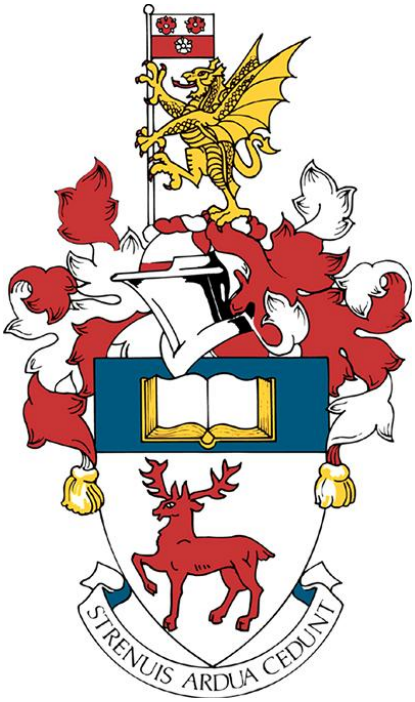
UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

BIOLOGICAL SCIENCES



Differences in Telomere Length Predicted from the Exome Data of Bladder Cancer and Matched Control Tissue Samples

by

PHIL PALMER

BSc Biology

Year 3

Supervisor: Dr Jane Gibson

Word count: 9311

Table of contents

I. Lay Summary

DNA, the molecule which plays a major role in determining who you are, is very long. Therefore, to fit inside the cell, DNA is tightly packed into structures called chromosomes which can form a characteristic X-shape. The end of the chromosome is referred to as a telomere. Telomeres are crucial in preventing all of the densely packed DNA from unravelling and so can be thought of in a similar way to the caps at the ends of shoelaces.

As a result of normal everyday damage and growth, the cells containing the DNA need to replicate themselves including their DNA. However, the cellular machinery that performs this cannot go all of the way to end of the chromosome/telomere. Therefore, with each consecutive cell division/duplication, the telomeres get shorter and shorter, like a fuse until eventually, “cell suicide” is committed or division stops. Hence, this limits the number of cell divisions that occur and is referred to as the Hayflick limit. While the Hayflick limit is one of the causes of ageing, because it causes cells to stop dividing and die as we get older, it also helps protect against cancer. When cells replicate their DNA, mistakes known as mutations can be made and so otherwise tightly controlled processes such as cell division may become uncontrolled. Telomeres however, limit the number of cell divisions and so reduce the chance of cells propagating these cancerous mutations.

Nevertheless, there are cells, such as rapidly dividing white blood cells, that need to divide lots of times. They have telomerase, an enzyme which increases telomere length. While most cells have the capacity to do this (as a gene in their DNA) it is turned off. However, when cancer cells accumulate mutations they can turn the telomerase gene on, rendering the cancer cells immortal. In this study, DNA sequence information was downloaded from a previous study which investigated bladder cancer. Computational techniques were used to analyse the data. It was found that there were more high-impact mutations in cancer samples compared to non-cancerous samples in telomere-related genes. It was possible to estimate telomere length because telomeres have a

specific and recognisable repeat sequence. Having compared telomere length in cancerous and non-cancerous samples, it was found that there was no difference in length between them. Another finding was that telomere length is not dependent upon age, contrary to the wider literature.

II. Abstract

Telomeres are regions of repetitive DNA sequences located at the end of chromosomes which are important in ageing and cancer. The aim of this study is to determine if there is a difference in telomere length between exomes of bladder cancer and matched control samples from the same patients and to investigate any mutations responsible for these differences. While the exome data used was from a previous study, the data is used here to answer a relatively unexplored research question using computational techniques. Whereas other studies have predominantly used experimental techniques to measure telomere length in bladder cancer, which is still poorly understood. The program TelSeq was used to estimate telomere length for bladder cancer and matched blood control samples, from 9 different patients, and the online tool VEP was used to generate variants in the samples. There were more high-impact variants in the cancer samples compared to the control samples than expected by chance and high-impact mutations were found in genes such as TP53, TERT and DKC1. No significant difference was found between telomere length of matched bladder cancer and control samples from the same patient. It was also found that telomere length does not depend on age. However, as this contradicts the wider literature more work needs to be done to confirm this. Overall, despite more genetic variation in bladder cancer tumours, this does not appear to affect the telomere length. Finally, improving the understanding of telomeres is important to help in beating cancer and increasing human longevity.

III. Acknowledgements

Many thanks to Dr Jane Gibson for helping with the project and Roshan Sood for writing some of the scripts used.

The use of the IRIDIS High-Performance Computing Facility, and associated support services at the University of Southampton is acknowledged in the completion of this work.

This study makes use of data collected in “Frequent mutations of chromatin remodelling genes in transitional cell carcinoma of the bladder.”[1]

IV. Abbreviations

ALT	Alternative Lengthening of Telomeres
ATM kinase pathway	Ataxia-Telangiectasia Mutated kinase pathway
ATR kinase pathway	Ataxia-Telangiectasia and Rad3-related protein kinase pathway
BAM	Binary Alignment Map
BED	Browser Extensible Data
bp	Base pairs
BWA	Burrows-Wheeler Aligner
DKC1	Dyskeratosis Congenita 1
EMBL-EBI	European Molecular Biology Laboratory – European Bioinformatics Institute
ENA	European Nucleotide Archive
GVCF	Genomic Variant Call File
GATK	Genome Analysis Toolkit
GC	Gas Chromatography or Guanine Cytosine
GRCh	Genome Reference Consortium human genome
gz	GNU ZIP
hg38	Human Genome 38
HR	Homologous Recombination
KB	Kilo Bases
MB	Mega bases
ncvTest	Non-Constant Variance Test
NHEJ	Non-Homologous End Joining
PCA	Principal Component Analysis
POT1	Protection of Telomeres Protein 1
Q-FISH	Quantitative Fluorescence In Situ Hybridisation
qPCR	quantitative Polymerase Chain Reaction

RAP1	Ras-Proximate-1
Rb	Retinoblastoma
ROS	Reactive Oxygen Species
SAM	Sequence Alignment Map
SLX4	Structure-specific endonuclease subunit
SVM	Support Vector Machine
TCC	Transitional Cell Carcinoma
TERC	Telomerase RNA Component
TERT	Telomerase Reverse Transcriptase
TIN2	TERF1-Interacting Nuclear factor 2
T-loop	Telomere loop
TNKS2	Tankyrase 2
TP53	Tumour Protein 53
TP53BP1	Tumour Protein 53 Binding Protein 1
TPP1	Tripeptidyl Peptidase 1
TRF southern blot	Terminal Restriction Fragment southern blot
TRF1	Telomere Repeat Factor 1
TRF2	Telomere Repeat Factor 2
VCF	Variant Call File
VEP	Variant Effect Predictor

V. List of tables

VI. List of figures

1. Literature Review

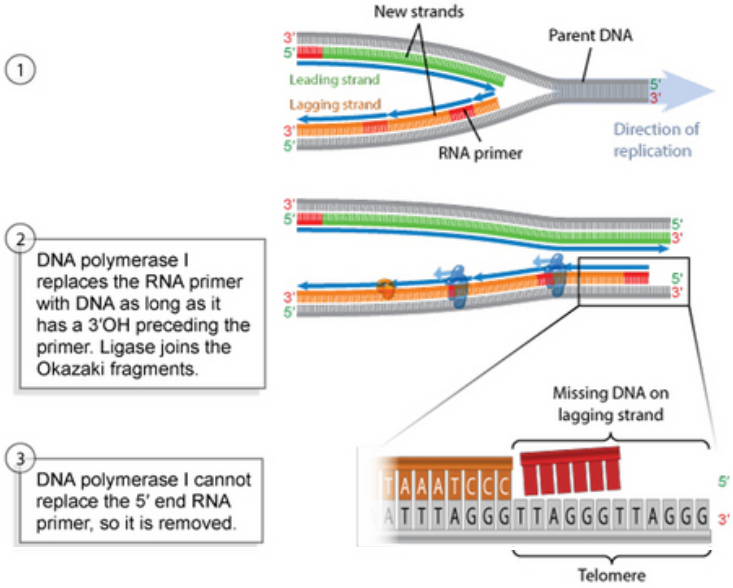
1.1 Telomeres background

Telomeres are regions of repetitive DNA sequences (often TTAGGG, for vertebrates), located at the end of chromosomes.[2] They are present in most eukaryotic organisms but not the majority of prokaryotic organisms.[3] In humans, telomere length is approximately 15kb at birth, however this decreases to 4kb in old age (although this varies for different cell types).[4] While telomeres are intronic (non-coding) DNA they are able to bind to proteins. Their role is to protect chromosomes and therefore DNA from deterioration, end-end fusion and recombination.[2]

In 1961, Leonard Hayflick observed that there is a finite limit to the number of times a cell is able to divide in culture. In 1972, James Watson determined the end replication problem of how chromosomes are unable to fully replicate. These ideas were then first connected by Alexei Olovnikov around the same time.[5] This is because it is the inability of chromosomes to fully replicate that is largely responsible for the limit in the number of cell divisions. Then in 1984 the mechanisms became better understood as Elizabeth Blackburn discovered telomerase, the enzyme which is able to overcome the end replication problem and lengthen the telomeres.[2] Afterwards, in 1993 Barbara McClintock theorised that the ends of chromosomes are protective and increase genomic stability by preventing end-end fusion.[5] However, it wasn't until 1999 that the importance of telomeres in (cellular) ageing and immortality, and therefore cancer, was recognised which was established at the biotech company Geron.[6]

1.2 Telomere shortening: how the end replication problem leads to replicative senescence and the Hayflick limit

The “end replication problem” is where DNA polymerase (which is the replicative machinery that replicates the DNA) cannot go all the way to end of the chromosome and so cannot replicate the chromatin fibres.[7] This is because DNA synthesis requires the binding of RNA primers in front of the 5’/ lagging strand, as shown in figure 1.[8]



This then means that with each successive division, some DNA at the end of the chromosome (part of the telomere) is lost. Therefore over time, as a part of normal cellular ageing, telomeres shorten.[2] This occurs so that telomeres prevent the genes themselves from being truncated.[9] However as another result of this telomere shortening, replicative senescence occurs,[2] which is a limit in the number of times a cell is able to divide. This is also known as the Hayflick limit[10] and is typically around 60-80 cell divisions for a population of normal healthy human cells and results in either apoptosis or cellular senescence (depending on p53 status).[11] The Hayflick limit may help prevent cells from becoming cancerous[10] because when telomeres get too short, cells are triggered to enter apoptosis. For example, through pathway activation of cell-cycle checkpoints by genes such as p53 (or cells are triggered to enter senescence by Retinoblastoma, which stops cell proliferation).[11] After many successive cell divisions, cells are also more likely to have accumulated mutations which could cause cancers. So by triggering apoptosis/senescence the propagation of these mutations and therefore the risk of developing cancers is reduced.[12]

Oxidative stress also decreases telomere length. This is mainly because oxidative stress inactivates repair proteins and damages DNA within the telomeres.[9] On average, this accounts for over double as much of the telomere shortening as the end replication problem.[13] However reactive oxygen species (ROS) also cause damage to cells and induce apoptosis (for example by inducing p53). This then means that these lost cells must be replaced and so further cells must be replicated.[14] This causes further telomere shortening as a result of the end replication problem mechanism.[2]

Lifestyle factors such as poor diet, physical inactivity, high body weight and smoking decrease telomere length.[15] This is likely as a result of inducing cellular damage leading to an increased number of cell divisions for repair and/or oxidative stress.[14] For example, obesity, a diet low in antioxidants, and smoking all increase oxidative stress therefore, decreasing telomere length.[9]

1.3 Telomere lengthening: ALT, telomerase and its regulation by the shelterin complex

Telomeres can be made longer, and the end replication problem overcome, by extending the telomere using the action of enzymes. This is mainly by the enzyme telomerase but also by another enzyme ALT (Alternative Lengthening of Telomeres). Telomerase is able to overcome the end replication by using an RNA template to synthesize DNA.[2] As shown by figure 2, this is done by binding to the RNA template and then extending the G-overhang. After sufficient lengthening, DNA polymerase is then able to extend the RNA primer to synthesize a complementary strand in the reverse direction.[8]

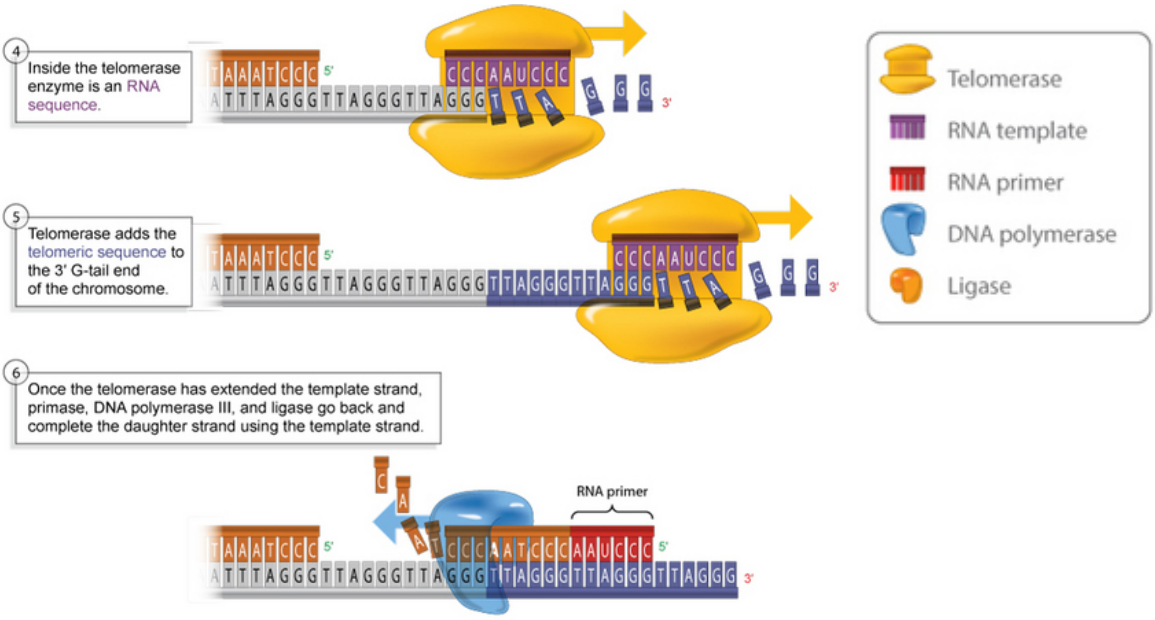


Figure 2: How telomerase overcomes the end replication problem by extending the telomere using the template RNA to synthesise DNA at the G-overhang[8]

Telomerase is active in some cells which divide rapidly such as, peripheral white blood cells. Telomerase is also active in germ cells and some stem cells, like embryonic stem cells.[16] The activity of telomerase is high in these cells in order to reset the telomeric 'clock',[15] as these cells are capable of giving rise to a whole new organism. Telomerase activity is important to increase telomere length, in these cells, because otherwise telomere length would be dramatically shortened, and this may result in a decreased lifespan.[16]

Telomerase is a holoenzyme consisting of three subunits, and it is formed by the combination of a catalytically active enzyme TERT (Telomerase Reverse Transcriptase) and co-activating proteins which regulate the activity of telomerase. The third subunit is TR (telomeric RNA, also known as TERC), which acts as the template for telomere elongation by associating with the TERT and proteins. The proteins form a complex called shelterin/telosome. This complex consists of six proteins in humans, these are POT1, TIN2, TPP1, TRF1, TRF2, and RAP1.[17]

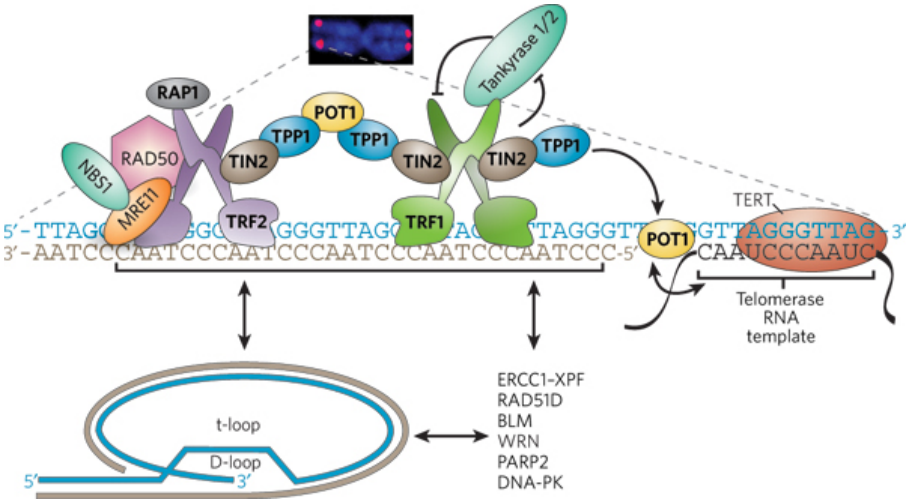


Figure 3: The mammalian telomeric complex including shelterin; the fluorescence microscopy image shows the position of telomeres within the chromosomes[18]

