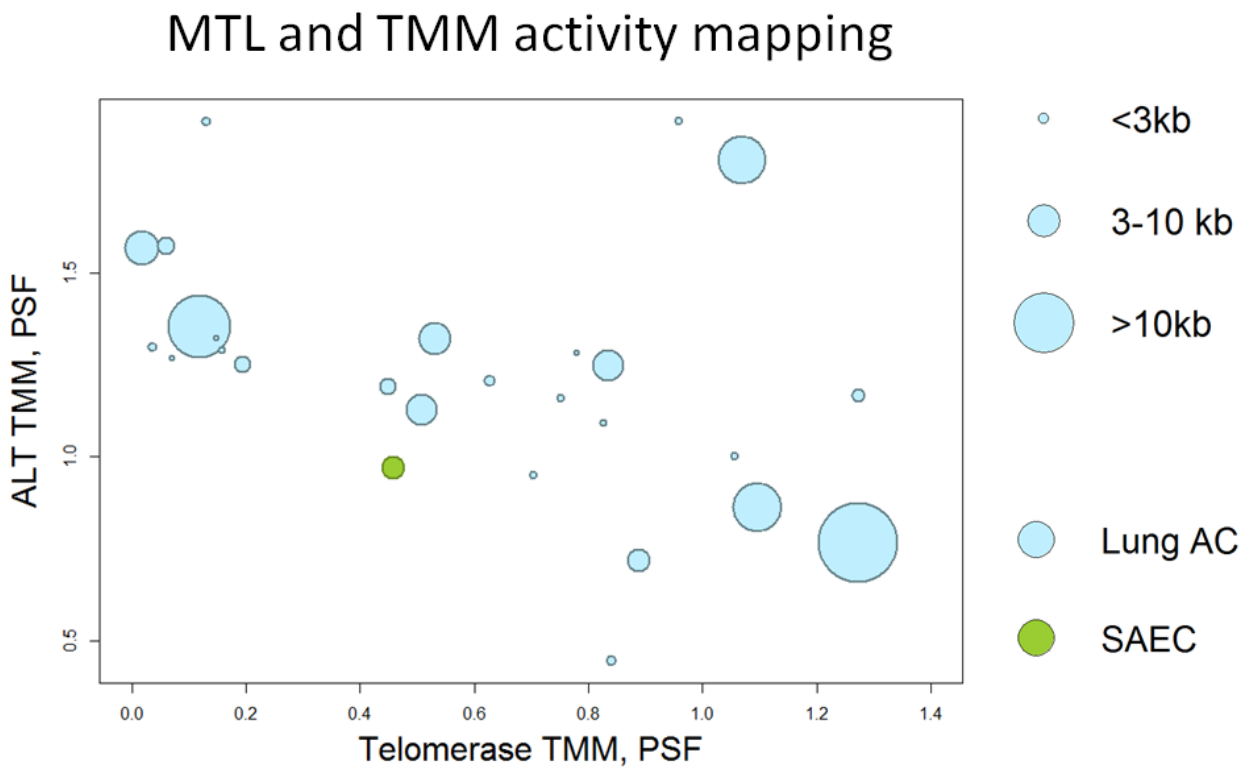


Figure 29. Map of the activities of Telomerase and ALT TMMs in Lung AC cell lines and their MTLs. The blue circles represent Lung AC cell lines, while the green one is the healthy lung cell line (SAEC). The size of the circles is proportional to the MTL – except for the very long and very short ones, which were constrained to maximum and minimum circle sizes to enhance visibility. The MTL for SAEC was not computed, since WGS data for it was not available – it was given an average circle size.

5.2.5 COMPARISON TO FUNCTIONAL ANNOTATION ANALYSIS

Functional annotation using over-representation (ORA) [239] or gene set enrichment analysis (GSEA) [240] is the usual pipeline applied to understand the phenotypic differences between two groups of samples under investigation. We have performed ORA on our test datasets (cell lines and liposarcoma group) to compare the power of that methodology to our TMM based approach. For this we have conducted gene-wise Kruskal-Walllis tests to obtained genes differentially expressed in ALT+ and Telomerase+ cell lines/tumors. Multiple test correction with Benjamini-Hochberg method did not revealed any significantly deregulated gene nor in cell lines, neither in tumors. We have thus taken top 100 genes with unadjusted p values of less than 0.01, and have supplied them to functional annotation packages David (https://david.ncifcrf.gov/) [241, 242] and

Webgestalt (http://webgestalt.org/) [243]. According to the results, no functional category was significantly enriched. Among the non-significantly enriched terms/categories, none was explicitly related to telomere maintenance mechanisms. Moreover, there was only one gene (*NANS*) common between the two lists of differentially expressed genes, which means that the lists are not identified by ALT^+ versus $telomerase^+$ phenotypes, but possibly by other phenotypic differences between the two sets of cells. Finally, none of the genes included in our TMM pathways was present in the lists of top 100 differentially regulated genes.

Previously, Lafferty-Whyte *et al* [232], had identified a set of 297 genes that were able to cluster ALT+ and telomerase+ cell lines and liposarcoma samples. We have obtained the list of the genes in this signature by request from the authors, and have submitted it to overrepresentation analysis by Webgestalt. According to the results, these genes were significantly enriched in GO terms associated with Golgi vesicle transport (GO:0048193, FDR = 0.16, 10 genes), Protein transport (GO:0015031, FDR = 0.16, 24 genes) and Establishment of protein localization (GO:0045184, FDR = 0.16, 25 genes). No category or pathway related to telomeres was enriched. Comparison of the signature set with the genes involved in our TMMs revealed three common genes: *TERT* and *PRKCA* from the Telomerase pathway, and *PRK* and *POLD3* from the ALT pathway.

5.3 DISCUSSION

Telomere maintenance and elongation mechanisms have long been investigated, and have gained particular attention in the last decade. Those are interesting to many labs focused on basic research in aging, stem cells, senescence etc., as well as to those engaged in translational research, particularly in development of TMM targeting anticancer therapies. While the TMM pathways are laborious to study experimentally, a lot of information has already been accumulated. However, the main obstacle for their efficient utilization is the absence of well-established and valid TMM models and poor utilization of the recent developments in high-throughput technologies, mainly because of the lack of respective computational methodologies.

The TMM pathways we have created here were intended to serve as a computational model to integrate the studies of telomere maintenance mechanisms into high-throughput data analysis pipelines. Gene expression-based assessment of TMM activities has allowed us to predict the TMM phenotypes of ten cell lines with 100% accuracy. The accuracy of predicting the phenotypes

of liposarcoma tumors and hMSCs, however, was lower: around 60%: with some misclassified ALT^+ and $Telomerase^+$ tumors, and with hMSCs all being classified as ALT^+. The misclassification of liposarcoma samples may be explained by several possible reasons, such probable incompleteness of the TMM pathways, as well as the sensitivity/specificity of TMM detection assays. In liposarcoma samples, telomerase activity was measured by TRAP assay [233] and ALT was detected by measuring the amount of APB bodies [233]. It is worth mentioning that, as other PML bodies, APBs may be generated because of cellular senescence, and are not necessarily indicative of ALT activity in the cell [221]. It is a question of further investigations whether the misclassified tumor samples are, in fact, misclassified by our approach, or they are misclassified by the experimental assays. On the other hand, the possibility of existence of ALT positive and telomerase positive cells within the same tumor should also not be excluded. In other words, it's possible that RNA-sequencing be performed on ALT positive cells and TMM detection – on $telomerase^+$ cells derived from the same tumor, and vice versa, thus leading to the observed misclassification.

The mesenchimal stem cells were classified by us as having considerable levels of ALT activity and very low levels of telomerase activity (Figure 27). As has already been mentioned, the mechanism of telomere length maintenance in hMSCs is still not established: many studies have identified very low or undetectable levels of telomerase activity [237], and some have performed ALT activity assays, but have not detected significant levels of ALT markers [238]. However, it is known that the majority of tumors with ALT activity originate from mesenchymal cells [201]. The fact that we have detected ALT activity in hMSCs tempts to speculate that mesenchymal cells normally have high expression levels of genes involved in the ALT pathway, which may ultimately foster these cells to more easily convert to ALT phenotype during malignant transformations.

The relatively high accuracy of our TMM based approach has allowed us to use it for investigation of TMM activation states of the lung adenocarcinoma cell lines, where we have identified four groups, three of which were described by high activation levels of either ALT or Telomerase pathways or both, and the other group by low-to-middle activity values for both TMM pathways (with the healthy cell line having the lowest TMM activities in this group). We have also observed that even though cell lines with extremely long telomeres (> 15kb) had high activity of either ALT or Telomerase pathways or both, there was no general association between MTL and TMM in the

rest of the cell lines. The absence of association between MTL and TMM is not surprising. Telomerase usually gets activated when telomeres become extremely short, and it usually elongates the shortest telomeres among the different chromosome ends [200]. Thus, since extremely short telomeres are powerful triggers of telomerase activation, cells with very short telomeres may have very high activity of Telomerase pathway. On the other hand, constant high activity of telomerase caused by other factors, such as genomic variations, may induce extreme lengthening of telomeres. In our dataset, the H1703 cell line has extremely long telomeres (32 kb) with very high activity of the Telomerase and very low activity of the ALT pathway. The same reasoning may be applied to the cells with ALT^+ phenotype. According to a common view, cells may undergo phenotype switching from telomerase to ALT, when internal or external factors do not permit for telomerase activation [203, 226]. In such cases, telomeres shorten and reach extremely low values, and afterwards trigger ALT activation in the absence of telomerase-permissive environment. At this stage there may be cells with high expression of genes related to ALT and with yet short telomeres. In all other cases, cells with persistent ALT activity are usually characterized with abnormally long telomeres at some chromosome ends, and very short telomeres on the rest of chromosomes. We cannot observe such length heterogeneity when measuring MTL only. Among the ALT^+ cell lines defined by our TMMs, PC.9 had abnormally long mean telomere length of >19kb. One of the two cell lines (PC.3) with ALT^+/$Telomerase^+$ phenotypes (predicted by TMMs) also had very high MTL of > 15kb. Overall, this indicates that while mere MTL computation is of great value by itself, detection of the active TMM should be performed separately, since there is no deterministic link between those two.