The role of the shelterin is to protect the telomere, which then in turn, protects the DNA/genes of the chromosome.[7,17] The shelterin complex is able to protect the telomere by forming a T-loop (Telomere-loop), as shown in figure 3. This is where the single-stranded DNA forms a circular coil which is stabilised by telomere-binding proteins such as TRF. At the end of the T-loop, the telomeric DNA binds onto one of the strands of DNA in the double helix while displacing the other. This is known as the D-loop (Displacement-loop) and forms a DNA triplex (triple-stranded helix).[7] It is also possible for the telomeric sequences to form a G-quadruplex (demonstrated by figure 4). This is a four-stranded DNA structure formed from G-rich telomeric sequences which consist of four bases held in a plane stacked on-top of one another.[19]

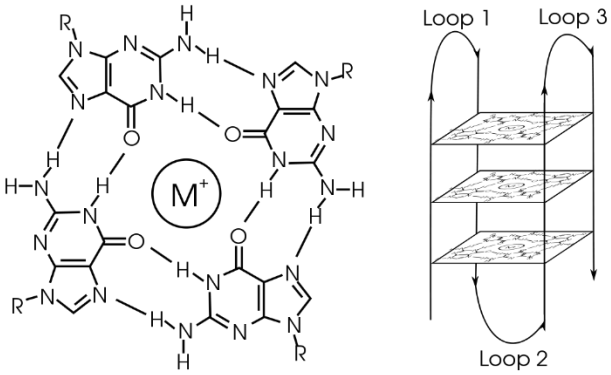


Figure 4: Structure of a G-quadruplex. The image on the left shows the view from above the G-quadruplex, while the image of the right demonstrates the structure from in-front the G-quadruplex [20]

The shelterin complex is able to protect the telomeres by preventing the telomere being recognised as double-stranded breaks. This is done by repressing one of the major DNA repair mechanisms, the ATR kinase pathway. The ATR kinase pathway involves the enzyme, ataxia telangiectasia and Rad3-related protein, which is activated in response to single-stranded DNA breaks. This is blocked by the POT1 proteins which form part of the shelterin. However, shelterin also represses the other major DNA repair mechanism, the ATM kinase pathway. ATM is a serine/ threonine kinase which is recruited to repair double-stranded breaks and is blocked by the shelterin protein TRF2.[21] Therefore the single-stranded DNA repair mechanisms such as non-homologous end joining (NHEJ) and homologous recombination (HR) are prevented.[22] Shelterin therefore prevents the ends of chromosomes being recognised as breaks in DNA as this may cause indefinite cell cycle arrest while the telomeres would mistakenly be attempted to be repaired.[23]

The other telomere-lengthening enzyme, ALT, works by transferring the telomeric repeats between sister-chromatids. ALT, unlike telomerase, therefore works via homologous recombination to increase telomere length.[24] While the precise mechanism by which this occurs is still unknown, it is thought that the recombination generates t-circles (circular loops of telomeric DNA that can be either single or double-stranded).[24] The t-circles then act as a substrate for telomere synthesis to form t-loops. It has been suggested that there may also be

efficient cycling between t-circles and t-loops in order to dynamically and rapidly increase or decrease telomere length.[25]

Telomere lengthening enzymes such as telomerase and ALT play a crucial role in regulating telomere length. This is clear from the fact that telomerase deficiency is associated with diseases such as aplastic anaemia, dyskeratosis congenita and pulmonary fibrosis.[26] However it is important to understand these mechanisms of telomere lengthening not only because of their involvement in diseases but also for useful applications. Better manipulation of these enzymes could be useful to culture enough healthy human tissue for biomedical repairs.[27] Factors that influence the activity of these enzymes and so affect telomere length; include lifestyle factors such as reducing stress. This is because stress decreases telomerase activity and therefore shortens telomeres.[15]

1.4 Telomere's role in ageing

Ageing is defined as time-dependent functional decline and is characterised by nine hallmarks of ageing. One of these hallmarks is telomere attrition[28] which clearly demonstrates that telomeres play a key role in ageing. This is mainly because a shortened telomere phenotype seems to cause increased genomic instability.[29] When telomeres become too short, which is often seen in old age, they can unfold from their closed structure.[30] This can be detected as DNA damage and induce either cellular senescence or apoptosis.[11] Unfolding telomeres can also be recognised as double-stranded breaks in DNA, which can then lead to chromosomal fusion (as a result of the ATM kinase pathway).[30] As a result of this, shortened telomeres are implicated in many age-related diseases due to organs degenerating as more and more cells die.[28]

One way in which we may be able to assess the role of telomeres in ageing is by looking at progeroid syndromes, which are rare genetic diseases that accelerate the ageing process. Such diseases include Blooms Syndrome, Hutchinson-Gilford syndrome and Werner's syndrome.[31] For example, individuals who suffer from

Werner's syndrome tend to have shorter telomeres when compared to healthy individuals of a similar age.[2] Werner's syndrome is characterised by the loss of the gene WRN which encodes a DNA helicase of the RecQ family, which unzips the DNA double-stranded helix. This causes genomic instability and shortens lifespan. As sufferers from Werner's syndrome are also more cancer-prone, this demonstrates the interlinked nature of telomeres, genomic instability, cancer and ageing.[32]

Premature ageing can also be a problem in cloned animals. For example, because Dolly the sheep was cloned she inherited shortened telomeres, as she was derived from a cell of a middle-aged sheep. However, this experiment may indicate that our knowledge of telomeres is still limited because the post-mortem did not support the theory that she had aged prematurely. However, she did seem more fragile when compared to naturally mated offspring.[33]

While telomere attrition may be one of the causes of ageing,[28] telomere lengthening enzymes can increase longevity in cells. This can be seen whereby they confer immortality in cancerous cells, for example, HeLa cells.[10] However, it may also be that telomerase could maintain/elongate telomeres and therefore help delay ageing in mammals, such as mice and therefore possibly humans.[34] Methods to manipulate telomerase and reduce this telomere shortening include drugs, gene therapy and metabolic suppression such as torpor or hibernation.[19,34,35] These methods such as gene therapy can be used to increase telomere activity and could foreseeably be used in the future to inhibit ageing and increase human lifespan.[34]

1.5 Telomere's in cancers

Genomic instability as a result of telomere shortening also causes cancer, similar to how its responsible for ageing/senescence. Shortened telomeres increase cancer risk, by approximately a factor of 1.4-3.0, as shown by meta-analyses.[36,37] This is largely due to genomic instability which causes both the one-way transfer of segments of chromosomes (i.e. non-reciprocal translocations) and the loss of whole

chromosomes in tumour cells.[38] When telomeres become critically short, further cell division usually results in cell death, due to the genomic instability and chromosomal rearrangements.[7] However, to reach this stage, cells often need to have had inactivating mutations in genes such as p53 and Rb.[36]

A common question is why, if shorter telomeres are associated with increased cancer risk, are longer telomeres not selected for. One explanation is that cancer often does not occur until later life.[44] This, therefore means that it is less likely to have had as strong an evolutionary selection pressure because later in life reproduction will likely have already occurred and so alleles will have already been passed on from parent to offspring. An alternate explanation is that longer telomeres may increase energy consumption. This is because DNA replication is energy dependent and so the longer telomeres are, the more energy is required for replication. Therefore, longer telomeres would be selected against. This is known as the "thrifty telomere" hypothesis[45] and is particularly important because for most of human evolutionary time, resources such as food would have been scarce and lifespans would have been short. Therefore, protection from cancer via metabolically expensive long telomeres would likely not have been advantageous.

1.5.1 Anti-cancer drugs

High telomerase activity is present in 90% of tumours because of the up-regulation of TERT[39] which helps maintain/increase telomere length. It is therefore these mutations that confer cancer cells immortality, one of the hallmarks of cancers.[40] This is due to the fact that the cells are able to divide beyond the Hayflick limit seen in most healthy cells without undergoing apoptosis or senescence.[10] Telomerase is therefore a potential anti-cancer drug target such as the drug Imetelstat, which is currently in clinical trials. As shown in figure 5, Imetelstat competitively inhibits telomerase by binding to telomeric RNA (TR).[41] This may work as an effective anti-cancer therapy because the role of telomeres in cancer may be a rate-limiting factor;[34] i.e. by using Imetelstat to inhibit telomerase, the growth of a tumour is reduced.[41]

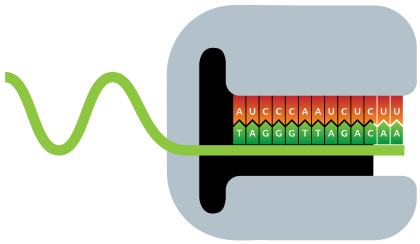


Figure 5: Imetelstat is shown in green (consisting of an oligonucleotide bound to a lipid group) which is then bound to the telomeric RNA of telomerase, shown in red[41]

Another possible anti-cancer drug is Telomestatin, which induces G-quadruplexes in the (G-rich) telomeres. This then prevents the telomerase from being able to access and lengthen the telomeres;[42] leading to normal telomere attrition for the cancer cells, and so preventing the otherwise limitless replication potential of cancer cells. Telomestatin may be a better anti-cancer drug, as it works on the telomere and not telomerase. This is because it will likely block telomerase and ALT (which is dependent upon double-stranded DNA due to its reliance on homologous recombination) whereas Imetelstat may only block telomerase. This is important because 10% of tumours have high ALT activity. Much like telomerase, this can lengthen the telomeres of cancer cells, leading to immortality.[43]

1.5.2 Bladder cancer

Bladder cancer is characterised by uncontrolled cell division in urinary bladder tissues and is the ninth most common cancer in the world.[1] In the UK, bladder cancer is both the seventh most common form of cancer and the seventh most common form of death from cancer. 10,400 people were diagnosed with bladder cancer and 5,200 people died from bladder cancer in 2011 and 2012 respectively.[46] While there are many different types (and stages) of bladder cancer, transitional cell carcinoma is the most common form and represents 90% of cases.[1] The different stages of bladder cancer can be characterised by the TNM classification.[47] This is where T indicates the size/invasion of the tumour, N indicates any lymph node metastasis and M refers to metastasis. Each letter is followed by a number

from one to three with higher numbers indicating a worse prognosis, for example T3N0M0.[1]

The single biggest (avoidable) risk factor for bladder cancer is smoking.[46] Other risk factors include being male, exposure to carcinogens such as benzidine or 2-Naphthylamine (which is in cigarette smoke), and a family history of bladder cancer.[48] Metastasis is also very common in bladder cancers. One of the most common symptoms is blood in urine, however, pain during urination is also another symptom commonly associated with bladder cancers. Diagnosis usually takes place via a cystoscopy; where a thin camera used to look inside the bladder, which has the capability to take samples, allowing for a biopsy to be carried out. Treatments include chemotherapy, immunotherapy and surgery.[47]

From a meta-analysis of population studies, including those with different types of cancers, it was found that cancers, such as bladder cancer, are associated with a shortened telomere length.[37] However, more studies need to be carried out to confirm this. This is despite telomerase being active in 98% of bladder cancers.[49]

1.6 Measuring/estimating telomere length

Telomere length can be measured experimentally using various techniques. These include southern blot, quantitative fluorescence in situ hybridisation (Q-FISH) and PCR based techniques such as real-time polymerase chain reaction (also called qPCR).[50]

The southern blot method used is referred to as Terminal Restriction Fragment (TRF) southern blot. This can be used to detect the specific telomeric sequence motif TTAGGG. This is done by separating the DNA fragments using gel electrophoresis with a nylon filter membrane, before detection using probe hybridisation. Q-FISH can also be used, however larger amounts of DNA are initially required. This can be overcome using qPCR[50] which looks at the ratio of telomeres compared to single copy genes in order to determine the telomere length.[50]

Telomere length can also be estimated computationally using software such as the TelSeq program (as used in this study). TelSeq estimates telomere length using the equation $l = t_k sc$, where l is an estimate of telomere length, t_k is the number of telomeric reads at threshold k (default is seven). Finally, s is the sequence depth and c is a constant for the genome length divided by the number of chromosome ends.[50]

While neither TelSeq nor experimental methods such as TRF are able to capture all of the available telomere length information, TelSeq seems to be able to capture more information. This is partly because TRF tends to overestimate the telomere length.[50] However, this was done for genome-wide sequencing data and it is unclear if the results would be the same for exome data.

One reason why using TelSeq may be better than using experimental methods, particularly compared to methods such as qPCR for measuring telomere length in cancer cells, is because cancer cells often suffer from aneuploidy which will then make it difficult to validate results (due to the reliance on normalisation compared to single copy genes).[50]

2. Aims of the project

The overall aim was to determine if there is a difference in telomere length between exomes of bladder cancer and matched control samples from the same patients and to investigate any genes/mutations responsible for these differences. To achieve this, principal components analysis (PCA) was used to indicate which variables may be best suited to this analysis. A linear regression was then carried out. For this, the null hypothesis (H_0) is that there is not a significant linear relationship between the patient age and telomere length estimate and that the slope is equal to zero. The alternative/test hypothesis (H_1) is that there is a significant linear relationship between the patient age and telomere length estimate and that the slope is not equal to zero. A paired t-test was then used. For the paired t-test, the null hypothesis is that there is no significant difference between the sample means and so the mean difference is equal to zero for telomere length estimate of bladder cancer and matched blood control tissue samples. The alternative hypothesis is that there is a significant difference between the sample means and so that the true difference in the means is not equal to zero for telomere length of bladder cancer and matched blood control tissue samples. A chi-squared test was also carried out with the null hypothesis that there is no significant difference between the observed and expected number of high-impact variants in telomere-related for the bladder cancer and control samples. The alternative hypothesis is that there is a significant difference between the observed and expected number of high-impact variants for the bladder cancer and control samples. The final aim was to use the data collected in this experiment to determine the extent to which the different machine learning classifiers can predict which of the tissues the samples came from (either bladder cancer or the blood control).

3. Materials & Methods

3.1 About the dataset

These dataset was downloaded from the ENA, which is part of EMBL-EBI. While the database stores both publicly and privately available data, the files that were downloaded are publicly available. The study accession for this data is SRP007205, which can be found at https://www.ebi.ac.uk.[51] The original publication is entitled "Frequent mutations of chromatin remodelling genes in transitional cell carcinoma of the bladder".[1] The data was download in fastq.gz file format. FASTQ files contain the raw sequence information. The .gz extension of the FASTQ files means that the files were gzipped which compresses the files. This is beneficial for storing the files and still allows analysis without needing to uncompress the files. In this study, paired testing was carried out whereby both bladder cancer and matched control samples were taken from each patient. There were nine patients in total. In total 42 fastq.gz files/runs were downloaded including files for both blood and cancer samples (see Appendix A for more details on files).

The files were downloaded onto a personal account on the Irdis4 supercluster. This is the supercomputer at the University of Southampton, which was used to process and securely store the files.

3.2 Quality control

Once the fastq.gz files were downloaded onto Iridis4 using the bash command wget, FASTQC could then be used to check the overall quality of the runs from the raw sequence data.[52] This is shown in figure 6. As there were multiple FASTQ files for some samples, the files were joined/merged and so the bash/Linux command "zcat" was used. It was necessary because the files were in their compressed format and so in order to join the files they were uncompressed using zcat. The combined output was then redirected to another file which could then be recompressed. Two fastq.gz files were made per patient (for blood and tumour respectively) even for cases where there were multiple runs.

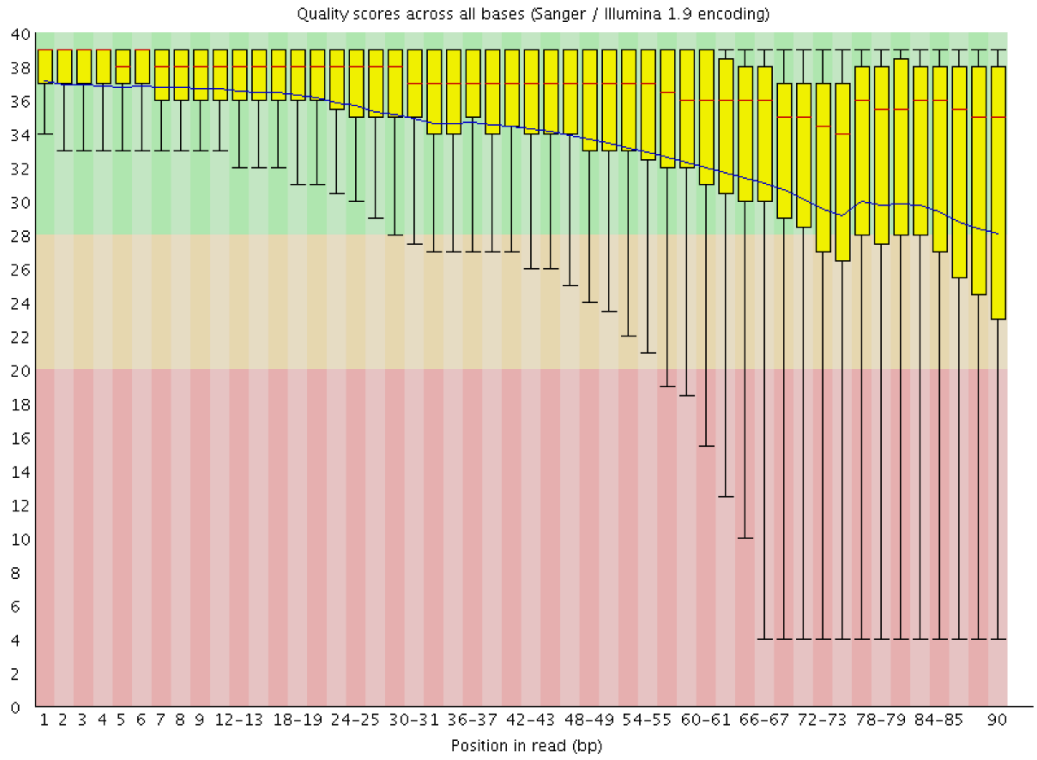


Figure 6: FastQC report plot for patient B2 blood sample SRR290592. Good per base sequence quality. Quality score decreases towards the end of the read for Illumina sequencing due to phasing problems, (when single molecules within a cluster become out of sync with one another)

3.3 Bioinformatic Pipeline

As shown in figure 7, BAM (Binary Alignment Map) files could then be made from the fastq.gz files. BAM files are binary versions of SAM files (Sequence Alignment Map), a type of text file containing tab-delimited sequence alignment information. This needed to be done for input into the TelSeq program which takes BAM files as input. To do this step, BWA was used (see the script in Appendix D for more details). BWA uses an algorithm (BWA-MEM) to map the fastq.gz sequences onto the human reference genome hg38 (GRCh Build 38), which was located in a directory on the Iridis server.[53] The BWA-MEM algorithm was used as it was designed for longer sequence reads, ranging from 70bp to 1Mb and because it is the latest algorithm, meaning it is faster/more accurate compared to alternatives such as BWA-SW.[53]

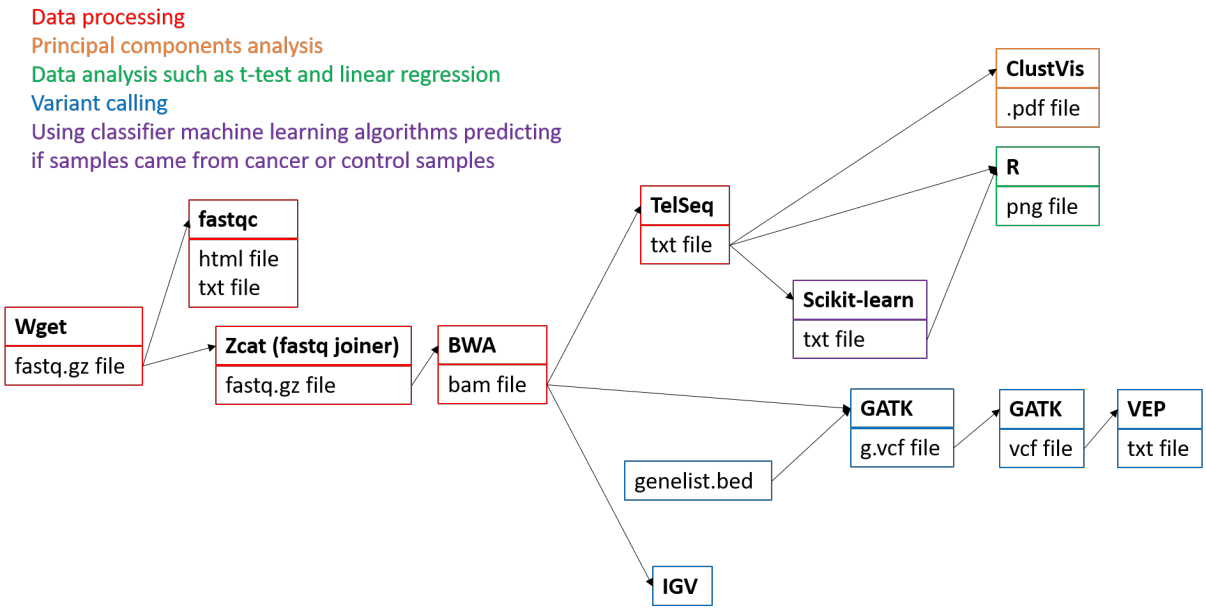


Figure 7: Flow diagram of the project. All data processing was carried out on the Iridis4 supercluster. The coloured text in the top-left refers to the processes that were carried out and the colour is used to correspond to the tools/programs used. The bold text is tools/programs that were carried out and often refers to scripts in appendix D. The boxes below are the output file for that program. The arrows connecting the boxes indicate that the output from one tool was used as input for the next

A script was then written for TelSeq (see the script in Appendix D for more details.) TelSeq is a program which is used to output an estimate for telomere length and other information such as GC content, this can be outputted in the form of a text file.[50] While the software is often used for genome-wide data it can also be used for exome data; as 10-50% of most exome data is non-exonic.[50] This non-exonic data is therefore likely to include telomeres which are intronic.[3]

3.4 Principal Components Analysis (PCA)

ClustVis was used to generate a principal components analysis (PCA) plot.[54] Data used to generate the PCA included the patient's age in years,[1] as well as the telomere length estimate (in kb) and percentage GC content (between 40-60%) of the total number of reads. These were generated from the output from TelSeq. To get the value for GC content, all of the GC values for reads between 40-60% were added together and then divided by the total number of reads and multiplied by

100 to convert it to a percentage. This was done because the absolute GC content was largely dependent upon the total number of reads, which were much greater for some samples than others. Options used for the data pre-processing were to keep “all annotations”, have “no collapse” for similar annotations, to use “unit variance scaling” and “SVD with imputation” for the PCA method. All other settings were default.

3.5 Linear regression predicting telomere length from patient age

A linear regression was performed to predict the response of the continuous factor (telomere length estimate) to one sample of a continuous factor (patient age) in R(version 3.4.2).[55]

3.6 Paired t-test comparing cancer and control samples telomere length

A paired t-test was performed to compare two population means with two groups (bladder cancer and matched control) in which the observation from one group can be paired with the other (due to the samples coming from the same patient). The paired t-test tested for a difference between samples means from one factor. This included testing one response variable (telomere length estimate) with two or more treatments (bladder cancer and matched control) of one categorical factor (tissue sample type). This was performed in R(version 3.4.2) where mu was set to zero to set if the difference between the means was zero and alt was set to “two.sided”, for a two-tailed test a difference could go in either direction.[55] Finally paired was set to true as the samples were collected from the same patients from either group and the confidence level was set to 95% to determine if the probability of seeing the results due to chance was greater or less than 5%.

3.7 Variant Calling

To carry out variant calling, a list of candidate genes with telomere-related functions was generated, (see Appendix C for the full list of genes), by performing

a literature search. Web of Science (available at http://wok.mimas.ac.uk) was used to generate the list using search terms such as “telomere”, “telomerase”, “related”, “important”, “structure”, “function” and “genes”. To make a BED file to use for variant calling the file had to be in the correct format (of the chromosome, start, stop position). To get the start and stop positions of the genes, the table browser of UCSC genome browser was used.[56] The candidate gene list was uploaded and gene names, chromosome, transcription start and transcription endpoint were selected for the output (all other options were default). The candidate gene list (now including the chromosome, start and stop positions) was then saved in excel as tab-delimited and renamed to a “.bed” file. GATK GenotypeGVCFs scripts (see Appendix D) were then run, using the human genome hg38 and the bed file of candidate genes to generate g.vcf files (which is a Genomic VCF, i.e. a VCF file with extra information) for each of the samples.[57] Another GATK GenotypeGVCFs script was then run to generate one VCF file for all of the samples. This VCF file was then annotated using the online tool VEP, and then downloaded as a text file and opened in excel.[58] CSN(p), 1000 Genomes global minor allele frequency and gnomAD (exomes) allele frequencies were all selected. Then dbNSFP(p) was enabled before CADD_raw and GERP++_NR scores were selected. All other options were default.

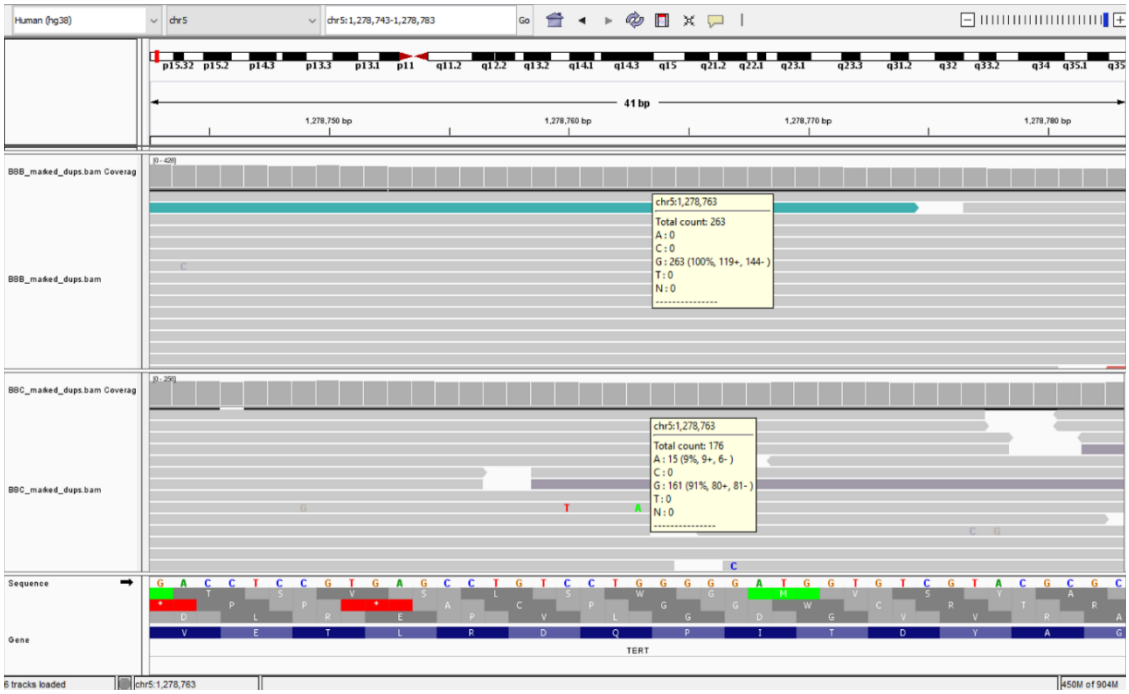


Figure 8: Integrated Genome Viewer (IGV version 2.4.9) screenshot of the variant at base pair location 1278763 on chromosome 5 in the TERT gene for patient B8. The blood control sample is shown on the top track which is homozygous with no variant present. The track below that is the cancer sample with the stop gain present making the genotype heterozygous with one copy each of the reference (G) and alternate allele (A).

As shown in figure 8, IGV(version 2.4.9) was used for some specific somatic variants to ensure that variants were present in the tumour sample but not the control sample.[59] The variants were then filtered for high-impact (frameshift and stop gain) variants. Before a one-way classification chi-square, a goodness-of-fit test was carried out in R(version 3.4.2)[55] to determine if there was a significant difference between the observed and expected number of high-impact variants in cancer and control samples. The expected number of variants for both cancer and control samples was half the total number of high-impact variants (149/2 = 74.5). Further filtering was then carried out to select only somatic variants. The remaining variants were then cross-referenced against the original publication supplementary reading for confirmed somatic mutations. Variants in the genes TERT and DKC1 which are of critical importance in telomeres and telomere-related disease were selected and appended to the variant shortlist. Finally, variants with a gnomAD score equal to zero for East Asian populations were selected and GERP++ neutral rate (NR) above five, indicating highly conserved genes were selected.

3.8 Classifier machine learning algorithms predicting if samples came from cancer or control samples

Using the python module "sciKit-learn", a script was written to implement six different machine learning classifier algorithms (see appendix D for script).[60] This was done to determine to what extent the different machine learning classifiers can predict which of the tissues the samples came from (either bladder cancer or the blood control), based upon the telomere length and GC content data from the experiment. The machine learning algorithms implemented were AdaBoost, decision tree, Gaussian Naïve Bayes, k-nearest neighbours, random forests and Support Vector Machine (SVM). The input for the machine learning algorithms were the telomere length estimate (in kb) and percentage GC content (between 40-60%) of the total number of reads. This was inputted in the form of an array. The features were then scaled to prevent GC content having more of an effect on the prediction because the values were larger. (Stratified) k-fold cross-validation was carried out in order to enable all of the data to be used for training and testing due to the small sample size. It was stratified because 50% of the data belonged to each of the classes (cancer and control). The classifiers could then be made and tested to determine their accuracy, recall and precision. Accuracy is the number of correct predictions over the total number of predictions, precision = true positives / (true positives + false positives) and recall = true positives / (true positives + false negatives).[61] The recall was optimised over the other metrics because reducing the number of false negatives, and so detecting lots of the cancerous samples, seems more important than reducing the number of false positives.

4. Results

4.1 Principal Components Analysis (PCA)

The PCA was used to determine if there was variance on more than one axis to allow for variable reduction. This is because if there was only variance on the x-axis, then it is likely that some of the information used for the PCA would be redundant (if it is measuring related properties) and so could be discarded. As shown in figure 9, this was not the case however, as the X and Y axes, which show principal component 1 and principal component 2, explain 43.6% and 31.7% of the total variance, respectively. This shows that there is a large amount of variation in both the axis and so the patient's age, telomere length and percentage GC content show different non-redundant information.

The PCA plot was also used to look for clustering of the data, to find variables that can differentiate the blood and cancer samples. While there may be some slight clustering of the blood and tumour samples, they are mostly overlapping. This suggests that it may be difficult to use the variables of the patient's age, telomere length and percentage GC content to discriminate between the blood and cancer sample types. There are however non-overlapping regions in the prediction ellipses. The prediction ellipses are made so that there is a 95% chance that a new observation from the same group would be inside the ellipse.[54] The prediction ellipse for the tumour sample type is also larger and so there is larger variation in the principal components, and therefore likely the variables used, for tumour sample types compared to blood sample types.