It is important to mention that the TMM activity values computed by PSF are strictly dependent on the phenotypic composition of the data, since the PSF algorithm takes as input the fold change values that are computed in reference to the mean expression value of each gene across the samples, as described in Methods. Thus, if the dataset we have used contained not one but more cell lines from healthy lungs, the predicted TMM phenotypes would probably have a different distribution (Figure 28).

In order to compare the power of our methodology to that of functional annotation analysis, we have used the lists of genes showing differential expression in ALT^+ versus $Telomerase^+$ samples,

as well as have used the list of genes generated by the Lafferty-Whyte *et al* [232] based on regression analysis, and submitted those sets to Overrepresentation analysis (ORA). None of the lists did return any functional category associated with telomere maintenance and lengthening mechanisms. It's important to take into consideration that larger sample sizes might produce lists more related to TMM mechanisms. However, while the ORA or GSEA approaches are valuable and easy-to-use methods for identification of functional categories able to distinguish between studied groups, they have two main disadvantages that make them invalid for use in TMM prediction or for identification of genes related to TMM. First, the lists of differentially expressed genes speak not only about the phenotype of interest, but also about other factors, such as batch effects, sampling and random noise. Thus, functional annotation analyses on different samples with the same phenotypic difference will, in most of the cases, reveal inconsistent results. On the other hand, they strictly depend on the quality and completeness of respective functional categories. So far, there is no publicly available complete functional category/pathway that has thoroughly collected genes related to ALT or Telomerase pathways. The most relevant functional category is the "Regulation of telomere maintenance via telomere lengthening" term available in the Gene Ontology database (GO:1904356). There are 75 *Homo sapiens* gene annotations to this term. Some of those are part of our TMM pathways, however most of them are genes that regulate expression of *TERT* or other genes; some indirectly regulate telomere maintenance; and some genes encode shelterin or telomere-interacting proteins. Of important consideration is also the fact that enrichment of DEG genes in one functional category does not imply whether that category is activated or repressed in the studied samples: some of the genes in the category may be under-expressed and some may be over-expressed, which in many cases will lead to an 'averaging out' effect.

Overall, data driven methods, such as functional annotation clustering, are powerful tools to detect unknown differences between two groups of samples. However, functional annotation analysis is not powerful in the particular case of detecting the TMM phenotypes, because of the lack of properly curated and annotated functional categories or pathways. Another shortcoming of data driven methods is their inconsistency across studies. Mining for differentially regulated genes between ALT^+ and $Telomerase^+$ samples will reveal varying sets of genes, as was the case in [232]. Moreover, the genes associated with the TMM phenotypes revealed by data driven methods are not necessarily involved in TMM activation pathways: they may have regulatory functions,

indirect effects, confounding factors and common oncogenes. The use of such gene sets, thus, does not provide the advantage of gaining knowledge and insight into the actual mechanisms of TMM pathway activation. We think that the TMM pathways constructed in this thesis and the PSF based computation of their activities is more valuable in this regard.

5.4 CONCLUSION

The TMM pathways that we have constructed allow for further extensions, in hand with accumulation of data and knowledge on TMM mechanisms. As new data come in, we may extend or modify the existing model to see whether its prediction accuracy increases. In other words, we have developed this methodology not only for phenotype prediction, but also for using the pathways to gain deeper insights into telomere maintenance mechanisms. Besides this, we have demonstrated that the results obtained by functional annotation analysis are inconsistent and are not powerful enough to identify the gene sets associated with TMMs. Thus, another major advantage of the pathway based approach is that the pathways are stable.

Even though we have used here only gene expression data to assess the activities of the TMM pathways, other layers of information would add to the accuracy of predictions, such as genomic variations, abundance of telomeric repeat variants, as well as DNA methylation and histone modification data.

Telomeres are critical players in normal developmental, regulation of cellular activities, as well as for initiation and progression of complex disorders, including age-related diseases and cancers. Telomere biology is investigated by a plethora of experimental methods in a number of model organisms and humans. It's also been investigated by groups involved in translational research that target telomeres and telomere biology to fight cancers and aging. Despite the importance of the topic and the recently increasing attention towards it, the experimental methodologies for studying telomere biology are still laborious and of low throughput in nature. This creates a huge impediment for integrative analysis of telomeres that would benefit from the power of high-throughput technologies.

The main aim of current study was to generate computational methods and models to study telomere length dynamics, their association with genomic, transcriptomic and epigenomic characteristics of a cell, and to study the mechanisms that regulate telomere length in health and disease. On one side, we have developed Computel for computation of mean telomere length from whole genome sequencing data. It has allowed us to utilize publicly available DNA-seq data to compute MTL, and to couple this information with RNA-seq, Bisulfite-seq and ChIP-seq data to study the association of MTL with gene expression and epigenetics. On the other side, we have constructed a pathway based model of telomere maintenance mechanisms (TMM) that has allowed us to study the activation states of Telomerase dependent and alternative lengthening mechanisms (ALT) based on high-throughput gene expression data. Our TMM based approach is not only intended for prediction of TMM activities in the cells, but most importantly for using it as an extendable model to help gain deeper insights into the mechanisms and factors leading to TMM activation in cancer cells.

Overall, we believe that these approaches are valuable for utilizing the existing multi-*omics* data to enhance understanding of telomere biology. Inclusion of additional layers of information, particularly in the TMM pathways, will add extra layer of accuracy and informative power to our models.

1. Hayflick L: **The limited in vitro lifetime of human diploid cell strains**. *Exp Cell Res* 1965, **37**:614–636.

2. Shay JW, Wright WE: **Hayflick, his limit, and cellular ageing.** *Nat Rev Mol Cell Biol* 2000, **1**(October):72–76.

3. Watson JD: **Origin of concatemeric T7 DNA.** *Nat New Biol* 1972, **239**:197–201.

4. McClintock B: **The Stability of Broken Ends of Chromosomes in Zea Mays.** *Genetics* 1941, **26**:234–282.

5. Baker WK: **Studies in Genetics. The selected papers of H. J. Muller. Indiana University Press, Bloomington, 1962. xiv + 618 pp. Illus**. *Science (80-)* 1963, **140**:285.

6. Blackburn EH, Gall JG: **A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena**. *J Mol Biol* 1978, **120**:33–53.

7. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, Meyne J, Ratcliff RL, Wu JR: **A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres\rof human chromosomes**. *Proc. Natl. Acad. Sci. U S A* 1988:6622–6626.

8. Greider CW, Blackburn EH: **Identification of a specific telomere terminal transferase activity in tetrahymena extracts**. *Cell* 1985, **43**(2 PART 1):405–413.

9. Orr-Weaver TL, Szostak JW, Rothstein RJ: **Yeast transformation: a model system for the study of recombination.** *Proc Natl Acad Sci U S A* 1981, **78**:6354–6358.

10. **The Nobel Prize in Physiology or Medicine 2009**
[http://www.nobelprize.org/nobel_prizes/medicine/laureates/2009/]

11. Xi H, Li C, Ren F, Zhang H, Zhang L: **Telomere, aging and age-related diseases**. *Aging Clinical and Experimental Research* 2013:139–146.

12. Greenberg R a: **Telomeres, crisis and cancer.** *Curr Mol Med* 2005, **5**:213–8.

13. Robin JD, Ludlow AT, Batten K, Magdinier F, Stadler G, Wagner KR, Shay JW, Wright WE: **Telomere position effect: Regulation of gene expression with progressive telomere shortening over long distances**. *Genes Dev* 2014, **28**:2464–2476.

14. Baur JA, Zou Y, Shay JW, Wright WE: **Telomere position effect in human cells.** *Science* 2001, **292**:2075–7.

15. Hern??ndez-Caballero E, Herrera-Gonz??lez NE, Salamanca-G??mez F, Arenas-Aranda DJ: **Role of telomere length in subtelomeric gene expression and its possible relation to cellular senescence**. *BMB Rep* 2009, **42**:747–751.

16. Blackburn EH: **Structure and function of telomeres.** *Nature* 1991, **350**:569–73.

17. Lansdorp PM, Verwoerd NP, Van De Rijke FM, Dragowska V, Little MT, Dirks RW, Raap AK, Tanke HJ: **Heterogeneity in telomere length of human chromosomes**. *Hum Mol Genet* 1996, **5**:685–691.

18. Monaghan P: **Telomeres and life histories: The long and the short of it**. *Annals of the New York Academy of Sciences* 2010:130–142.

19. Samassekou O, Gadji M, Drouin R, Yan J: **Sizing the ends: Normal length of human telomeres**. *Ann Anat* 2010, **192**:284–291.

20. Meyne J, Ratliff RL, Moyzis RK: **Conservation of the human telomere sequence (TTAGGG)n among vertebrates.** *Proc Natl Acad Sci U S A* 1989, **86**:7049–53.

21. Okazaki S, Tsuchida K, Maekawa H, Ishikawa H, Fujiwara H: **Identification of a pentanucleotide telomeric sequence, (TTAGG)n, in the silkworm Bombyx mori and in other insects.** *Mol Cell Biol* 1993, **13**:1424–1432.

22. Richards EJ, Ausubel FM: **Isolation of a higher eukaryotic telomere from Arabidopsis thaliana**. *Cell* 1988, **53**:127–136.

23. Zakian VA: **Telomeres: Beginning to Understand the End**. *Science (80-)* 1995, **270**:1601–1607.

24. Wellinger RJ, Zakian VA: **Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: Beginning to end**. *Genetics* 2012:1073–1105.

25. Rahman R, Forsyth NR, Cui W: **Telomeric 3???-overhang length is associated with the size of telomeres**. *Exp Gerontol* 2008, **43**:258–265.

26. Doksani Y, Wu JY, Lange T De, Zhuang X: **Super-Resolution Fluorescence Imaging of Telomeres Reveals TRF2- Dependent T-loop Formation**. *Cell* 2013, **155**:345–356.

27. Palm W, de Lange T: **How shelterin protects mammalian telomeres.** *Annu Rev Genet* 2008, **42**:301–34.

28. de Lange T: **How telomeres solve the end-protection problem.** *Science (80-)* 2009, **326**:948–52.

29. Lei M, Podell ER, Cech TR: **Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection**. *Nat Struct & Mol Biol* 2004, **11**:1223–1229.

30. Chen Y, Yang Y, van Overbeek M, Donigian JR, Baciu P, de Lange T, Lei M: **A shared docking motif in TRF1 and TRF2 used for differential recruitment of telomeric proteins.** *Science* 2008, **319**:1092–6.

31. Blasco MAM a M: **The epigenetic regulation of mammalian telomeres.** *Nat Rev Genet* 2007, **8**:299–309.

32. Bandaria JN, Qin P, Berk V, Chu S, Yildiz A: **Shelterin Protects Chromosome Ends by Compacting Telomeric Chromatin**. *Cell* 2016, **164**:735–746.

33. García-Cao M, O'Sullivan R, Peters AHFM, Jenuwein T, Blasco M a: **Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases.** *Nat Genet* 2004, **36**:94–99.

34. Gonzalo S, Blasco MA, Jaco I, Fraga MF, Chen T, Li E, Esteller M, Blasco M: **DNA**

methyltransferases control telomere length and telomere recombination in mammalian cells. *Nat Cell Biol* 2006, **8**:416–24.

35. Blasco M, Schoeftner S: **A "higher order" of telomere regulation: telomere heterochromatin and telomeric RNAs.** *EMBO J* 2009, **28**:2323–36.

36. García-Cao M, O'Sullivan R, Peters AHFM, Jenuwein T, Blasco MA: **Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases**. *Nat Genet* 2003, **36**:94–99.

37. Marion RM, Strati K, Li H, Tejera A, Schoeftner S, Ortega S, Serrano M, Blasco MA: **Telomeres Acquire Embryonic Stem Cell Characteristics in Induced Pluripotent Stem Cells**. *Cell Stem Cell* 2009, **4**:141–154.

38. Flores I, Canela A, Vera E, Tejera A, Cotsarelis G, Blasco MA: **The longest telomeres: A general signature of adult stem cell compartments**. *Genes Dev* 2008, **22**:654–667.

39. Jackson SP, Bartek J: **The DNA-damage response in human biology and disease**. *Nature* 2009, **461**:1071–8.

40. Meena J, Rudolph KL, Günes C: **Telomere dysfunction, chromosomal instability and cancer**. In *Chromosomal Instability in Cancer Cells*; 2015:61–79.

41. Mondoux M a., Zakian V a.: *Telomere Position Effect: Silencing Near the End*. *Volume 45*; 2006.

42. Tennen RI, Bua DJ, Wright WE, Chua KF: **SIRT6 is required for maintenance of telomere position effect in human cells**. *Nat Commun* 2011, **2**:433.

43. Kitada T, Kuryan BG, Tran NNH, Song C, Xue Y, Carey M, Grunstein M: **Mechanism for epigenetic variegation of gene expression at yeast telomeric heterochromatin**. *Genes Dev* 2012, **26**:2443–2455.

44. Gottschling DE, Aparicio OM, Billington BL, Zakian VA: **Position effect at S. cerevisiae telomeres: Reversible repression of Pol II transcription**. *Cell* 1990, **63**:751–762.

45. Mason JM, Konev AY, Biessmann H: **Telomeric position effect in Drosophila melanogaster reflects a telomere length control mechanism**. *Genetica* 2003:319–325.

46. Pedram M, Sprung CN, Gao Q, Lo AWI, Reynolds GE, Murnane JP: **Telomere position effect and silencing of transgenes near telomeres in the mouse.** *Mol Cell Biol* 2006, **26**:1865–1878.

47. Palmer JM, Mallaredy S, Perry DW, Sanchez JF, Theisen JM, Szewczyk E, Oakley BR, Wang CCC, Keller NP, Mirabito PM: **Telomere position effect is regulated by heterochromatin-associated proteins and NkuA in Aspergillus nidulans**. *Microbiology* 2010, **156**:3522–3531.

48. Baranasic D, Oppermann T, Cheaib M, Cullum J, Schmidt H, Simon M: **Genomic characterization of variable surface antigens reveals a telomere position effect as a prerequisite for RNA interference-mediated silencing in Paramecium tetraurelia**. *MBio* 2014, **5**.

49. Ottaviani A, Gilson E, Magdinier F: **Telomeric position effect: From the yeast paradigm to human pathologies?** *Biochimie* 2008, **90**:93–107.

50. Koering CE, Pollice A, Zibella MP, Bauwens S, Puisieux A, Brunori M, Brun C, Martins L, Sabatier L, Pulitzer JF, Gilson E: **Human telomeric position effect is determined by chromosomal context and telomeric chromatin integrity**. *EMBO Rep* 2002, **3**:1055–1061.

51. Baur JA, Zou Y, Shay JW, Wright WE: **Telomere position effect in human cells.** *Science* 2001, **292**:2075–7.

52. Stadler G, Rahimov F, King OD, Chen JC, Robin JD, Wagner KR, Shay JW, Emerson Jr. CP, Wright WE: **Telomere position effect regulates DUX4 in human facioscapulohumeral muscular dystrophy**. *Nat Struct Mol Biol* 2013, **20**:671–678.

53. Robin JD, Ludlow AT, Batten K, Gaillard MC, Stadler G, Magdinier F, Wright WE, Shay JW: **SORBS2 transcription is activated by telomere position effect-over long distance upon telomere shortening in muscle cells from patients with facioscapulohumeral dystrophy**. *Genome Res* 2015, **25**:1781–1790.

54. Romina Burla, Mattia La Torre IS: **Mammalian telomeres and their partnership with lamins**. *J Chem Inf Model* 2013, **53**:1689–1699.

55. Lou Z, Wei J, Riethman H, Baur JA, Voglauer R, Shay JW, Wright WE: **Telomere length regulates ISG15 expression in human cells.** *Aging (Albany NY)* 2009, **1**:608–621.

56. Ning Y, Xu JF, Li Y, Chavez L, Riethman HC, Lansdorp PM, Weng NP: **Telomere length and the expression of natural telomeric genes in human fibroblasts**. *Hum Mol Genet* 2003, **12**:1329–1336.

57. Kim W, Ludlow AT, Min J, Robin JD, Stadler G, Mender I, Lai T-P, Zhang N, Wright WE, Shay JW: **Regulation of the Human Telomerase Gene TERT by Telomere Position Effect-Over Long Distances (TPE-OLD): Implications for Aging and Cancer.** *PLoS Biol* 2016, **14**:e2000016.

58. Blasco M, Schoeftner S: **Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II.** *Nat Cell Biol* 2008, **10**:228–36.

59. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J: **Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends.** *Science* 2007, **318**:798–801.

60. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J: **Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends.** *Science* 2007, **318**:798–801.

61. Blasco M, Schoeftner S: **Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II.** *Nat Cell Biol* 2008, **10**:228–36.

62. Blasco M, Schoeftner S: **A "higher order" of telomere regulation: telomere heterochromatin and telomeric RNAs.** *EMBO J* 2009, **28**:2323–36.

63. Martínez P, Blasco MA: **Replicating through telomeres: A means to an end**. *Trends in Biochemical Sciences* 2015:504–515.

64. Chow TT, Zhao Y, Mak SS, Shay JW, Wright WE: **Early and late steps in telomere overhang processing in normal human cells: The position of the final RNA primer drives telomere shortening**. *Genes Dev* 2012, **26**:1167–1178.

65. Sfeir AJ, Chai W, Shay JW, Wright WE: **Telomere-end processing the terminal nucleotides of human chromosomes.** *Mol Cell* 2005, **18**:131–8.

66. Trusina A: **Stress induced telomere shortening: longer life with less mutations?** *BMCSystBiol* 2014, **8**:27.

67. Hug N, Lingner J: **Telomere length homeostasis**. *Chromosoma* 2006:413–425.

68. Nittis T, Guittat L, Stewart SA: **Alternative lengthening of telomeres (ALT) and chromatin: Is there a connection?** *Biochimie* 2008:5–12.

69. Slagboom PE, Droog S, Boomsma DI: **Genetic determination of telomere size in humans: a twin study of three age groups.** *Am J Hum Genet* 1994, **55**:876–82.

70. Al-Attas OS, Al-Daghri NM, Alokail MS, Alkharfy KM, Alfadda AA, McTernan P, Gibson GC, Sabico SB, Chrousos GP: **Circulating leukocyte telomere length is highly heritable among families of Arab descent.** *BMC Med Genet* 2012, **13**:38.

71. Honig LS, Kang MS, Cheng R, Eckfeldt JH, Thyagarajan B, Leiendecker-Foster C, Province MA, Sanders JL, Perls T, Christensen K, Lee JH, Mayeux R, Schupf N: **Heritability of telomere length in a study of long-lived families.** *Neurobiol Aging* 2015, **36**:2785–90.

72. Ju Z, Rudolph KL: **Telomeres and telomerase in stem cells during aging and disease.** *Genome Dyn* 2006, **1**:84–103.

73. Blasco M a: **Telomere length, stem cells and aging.** *Nat Chem Biol* 2007, **3**:640–649.

74. Kappei D, Londoño-Vallejo JA: **Telomere length inheritance and aging**. *Mech Ageing Dev* 2008, **129**:17–26.

75. Mather KA, Jorm AF, Parslow RA, Christensen H: **Is telomere length a biomarker of aging? A review**. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 2011:202–213.

76. M??ezzinler A, Zaineddin AK, Brenner H: **A systematic review of leukocyte telomere length and age in adults**. *Ageing Research Reviews* 2013:509–519.

77. Roux A V, Ranjit N, Jenny NS, Shea S, Cushman M, Fitzpatrick A, Seeman T: **Race/ethnicity and telomere length in the Multi-Ethnic Study of Atherosclerosis**. *Aging Cell* 2009, **8**:251–257.

78. Hakobyan A, Nersisyan L, Arakelyan A: **Quantitative trait association study for mean telomere length in the South Asian genomes**. *Bioinformatics* 2016, **32**:1697–1700.

79. Lansdorp PM: **Telomeres and disease**. *EMBO J* 2009, **28**:2532–2540.

80. Savage SA, Alter BP: **Dyskeratosis Congenital**. *Hematol Clin NORTH Am* 2009, **23**:215+.

81. Yamaguchi H, Calado RT, Ly H, Kajigaya S, Baerlocher GM, Chanock SJ, Lansdorp PM, Young NS: **Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia.** *N Engl J Med* 2005, **352**:1413–24.