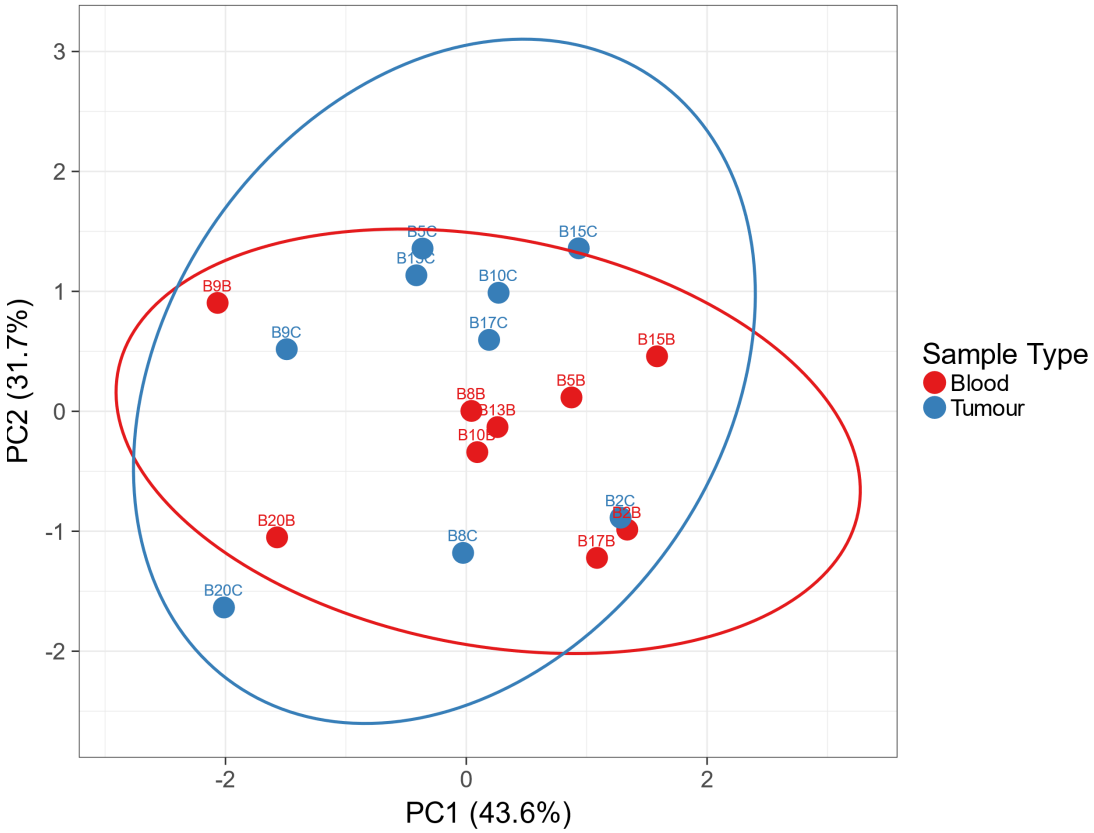


Figure 9: Principal Components Analysis (PCA) using patients age, telomere length estimate and percentage GC content for reads with content between 40-60% Each sample type for each patient is represented by a dot and so there are two dots (one for blood in blue and one for cancer in red) for each patient. In order to generate the principal components (1 and 2) on the X and Y axis respectively the inputted variables (patients age, telomere length estimate and percentage GC content) were combined among all conceivable linear combinations to show as much variation as possible while still being able to predict or "reconstruct" the original variables. Unit variance scaling was applied the rows and SVD with imputation was used to calculate the principal components. N = 18 data points.

4.2 Linear regression predicting telomere length from patient age

The linear regression indicated that there was no significant change in telomere length with age ($F_{1,7}$=0.88, P > 0.05). As the p-value is greater than 0.05 the regression slope (as shown in figure 10) does not differ significantly from zero. This means that it is unlikely for there to be a linear relationship in the data and so it is not represented well by a straight line. This was also clear from the r^2 value of 0.112 which means that only 11.2% of the variation in telomere length can be explained by age. The line was also very flat with the equation of the line being y = 0.066x - 2.04 indicating that telomere length does not change with age.

The assumptions were tested, all of which the linear regression passed. The p values for the Shapiro-Wilk normality test and ncvTest were 0.0955 and 0.451. Visual inspection was passed for the residuals vs fitted, scale-location, Q-Q plot and residuals vs leverage plots.

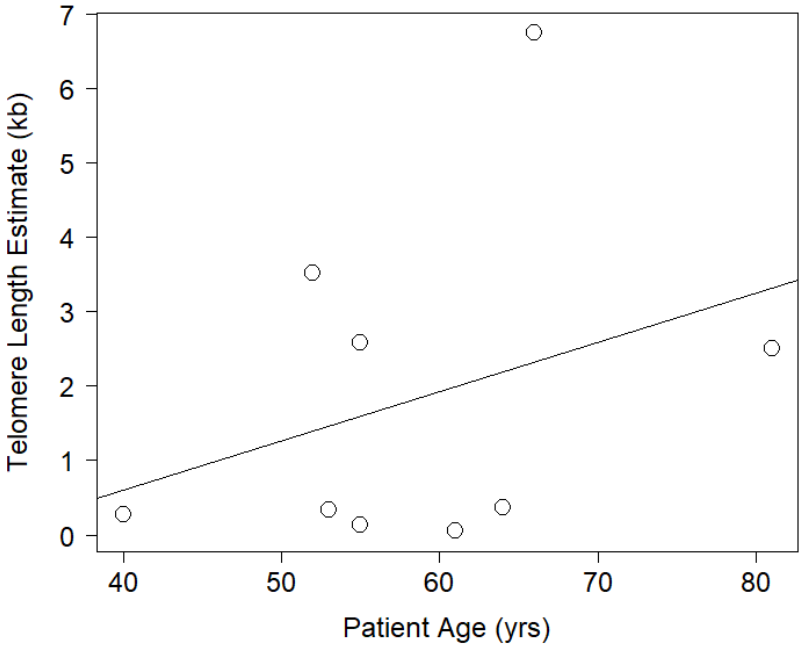


Figure 10: The effect of patient age on telomere length estimate for nine bladder cancer patients. The equation of the line is y = 0.066 - 2.04 and R^2 value is 0.11. Telomere length was estimated using the program TelSeq from exome data of blood samples, taken from each of the nine patients

4.3 Paired t-test comparing cancer and control samples telomere length

The paired t-test indicated that telomere length does not depend on bladder cancer (t_8=0.735, P > 0.05). As shown in figure 11, median telomere length (shown by the thick black line) in the blood control samples is lower than the median telomere length estimate for cancer samples. As the boxes (interquartile range) are overlapping, this suggests that the difference is not significant. For the blood tissue sample type, the interquartile range was also much larger than that of the cancer sample type. The median for the blood tissue sample type was also situated very low in the interquartile range indicating that lots of patients had a very low telomere length for their blood samples. Finally, the dotted line and error bars show the range which was greater for the cancer sample type compared to the blood sample type.

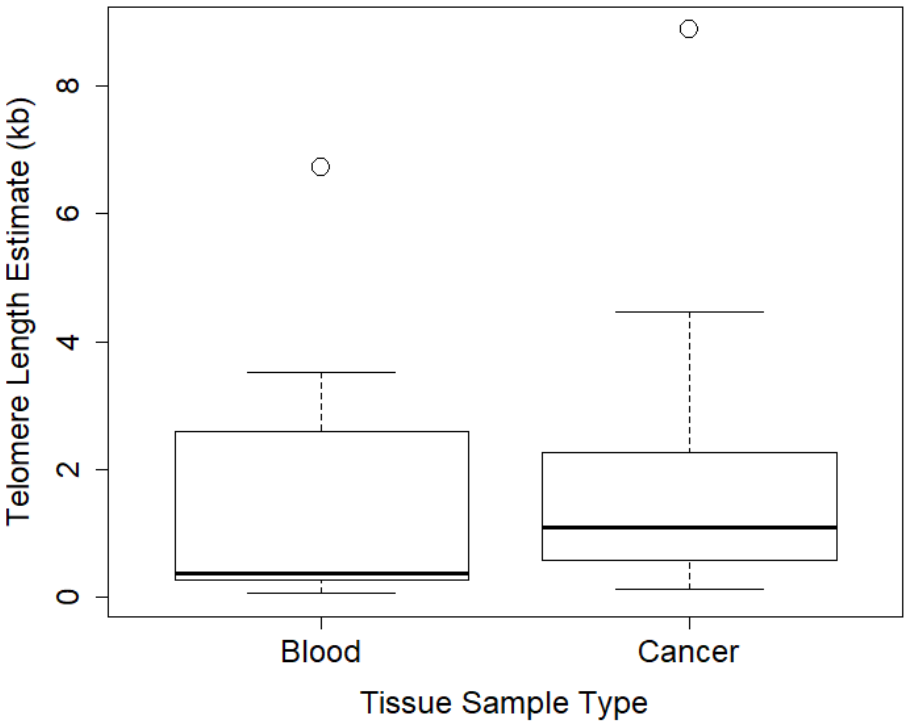


Figure 11: Boxplot of the effect of matched blood control and bladder cancer tissues samples type on telomere length estimated using the TelSeq program from exome data from the same patients. Both the bladder cancer and matched blood control samples were taken from the same patient, and so the test conducted was a paired t-test. A two-sided test was used. Number of patients, N = 9

As shown in figure 12, seven of the nine patients had longer telomeres for the cancer samples type compared to the blood sample type, while two of the nine patients had longer telomeres for the blood sample type compared to the cancer sample type. The figure also demonstrates that patient B20 was an outlier.

The p-value for the Shapiro-Wilk normality test was 0.923. This also passed visual inspection of a Q-Q plot. A Bartlett test was also performed in R and had a p-value of 0.53.

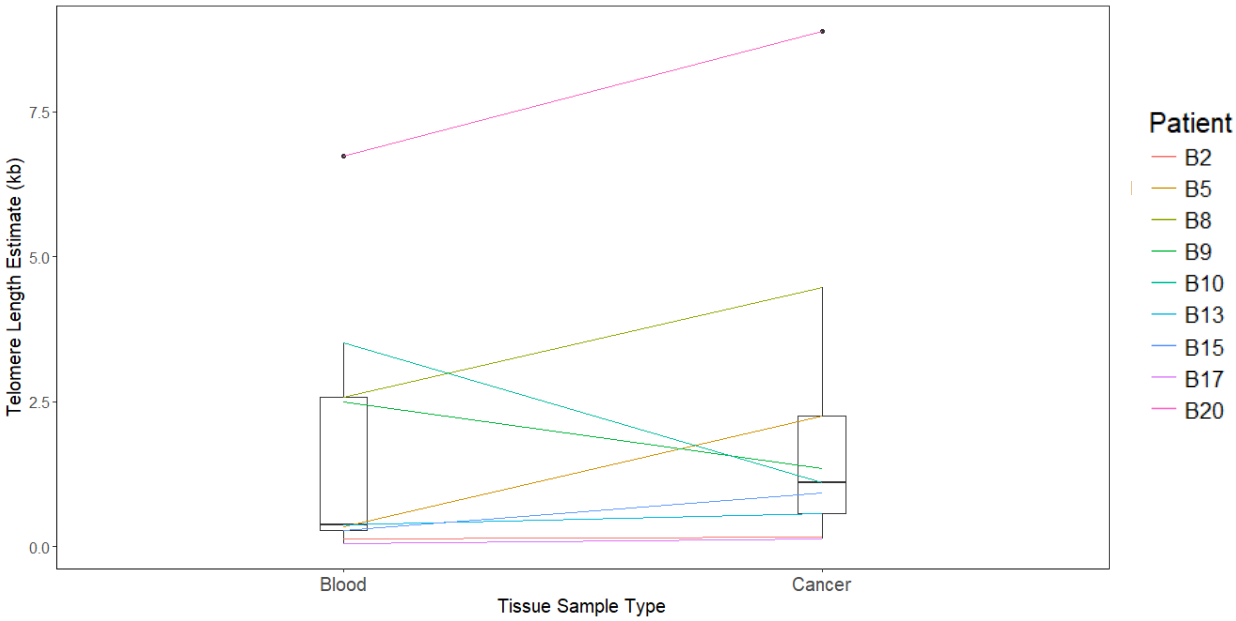


Figure 12: Boxplot of the effect of matched blood control and bladder cancer tissues samples type on telomere length estimated using the TelSeq program from exome data from the same patients with lines for patients plotted. Number of patients, N = 9

4.4 Variant Calling

There was a significant difference between observed and expected frequencies of high-impact variants for cancer and control samples (X^2 = 4.19, degrees of freedom = 1, $p<0.05$). As shown in figure 13, the observed frequency of high impact variants for cancer samples was higher than expected while the observed frequency of high-impact variants for control samples was lower than expected.

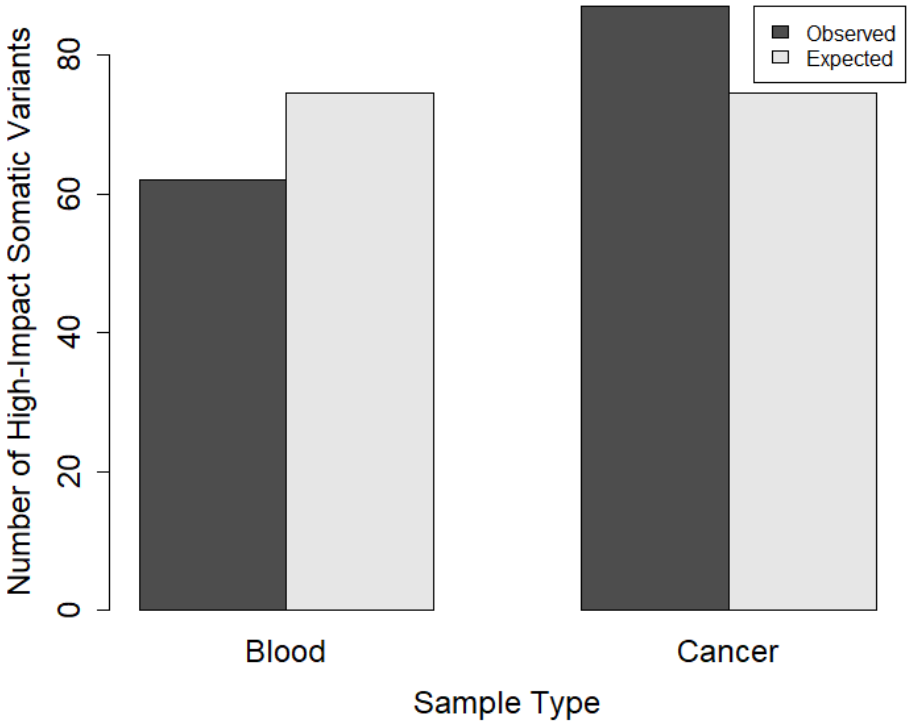


Figure 13: Observed and expected number of high-impact (stop-gain and frameshift) somatic variants in telomere-related genes for blood (control) and bladder cancer samples

As shown in table 1 the original variant list was filtered down to eight key variants of interest, found in five different patients. Cross-referencing against the original publication highlighted the one variant (in TP53). Four variants were selected in the genes TERT and DKC1, due to genes critical functions in telomeres and telomere-related disease. Finally, three other variants were selected due to having a gnomAD score equal to zero for East Asian populations or GERP++ neutral rate (NR) above five indicating highly conserved genes.

Table 1: Shortlist of high impact (frameshift and stop gain) somatic variants of interest in telomere-related genes. All variants are predicted to have a protein-encoding effect and for all variants genotype was heterozygous, i.e. carrying one copy each of the reference and alternate allele. Key cells used for filtering have been highlighted in yellow.

Patient	Location	Consequence	SYMBOL	Protein position	Amino acids	Existing variation	gnomAD EAS AF	CADD raw	GERP++ NR
B9	5:1278703	Frameshift variant	TERT	741-742	-/IQNSR AREFWIX	-	-	-	-
B8	5:1278763	Stop gained	TERT	722	Q/*	-	-	12.45737	4.63
B20	10:91842188-91842189	Frameshift variant	TNKS2	619	T/X	rs749004712, COSM294902	0	-	-
B2	15:43455904	Stop gained	TP53BP1	902	E/*	COSM253866, COSM4908986	-	12.46278	5.37
B8	16:3589452	Stop gained	SLX4	1396	G/*	-	-	9.131425	5.77
B9	17:7675194-7675208	Frameshift variant	TP53	135-139	CQLAK/X	COSM1745385, COSM255062	-	-	-
B10	X:154766295	Frameshift variant	DKC1	115	V/DPEFX	-	-	-	-
B8	X:154776314-154776316	Frameshift variant	DKC1	489-490	PG/PX	-	-	-	-

4.5 Classifier machine learning algorithms predicting if samples came from cancer or control samples

As shown in figure 14, the SVM (Support Vector Machine) classifier algorithm had the highest scores across the three metrics (accuracy, precision and recall). For scores across the three metrics was then followed by AdaBoost, then k-nearest neighbours, then the decision tree, then random forests and finally by Gaussian Naive Bayes which had the lowest scores across the three metrics. The error bars plotted are for standard error. The plotted standard error bars are quite large which indicates that the scores could be inaccurate. Also, as many of standard error bars overlap this suggests that difference (between many of the different algorithms and metrics) may not be significant. While the accuracy and recall often interchanged between being the highest, the precision was consistently the lowest (or joint lowest). As stated in the methods, this was by design because it seems more important to correctly predict if a sample came from cancerous tissue (or be able to make correct predictions in general), than to correctly predict if samples came from blood and was from the control. This was therefore taken into account during optimisation of the algorithms.