82. Tsakiri KD, Cronkhite JT, Kuan PJ, Xing C, Raghu G, Weissler JC, Rosenblatt RL, Shay JW, Garcia CK: **Adult-onset pulmonary fibrosis caused by mutations in telomerase.** *Proc Natl Acad Sci U S A* 2007, **104**:7552–7.

83. Takai H, Smogorzewska A, De Lange T: **DNA damage foci at dysfunctional telomeres**. *Curr Biol* 2003, **13**:1549–1556.

84. Kota LN, Bharath S, Purushottam M, Moily NS, Sivakumar PT, Varghese M, Pal PK, Jain S: **Reduced telomere length in neurodegenerative disorders may suggest shared biology.** *J Neuropsychiatry Clin Neurosci* 2015, **27**:e92-6.

85. Farzaneh-Far R, Cawthon RM, Na B, Browner WS, Schiller NB, Whooley MA: **Prognostic value of leukocyte telomere length in patients with stable coronary artery disease: Data from the heart and soul study**. *Arterioscler Thromb Vasc Biol* 2008, **28**:1379–1384.

86. Martin-Ruiz C, Dickinson HO, Keys B, Rowan E, Kenny RA, Von Zglinicki T: **Telomere length predicts poststroke mortality, dementia, and cognitive decline**. *Ann Neurol* 2006, **60**:174–180.

87. Salpea KD, Talmud PJ, Cooper JA, Maubaret CG, Stephens JW, Abelak K, Humphries SE: **Association of telomere length with type 2 diabetes, oxidative stress and UCP2 gene variation**. *Atherosclerosis* 2010, **209**:42–50.

88. Huzen J, van Veldhuisen DJ, van Gilst WH, van der Harst P: **Telomeres and biological ageing in cardiovascular disease**. *Ned Tijdschr Geneeskd* 2008, **152**:1265–1270.

89. Cawthon RM, Smith KR, O'Brien E, Sivatchenko A, Kerber RA: **Association between telomere length in blood and mortality in people aged 60 years or older**. *Lancet* 2003, **361**:393–395.

90. Andrews NP, Fujii H, Goronzy JJ, Weyand CM: **Telomeres and immunological diseases of aging**. *Gerontology* 2010:390–403.

91. Georgin-Lavialle S, Aouba A, Mouthon L, Londono-Vallejo JA, Lepelletier Y, Gabet AS, Hermine O: **The telomere/telomerase system in autoimmune and systemic immune-mediated diseases**. *Autoimmunity Reviews* 2010:646–651.

92. Campbell PJ: **Telomeres and cancer: From crisis to stability to crisis to stability**. *Cell* 2012, **148**:633–635.

93. Shay JW, Wright WE: **Role of telomeres and telomerase in cancer**. *Seminars in Cancer Biology* 2011:349–353.

94. Reddel RR: **Alternative lengthening of telomeres, telomerase, and cancer**. *Cancer Lett* 2003, **194**:155–162.

95. Bertorelle R, Rampazzo E, Pucciarelli S, Nitti D, De Rossi A: **Telomeres, telomerase and colorectal cancer.** *World J Gastroenterol* 2014, **20**:1940–50.

96. Ruden M, Puri N: **Novel anticancer therapeutics targeting telomerase**. *Cancer Treatment Reviews* 2013:444–456.

97. Fernndez-Marcelo T, Morn A, De Juan C, Pascua I, Head J, Gmez A, Hernando F, Lpez-Asenjo JA, Hernndez S, Snchez-Pernaute A, Torres AJ, Benito M, Iniesta P: **Differential expression of senescence and cell death factors in non-small cell lung and colorectal tumors showing telomere attrition**. *Oncology* 2012, **82**:153–164.

98. Aubert G, Hills M, Lansdorp PM: **Telomere length measurement-Caveats and a critical assessment of the available technologies and tools**. *Mutation Research - Fundamental and*

Molecular Mechanisms of Mutagenesis 2012:59–67.

99. Kimura M, Stone RC, Hunt SC, Skurnick J, Lu X, Cao X, Harley CB, Aviv A: **Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths.** *Nat Protoc* 2010, **5**:1596–607.

100. Göhring J, Fulcher N, Jacak J, Riha K: **TeloTool: A new tool for telomere length measurement from terminal restriction fragment analysis with improved probe intensity correction**. *Nucleic Acids Res* 2014, **42**.

101. Eisenberg DTA, Kuzawa CW, Hayes MG: **Improving qPCR telomere length assays: Controlling for well position effects increases statistical power**. *American Journal of Human Biology* 2015.

102. Cawthon RM: **Telomere length measurement by a novel monochrome multiplex quantitative PCR method**. *Nucleic Acids Res* 2009, **37**.

103. Baird DM, Rowson J, Wynford-Thomas D, Kipling D: **Extensive allelic variation and ultrashort telomeres in senescent human cells.** *Nat Genet* 2003, **33**:203–7.

104. Britt-Compton B, Rowson J, Locke M, Mackenzie I, Kipling D, Baird DM: **Structural stability and chromosome-specific telomere length is governed by cis-acting determinants in humans**. *Hum Mol Genet* 2006, **15**:725–733.

105. O'Sullivan JN, Finley JC, Risques R-A, Shen W-T, Gollahon K a, Rabinovitch PS: **Quantitative fluorescence in situ hybridization (QFISH) of telomere lengths in tissue and cells.** *Curr Protoc Cytom* 2005, **Chapter 12**:Unit 12.6.

106. Verhulst S, Susser E, Factor-Litvak PR, Simons MJP, Benetos A, Steenstrup T, Kark JD, Aviv A: **Commentary: The reliability of telomere length measurements**. *Int J Epidemiol* 2015, **44**:1683–1686.

107. Aviv A: **Commentary: Raising the bar on telomere epidemiology**. *Int J Epidemiol* 2009, **38**:1735–1736.

108. Buermans HPJ, den Dunnen JT: **Next generation sequencing technology: Advances and applications.** *Biochim Biophys Acta* 2014, **1842**:In Press.

109. Wang Y, Navin NE: **Advances and Applications of Single-Cell Sequencing Technologies**. *Molecular Cell* 2015:598–609.

110. Bick D, Dimmock D: **Whole exome and whole genome sequencing**. *Curr Opin Pediatr* 2011, **23**:594–600.

111. Leinonen R, Sugawara H, Shumway M: **The sequence read archive**. *Nucleic Acids Res* 2011, **39**(SUPPL. 1).

112. Kodama Y, Shumway M, Leinonen R: **The sequence read archive: Explosive growth of sequencing data**. *Nucleic Acids Res* 2012, **40**.

113. Parker M, Chen X, Bahrami A, Dalton J, Rusch M, Wu G, Easton J, Cheung N-K, Dyer M, Mardis ER, Wilson RK, Mullighan C, Gilbertson R, Baker SJ, Zambetti G, Ellison DW, Downing JR, Zhang J: **Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data.**

Genome Biol 2012, **13**:R113.

114. Ding Z, Mangino M, Aviv A, Spector T, Durbin R: **Estimating telomere length from whole genome sequence data**. *Nucleic Acids Res* 2014, **42**.

115. Cai N, Chang S, Li Y, Li Q, Hu J, Liang J, Song L, Kretzschmar W, Gan X, Nicod J, Rivera M, Deng H, Du B, Li K, Sang W, Gao J, Gao S, Ha B, Ho HY, Hu C, Hu J, Hu Z, Huang G, Jiang G, Jiang T, Jin W, Li G, Li K, La Y, Li Y, et al.: **Molecular signatures of major depression**. *Curr Biol* 2015, **25**:1146–1156.

116. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S, Wang J, Manyam G, Zoppoli P, Ling S, Rao AA, Grifford M, Cherniack AD, Zhang H, Poisson L, Carlotti Jr. CG, Tirapelli DP da C, Rao A, Mikkelsen T, Lau CC, Yung WKA, Rabadan R, Huse J, Brat DJ, Lehman NL, et al.: **Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma**. *Cell* 2016, **164**:550–563.

117. Nersisyan L, Arakelyan A: **Computel: Computation of mean telomere length from whole-genome next-generation sequencing data**. *PLoS One* 2015, **10**.

118. Azzalin CM, Nergadze SG, Giulotto E: **Human intrachromosomal telomeric-like repeats: sequence organization and mechanisms of origin.** *Chromosoma* 2001, **110**:75–82.

119. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357–359.

120. Huang W, Li L, Myers JR, Marth GT: **ART: A next-generation sequencing read simulator**. *Bioinformatics* 2012, **28**:593–594.

121. Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, Mudunuri U, Paul S, Wei J: **Mapping and initial analysis of human subtelomeric sequence assemblies**. *Genome Res* 2004, **14**:18–28.

122. Castle JC, Biery M, Bouzek H, Xie T, Chen R, Misura K, Jackson S, Armour CD, Johnson JM, Rohl C a, Raymond CK: **DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing.** *BMC Genomics* 2010, **11**:244.

123. Montpetit AJ, Alhareeri AA, Montpetit M, Starkweather AR, Elmore LW, Filler K, Mohanraj L, Burton CW, Menzies VS, Lyon DE, Jackson-Cook CK: **Telomere length: a review of methods for measurement.** *Nurs Res* 2014, **63**:289–299.

124. Ishizuka T, Xu Y, Komiyama M: **A chemistry-based method to detect individual telomere length at a single chromosome terminus**. *J Am Chem Soc* 2013, **135**:14–17.

125. Beirne C, Delahay R, Hares M, Young A: **Age-related declines and disease-associated variation in immune cell telomere length in a wild mammal**. *PLoS One* 2014, **9**.

126. Balistreri CR, Pisano C, Merlo D, Fattouch K, Caruso M, Incalcaterra E, Colonna-Romano G, Candore G: **Is the mean blood leukocyte telomere length a predictor for sporadic thoracic aortic aneurysm? Data from a preliminary study.** *Rejuvenation Res* 2012, **15**:170–173.

127. Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, Hottenga JJ, Fischer K, Esko T, Surakka I, Broer L, Nyholt DR, Mateo Leach I, Salo P, Hägg S, Matthews MK, Palmen J, Norata GD, O'Reilly PF, Saleheen D, Amin N, Balmforth AJ, Beekman M, de Boer RA, Böhringer S, Braund PS,

Burton PR, de Craen AJM, Denniff M, Dong Y, et al.: **Identification of seven loci affecting mean telomere length and their association with disease.** *Nat Genet* 2013, **45**:422–7, 427–2.

128. Guan J-Z, Guan W-P, Maeda T, Guoqing X, GuangZhi W, Makino N: **Patients with multiple sclerosis show increased oxidative stress markers and somatic telomere length shortening.** *Mol Cell Biochem* 2015, **400**:183–7.

129. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.

130. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: An improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**:1966–1967.

131. Li R, Li Y, Kristiansen K, Wang J: **SOAP: Short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**:713–714.

132. Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, Hottenga JJ, Fischer K, Esko TT, Surakka I, Broer L, Nyholt DR, Mateo Leach I, Salo P, Hägg S, Matthews MK, Palmen J, Norata GD, O'Reilly PF, Saleheen D, Amin N, Balmforth AJ, Beekman M, de Boer RA, Böhringer S, Braund PS, Burton PR, de Craen AJM, Denniff M, Dong Y, et al.: **Identification of seven loci affecting mean telomere length and their association with disease**. *Nat Genet* 2013, **45**:422.

133. Pooley KA, Bojesen SE, Weischer M, Nielsen SF, Thompson D, Amin Al Olama A, Michailidou K, Tyrer JP, Benlloch S, Brown J, Audley T, Luben R, Khaw KT, Neal DE, Hamdy FC, Donovan JL, Kote-Jarai Z, Baynes C, Shah M, Bolla MK, Wang Q, Dennis J, Dicks E, Yang R, Rudolph A, Schildkraut J, Chang-Claude J, Burwinkel B, Chenevix-Trench G, Pharoah PDP, et al.: **A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk**. *Hum Mol Genet* 2013, **22**:5056–5064.

134. Lee JH, Cheng R, Honig LS, Feitosa M, Kammerer CM, Kang MS, Schupf N, Lin SJ, Sanders JL, Bae H, Druley T, Perls T, Christensen K, Province M, Mayeux R: **Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: The long life family study**. *Front Genet* 2013, **4**(JAN).

135. Chambers JC, Abbott J, Zhang W, Turro E, Scott WR, Tan ST, Afzal U, Afaq S, Loh M, Lehne B, O'Reilly P, Gaulton KJ, Pearson RD, Li X, Lavery A, Vandrovcova J, Wass MN, Miller K, Sehmi J, Oozageer L, Kooner IK, Al-Hussaini A, Mills R, Grewal J, Panoulas V, Lewin AM, Northwood K, Wander GS, Geoghegan F, Li Y, et al.: **The South Asian genome**. *PLoS One* 2014, **9**.

136. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: An R library for genome-wide association analysis**. *Bioinformatics* 2007, **23**:1294–1296.

137. Price A, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904–9.

138. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**:559–575.

139. Bhupatiraju C, Saini D, Patkar S, Deepak P, Das B, Padma T: **Association of shorter telomere length with essential hypertension in Indian population**. *Am J Hum Biol* 2012, **24**:573–578.

140. Wong AM, Kong KL, Chen L, Liu M, Zhu C, Tsang JW, Guan XY: **Characterization of CACNA2D3 as a putative tumor suppressor gene in the development and progression of nasopharyngeal carcinoma**. *Int J Cancer* 2013.

141. Rode L, Nordestgaard BG, Bojesen SE: **Peripheral Blood Leukocyte Telomere Length and Mortality Among 64 637 Individuals From the General Population**. *JNCI J Natl Cancer Inst* 2015, **107**:djv074-djv074.

142. Li Z, Tang J, Li H, Chen S, He Y, Liao Y, Wei Z, Wan G, Xiang X, Xia K, Chen X: **Shorter telomere length in peripheral blood leukocytes is associated with childhood autism.** *Sci Rep* 2014, **4**:7073.

143. Hochstrasser T, Marksteiner J, Humpel C: **Telomere length is age-dependent and reduced in monocytes of Alzheimer patients**. *Exp Gerontol* 2012, **47**:160–163.

144. Benetos a, Okuda K, Lajemi M, Kimura M, Thomas F, Skurnick J, Labat C, Bean K, Aviv a: **Telomere length as an indicator of biological aging: the gender effect and relation with pulse pressure and pulse wave velocity.** *Hypertension* 2001, **37**(2 Part 2):381–385.

145. Zhu H, Wang X, Gutin B, Davis CL, Keeton D, Thomas J, Stallmann-Jorgensen I, Mooken G, Bundy V, Snieder H, Van Der Harst P, Dong Y: **Leukocyte telomere length in healthy caucasian and african-american adolescents: Relationships with race, sex, adiposity, adipokines, and physical activity**. *J Pediatr* 2011, **158**:215–220.

146. Diez Roux A V., Ranjit N, Jenny NS, Shea S, Cushman M, Fitzpatrick A, Seeman T: **Race/ethnicity and telomere length in the Multi-Ethnic Study of Atherosclerosis**. *Aging Cell* 2009, **8**:251–257.

147. Cassidy A, De Vivo I, Liu Y, Han J, Prescott J, Hunter DJ, Rimm EB: **Associations between diet, lifestyle factors, and telomere length in women**. *Am J Clin Nutr* 2010, **91**:1273–1280.

148. Hjelmborg JB, Dalgard C, Moller S, Steenstrup T, Kimura M, Christensen K, Kyvik KO, Aviv a.: **The heritability of leucocyte telomere length dynamics**. *J Med Genet* 2015, **52**:297–302.

149. Sebastiani P, Montano M, Puca A, Solovieff N, Kojima T, Wang MC, Melista E, Meltzer M, Fischer SEJ, Andersen S, Hartley SH, Sedgewick A, Arai Y, Bergman A, Barzilai N, Terry DF, Riva A, Anselmi CV, Malovini A, Kitamoto A, Sawabe M, Arai T, Gondo Y, Steinberg MH, Hirose N, Atzmon G, Ruvkun G, Baldwin CT, Perls TT: **RNA editing genes associated with extreme old age in humans and with lifespan in C. elegans**. *PLoS One* 2009, **4**.

150. Skinner HG, Gangnon RE, Litzelman K, Johnson RA, Chari ST, Petersen GM, Boardman LA: **Telomere length and pancreatic cancer: a case-control study**. *Cancer Epidemiol Biomarkers Prev* 2012, **21**:2095–2100.

151. Naing C, Aung K, Lai PK, Mak JW: **Association between telomere length and the risk of colorectal cancer: a meta-analysis of observational studies**. *BMC Cancer* 2017, **17**:24.

152. Shen M, Cawthon R, Rothman N, Weinstein SJ, Virtamo J, Hosgood HD, Hu W, Lim U, Albanes D, Lan Q: **A prospective study of telomere length measured by monochrome multiplex quantitative PCR and risk of lung cancer**. *Lung Cancer* 2011, **73**:133–137.

153. Weischer M, Nordestgaard BG, Cawthon RM, Freiberg JJ, Tybjærg-Hansen A, Bojesen SE: **Short telomere length, cancer survival, and cancer risk in 47102 individuals**. *J Natl Cancer Inst* 2013,

105:459–468.

154. Rode L, Nordestgaard BG, Bojesen SE: **Long telomeres and cancer risk among 95 568 individuals from the general population**. *Int J Epidemiol* 2016, **45**:1634–1643.

155. The Cancer Genome Atlas Research Network: **Comprehensive molecular profiling of lung adenocarcinoma**. *Nature* 2014, **advance on**:543–550.

156. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C, Auclair D, Lawrence MS, Stojanov P, Cibulskis K, Choi K, De Waal L, Sharifnia T, Brooks A, Greulich H, Banerji S, Zander T, Seidel D, Leenders F, Ans??n S, Ludwig C, Engel-Riedel W, Stoelben E, Wolf J, Goparju C, et al.: **Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing**. *Cell* 2012, **150**:1107–1120.

157. Zhang C, Doherty JA, Burgess S, Hung RJ, Lindström S, Kraft P, Gong J, Amos CI, Sellers TA, Monteiro ANA, Chenevix-Trench G, Bickeböller H, Risch A, Brennan P, Mckay JD, Houlston RS, Landi MT, Timofeeva MN, Wang Y, Heinrich J, Kote-Jarai Z, Eeles RA, Muir K, Wiklund F, Grönberg H, Berndt SI, Chanock SJ, Schumacher F, Haiman CA, Henderson BE, et al.: **Genetic determinants of telomere length and risk of common cancers: A Mendelian randomization study**. *Hum Mol Genet* 2015, **24**:5356–5366.

158. Fernández-Marcelo T, Gómez A, Pascua I, de Juan C, Head J, Hernando F, Jarabo J-R, Calatayud J, Torres-García A-J, Iniesta P: **Telomere length and telomerase activity in non-small cell lung cancer prognosis: clinical usefulness of a specific telomere status.** *J Exp Clin Cancer Res* 2015, **34**:78.

159. Suzuki A, Makinoshima H, Wakaguri H, Esumi H, Sugano S, Kohno T, Tsuchihara K, Suzuki Y: **Aberrant transcriptional regulations in cancers: Genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines**. *Nucleic Acids Res* 2014, **42**:13557–13572.

160. Schäfer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics* 2005, **21**:754–764.

161. Opgen-Rhein R, Strimmer K: **From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.** *BMC Syst Biol* 2007, **1**:37.

162. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**:139–140.

163. Komai K, Niwa Y, Sasazawa Y, Simizu S: **Pirin regulates epithelial to mesenchymal transition independently of Bcl3-Slug signaling**. *FEBS Lett* 2015, **589**:738–743.

164. Kastner S, Voss T, Keuerleber S, Glöckel C, Freissmuth M, Sommergruber W: **Expression of G protein-coupled receptor 19 in human lung cancer cells is triggered by entry into S-phase and supports G(2)-M cell-cycle progression.** *Mol Cancer Res* 2012, **10**:1343–58.

165. Jodoin JN, Sitaram P, Albrecht TR, May SB, Shboul M, Lee E, Reversade B, Wagner EJ, Lee LA: **Nuclear-localized Asunder regulates cytoplasmic dynein localization via its role in the Integrator complex**. *Mol Biol Cell* 2013, **24**:2954–2965.