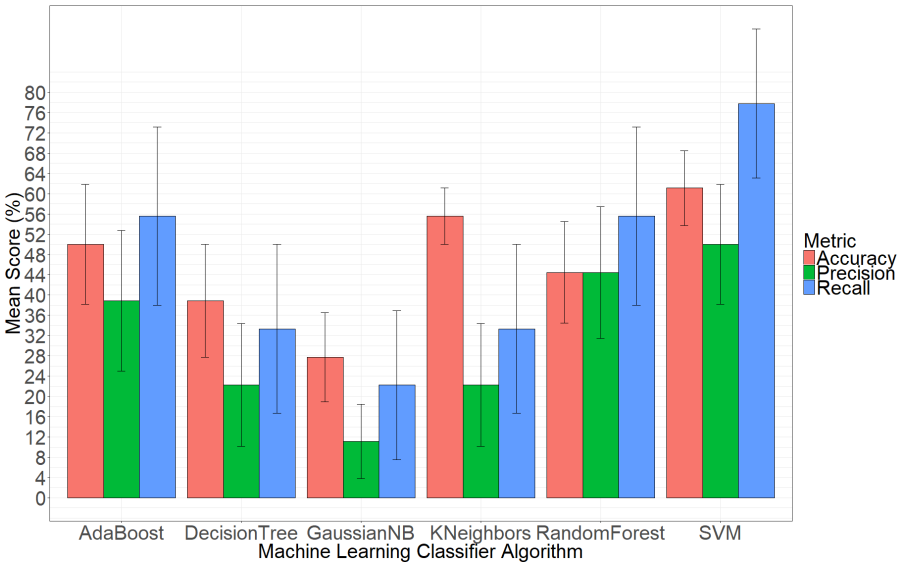


Figure 14: The effect of implementing different machine learning classifier algorithms on the scores of three different metrics (accuracy, precision and recall). The samples were classified between either being blood control or bladder cancer samples. The error bars are for standard deviation. The python module Scikit-learn was used with input as telomere length estimate (in kb) and percentage GC content (between 40-60%) of the total number of reads for all nine patient's samples

5. Discussion

5.1 Telomere's role in ageing

In order to determine whether telomere length depends on age, a linear regression was used. It was found that the p-value for the linear regression was greater than 0.05, therefore the null hypothesis of there being no significant linear relationship between the patient age and telomere length estimate (with the slope equal to zero) is therefore accepted and the alternative hypothesis is rejected. This means that telomere length does not depend on bladder cancer as the slope does not differ significantly from zero.

It was clear that the linear model did not fit the data well because the gradient was (slightly) positive, indicating that telomere length increases with age. This is not the case because telomere attrition is one of the hallmarks of ageing.[28] Another problem is that the length at birth (shown by the y-intercept, where age = 0) is negative. This would mean that at birth the genes would already be truncated as opposed to being at a peak length of 15kb, before slowly shortening over time.[4]

One reason why this may have occurred is that patients in this age range were not tested and the model cannot extrapolate well to very young ages. Evidence for this is that the youngest patient in this study was 40 years old and eight of the nine patients were 66 years old or younger.[1] Therefore all of the patients were of similar age in comparison to the actual biological age range seen in humans. This also may have made it difficult to discriminate between the different ages and so may have been responsible for not finding a significant result. The reason why all of the patients were likely middle-aged to old-aged is that age is a risk factor for cancers such as bladder cancer and so bladder cancer is less common in younger individuals.[44] Therefore to improve the experimental design, to better answer the question of whether telomere length depends on age, a more representative sample of individuals of more varying ages could have been used. This could be done using stratified sampling to include younger individuals as well as middle and old aged individuals.

The model seems to work better for older ages, where (using the equation for the line) the telomere length of an 80yr old individual is predicted to be 3.26kb (0.0662 * 80 – 2.04). This is far closer to reality as telomere length seen in old age is roughly around 4kb.[4] However this still seems lower than expected and this may be because the patients were all individuals who had been diagnosed with bladder cancer and so the data cannot be used to infer what the telomere length of a healthy individual at a similar age would be. While there is contradicting evidence for this, another study did find that shortened telomere length in blood tissues is present in bladder cancer patients and so this may explain the shorter predictions for telomere length compared to healthy individuals.[62]

An alternative explanation for the finding of this study, that telomere length does depend on age, could be due to confounding variables. Confounding variables are also likely to have a bigger effect with a small sample size, such as the one used in this study with only nine patients, as the sample may not be representative. Examples of possibly confounding variables include other factors, for example, genetics or lifestyle factors such as smoking that may have had a larger impact in determining telomere length than patient age. This is likely to be the case with smoking, as smoking increases oxidative stress which is responsible for double the amount of telomere shortening as caused by the end replication problem on average.[13] Using information from the original publication's supplementary reading, it can be calculated that the average telomere length for patients who had never smoked was 2.46kb compared to 0.139kb for the patient that did smoke.[1] However this may be unreliable because only one of the nine patients had a reported status as being a smoker (patient B2). The reason why oxidative stress (and therefore smoking) accounts for so much telomere attrition is largely as a result of damaging DNA repair proteins.[9] Further evidence for this is that patient B2 had a stop gain mutation in the gene TP53BP1, which is involved in (double-stranded break) DNA repair in response to damage.[63]

5.2 Telomere's role in cancer

In order to determine whether telomere length was different between the bladder cancer and blood control samples, a paired t-test was used. The result from the paired t-test was that the p-value was greater than 0.05. Therefore, the null hypothesis, that there is no significant difference between the sample means and so the mean difference is equal to zero, is accepted and the alternative hypothesis rejected. The result indicates that telomere length does not depend on bladder cancer. This may have simply been because telomere length is very dynamic in nature.[17] Therefore as the telomere length is constantly changing it was difficult to find a significant difference between the two sample groups (cancer and control). By measuring telomere length across many different samples/patients this should be accounted for, as these differences should average out; however, it is not clear if nine patients would be enough to account for this effect. However, other studies suggest that this is unlikely to be a problem because telomere length measurements taken over a short period of time are reliable.[64]

An explanation for there not being a difference between telomere length in bladder cancer and control samples is that, while telomere length is decreased by the high number of cell divisions seen in cancerous cells causing telomere shortening (due to the end replication problem), telomerase may be increasing telomere length. One indicator that this may be true is that telomerase is active in 98% of bladder cancers.[49] Therefore, there could be an antagonistic effect of the increased number of cell divisions (decreasing telomere length) and active telomerase (increase telomere length) and so overall no effect is seen. This does provide a plausible explanation for no difference being observed, but it is strange that telomere length was greater for cancer samples for seven of the nine patients because telomerase tends to maintain, not increase telomere length in tumours such as bladder cancer.[39] This is because it is common for telomerase to become active only once tumour cells have bypassed the cell-cycle checkpoint which leads to the crisis phase.[2] Therefore cancerous cells often contain short and stable telomeres compared to healthy cells.[2]

In order to help determine if telomerase is having an effect, variants in the TERT gene (the catalytic reverse transcriptase component of telomerase) can be observed.[17] In total four high-impact somatic variants were observed for three different patients. Not only does this seem too few mutations to explain to the high activity of telomerase (98%) in bladder cancers but none of these mutations are near the promoter with many of these mutations being more likely to be inactivating.[49] For example, the frameshift and stop gain variants in patients B9 and B8 respectively occurred at similar positions 741-742 and 722 respectively in the TERT protein. Both of these mutations would be likely to prevent or alter the downstream translation for the rest of the protein. As this includes a site required for nucleotide incorporation and two sites required for magnesium binding, affecting the catalytic activity of TERT, it's likely that the protein would be inactive and so decrease the activity of telomerase, as opposed to increasing it.[65] An explanation for how telomerase activity could still be higher in the cancer samples compared to the control could be that the reason why telomerase is barely detectable in most adult cells is due to transcriptional inactivation.[39] Therefore, changes that occur increasing telomerase activity are most likely to be transcriptional/epigenetic, such as demethylation. These changes will not be detectable by exome sequencing and so other transcriptomics techniques such as microarrays may be better suited to answering this question, also enabling a more systems biology/holistic approach.

Patient B9 was one of only two patients to have shorter telomeres for the cancer samples compared to the control sample. The telomere length estimate for the tumour of patient B9 was 1.34kb. The telomeres would, therefore, be described as critically short (being less than 3kb).[66] At this critically short length, telomeres can cause a DNA damage response, which can lead to activation of TP53.[67] TP53 is the human homolog for p53, often referred to as the master tumour suppressor gene as TP53 is the most frequently mutated gene in human cancers, with mutations occurring in over 50% of cancers.[68] In response to cellular stress such as critically

short telomeres[67], TP53 causes cell-cycle arrest (for example through p-21 waf-1), apoptosis and senescence via its role as a trans-activator transcription factor.[68]

However patient B9 had a frameshift variant in TP53. This means that TP53 would not have been activated and caused apoptosis or senescence. Evidence for this is that as the variant is at position 135-139 is the protein; four downstream zinc binding sites are likely to be out of frame causing the zinc-dependent TP53 to be inactive.[69]

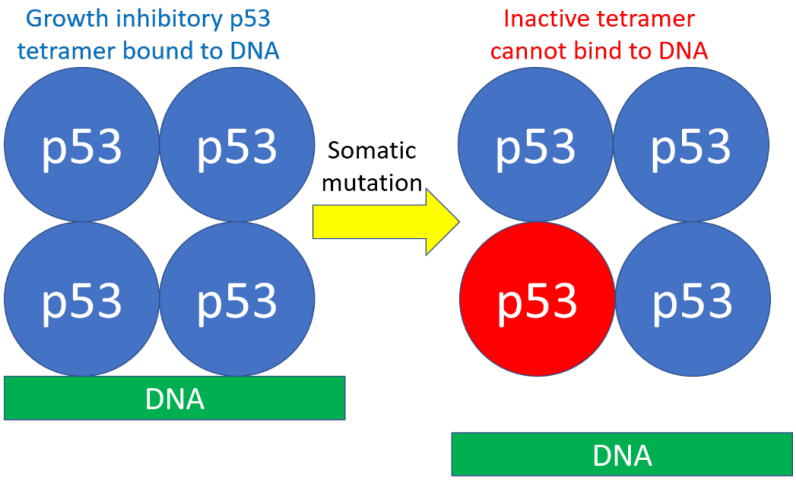


Figure 15: How one inactive p53 molecule, as a result of a somatic mutation, inactivates the whole tetramer preventing binding to DNA. This then removes the growth inhibition and so increases cancer susceptibility

While most tumour suppressor genes require two recessive copies of a gene to have an effect (known as the two hit hypothesis), TP53 is a dominant negative tumour suppressor gene. This is because (as shown by figure 15) the TP53 transcription factor binds to DNA as a tetramer and only one non-functional copy of TP53 is enough to inactivate the whole tetramer.[68] Therefore although patient B9 did only have the one mutant allele present this would be enough to promote cancer. The critically short telomeres, which would normally act as a stimulus inducing TP53 and apoptosis/senescence, would likely be unable to activate TP53. Therefore, the tumour would continue dividing and accumulating mutations. This may also be responsible for patient B9 having a large tumour (of stage T3N0M0)

bladder cancer because mutations in TP53 are associated with high-grade tumours.[1] This information may also be useful for treating the patient, as TP53 mutations are associated with altered sensitivity to drugs such as 5-Fluorouracil, Dabrafenib and (5Z)-7-Oxozeanol.[70]

5.3 Genes involved in telomeres

The chi-squared test was used to investigate the numbers of high-impact somatic variants in the bladder cancer and blood control samples. As the p-value was less than 0.05, there was a significant difference between the observed and expected number of high-impact variants for the bladder cancer and control samples. Therefore, the null hypothesis is accepted and the alternative rejected. This was because more high-impact somatic variants were found in the cancer samples compared to the control. This included genes such as DKC1, in which some genotypes/mutations have an impact on telomere length regulation which is first detectable at older ages.[71]

The more high-impact variants (seen more in cancer samples than control samples) did not have increased telomere length, as demonstrated by the paired t-test; this may however increase the variation in telomere length. Evidence for this is that the range for telomere length was greater for cancer samples compared to the control. This may have also been represented in the PCA, which showed more variation was seen in the cancer principal components compared to the blood samples (as shown by the larger predicted ellipse). This is similar to the findings of a case-control study which showed that increased telomere length variation is associated with elevated cancer risk.[72]

5.5 Limitations & further work

A limitation is that the output from TelSeq, and therefore the telomere length estimate, may be inaccurate. This is because it is difficult to map reads to the human reference genome at the ends of the chromosomes, as they are often sequences of unknown nucleotides. This is due to the ends of the reads often being very repetitive which makes mapping the reads to their original position difficult.[50]

Therefore, as the (telomeric) reads are repetitive/similar, different reads may be mistaken for the same one. Thus, the value for the number of telomeric reads and therefore the telomere length estimate may be lower, giving an inaccurate estimate of the telomere length. This is according to the equation $l=t_k sc$, which is used by the TelSeq program (where l is the telomere length and t is the number of telomeric reads).[50] Further evidence for this is that eight of the nine patients had shorter telomeres than would be expected (even taking into account their age). As eight patients' telomere lengths were below 4kb, which is expected in old age,[4] and seven of the nine patients had critically short telomere lengths (below 3kb).[66] While this could be the result of using exome sequencing, as this further restricts the available data and crucially the number of telomeric reads,[50] it could also be the case that patients with bladder cancer are more likely to have short telomeres (in blood tissues).[62] Another explanation is that this may be an error. Therefore it may have been better to use another program to estimate telomere length, such as Computel, which is less influenced by errors.[73]

Due to the low number of degrees of freedom, it was difficult to find a significant result for the differences in telomere length for both ageing and bladder cancer. Therefore, more repeats could be done with more data for patients with bladder cancer. The same technique used in this study could also be used for different types of cancers and so the results (such as the telomere length) between different types of cancer could then be compared. As all of the patients were male,[1] further repeats could also be carried out (ensuring to include women), as this would increase the scope of the experiment; otherwise, we cannot be sure that the results of experiments apply to both men and women with bladder cancer.

There were also many limitations for implementing the machine learning classifiers. These include that the features may not be independent and that the models (for example, the SVM) may be overfitted to the data. This is where the algorithm essentially learns where the data is and so does not extrapolate well to new data. Therefore, when tested on a new dataset the metrics are likely to decrease. The implementation is also unlikely to be useful because of the

processing of the data/samples. This is because the sample must be from the potentially cancerous cell (which is likely to require surgery) and the data must be in the format of TelSeq output, which requires a time consuming and likely expensive process (due to the exome sequencing and computational processing). When this is taken into account (as well as the reasonably low percentage scores for the different metrics) the classifier has little/no utility. This is especially true considering the importance of correctly identifying if a sample came from cancerous tissue, meaning that the metrics (for example, the accuracy) would need to be very high.

Overall, it was found that telomere length does not depend on age for bladder cancer patients. However, as telomere attrition is well established as one of the causes of ageing, more work needs to be done to confirm this. It is more likely that this is the result of a small sample size and poor experimental design rather than the underlying biology. Another finding was that telomere length does not depend on bladder cancer. Again, this may have been a result of the small sample size making it difficult to find a conclusive result. I suspect, given a sufficiently large sample size, that telomere length would be found to be shorter in tumours such as bladder cancers. However, there is conflicting evidence for this across different studies. More high-impact variants were seen in telomere-related genes for the cancer samples compared to the control samples than expected by chance. This included mutations in important genes such as TP53, TERT and DKC1. This is consistent with the wider field which would predict there to be more variants in tumour samples. Finally, while telomere length and GC content can be used, to some extent, to predict if a sample came from bladder cancer or blood control tissue, it is unlikely to be of any clinical utility. However, it remains necessary to further telomere research and broaden the body of knowledge so to better understand their role in ageing, cancer and stem cells; which will likely increase human longevity and contribute to beating cancer (for example, by using telomeres as anti-cancer drug targets).