166. Sahin E, Colla S, Liesa M, Moslehi J, Müller FL, Guo M, Cooper M, Kotton D, Fabian AJ, Walkey C, Maser RS, Tonon G, Foerster F, Xiong R, Wang YA, Shukla SA, Jaskelioff M, Martin ES, Heffernan TP,

Protopopov A, Ivanova E, Mahoney JE, Kost-Alimova M, Perry SR, Bronson R, Liao R, Mulligan R, Shirihai OS, Chin L, DePinho RA: **Telomere dysfunction induces metabolic and mitochondrial compromise.** *Nature* 2011, **470**:359–65.

167. Sahin E, Depinho R a: **Linking functional decline of telomeres, mitochondria and stem cells during ageing.** *Nature* 2010, **464**:520–528.

168. Dieckmann AK, Babin V, Harari Y, Eils R, König R, Luke B, Kupiec M: **Role of the ESCRT Complexes in Telomere Biology.** *MBio* 2016, **7**.

169. Taira M, Iizasa T, Shimada H, Kudoh J, Shimizu N, Tatibana M: **A human testis-specific mRNA for phosphoribosylpyrophosphate synthetase that initiates from a non-AUG codon**. *J Biol Chem* 1990, **265**:16491–16497.

170. Giannone RJ, McDonald HW, Hurst GB, Shen RF, Wang Y, Liu Y: **The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1**. *PLoS One* 2010, **5**.

171. Tanaka H, Maeda R, Shoji W, Wada H, Masai I, Shiraki T, Kobayashi M, Nakayama R, Okamoto H: **Novel mutations affecting axon guidance in zebrafish and a role for plexin signalling in the guidance of trigeminal and facial nerve axons**. *Development* 2007, **134**.

172. Schwarz Q, Waimey KE, Golding M, Takamatsu H, Kumanogoh A, Fujisawa H, Cheng H-J, Ruhrberg C: **Plexin A3 and plexin A4 convey semaphorin signals during facial nerve development.** *Dev Biol* 2008, **324**:1–9.

173. Wang C, Gu Y, Zhang K, Xie K, Zhu M, Dai N, Jiang Y, Guo X, Liu M, Dai J, Wu L, Jin G, Ma H, Jiang T, Yin R, Xia Y, Liu L, Wang S, Shen B, Huo R, Wang Q, Xu L, Yang L, Huang X, Shen H, Sha J, Hu Z: **Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types.** *Nat Commun* 2016, **7**:10499.

174. He Y, Sun S, Sha H, Liu Z, Yang L, Xue Z, Chen H, Qi L: **Emerging roles for XBP1, a sUPeR transcription factor**. *Gene Expression* 2010:13–25.

175. Zeng L, Xiao Q, Chen M, Margariti A, Martin D, Ivetic A, Xu H, Mason J, Wang W, Cockerill G, Mori K, Yi-Shuan Li J, Chien S, Hu Y, Xu Q: **Vascular endothelial cell growth-activated XBP1 splicing in endothelial cells is crucial for angiogenesis**. *Circulation* 2013, **127**:1712–1722.

176. Hiyama E, Hiyama K, Yokoyama T, Matsuura Y, Piatyszek MA, Shay JW: **Correlating telomerase activity levels with human neuroblastoma outcomes.** *Nat Med* 1995, **1**:249–55.

177. Neumann AA, Watson CM, Noble JR, Pickett HA, Tam PPL, Reddel RR: **Alternative lengthening of telomeres in normal mammalian somatic cells**. *Genes Dev* 2013, **27**:18–23.

178. Cesare AJ, Reddel RR: **Telomere uncapping and alternative lengthening of telomeres**. *Mech Ageing Dev* 2008, **129**:99–108.

179. Zhao Z, Pan X, Liu L, Liu N: **Telomere length maintenance, shortening, and lengthening**. *J Cell Physiol* 2014, **229**:1323–1329.

180. Cohen SB, Graham ME, Lovrecz GO, Bache N, Robinson PJ, Reddel RR: **Protein composition of catalytically active human telomerase from immortal cells.** *Science* 2007, **315**:1850–3.

181. Nicholls C, Li H, Wang JQ, Liu JP: **Molecular regulation of telomerase activity in aging**. *Protein and Cell* 2011:726–738.

182. Jafri MA, Ansari SA, Alqahtani MH, Shay JW: **Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies.** *Genome Med* 2016, **8**:69.

183. Theimer CA, Feigon J: **Structure and function of telomerase RNA**. *Current Opinion in Structural Biology* 2006:307–318.

184. Tseng C-K, Wang H-F, Burns AM, Schroeder MR, Gaspari M, Baumann P: **Human Telomerase RNA Processing and Quality Control**. *Cell Rep* 2015, **13**:2232–2243.

185. Montanaro L: **Dyskerin and cancer: More than telomerase. the defect in mRNA translation helps in explaining how a proliferative defect leads to cancer**. *Journal of Pathology* 2010:345–349.

186. Venteicher AS, Meng Z, Mason PJ, Veenstra TD, Artandi SE: **Identification of ATPases Pontin and Reptin as Telomerase Components Essential for Holoenzyme Assembly**. *Cell* 2008, **132**:945–957.

187. Schmidt JC, Cech TR: **Human telomerase: biogenesis, trafficking, recruitment, and activation.** *Genes Dev* 2015, **29**:1095–105.

188. Ramlee M, Wang J, Toh W, Li S: **Transcription Regulation of the Human Telomerase Reverse Transcriptase (hTERT) Gene**. *Genes (Basel)* 2016, **7**:50.

189. Jeong SA, Kim K, Lee JH, Cha JS, Khadka P, Cho H-S, Chung IK: **Akt-mediated phosphorylation increases the binding affinity of hTERT for importin α to promote nuclear translocation.** *J Cell Sci* 2015, **128**:2287–301.

190. Alessi DR, James SR, Downes CP, Holmes AB, Gaffney PR, Reese CB, Cohen P: **Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase Balpha**. *Curr Biol* 1997, **7**:261–269.

191. Hers I, Vincent EE, Tavaré JM: **Akt signalling in health and disease**. *Cellular Signalling* 2011:1515–1527.

192. Chang JT, Lu Y-C, Chen Y-J, Tseng C-P, Chen Y-L, Fang C-W, Cheng a-J: **hTERT phosphorylation by PKC is essential for telomerase holoprotein integrity and enzyme activity in head neck cancer cells.** *Br J Cancer* 2006, **94**:870–8.

193. Kharbanda S, Kumar V, Dhar S, Pandey P, Chen C, Majumder P, Yuan ZM, Whang Y, Strauss W, Pandita TK, Weaver D, Kufe D: **Regulation of the hTERT telomerase catalytic subunit by the c-Abl tyrosine kinase**. *Curr Biol* 2000, **10**:568–575.

194. Penzo M, Ludovini V, Treré D, Siggillino A, Vannucci J, Bellezza G, Crinò L, Montanaro L: **Dyskerin and TERC expression may condition survival in lung cancer patients.** *Oncotarget* 2015, **6**:21755–60.

195. Cayuela ML, Flores JM, Blasco MA: **The telomerase RNA component Terc is required for the tumour-promoting effects of Tert overexpression.** *EMBO Rep* 2005, **6**:268–74.

196. Shukla S, Schmidt JC, Goldfarb KC, Cech TR, Parker R: **Inhibition of telomerase RNA decay**

rescues telomerase deficiency caused by dyskerin or PARN defects. *Nat Struct Mol Biol* 2016, **23**:286–92.

197. Schmidt JC, Cech TR: **Human telomerase: biogenesis, trafficking, recruitment, and activation.** *Genes Dev* 2015, **29**:1095–105.

198. Zhu Y, Tomlinson RL, Lukowiak AA, Terns RM, Terns MP: **Telomerase RNA Accumulates in Cajal Bodies in Human Cancer Cells**. *Mol Biol Cell* 2004, **15**(April):3751–3737.

199. Hockemeyer D, Collins K: **Control of telomerase action at human telomeres.** *Nat Struct Mol Biol* 2015, **22**:848–52.

200. Britt-Compton B, Capper R, Rowson J, Baird DM: **Short telomeres are preferentially elongated by telomerase in human cells**. *FEBS Lett* 2009, **583**:3076–3080.

201. Henson JD, Hannay JA, McCarthy SW, Royds JA, Yeager TR, Robinson RA, Wharton SB, Jellinek DA, Arbuckle SM, Yoo J, Robinson BG, Learoyd DL, Stalley PD, Bonar SF, Yu D, Pollock RE, Reddel RR: **A robust assay for alternative lengthening of telomeres in tumors shows the significance of alternative lengthening of telomeres in sarcomas and astrocytomas**. *Clin Cancer Res* 2005, **11**:217–225.

202. Lee Y-K, Park N-H, Lee H: **Prognostic value of alternative lengthening of telomeres-associated biomarkers in uterine sarcoma and uterine carcinosarcoma.** *Int J Gynecol Cancer* 2012, **22**:434–41.

203. Hu Y, Shi G, Zhang L, Li F, Jiang Y, Jiang S, Ma W, Zhao Y, Songyang Z, Huang J: **Switch telomerase to ALT mechanism by inducing telomeric DNA damages and dysfunction of ATRX and DAXX.** *Sci Rep* 2016, **6**(April):32280.

204. Perrem K, Colgin LM, Neumann AA, Yeager TR, Reddel RR: **Coexistence of alternative lengthening of telomeres and telomerase in hTERT-transfected GM847 cells.** *Mol Cell Biol* 2001, **21**:3862–75.

205. Hu J, Hwang SS, Liesa M, Gan B, Sahin E, Jaskelioff M, Ding Z, Ying H, Boutin AT, Zhang H, Johnson S, Ivanova E, Kost-Alimova M, Protopopov A, Wang YA, Shirihai OS, Chin L, Depinho RA: **Antitelomerase therapy provokes ALT and mitochondrial adaptive mechanisms in cancer**. *Cell* 2012, **148**:651–663.

206. Bryan TM, Englezou A, Gupta J, Bacchetti S, Reddel RR: **Telomere elongation in immortal human cells without detectable telomerase activity**. *EMBO J* 1995, **14**:4240–4248.

207. Henson JD, Reddel RR: **Assaying and investigating Alternative Lengthening of Telomeres activity in human cells and cancers**. *FEBS Letters* 2010:3800–3811.

208. Lovejoy CA, Li W, Reisenweber S, Thongthip S, Bruno J, de Lange T, De S, Petrini JHJ, Sung PA, Jasin M, Rosenbluh J, Zwang Y, Weir BA, Hatton C, Ivanova E, Macconaill L, Hanna M, Hahn WC, Lue NF, Reddel RR, Jiao Y, Kinzler K, Vogelstein B, Papadopoulos N, Meeker AK: **Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of Telomeres pathway**. *PLoS Genet* 2012, **8**.

209. Cesare AJ, Reddel RR: **Alternative lengthening of telomeres: models, mechanisms and implications.** *Nat Rev Genet* 2010, **11**:319–30.

210. Pickett HA, Reddel RR: **Molecular mechanisms of activity and derepression of alternative lengthening of telomeres**. *Nat Struct Mol Biol* 2015, **22**:875–880.

211. Krejci L, Altmannova V, Spirek M, Zhao X: **Homologous recombination and its regulation**. *Nucleic Acids Research* 2012:5795–5818.

212. Nosek J, Rycovska A, Makhov AM, Griffith JD, Tomaska L: **Amplification of telomeric arrays via rolling-circle mechanism**. *J Biol Chem* 2005, **280**:10840–10845.

213. Muntoni A, Neumann AA, Hills M, Reddel RR: **Telomere elongation involves intra-molecular DNA replication in cells utilizing alternative lengthening of telomeres**. *Hum Mol Genet* 2009, **18**:1017–1027.

214. Cho NW, Dilley RL, Lampson MA, Greenberg RA: **Interchromosomal homology searches drive directional ALT telomere movement and synapsis**. *Cell* 2014, **159**:108–121.

215. Dilley RL, Verma P, Cho NW, Winters HD, Wondisford AR, Greenberg RA: **Break-induced telomere synthesis underlies alternative telomere maintenance.** *Nature* 2016, **539**:54–58.

216. Blasco M, Benetti R, Gonzalo S, Jaco I, Muñoz P, Gonzalez S, Schoeftner S, Murchison E, Andl T, Chen T, Klatt P, Li E, Serrano M, Millar S, Hannon G: **A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases.** *Nat Struct Mol Biol* 2008, **15**:268–79.

217. Benetti R, Blasco MA, García-Cao M, Blasco M: **Telomere length regulates the epigenetic status of mammalian telomeres and subtelomeres.** *Nat Genet* 2007, **39**:243–50.

218. Gonzalo S, Blasco MA, Jaco I, Fraga MF, Chen T, Li E, Esteller M, Blasco M: **DNA methyltransferases control telomere length and telomere recombination in mammalian cells.** *Nat Cell Biol* 2006, **8**:416–24.

219. Conomos D, Reddel RR, Pickett HA: **NuRD–ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination**. *Nat Struct Mol Biol* 2014, **21**:760–770.

220. Lai AY, Wade PA: **NuRD: A multi-faceted chromatin remodeling complex in regulating cancer biology**. *Nat Rev Cancer* 2011, **11**:588–596.

221. Jiang WQ, Zhong ZH, Nguyen A, Henson JD, Toouli CD, Braithwaite AW, Reddel RR: **Induction of alternative lengthening of telomeres-associated PML bodies by p53/p21 requires HP1 proteins**. *J Cell Biol* 2009, **185**:797–810.

222. Draskovic I, Arnoult N, Steiner V, Bacchetti S, Lomonte P, Londoño-Vallejo A: **Probing PML body function in ALT cells reveals spatiotemporal requirements for telomere recombination.** *Proc Natl Acad Sci U S A* 2009, **106**:15726–31.

223. Jiang W-Q, Zhong Z-H, Henson JD, Neumann AA, Chang AC-M, Reddel RR: **Suppression of alternative lengthening of telomeres by Sp100-mediated sequestration of the MRE11/RAD50/NBS1 complex.** *Mol Cell Biol* 2005, **25**:2708–21.

224. Zhong ZH, Jiang WQ, Cesare AJ, Neumann AA, Wadhwa R, Reddel RR: **Disruption of telomere maintenance by depletion of the MRE11/RAD50/NBS1 complex in cells that use alternative lengthening of telomeres**. *J Biol Chem* 2007, **282**:29314–29322.

225. Reddel RR: **A SUMO ligase for ALT**. *Nat Struct Mol Biol* 2007, **14**:570–571.

226. Gocha ARS, Acharya S, Groden J: **WRN loss induces switching of telomerase-independent mechanisms of telomere elongation**. *PLoS One* 2014, **9**.

227. Conomos D, Pickett HA, Reddel RR: **Alternative lengthening of telomeres: remodeling the telomere architecture.** *Front Oncol* 2013, **3**(February):27.

228. Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, Wen D, Chapgier A, DeKelver RC, Miller JC, Lee YL, Boydston EA, Holmes MC, Gregory PD, Greally JM, Rafii S, Yang C, Scambler PJ, Garrick D, Gibbons RJ, Higgs DR, Cristea IM, Urnov FD, Zheng D, Allis CD: **Distinct Factors Control Histone Variant H3.3 Localization at Specific Genomic Regions**. *Cell* 2010, **140**:678–691.

229. Wong LH, McGhie JD, Sim M, Anderson MA, Ahn S, Hannan RD, George AJ, Morgan KA, Mann JR, Choo KHA: **ATRX interacts with H3.3 in maintaining telomere structural integrity in pluripotent embryonic stem cells**. *Genome Res* 2010, **20**:351–360.

230. Teasley DC, Parajuli S, Nguyen M, Moore HR, Alspach E, Lock YJ, Honaker Y, Saharia A, Piwnica-Worms H, Stewart SA: **Flap endonuclease 1 limits telomere fragility on the leading strand**. *J Biol Chem* 2015, **290**:15133–15145.

231. Saharia A, Stewart SA: **FEN1 contributes to telomere stability in ALT-positive tumor cells**. *Oncogene* 2009, **28**:1162–1167.

232. Lafferty-Whyte K, Cairney CJ, Will MB, Serakinci N, Daidone M-G, Zaffaroni N, Bilsland A, Keith WN: **A gene expression signature classifying telomerase and ALT immortalization reveals an hTERT regulatory network and suggests a mesenchymal stem cell origin for ALT.** *Oncogene* 2009, **28**:3765–74.

233. Cairney CJ, Hoare SF, Daidone M-G, Zaffaroni N, Keith WN: **High level of telomerase RNA gene expression is associated with chromatin modification, the ALT phenotype and poor prognosis in liposarcoma.** *Br J Cancer* 2008, **98**:1467–1474.

234. Nersisyan L, Johnson G, Riel-Mehan M, Pico A, Arakelyan A: **PSFC: a Pathway Signal Flow Calculator App for Cytoscape [v1; ref status: approved 1]**. *F1000Research* 2015, **4**.

235. Arakelyan A, Aslanyan L, Boyajyan A: **High-throughput Gene Expression Analysis Concepts and Applications**. In *Genomics II - Bacteria, Viruses and Metabolic Pathways*. 1st Editio. iConcept Press Ltd.; 2013:71–95.

236. Nersisyan L, Löffler-Wirth H, Arakelyan A, Binder H: **Gene Set- and Pathway- Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: A Bioinformatics Perspective**. *Int J Knowl Discov Bioinforma* 2016, **4**:24.

237. Zimmermann S, Voss M, Kaiser S, Kapp U, Waller CF, Martens UM: **Lack of telomerase activity in human mesenchymal stem cells.** *Leuk Off J Leuk Soc Am Leuk Res Fund, UK* 2003, **17**:1146–9.

238. Bernardo ME, Zaffaroni N, Novara F, Cometa AM, Avanzini MA, Moretta A, Montagna D, Maccario R, Villa R, Daidone MG, Zuffardi O, Locatelli F: **Human bone marrow derived mesenchymal stem cells do not undergo transformation after long-term in vitro culture and do not exhibit telomere maintenance mechanisms.** *Cancer Res* 2007, **67**:9142–9.

239. Bauer S: **Gene-Category Analysis**. In *Methods in molecular biology (Clifton, N.J.)*. *Volume 1446*; 2017:175–188.

240. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette M a, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545–50.

241. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 2009, **37**:1–13.

242. Huang DW, Lempicki R a, Sherman BT: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.

243. Wang J, Duncan D, Shi Z, Zhang B: **WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.** *Nucleic Acids Res* 2013, **41**(Web Server issue).

Lilit Nersisyan

Date of Birth: 19 February 1990

Place of Birth: Yerevan, Armenia

Academic appointments

Phd Student — February 2014 – present

Bioinformatics Group, Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia,

Supervisor: Arsen Arakelyan

Topic: Telomere analysis based on high-throughput multI -*omics* data

Junior Researcher — January 2011 – present

Bioinformatics Group, Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia

Research Assistant — September 2009 – 2011

Group of Immunoregulation, Institute of Molecular Biology, National Academy of Sciences of the Republic of Armenia,

Research fellowships

Research stay at IZBI — July – November 2015

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

Sponsored by: DAAD (91572203) Short Term Research Fellowship

Topic: Bioinformatics pathway activity analysis of epigenetic regulation of gene expression in cancer and pulmonary diseases

Research stays

Research stay at IZBI — July-August 2016

Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

Sponsored by: BMBF MycSys project

Topic: Aberrant pathway function and telomerelengths in Lymphoma

Research grants

Participant, DFG-Project (LO 2242/1-1)

Initiation of collaboration between the Institute of Molecular Biology and the Interdisciplinary Centre for Bioinformatics, University of Leipzig, Leipzig, Germany

Topic: Aberrant epigenetic regulation paths in cancer – a combined pathway flow and machine learning analysis of molecular omics data

PI, Research Grant molbio-3818 January-December 2015

Fund: The Armenian National Science and Education Fund (ANSEF)

Topic: PSFC: pathway signal flow calculator

Student, Google Summer of Code May – August 2014

Fund: Google, Inc

Topic: PSFC: a Cytoscape app for calculating pathway signal flows

Education

American University of Armenia

Masters in Engineering, Computer Science specialization

Thesis topic: A software for KEGG pathway based analysis in Cytoscape

Advisers: Arsen Arakelyan, Suren Khachatryan

Department of Biotechnology, International Scientific Education Center of the National Academy of Sciences of Armenia

Masters in Engineering, Biotechnology specialization

Thesis topic: *In silico* structure characterization of Familial Mediterranean Fever protein pyring

Supervisor: Arsen Arakelyan

Department of Biology, Yerevan State University

Bachelor's degree in Biosciences, Biophysics specialization

Focus: Biophysics

Thesis topic: Expression of CD4 and CD14 receptors on the surface of human peripheral blood mononuclear cells under different ways of stimulation

Supervisor: David Poghosyan

Publications for Dissertation

Journals

1. Nersisyan L, Arakelyan A: **Computel: Computation of mean telomere length from whole-genome next-generation sequencing data.** *PLoS One* 2015, **10**.
2. Nersisyan L: **Integration of telomere length dynamics into systems biology framework: A review.** *Gene Regul Syst* Bio 2016, **10**.
3. Hakobyan A, Nersisyan L, Arakelyan A: **Quantitative trait association study for mean telomere length in the South Asian genomes.** *Bioinformatics* 2016, **32(11).**
4. Nersisyan L, Hakobyan A, Arakelyan A: **Telomere-associated gene network in lung adenocarcinoma.** *Eur Respir J* 2015, **46(suppl 59)**.