6. References

1. Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, et al. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nature Genetics. 2011 Sep 7;43(9):875–8.
2. Wai LK. Telomeres, telomerase, and tumorigenesis--a review. MedGenMed: Medscape general medicine. 2004 Jul 26;6(3):19.
3. Zakian VA. Telomeres: The beginnings and ends of eukaryotic chromosomes. Vol. 318, Experimental Cell Research. NIH Public Access; 2012. p. 1456–60.
4. Bischoff C, Graakjaer J, Petersen HC, Jeune B, Bohr VA, Koelvraa S, et al. Telomere Length Among the Elderly and Oldest-Old. Twin Research and Human Genetics. 2005 Oct 1;8(5):425–32.
5. Yegorov YE, Zelenin A V. Racing for cell immortality, telomeres, telomerase, and the measure of health. Russian Journal of Developmental Biology. 2011 Jan 13;42(1):53–6.
6. Shay JW. Toward Identifying a Cellular Determinant of Telomerase Repression. JNCI Journal of the National Cancer Institute. 1999 Jan 6;91(1): 4–6.
7. Lu W, Zhang Y, Liu D, Songyang Z, Wan M. Telomeres—structure, function, and regulation. Experimental Cell Research. 2013 Jan 15;319(2):133–41.
8. Telomerase solves the end replication problem; | Learn Science at Scitable [Internet]. 2014 [cited 2018 Mar 11]. Available from: https://www.nature.com/scitable/content/telomerase-solves-the-end-replication-problem-113457928
9. Prasad KN, Wu M, Bondy SC. Telomere shortening during aging: Attenuation by antioxidants and anti-inflammatory agents. Vol. 164, Mechanisms of Ageing and Development. 2017. p. 61–6.
10. Shay JW, Wright WE. Hayflick, His Limit, and Cellular Ageing. Nature reviews Molecular cell biology. 2000 Oct 1;1(October):72–6.
11. Sherr CJ, DePinho RA. Cellular senescence: mitotic clock or culture shock? Cell. 2000 Aug 18;102(4):407–10.
12. Campisi J. Cellular senescence as a tumor-suppressor mechanism. Vol. 11, Trends in Cell Biology. Elsevier Current Trends; 2001. p. S27–31.
13. Shen J, Gammon MD, Terry MB, Wang Q, Bradshaw P, Teitelbaum SL, et al. Telomere length, oxidative damage, antioxidants and breast cancer risk. International Journal of Cancer. 2009 Apr 1;124(7):1637–43.
14. Wiseman H, Halliwell B. Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. The Biochemical journal. 1996 Jan 1;313 (Pt 1(2):17–29.
15. Shammas MA. Telomeres, lifestyle, cancer, and aging. Current Opinion in

Clinical Nutrition and Metabolic Care. 2011 Jan;14(1):28–34.

16. Wang LQ, Liu DW. Telomere, telomerase and stem cells. Vol. 13, Journal of Clinical Rehabilitative Tissue Engineering Research. Nature Publishing Group; 2009. p. 1985–8.

17. Sexton AN, Youmans DT, Collins K. Specificity requirements for human telomere protein interaction with telomerase holoenzyme. Journal of Biological Chemistry. 2012 Oct 5;287(41):34455–64.

18. The mammalian telomeric complex [Internet]. [cited 2018 Mar 12]. Available from: https://www.researchgate.net/profile/Ramiro_Verdun/publication/6254858/figure/fig3/AS:267669240807440@1440828728504/The-mammalian-telomeric-complexThe-fluorescence-image-shows-the-location-of-a-telomere.png

19. Han H, Hurley LH. G-quadruplex DNA: A potential target for anti-cancer drug design. Vol. 21, Trends in Pharmacological Sciences. Elsevier Current Trends; 2000. p. 136–42.

20. Structure of a G-quadruplex [Internet]. [cited 2018 Mar 12]. Available from: https://www.intechopen.com/source/html/42297/media/image1.jpeg

21. Denchi EL, De Lange T. Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. Nature. 2007 Aug 30;448(7157):1068–71.

22. Maréchal A, Zou L. DNA damage sensing by the ATM and ATR kinases. Cold Spring Harbor Perspectives in Biology. 2013 Sep 1;5(9):a012716.

23. De Lange T. How telomeres solve the end-protection problem. Vol. 326, Science. American Association for the Advancement of Science; 2009. p. 948–52.

24. Cesare AJ, Griffith JD. Telomeric DNA in ALT cells is characterized by free telomeric circles and heterogeneous t-loops. Mol Cell Biol. 2004 Nov 15;24(22):9948–57.

25. Tomaska L, Nosek J, Kramara J, Griffith JD. Telomeric circles: Universal players in telomere maintenance. Vol. 16, Nature Structural and Molecular Biology. Nature Publishing Group; 2009. p. 1010–5.

26. Alder JK, Chen JJ-L, Lancaster L, Danoff S, Su S -c., Cogan JD, et al. Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proceedings of the National Academy of Sciences. 2008 Sep 2;105(35):13051–6.

27. De Bari C, Dell'Accio F, Tylzanowski P, Luyten FP. Multipotent mesenchymal stem cells from adult human synovial membrane. Arthritis and Rheumatism. 2001 Aug 1;44(8):1928–42.

28. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Vol. 153, Cell. Cell Press; 2013. p. 1194–217.

29. Paeschke K, McDonald KR, Zakian VA. Telomeres: Structures in need of

unwinding. Vol. 584, FEBS Letters. 2010. p. 3760–72.

30. O'Sullivan RJ, Karlseder J. Telomeres: Protecting chromosomes against genome instability. Vol. 11, Nature Reviews Molecular Cell Biology. Nature Publishing Group; 2010. p. 171–81.

31. Martin GM, Oshima J. Lessons from human progeroid syndromes. Vol. 408, Nature. 2000. p. 263–6.

32. Ding SL, Shen CY. Model of human aging: Recent findings on Werner's and Hutchinson-Gilford progeria syndromes. Vol. 3, Clinical Interventions in Aging. Dove Press; 2008. p. 431–44.

33. Xu J, Yang X. Will cloned animals suffer premature aging - The story at the end of clones' chromosomes. Vol. 1, Reproductive Biology and Endocrinology. BioMed Central; 2003. p. 105.

34. Bernardes de Jesus B, Vera E, Schneeberger K, Tejera AM, Ayuso E, Bosch F, et al. Telomerase gene therapy in adult and old mice delays aging and increases longevity without increasing cancer. EMBO Molecular Medicine. 2012 Aug 3;4(8):691–704.

35. Turbill C, Smith S, Deimel C, Ruf T. Daily torpor is associated with telomere length change over winter in Djungarian hamsters. Biology Letters. 2012 Apr 23;8(2):304–7.

36. Wentzensen IM, Mirabello L, Pfeiffer RM, Savage SA. The association of telomere length and cancer: A meta-analysis. Cancer Epidemiology Biomarkers and Prevention. 2011 Jun 17;20(6):1238–50.

37. Ma H, Zhou Z, Wei S, Liu Z, Pooley KA, Dunning AM, et al. Shortened Telomere length is associated with increased risk of cancer: A meta-analysis. Toland AE, editor. PLoS ONE. 2011 Jun 10;6(6):e20466.

38. Zhu X, Han W, Xue W, Zou Y, Xie C, Du J, et al. The association between telomere length and cancer risk in population studies. Scientific Reports. 2016 Apr 26;6(1):22243.

39. Theodorescu D, Cech TR. Telomerase in bladder cancer: Back to a better future? European Urology. 2014 Feb;65(2):370–1.

40. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Vol. 144, Cell. Elsevier; 2011. p. 646–74.

41. Imetelstat [Internet]. [cited 2018 Mar 12]. Available from: http://www.geron.com/r-d/imetelstat/

42. Kim MY, Vankayalapati H, Shin-Ya K, Wierzba K, Hurley LH. Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular G-quadruplex. Journal of the American Chemical Society. 2002 Mar 13;124(10):2098–9.

43. Dilley RL, Greenberg RA. ALTernative Telomere Maintenance and Cancer. Vol.

1, Trends in Cancer. NIH Public Access; 2015. p. 145–56.

44. Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet. 2005 Nov 19;366(9499):1784–93.

45. Eisenberg DTA. An evolutionary review of human telomere biology: The thrifty telomere hypothesis and notes on potential adaptive paternal effects. American Journal of Human Biology. 2011 Mar;23(2):149–67.

46. UK C research. Cancer research UK cancer statistics [Internet]. 2014 [cited 2018 Mar 10]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bladder-cancer

47. RAMAKRISHNAN K. American family physician. Vol. 80, American Family Physician. American Academy of Family Physicians; 2008. 717-723 p.

48. Johansson SL, Cohen SM. Epidemiology and etiology of bladder cancer. Vol. 13, Seminars in Surgical Oncology. John Wiley & Sons, Inc.; 1997. p. 291–8.

49. Belair CD, Yeager TR, Lopez PM, Reznikoff CA. Telomerase activity: A biomarker of cell proliferation, not malignant transformation. Journal of Urology. 1998 Dec 9;160(2):620–1.

50. Ding Z, Mangino M, Aviv A, Spector T, Durbin R. Estimating telomere length from whole genome sequence data. Nucleic Acids Research. 2014 May; 42(9):e75.

51. Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, et al. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder [Internet]. Nature Genetics. 2011 [cited 2018 Mar 10]. Available from: https://www.ebi.ac.uk/ena/data/view/SRP007205

52. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. FastQC: a quality control tool for high throughput sequence data. 2015 [cited 2018 Mar 10]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

53. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010 Mar 1;26(5):589–95.

54. Metsalu T, Vilo J. ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research. 2015 Jul 1;43(W1):W566–70.

55. Team RC. The R Project for Statistical Computing. Http://WwwR-ProjectOrg/. 2013;1–12.

56. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Research. 2004 Jan 1;32(90001):493D–496.

57. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-

generation DNA sequencing data. Genome Research. 2010 Sep;20(9):1297–303.

58. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016 Dec 6;17(1):122.

59. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013 Mar 1;14(2):178–92.

60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2012;12(Oct):2825–30.

61. Olson DL, Delen D. Advanced Data Mining Techniques. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. 138 p.

62. Pavanello S, Carta A, Mastrangelo G, Campisi M, Arici C, Porru S. Relationship between telomere length, genetic traits and environmental/occupational exposures in bladder cancer risk by structural equation modelling. International Journal of Environmental Research and Public Health. 2018 Dec 21;15(1):5.

63. TP53BP1 in Cancer [Internet]. [cited 2018 Mar 12]. Available from: http://cis.hku.hk/CR2Cancer/browse.php?gene=TP53BP1

64. Kim S, Sandler DP, Carswell G, Weinberg CR, Taylor JA. Reliability and short-term intra-individual variability of telomere length measurement using monochrome multiplexing quantitative PCR. Lustig AJ, editor. PLoS ONE. 2011 Sep 30;6(9):e25774.

65. UniProt TERT Human Telomerase reverse transcriptase [Internet]. Available from: https://www.uniprot.org/uniprot/O14746

66. Quintela-Fandino M, Soberon N, Lluch A, Manso L, Calvo I, Cortes J, et al. Critically short telomeres and toxicity of chemotherapy in early breast cancer. Oncotarget. 2017 Mar 28;8(13):21472–82.

67. Artandi SE, Attardi LD. Pathways connecting telomeres and p53 in senescence, apoptosis, and cancer. Vol. 331, Biochemical and Biophysical Research Communications. 2005. p. 881–90.

68. Ozaki T, Nakagawara A. Role of p53 in cell death and human cancers. Vol. 3, Cancers. Multidisciplinary Digital Publishing Institute (MDPI); 2011. p. 994–1013.

69. UniProt P53 Cellular tumour antigen p53 TP53 [Internet]. [cited 2018 Mar 13]. Available from: https://www.uniprot.org/uniprot/P04637

70. TP53 Gene - Somatic Mutations in Cancer [Internet]. [cited 2018 Mar 13]. Available from: http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TP53

71. Melin BS, Nordfjäll K, Andersson U, Roos G. HTERT cancer risk genotypes are

associated with telomere length. Genetic Epidemiology. 2012 May 1;36(4): 368–72.

72. Wang H, Wang Y, Kota KK, Kallakury B, Mikhail NN, Sayed D, et al. Strong association between long and heterogeneous telomere length in blood lymphocytes and bladder cancer risk in Egyptian. Carcinogenesis. 2015 Nov; 36(11):1284–90.

73. Nersisyan L, Arakelyan A. Computel: Computation of mean telomere length from whole-genome next-generation sequencing data. PLoS ONE. 2015;10(4):e0125201.