Conference proceedings

1. Nersisyan L., Hakobyan A., Arakelyan A. **Association of telomere length with epigenetic regulation of gene expression.** *F1000Research 2016,* **5**:2159 (poster) (doi: 10.7490/f1000research.1112996.1).
2. Nersisyan L., Hakobyan A., Arakelyan A. **Telomere-associated gene network in lung adenocarcinoma.** *European Respiratory Journal* Sep 2015, 46 (suppl 59) doi: 10.1183/13993003.congress-2015.OA3493.
3. Nersisyan L, Wirth H, Gevorgyan A, Binder H, Arakelyan A. **Methylation associated pathway activity deregulation in lung adenocarcinoma.** *Eur Respir J* 2014; 44: Suppl. 58, 403.

Publications

Book chapters

1. Arakelyan A, Nersisyan L, Hakobyan A. **Application of MATLAB in -Omics and Systems Biology**. Applications from Engineering with MATLAB Concepts, Associate Prof. Jan Valdman (Ed.), ISBN: 978-953-51-2459-7. InTech, Croatia, 2016. DOI: 10.5772/62847.

Journals

1. Arakelyan A, Nersisyan L, Petrek M, Löffler-Wirth H, Binder H. **Cartography of pathway signal perturbations identifies distinct molecular pathomechanisms in malignant and chronic lung diseases.** *Front Genet.* 2016, 7:79.
2. Nersisyan L, Löffler-Wirth H, Arakelyan A, Binder H. **Gene Set- and Pathway- Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: A Bioinformatics Perspective.** *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)* 2016, 4(2):46-49.
3. Hopp L., Nersisyan L., Löffler-Wirth H. *et al.* **Epigenetic heterogeneity of B-cell lymphoma: Chromatin modifiers.** *Genes* 2015, **6**(4):1076-1112.
4. Nersisyan L, Johnson G, Riel-Mehan M *et al.* **PSFC: a Pathway Signal Flow Calculator App for Cytoscape** [version 1; referees: 1 approved] *F1000Research* 2015, **4**:480.
5. Arakelyan A, Nersisyan L, Gevorgyan A, Boyajyan A. **Geometric Approach for Gaussian-Kernel Bolstered Error Estimation for Linear Classification in Computational Biology.** *International Journal of Information theories & Applications* 2014, 21: 170-182.
6. Nersisyan L, Samsonyan R, Arakelyan A. **CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows** [v2; ref status: indexed, http://f1000r.es/45p]. *F1000Research* 2014, 3:145.
7. Nersisyan L, Arakelyan A. **MEFV expression during macrophage activation.** *National Electronic Journal of Natural Sciences*, 2(21): 77-81, 2013.
8. Arakelyan A, Nersisyan L. **KEGGParser: parsing and editing KEGG pathway maps in Matlab.** *Bioinformatics* 2013, 29(4): 518-9.
9. Boyajyan A, Arakelyan A, Nersisyan L, Avetisyan N, Martirosyan G. **Evidence-Based Validation of Pyrin Structural Models.** *Protein Science* 2012, 21 Special Issue, Suppl. 1: 196-196.
10. Nersisyan L, Arakelyan A. **In silico structure characterization of Familial Mediterranean fever gene product (pyrin).** *IPCBEE* 2012; 29: 40-44.
11. Karapetyan D, Arakelyan A, Nersisyan L, Aslanyan L, Boyajyan A. **GOMESH: A tool for analyzing gene ontology data.** *Biological Journal of Armenia* 2010, Suppl. 1(62): 41-43.
12. Poghosyan D, Tadevosyan G, Nersisyan L, Arakelyan A. **CD4 expression on activated human monocytes after different ways of stimulation.** *The New Armenian Medical Journal* 2010, 4(1): 111-112.

Conference proceedings

1. Nersisyan L, Hakobyan A, Löffler-Wirth H, Binder Hans, Arakelyan A. **Association of mean telomere length with biomolecular pathway deregulations in lung adenocarcinoma.** *F1000Research 2015,* 4(ISCB Comm J):608 (poster) (doi: 10.7490/f1000research.1110351.1).
2. Nersisyan L, Lusine K, Hakobyan A et al. **A systems view on mining common pathway deregulation profiles in autoimmunity, autoinflammation and inflammation.** *F1000Research 2015*, 4(ISCB Comm J):604 (poster) (doi: 10.7490/f1000research.1110348.1).

3. Arakelyan A, Nersisyan L, Wirth H, Binder H. **Mining common pathway deregulation profiles in lung diseases.** *Eur Respir J* 2014; 44: Suppl. 58, 2021.
4. Nersisyan L, Arakelyan A. **3D structure prediction of pyrin-d2 isoform.** *2nd International Conference "Postgenomic methods of analyses in biology, laboratory and clinical medicine: genomics, proteomics, bioinformatics"* 2011 (p. 112), Novosibirsk, Russia.
5. Arakelyan A, Boyajian A, Aslanyan L, Nersisyan L, Sahakyan H. **Growing Support Sets For Pathway Specific Microarray Gene Expression Analysis.** *"8th International Conference on Computer Science and Information Technologies"* 2011 (p.207-210), Yerevan, Armenia.
6. Poghosyan D, Tadevosyan G, Nersisyan L, Arakelyan A**. The effect of Id 1F7+ Antibodies on LPS-induced Cytokine Secretion by Monocytes.** *Biological Journal of Armenia* 2010, Suppl. 1(62): 83-87.
7. Arakelyan A, Boyajyan A, Aslanyan L, Muradyan D, Chavushyan A, Hovsepyan T, Nersisyan L. **Functional gene sets in posttraumatic stress disorder: analysis of disease related gene expression.** *International Conference "Biotechnology and health-3"* 2009 (p. 57-60), Yerevan, Armenia.

Talks

Oral presentations

ERS INTERNATIONAL CONGRESS

26-30 Sept 2015, Amsterdam, Netherlands

Title: Telomere-associated gene network in lung adenocarcinoma

SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING IN BIOMEDICAL INFORMATICS AND DIGITAL HEALTH, ITA

29-10 July, 2015, Varna, Bulgaria

Title: Pathway signal flow in looped networks

ERS INTERNATIONAL CONGRESS

6-10 Sep, 2014, Munich, Germany

Title: Methylation associated pathway activity deregulation in lung adenocarcinoma

EMBO PRACTICAL COURSE “BIOINFORMATICS AND GENOMES ANALYSES”

5-17 May, 2014, Athens, Greece

Title: Estimation of mean telomere length from whole genome Next Generation Sequencing data

10TH NETWORK BIOLOGY SYMPOSIUM AND CYTOSCAPE WORKSHOP

9-10 Oct, 2013, Paris, France

Title: A Cytoscape plugin/app for KEGG pathway map parsing, automatic corrections, and reduction of abstractions

Poster presentations

EMBO CONFERENCE “FROM FUNCTIONAL GENOMICS TO SYSTEMS BIOLOGY”

12-15 Nov, 2016, Heidelberg, Germany

Title: A pathway based approach for classification of telomerase positive and ALT cancer cells

EUROPEAN CONFERENCE ON COMPUTATIONAL BIOLOGY (ECCB)

3-7 Sept, 2016, The Hague, Netherlands

Title: Association of telomere length dynamics with epigenetic regulation of gene expression

14TH ANNUAL CONFERENCE ISMB/ECCB

11-14 July, 2015, Dublin, Ireland

Title: Association of mean telomere length with biomolecular pathway deregulations in lung adenocarcinoma

FEBS PRACTICAL COURSE “BIOINFORMATICS FOR THE BENCH BIOLOGIST”

3-8 Sep, 2012, Dubrovnik, Croatia

Title: In silico structure and function prediction of pyrin – the protein of Familial Mediterranean fever

CSBE PRACTICAL WORKSHOP “FORMAL APPROACHES TO MODELLING BIO-MOLECULAR NETWORKS”

24-27 Apr, 2012, Edinburgh, UK

Title: Molecular modeling and docking of pyrin – the protein of Familial Mediterranean fever

Teaching experience

2016 -present	**Adjunct lecturer** *CSE 162 Introduction to Biosciences* American University of Armenia
2014 - 2015	**Teaching Associate** *American University of Armenia* *Introduction to Bioscience (2014-2015), Quantitative Biology (2015), Introduction to Algorithms (2015)*

Language skills and qualifications

Languages:

Armenian (native), English (fluent), Russian (fluent), German (beginner)

Programming languages:

Java, R, bash scripting

Bioinformatics skills:

Gene expression, NGS data analysis, Pathway and network analysis, High dimensional data analysis, previous experience in protein structural modeling

Wet lab experience:

DNA Isolation, PCR-SSP, qrt-PCR, ELISA, Flow cytometry.

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Eriwan, den 12 Feb 2017

Lilit Nersisyan