7. Appendices

A: Table of files

File Name	File Type	Input/Run Accession	Description
SRR290592.fastq.gz	FASTQ	SRR290592	Stores nucleotide sequence and corresponding quality scores for patient B2 blood sample
SRR290593.fastq.gz	FASTQ	SRR290593	Stores nucleotide sequence and corresponding quality scores for patient B2 tumour sample
SRR290581.fastq.gz	FASTQ	SRR290581	Stores nucleotide sequence and corresponding quality scores for patient B5 blood sample pair 1
SRR290589.fastq.gz	FASTQ	SRR290589	Stores nucleotide sequence and corresponding quality scores for patient B5 blood sample pair 2
SRR645261.fastq.gz	FASTQ	SRR645261	Stores nucleotide sequence and corresponding quality scores for patient B5 tumour sample pair 1
SRR645563.fastq.gz	FASTQ	SRR645563	Stores nucleotide sequence and corresponding quality scores for patient B5 tumour sample pair 2
SRR645564.fastq.gz	FASTQ	SRR645564	Stores nucleotide sequence and corresponding quality scores for patient B5 tumour sample pair 3
SRR290582.fastq.gz	FASTQ	SRR290582	Stores nucleotide sequence and corresponding quality scores for patient B5 tumour sample pair 4
SRR290590.fastq.gz	FASTQ	SRR290590	Stores nucleotide sequence and corresponding quality scores for patient B5 tumour sample pair 5
SRR290594.fastq.gz	FASTQ	SRR290594	Stores nucleotide sequence and corresponding quality scores for patient B8 blood sample
SRR290597.fastq.gz	FASTQ	SRR290597	Stores nucleotide sequence and corresponding quality scores for patient B8 tumour sample
SRR290595.fastq.gz	FASTQ	SRR290595	Stores nucleotide sequence and corresponding quality scores for patient B9 blood sample
SRR645373.fastq.gz	FASTQ	SRR645373	Stores nucleotide sequence and corresponding quality scores for patient B9 tumour sample pair 1
SRR645767.fastq.gz	FASTQ	SRR645767	Stores nucleotide sequence and corresponding quality scores for patient B9 tumour sample pair 2
SRR650205.fastq.gz	FASTQ	SRR650205	Stores nucleotide sequence and corresponding quality scores for patient B9 tumour sample pair 3
SRR290598.fastq.gz	FASTQ	SRR290598	Stores nucleotide sequence and corresponding quality scores for patient B9 tumour sample pair 4
SRR290599.fastq.gz	FASTQ	SRR290599	Stores nucleotide sequence and corresponding quality scores for patient B10 blood sample
SRR645188.fastq.gz	FASTQ	SRR645188	Stores nucleotide sequence and corresponding quality scores for patient B10 tumour sample pair 1
SRR645421.fastq.gz	FASTQ	SRR645421	Stores nucleotide sequence and corresponding quality scores for patient B10 tumour sample pair 2
SRR650134.fastq.gz	FASTQ	SRR650134	Stores nucleotide sequence and corresponding quality scores for patient B10 tumour sample pair 3
SRR290600.fastq.gz	FASTQ	SRR290600	Stores nucleotide sequence and corresponding quality scores for patient B10 tumour sample pair 4
SRR290583.fastq.gz	FASTQ	SRR290583	Stores nucleotide sequence and corresponding quality scores for patient B13 blood sample pair 1
SRR290585.fastq.gz	FASTQ	SRR290585	Stores nucleotide sequence and corresponding quality scores for patient B13 blood sample pair 2
SRR645201.fastq.gz	FASTQ	SRR645201	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 1
SRR645443.fastq.gz	FASTQ	SRR645443	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 2
SRR645444.fastq.gz	FASTQ	SRR645444	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 3
SRR650135.fastq.gz	FASTQ	SRR650135	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 4
SRR290584.fastq.gz	FASTQ	SRR290584	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 5
SRR290586.fastq.gz	FASTQ	SRR290586	Stores nucleotide sequence and corresponding quality scores for patient B13 tumour sample pair 6
SRR290579.fastq.gz	FASTQ	SRR290579	Stores nucleotide sequence and corresponding quality scores for patient B15 blood sample pair 1
SRR290587.fastq.gz	FASTQ	SRR290587	Stores nucleotide sequence and corresponding quality scores for patient B15 blood sample pair 2
SRR645205.fastq.gz	FASTQ	SRR645205	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 1
SRR645451.fastq.gz	FASTQ	SRR645451	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 2
SRR645452.fastq.gz	FASTQ	SRR645452	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 3
SRR650136.fastq.gz	FASTQ	SRR650136	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 4
SRR290580.fastq.gz	FASTQ	SRR290580	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 5
SRR290588.fastq.gz	FASTQ	SRR290588	Stores nucleotide sequence and corresponding quality scores for patient B15 tumour sample pair 6
SRR290601.fastq.gz	FASTQ	SRR290601	Stores nucleotide sequence and corresponding quality scores for patient B17 blood sample
SRR290577.fastq.gz	FASTQ	SRR290577	Stores nucleotide sequence and corresponding quality scores for patient B17 tumour sample pair 1
SRR290578.fastq.gz	FASTQ	SRR290578	Stores nucleotide sequence and corresponding quality scores for patient B17 tumour sample pair 2
SRR290596.fastq.gz	FASTQ	SRR290596	Stores nucleotide sequence and corresponding quality scores for patient B20 blood sample
SRR290591.fastq.gz	FASTQ	SRR290591	Stores nucleotide sequence and corresponding quality scores for patient B20 tumour sample
B5B.fastq.gz	FASTQ	SRR290581.fastq.gz, SRR290589.fastq.gz	Multiple pair files from patient B5 blood samples were joined
B5C.fastq.gz	FASTQ	SRR645261.fastq.gz, SRR645563.fastq.gz, SRR645564.fastq.gz, SRR290582.fastq.gz, SRR290590.fastq.gz	Multiple pair files from patient B5 tumour samples were joined
B9C.fastq.gz	FASTQ	SRR645373.fastq.gz, SRR645767.fastq.gz, SRR650205.fastq.gz, SRR290598.fastq.gz	Multiple pair files from patient B9 tumour samples were joined
B10C.fastq.gz	FASTQ	SRR645188.fastq.gz, SRR645421.fastq.gz, SRR650134.fastq.gz, SRR290600.fastq.gz	Multiple pair files from patient B10 tumour samples were joined
B13B.fastq.gz	FASTQ	SRR290583.fastq.gz, SRR290585.fastq.gz	Multiple pair files from patient B13 blood samples were joined
B13C.fastq.gz	FASTQ	SRR645201.fastq.gz, SRR645443.fastq.gz, SRR645444.fastq.gz, SRR650135.fastq.gz, SRR290584.fastq.gz, SRR290586.fastq.gz	Multiple pair files from patient B13 tumour samples were joined
B15B.fastq.gz	FASTQ	SRR290579.fastq.gz, SRR290587.fastq.gz	Multiple pair files from patient B15 blood samples were joined
B15C.fastq.gz	FASTQ	SRR645205.fastq.gz, SRR645451.fastq.gz, SRR645452.fastq.gz, SRR650136.fastq.gz, SRR290580.fastq.gz, SRR290588.fastq.gz	Multiple pair files from patient B15 tumour samples were joined
B17C.fastq.gz	FASTQ	SRR290577.fastq.gz, SRR290578.fastq.gz	Multiple pair files from patient B17 tumour samples were joined
B2B_marked_dups.bam	BAM	SRR290592.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B2 blood sample
B2C_marked_dups.bam	BAM	SRR290593.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B2 tumour sample
B5B_marked_dups.bam	BAM	B5B.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B5 blood sample
B5C_marked_dups.bam	BAM	B5C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B5 tumour sample
B8B_marked_dups.bam	BAM	SRR290594.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B8 blood sample
B8C_marked_dups.bam	BAM	SRR290597.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B8 tumour sample
B9B_marked_dups.bam	BAM	SRR290595.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B9 blood sample
B9C_marked_dups.bam	BAM	B9C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B9 tumour sample
B10B_marked_dups.bam	BAM	SRR290599.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B10 blood sample
B10C_marked_dups.bam	BAM	B10C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B10 tumour sample
B13B_marked_dups.bam	BAM	B13B.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B13 blood sample
B13C_marked_dups.bam	BAM	B13C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B13 tumour sample
B15B_marked_dups.bam	BAM	B15B.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B15 blood sample
B15C_marked_dups.bam	BAM	B15C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B15 tumour sample
B17B_marked_dups.bam	BAM	SRR290601.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B17 blood sample
B17C_marked_dups.bam	BAM	B17C.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B17 tumour sample
B20B_marked_dups.bam	BAM	SRR290596.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B20 blood sample
B20C_marked_dups.bam	BAM	SRR290591.fastq.gz	Binary version of a tab-delimited text file containing sequence alignment information for patient B20 tumour sample
telSeq.xlsx	XLSX	BAM files	TelSeq output including estimates for telomere lengths and GC content from samples
telSeq.csv	CSV	telSeq.xlsx	For ClustVis input to generate a Principal Components Analysis (PCA) plot and heatmap
genelist.bed	BED		Candidate gene list conatinng gene symbols for telomere-related genes
B2B.g.vcf	GVCF	B2B_marked_dups.bam	Contains information about positions in the genome for patient B2 blood sample
B2C.g.vcf	GVCF	B2C_marked_dups.bam	Contains information about positions in the genome for patient B2 tumour sample
B5B..gvcf	GVCF	B5B_marked_dups.bam	Contains information about positions in the genome for patient B5 blood sample
B5C.g.vcf	GVCF	B5C_marked_dups.bam	Contains information about positions in the genome for patient B5 tumour sample
B8B..gvcf	GVCF	B8B_marked_dups.bam	Contains information about positions in the genome for patient B8 blood sample
B8C.g.vcf	GVCF	B8C_marked_dups.bam	Contains information about positions in the genome for patient B8 tumour sample
B9B.g.vcf	GVCF	B9B_marked_dups.bam	Contains information about positions in the genome for patient B9 blood sample
B9C..gvcf	GVCF	B9C_marked_dups.bam	Contains information about positions in the genome for patient B9 tumour sample
B10B.g.vcf	GVCF	B10B_marked_dups.bam	Contains information about positions in the genome for patient B10 blood sample
B10C.g.vcf	GVCF	B10C_marked_dups.bam	Contains information about positions in the genome for patient B10 tumour sample
B13B.g.vcf	GVCF	B13B_marked_dups.bam	Contains information about positions in the genome for patient B13 blood sample
B13C.g.vcf	GVCF	B13C_marked_dups.bam	Contains information about positions in the genome for patient B13 tumour sample
B15B.g.vcf	GVCF	B15B_marked_dups.bam	Contains information about positions in the genome for patient B15 blood sample
B15C.g.vcf	GVCF	B15C_marked_dups.bam	Contains information about positions in the genome for patient B15 tumour sample
B17B.g.vcf	GVCF	B17B_marked_dups.bam	Contains information about positions in the genome for patient B17 blood sample
B17C.g.vcf	GVCF	B17C_marked_dups.bam	Contains information about positions in the genome for patient B17 tumour sample
B20B.g.vcf	GVCF	B20B_marked_dups.bam	Contains information about positions in the genome for patient B20 blood sample
B20C.g.vcf	GVCF	B20C_marked_dups.bam	Contains information about positions in the genome for patient B20 tumour sample
all.vcf	VCF	GVCF files	Contains information about positions in the genome for all samples
variants.txt	TXT	all.vcf	List of variants in telomere-related genes outputted from VEP for all samples

B: Data management plan

The data was downloaded from EMBL-EBI European Nucleotide Archive (ENA). This was downloaded from the web address: https://www.ebi.ac.uk/ena/data/view/SRP007205 and the study accession is SRP007205. I downloaded publicly available data to use for this project from this database.
The data type is exomic data. I.e. the protein-coding compartment of the DNA/genome and is from humans, the species, therefore, being *Homo sapiens*. 42 files from 9 different patients (multiple samples from some patients) were originally downloaded in FASTQ file format. This is a text-based format used for storing nucleotide sequence and corresponding quality scores. To encode the sequence letter and quality score respectively an ASCII character is used. Having downloaded the data BAM files were then made from it. This is a binary version of a tab-delimited text file containing sequence alignment information. Following this the data was used to generate output from the TelSeq program in the form of tabular data, This was stored on my personal laptop and the third party Dropbox in the excel/CSV file type. For more information on files see the table of file types.
The data is anonymous and the nine patients were given patient IDs. However, there is (a very small) possibility that this data could be traced back to the individuals involved. This is because the Y chromosome is always inherited from the father, and typically surnames are inherited from the father. There are distinctive short-tandem repetitive regions on the Y chromosome which are often shared between closely related men. Algorithms can, therefore, be used to mine the exome samples for tell-tale variations in the Y chromosome. Then commercial genealogy databases can be searched for close matches to identify potential surnames. Additional information on the individuals could also be used. For example from the data I used the patient's age and sex is also published in the accompanying journal article and so could be used to help track them down. While it is unlikely that all patients could identified, it may be possibly to identify the minority. This method was successful to identify 5 of 10 patients (and their entire families) in a study (available at the web address: http://science.sciencemag.org/content/339/6117/321.full) which details the same method as described above.
If the data were to end up in the wrong hands this could result in anything from discrimination from unscrupulous employers/insurance companies to DNA samples falsely being planted at a crime scene. However, as this data is already in the public domain, I cannot make the information on the patients anymore/less accessible.
The sequencing data is stored via the university's supercluster Iridis4. Where I have been an allocated a total quota of 990GB. Having completed the project the files will be deleted. I am currently storing the FASTQ, BAM and BAI (BAM files index) on my scratch directory on iridis. However, this is not backed up and so I am storing any scripts in my home directory which is backed up. That way if Iridis were to lose the data stored on Iridis (on my scratch directory) I would able to re-download the data from ENA and use the scripts to generate any required files again.

C: Candidate gene list

gene symbol		
ABL1	MSH2	RB1
ACD	MSH3	RFC1
AKT1	MUS81	RIF1
ATM	MYC	RTEL1
ATP5C1	NBN (NBS1)	SART1
BCL2	NCL	SIRT2
BLM	NHP2	SIRT6
CDK2	NOP10	SLX4
CHEK1	OBFC1	SMAD3
CHEK2 (RAD53)	PARP1 (ADPRT1)	SMG6
DCLRE1B	PAX8	SP1
DCLRE1C	PIF1	SSB
DKC1	PINX1	SUN1
EGF	PLK1	TERC
EME1	POT1	TEP1
ERCC1	PPARG	TERF1
ERCC4	PPP2R1A	TERF2
GAR1	PPP2R1B	TERF2IP
HAT1	PRKCA	TERT
HNRNPA2B1	PRKCB	TGFB1
HNRNPD	PRKDC	TINF2
HSP90AA1	PTGES3	TNKS (TIN1)
HSPA1L	PURA	TNKS2
IGF1	RAD17	TP53
KRAS	RAD50	TP53BP1

KRIT1	RAP1A	TPP1
MEN1	RAPGEF1	XRCC5
MRE11A	RASSF1	XRCC6 (G22P1)

D: Scripts

Fastqc

```
#PBS -l nodes=1:ppn=16
#PBS -l walltime=04:00:00
#PBS -l mem=20gb
cd $PBS_O_WORKDIR

module load FastQC

#B2Blood
fastqc /scratch/pp5g15/fastqs/SRR290592_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/SRR290592_2.fastq.gz

#B2Cancer
fastqc /scratch/pp5g15/fastqs/B2Cancer/SRR290593_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B2Cancer/SRR290593_2.fastq.gz

#B5Blood
fastqc /scratch/pp5g15/fastqs/B5Blood/Pair1/SRR290581_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Blood/Pair1/SRR290581_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Blood/Pair2/SRR290589_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Blood/Pair2/SRR290589_2.fastq.gz

#B5Cancer
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair1/SRR645261_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair1/SRR645261_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair2/SRR645563_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair2/SRR645563_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair3/SRR645564_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair3/SRR645564_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair4/SRR290582_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair4/SRR290582_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair5/SRR290590_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B5Cancer/Pair5/SRR290590_2.fastq.gz

#B8Blood
fastqc /scratch/pp5g15/fastqs/B8Blood/SRR290594_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B8Blood/SRR290594_2.fastq.gz

#B8Cancer
fastqc /scratch/pp5g15/fastqs/B8Cancer/SRR290597_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B8Cancer/SRR290597_2.fastq.gz

#B9Blood
fastqc /scratch/pp5g15/fastqs/B9Blood/SRR290595_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Blood/SRR290595_2.fastq.gz

#B9Cancer
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair1/SRR645373_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair1/SRR645373_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair2/SRR645767_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair2/SRR645767_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair3/SRR650205_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair3/SRR650205_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair4/SRR290598_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B9Cancer/Pair4/SRR290598_2.fastq.gz

#B10Blood
```

```
fastqc /scratch/pp5g15/fastqs/B10Blood/SRR290599_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Blood/SRR290599_2.fastq.gz

#B10Cancer
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair1/SRR645188_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair1/SRR645188_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair2/SRR645421_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair2/SRR645421_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair3/SRR650134_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair3/SRR650134_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair4/SRR290600_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B10Cancer/Pair4/SRR290600_2.fastq.gz

#B13Blood
fastqc /scratch/pp5g15/fastqs/B13Blood/Pair1/SRR290583_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Blood/Pair1/SRR290583_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Blood/Pair2/SRR290585_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Blood/Pair2/SRR290585_2.fastq.gz

#B13Cancer
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair1/SRR645201_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair1/SRR645201_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair2/SRR645443_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair2/SRR645443_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair3/SRR645444_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair3/SRR645444_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair4/SRR650135_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair4/SRR650135_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair5/SRR290584_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair5/SRR290584_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair6/SRR290586_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B13Cancer/Pair6/SRR290586_2.fastq.gz

#B15Blood
fastqc /scratch/pp5g15/fastqs/B15Blood/Pair1/SRR290579_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Blood/Pair1/SRR290579_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Blood/Pair2/SRR290587_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Blood/Pair2/SRR290587_2.fastq.gz

#B15Cancer
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair1/SRR645205_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair1/SRR645205_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair2/SRR645451_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair2/SRR645451_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair3/SRR645452_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair3/SRR645452_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair4/SRR650136_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair4/SRR650136_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair5/SRR290580_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair5/SRR290580_2.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair6/SRR290588_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B15Cancer/Pair6/SRR290588_2.fastq.gz

#B17Blood
fastqc /scratch/pp5g15/fastqs/B17Blood/SRR290601_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B17Blood/SRR290601_2.fastq.gz

#B17Cancer
fastqc /scratch/pp5g15/fastqs/B17Cancer/Pair1/SRR290577_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B17Cancer/Pair1/SRR290577_2.fastq.gz
```

```
fastqc /scratch/pp5g15/fastqs/B17Cancer/Pair2/SRR290578_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B17Cancer/Pair2/SRR290578_2.fastq.gz

#B20Blood
fastqc /scratch/pp5g15/fastqs/B20Blood/SRR290596_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B20Blood/SRR290596_2.fastq.gz

#B20Cancer
fastqc /scratch/pp5g15/fastqs/B20Cancer/SRR290591_1.fastq.gz
fastqc /scratch/pp5g15/fastqs/B20Cancer/SRR290591_2.fastq.gz
rm "$sampleID"_aligned.sam "$sampleID"_aligned.bam
```

Zcat

```
#PBS -l walltime=02:00:00
cd $PBS_O_WORKDIR

zcat /scratch/pp5g15/fastqs/B5Blood/SRR290581_1.fastq.gz /scratch/pp5g15/fastqs/B5Blood/
SRR290589_1.fastq.gz | gzip > B5B1.fastqc.gz
zcat /scratch/pp5g15/fastqs/B5Blood/SRR290581_2.fastq.gz /scratch/pp5g15/fastqs/B5Blood/
SRR290589_2.fastq.gz | gzip > B5B2.fastqc.gz

zcat /scratch/pp5g15/fastqs/B5Cancer/Pair1/SRR645261_1.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair2/
SRR645563_1.fastq.gz \
/scratch/pp5g15/fastqs/B5Cancer/Pair3/SRR645564_1.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair4/
SRR290582_1.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair5/SRR290590_1.fastq.gz \
| gzip > B5C1.fastqc.gz
zcat /scratch/pp5g15/fastqs/B5Cancer/Pair1/SRR645261_2.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair2/
SRR645563_2.fastq.gz \
/scratch/pp5g15/fastqs/B5Cancer/Pair3/SRR645564_2.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair4/
SRR290582_2.fastq.gz /scratch/pp5g15/fastqs/B5Cancer/Pair5/SRR290590_2.fastq.gz \
| gzip > B5C2.fastqc.gz

zcat /scratch/pp5g15/fastqs/B10Cancer/Pair1/SRR645188_1.fastq.gz /scratch/pp5g15/fastqs/B10Cancer/Pair2/
SRR645421_1.fastq.gz \
/scratch/pp5g15/fastqs/B10Cancer/Pair3/SRR650134_1.fastq.gz /scratch/pp5g15/fastqs/B10Cancer/Pair4/
SRR290600_1.fastq.gz \
| gzip > B10C_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B10Cancer/Pair1/SRR645188_2.fastq.gz /scratch/pp5g15/fastqs/B10Cancer/Pair2/
SRR645421_2.fastq.gz \
/scratch/pp5g15/fastqs/B10Cancer/Pair3/SRR650134_2.fastq.gz /scratch/pp5g15/fastqs/B10Cancer/Pair4/
SRR290600_2.fastq.gz \
| gzip > B10C_2.fastq.gz

zcat /scratch/pp5g15/fastqs/B13Blood/Pair1/SRR290583_1.fastq.gz /scratch/pp5g15/fastqs/B13Blood/Pair2/
SRR290585_1.fastq.gz | gzip > B13B_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B13Blood/Pair1/SRR290583_2.fastq.gz /scratch/pp5g15/fastqs/B13Blood/Pair2/
SRR290585_2.fastq.gz | gzip > B13B_2.fastq.gz

zcat /scratch/pp5g15/fastqs/B13Cancer/Pair1/SRR645201_1.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair2/
SRR645443_1.fastq.gz \
/scratch/pp5g15/fastqs/B13Cancer/Pair3/SRR645444_1.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair4/
SRR650135_1.fastq.gz \
/scratch/pp5g15/fastqs/B13Cancer/Pair5/SRR290584_1.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair6/
SRR290586_1.fastq.gz \
| gzip > B13C_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B13Cancer/Pair1/SRR645201_2.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair2/
SRR645443_2.fastq.gz \
/scratch/pp5g15/fastqs/B13Cancer/Pair3/SRR645444_2.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair4/
SRR650135_2.fastq.gz \
```

```
/scratch/pp5g15/fastqs/B13Cancer/Pair5/SRR290584_2.fastq.gz /scratch/pp5g15/fastqs/B13Cancer/Pair6/
SRR290586_2.fastq.gz \
| gzip > B13C_2.fastq.gz

zcat /scratch/pp5g15/fastqs/B15Blood/Pair1/SRR290579_1.fastq.gz /scratch/pp5g15/fastqs/B15Blood/Pair2/
SRR290587_1.fastq.gz | gzip > B13B_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B15Blood/Pair1/SRR290579_2.fastq.gz /scratch/pp5g15/fastqs/B15Blood/Pair2/
SRR290587_2.fastq.gz | gzip > B13B_2.fastq.gz

zcat /scratch/pp5g15/fastqs/B15Cancer/Pair1/*_1.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair2/*_1.fastq.gz
\
/scratch/pp5g15/fastqs/B15Cancer/Pair3/*_1.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair4/*_1.fastq.gz \
/scratch/pp5g15/fastqs/B15Cancer/Pair5/*_1.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair6/*_1.fastq.gz \
| gzip > B15C_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B15Cancer/Pair1/*_2.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair2/*_2.fastq.gz
\
/scratch/pp5g15/fastqs/B15Cancer/Pair3/*_2.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair4/*_2.fastq.gz \
/scratch/pp5g15/fastqs/B15Cancer/Pair5/*_2.fastq.gz /scratch/pp5g15/fastqs/B15Cancer/Pair6/*_2.fastq.gz \
| gzip > B15C_2.fastq.gz

zcat /scratch/pp5g15/fastqs/B17Cancer/Pair1/SRR290577_1.fastq.gz /scratch/pp5g15/fastqs/B17Cancer/Pair2/
SRR290578_1.fastq.gz | gzip > B17C_1.fastq.gz
zcat /scratch/pp5g15/fastqs/B17Cancer/Pair1/SRR290577_2.fastq.gz /scratch/pp5g15/fastqs/B17Cancer/Pair2/
SRR290578_2.fastq.gz | gzip > B17C_2.fastq.gz
```

Bwa

```
#PBS -l nodes=1:ppn=16
#PBS -l walltime=04:00:00
#PBS -l mem=20gb
cd $PBS_O_WORKDIR

module load samtools/1.2.1
module load picard/1.97
module load bwa

##############
#Enter the sampleID for the sample
# e.g. For the file "abc123_1.fastq.gz" put sampleID=abc123
#############

#B2Blood
sampleID=SRR290592

############################
#Concatentate Multilane FQ Files then BWA mem Align
# aligning fastq files  with the reference genome using bwa-mem
# -t = Threads -M = flag shorter split hits as secondary
# -R = Readgroups -O = Gap open penalty -E =Gap extension penalty
############################

bwa mem -t 16 -M -R '@RG\tID:'$sampleID'_lane1\tSM:'$sampleID'\tPL:ILLUMINA\tLB:Library' \
-O 65 -E 7 /scratch/pp5g15/refGenome/hs38.fa \
/scratch/pp5g15/fastqs/"$sampleID"_1.fastq.gz /scratch/pp5g15/fastqs/"$sampleID"_2.fastq.gz >
"$sampleID"_aligned.sam

######################
# convert SAM to BAM
######################
```

```
samtools view -bS -o "$sampleID"_aligned.bam "$sampleID"_aligned.sam
# Index BAM file
picard BuildBamIndex INPUT="$sampleID"_aligned.bam TMP_DIR=tmp3
VALIDATION_STRINGENCY=SILENT

# sort bam file
picard SortSam INPUT="$sampleID"_aligned.bam OUTPUT="$sampleID"_aligned_sorted.bam
SORT_ORDER=coordinate TMP_DIR=tmp1 VALIDATION_STRINGENCY=SILENT
MAX_RECORDS_IN_RAM=2000000

# mark duplicates
picard MarkDuplicates INPUT="$sampleID"_aligned_sorted.bam
METRICS_FILE="$sampleID"_dup_metrics OUTPUT="$sampleID"_marked_dups.bam TMP_DIR=tmp2
VALIDATION_STRINGENCY=SILENT

# Index BAM file
picard BuildBamIndex INPUT="$sampleID"_aligned_sorted.bam TMP_DIR=tmp3
VALIDATION_STRINGENCY=SILENT

#########################
#Remove SAM, Intermediate BAMs
#########################
rm "$sampleID"_aligned.sam "$sampleID"_aligned.bam
```

GATK

```
#!/bin/bash
##PBS -l walltime=05:00:00
##PBS -l mem=20gb
#cd $PBS_O_WORKDIR

module load GATK/3.7

########################
# copy index files
########################

cp /scratch/pp5g15/refGenome/hs38.* .
ref_genome=hs38.fa
sampleID=SRR290599
bamfile=SRR290599_marked_dups.bam

#########################
# GATK variants calling
#########################

java -jar /local/software/GATK/3.6/source/GenomeAnalysisTK.jar \
        -T HaplotypeCaller \
        --emitRefConfidence GVCF \
        -R $ref_genome \
        -I $bamfile \
        -L genelist.bed \
        -o ${sampleID}.g.vcf

#################################################################################
# Get alternative genotypes (include multiple input GVCFs here if needed)
#################################################################################

java -jar /local/software/GATK/3.6/source/GenomeAnalysisTK.jar \
        -T GenotypeGVCFs \
```

```
        -R $ref_genome \
        -V ${sampleID}.g.vcf \
        -L genelist.bed \
        -o ${sampleID}.vcf

########################
# Might want to add some commands to remove the files you no longer need
###################################
# You should be able to annotate the vcf file produced using VEP
```

GATK2

```
##PBS -l walltime=05:00:00
##PBS -l mem=20gb
#cd $PBS_O_WORKDIR

module load GATK/3.7

#######################
# copy index files
#######################

cp /scratch/pp5g15/refGenome/hs38.* .
ref_genome=hs38.fa
sampleID=SRR290599
bamfile=SRR290599_marked_dups.bam

#############################################################################
# Get alternative genotypes (include multiple input GVCFs here if needed)
#############################################################################

java -jar /local/software/GATK/3.6/source/GenomeAnalysisTK.jar \
        -T GenotypeGVCFs \
        -R $ref_genome \
    -V /scratch/pp5g15/fastqs/B2Blood/SRR290592.g.vcf \
        -V /scratch/pp5g15/fastqs/B2C/SRR290593.g.vcf \
        -V /scratch/pp5g15/fastqs/B5Blood/B5B.g.vcf \
        -V /scratch/pp5g15/fastqs/B5Cancer/B5C.g.vcf \
        -V /scratch/pp5g15/fastqs/B8B/SRR290594.g.vcf \
        -V /scratch/pp5g15/fastqs/B8C/SRR290597.g.vcf \
        -V /scratch/pp5g15/fastqs/B9B/SRR290595.g.vcf \
        -V /scratch/pp5g15/fastqs/B9C/B9C.g.vcf \
        -V /scratch/pp5g15/fastqs/B10B/SRR290599.g.vcf \
        -V /scratch/pp5g15/fastqs/B10C/B10C.g.vcf \
        -V /scratch/pp5g15/fastqs/B13Blood/B13B.g.vcf \
        -V /scratch/pp5g15/fastqs/B13C/B13C.g.vcf \
        -V /scratch/pp5g15/fastqs/B15Blood/B15B.g.vcf \
        -V /scratch/pp5g15/fastqs/B15Cancer/B15C.g.vcf \
        -V /scratch/pp5g15/fastqs/B17Blood/SRR290601.g.vcf \
        -V /scratch/pp5g15/fastqs/B17Cancer/B17C.g.vcf \
        -V /scratch/pp5g15/fastqs/B20Blood/SRR290596.g.vcf \
        -V /scratch/pp5g15/fastqs/B20Cancer/SRR290591.g.vcf \
    -L genelist.bed \
        -o all6.vcf


########################
# Might want to add some commands to remove the files you no longer need
###################################
# You should be able to annotate the vcf file produced using VEP
```

TelSeq

```
#!/bin/bash
   #PBS -l walltime=02:00:00
   cd $PBS_O_WORKDIR

   module load bamtools/2.3.0
   module load telseq/0.0.1
   module load gcc/6.1.0
   telseq /scratch/pp5g15/fastqs/B2Blood/SRR290592_aligqned_sorted.bam >> /home/pp5g15/telseq.txt
```

Sci-kit learn

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_predict
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import RandomForestClassifier
import numpy as np

#features telomere length estimate and % GC content for reads with GC between 40%-60% of total read number
telomere_length = [[0.139055], [0.339294], [2.58206], [2.49741], [3.51064], [0.372471], [0.273935],
[0.0504783], [6.73724], [0.162185], [2.25344], [4.47139], [1.3383], [1.0944], [0.575941], [0.927992],
[0.130705], [8.8823]]
gc_content = [[68.06475385], [67.36651112], [60.54493243], [49.76380211], [58.29520422], [63.81705136],
[50.64077184], [54.71576192], [59.60902671], [53.78914613], [61.43470356], [52.68115526], [59.18727428],
[52.30982409], [69.36167265], [57.063573], [59.31693188], [60.49230117]]
#labels (Normal = 0, Cancer = 1)
y = np.array([0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1])

#feature scaling
scaler = MinMaxScaler()
rescaled_telomere_length = scaler.fit_transform(telomere_length)
rescaled_gc_content = scaler.fit_transform(gc_content)
X = np.concatenate((rescaled_telomere_length, rescaled_gc_content), axis=1)

#stratified k-fold cross-validation
skf = StratifiedKFold(n_splits=9)
skf.get_n_splits(X, y)
#print(skf)
for train_index, test_index in skf.split(X, y):
  #print("TRAIN:", train_index, "TEST:", test_index)
  X_train, X_test = X[train_index], X[test_index]
  y_train, y_test = y[train_index], y[test_index]

"""#Optimizing C
for i in range(1,5):
   print "C = " + str(10**i)

   clf = SVC(kernel='rbf', C=10**i)

   t0 = time()
   clf.fit(X, y)
```

```
    print "training time:", round(time()-t0, 3), "s"

    t1 = time()
    pred = clf.predict(X_test)
    print "predicting time:", round(time()-t1, 3), "s"

    print clf.score(X, y)

#Optimizing  kernel
kernels = ["linear", "poly", "rbf", "sigmoid"]
for i in kernels:
    print "Kernel =", i

    clf = SVC(kernel=i, C=10000)
    accuracy = cross_val_score(clf, X, y, cv=9,)
    recall = cross_val_score(clf, X, y, cv=9, scoring='recall')
    print "Accuracy =", accuracy,"%"
    print "Recall =", recall,"%"
"""
#create classifiers
nb = GaussianNB()
dt = DecisionTreeClassifier()
clf = SVC(kernel='sigmoid', C=10000)
neigh = KNeighborsClassifier(n_neighbors=2)
ab = AdaBoostClassifier()
rf = RandomForestClassifier(n_estimators=1)


#GaussianNB
nb_accuracy = (cross_val_score(nb, X, y, cv=9,))
nb_recall = (cross_val_score(nb, X, y, cv=9, scoring='recall'))
nb_precision = (cross_val_score(nb, X, y, cv=9, scoring="precision"))
nb_prediction = 0
for i in cross_val_predict(nb, X, y, cv=9):
    if i == 1:
        nb_prediction+=1
print "GaussianNB:"
print "Predicted {}/18 samples to be cancerous".format(nb_prediction)
print "Accuracy = {}%".format(nb_accuracy)
print "Recall = {}%".format(nb_recall)
print "Precision = {}%".format(nb_precision)


#SVM output
accuracy = (cross_val_score(clf, X, y, cv=9,))
recall = (cross_val_score(clf, X, y, cv=9, scoring='recall'))
precision = (cross_val_score(clf, X, y, cv=9, scoring="precision"))
prediction = 0
for i in cross_val_predict(clf, X, y, cv=9):
    if i == 1:
        prediction+=1
print ""
print "SVM:"
print "Predicted {}/18 samples to be cancerous".format(prediction)
print "Accuracy = {}%".format(accuracy)
print "Recall = {}%".format(recall)
print "Precision = {}%".format(precision)


#Decision Tree
```

```
dt_accuracy = (cross_val_score(dt, X, y, cv=9,))
dt_recall = (cross_val_score(dt, X, y, cv=9, scoring='recall'))
dt_precision = (cross_val_score(dt, X, y, cv=9, scoring="precision"))
dt_prediction = 0
for i in cross_val_predict(dt, X, y, cv=9):
    if i == 1:
        dt_prediction+=1
print ""
print "Decision Tree:"
print "Predicted {}/18 samples to be cancerous".format(dt_prediction)
print "Accuracy = {}%".format(dt_accuracy)
print "Recall = {}%".format(dt_recall)
print "Precision = {}%".format(dt_precision)

#K-Nearest Neighbours
#Optimizing
"""
algorithms = ["auto", "ball_tree", "kd_tree", "brute"]
for i in algorithms:
    print "algorithms =", i
    neigh = KNeighborsClassifier(algorithm=i)
    neigh_accuracy = cross_val_score(neigh, X, y, cv=9,)
    neigh_recall = cross_val_score(neigh, X, y, cv=9, scoring='recall')
    print "Accuracy = {}%".format(neigh_accuracy)
    print "Recall = {}%".format(neigh_recall)
"""
neigh_accuracy = (cross_val_score(neigh, X, y, cv=9))
neigh_recall = (cross_val_score(neigh, X, y, cv=9, scoring='recall'))
neigh_precision = (cross_val_score(neigh, X, y, cv=9, scoring='precision'))
neigh_prediction=0
for i in cross_val_predict(neigh, X, y, cv=9):
    if i == 1:
        neigh_prediction+=1
print ""
print "K Nearest Neighbours: "
print "Predicted {}/18 samples to be cancerous".format(neigh_prediction)
print "Accuracy = {}%".format(neigh_accuracy)
print "Recall = {}%".format(neigh_recall)
print "Precision = {}%".format(neigh_precision)

#AdaBoost
ab_accuracy = (cross_val_score(ab, X, y, cv=9))
ab_recall = (cross_val_score(ab, X, y, cv=9, scoring='recall'))
ab_precision = (cross_val_score(ab, X, y, cv=9, scoring='precision'))
ab_prediction=0
for i in cross_val_predict(rf, X, y, cv=9):
    if i == 1:
        ab_prediction+=1
print ""
print "AdaBoost: "
print "Predicted {}/18 samples to be cancerous".format(ab_prediction)
print "Accuracy = {}%".format(ab_accuracy)
print "Recall = {}%".format(ab_recall)
print "Precision = {}%".format(ab_precision)

#Random Forests
rf_accuracy = (cross_val_score(rf, X, y, cv=9))
rf_recall = (cross_val_score(rf, X, y, cv=9, scoring='recall'))
rf_precision = (cross_val_score(rf, X, y, cv=9, scoring='precision'))
rf_prediction=0
```

```
for i in cross_val_predict(rf, X, y, cv=9):
    if i == 1:
        rf_prediction+=1
print ""
print "Random Forests: "
print "Predicted {}/18 samples to be cancerous".format(rf_prediction)
print "Accuracy = {}%".format(rf_accuracy)
print "Recall = {}%".format(rf_recall)
print "Precision = {}%".format(rf_precision)
```