**Computational estimation of telomere length from routinely collected clinical NGS data.**

**Christopher Pyatt**

**A dissertation submitted to The University of Manchester for the degree of Master of Clinical Science in the Faculty of Biology Medicine and Health**

**2022**

**Student ID: 9709086**

**School of Biological Science**

Table of Contents

Table of Contents

**Abstract (TOO LONG)**

**Background**

Telomeres are the protective cap that shields chromosomes from replicative damage and gradually reduce in length over an individual’s lifetime. Prematurely shortened telomeres therefore pose of problem to affected individuals as this may leave them exposed to risk of DNA damage, cancer, or other health problems. Telomere length as a phenotypic trait may be estimated from routinely gathered NGS data as part of an NHS clinical service, aiding diagnostic services and clinical variant interpretation. At present, computational telomere length estimation tools have been tested on Whole Genome Sequencing (WGS) data due to the need for telomeric sequence reads that are not the intended targets of most capture based assays. WGS is not yet part of routine clinical service, but it may be possible to use these methods on other data. Current practice at the East Genomics Laboratory Hub (EGLH) includes Whole Exome Sequencing (WES), which may provide sufficient telomeric reads if off-target capture occurs. Here I test whether a published tool, Telomerecat, can attain acceptable results from WES and other currently collected datasets.

**Methods**

Telomere estimates were obtained using Telomerecat from an array of sequencing runs carried out as part of the routine genomic testing service at EGLH, including WES and clinical exome data. These estimates were correlated with patient age to determine whether they are likely to be accurate without independent measurement. Patients with lower than expected telomere length were investigated for the presence of pathogenic genomic variants in a panel of telomere length related genes.

**Results**

A negative correlation was observed between estimates of telomere length and patient age, as would be expected if the estimates follow previously established biological patterns. There was significant variation in the quality of estimates obtained, perhaps due to the relatively low numbers of informative reads present in the capture-based data. Findings related to the pathogenic variants in TL related genes were inconclusive, with few relevant variants found in the patient cohort.

**Conclusion**

Telomere length estimate from NGS data appears to be a promising avenue, though given the natural variation in telomere length, it may be that a larger sample size is required to be fully confident of the clinical utility.

**Declaration**

This dissertation is the student’s original work unless referenced clearly to the contrary, and no portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

**Intellectual Property Statement**

i. The author of this dissertation (including any appendices and/or schedulesto this dissertation) owns certain copyright or related rightsin it (the “Copyright”) and s/he has given The University of Manchester certain rightsto use such Copyright, including for administrative purposes.

ii. Copies of this dissertation, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright worksin the dissertation, for example graphs and tables (“Reproductions”), which may be described in this dissertation, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Furtherinformation on the conditions under which disclosure, publication and commercialisation of this dissertation, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy, inany relevant Dissertation restriction declarations deposited in the University Library, and The University Library’s regulations.

**Introduction**

**Research Question**

This project intends to answer the questions of whether telomere length measurement from NGS data is useful from a clinical perspective, whether it can be made into a validated service in a genomics laboratory setting, and whether a published computational tool, Telomerecat, can be of use in achieving this.

**Background information regarding telomeres**

Telomeres are the repetitive regions at the very ends of chromosome arms (‘telos’ being the ancient greek word for ‘end’). In humans they are made up of TTAGGG repeats approximately 5-15kb in length (1)⁠. This repeating pattern runs in the 5’ to 3’ direction away from the centromere while the returning strand reads CCCTAA 5’ to 3’. This is an important feature that will become clear when we look at how to measure the telomeres. The former strand is sometimes known as the ‘G-rich’ strand, for obvious reasons, and typically overhangs the other by up to several hundred bases. The reason for the variation is unknown but the overhanging ‘tail’ itself is a result of the end-replication problem, namely the fact that DNA can only be synthesised in a 5’ to 3’ direction (see below). The overhang is left at the 5’ end of the lagging strand after the RNA primer has fallen off. This overhang plays an important role in the formation of the t-loop structure, in which the free end of the telomere is bound within the condensed chromosome, hiding it from the DNA repair machinery. Telomeres act as a buffer to prevent replicative damage to the gene-containing regions closer to the centromere. They, along with the specialised proteins that bind to them, also prevent the nuclear machinery from recognising the ends of chromosomes as double stranded breaks and stitching them together (2)⁠. The boundary between the telomere and the rest of the chromosome is not well defined, and difficult to resolve. This is because the sequence here (the subtelomere) is made up of a combination of telomeric repeat units and other motifs, i.e. there is no sharp delineation where the repeat structure begins.

Progressive shortening of telomeres with successive DNA replication cycles has been observed since the 1970s (3)⁠ and is linked to the eventual cessation of cell division (senescence) associated with ageing. This progressive shortening is due to the cell’s DNA replication machinery being unable to replicate the sequence at the 3’ ends as a result of the obligate unidirectional progression of DNA polymerase. This is known as the ‘end replication problem’. See Figure 1, below, for a visual explanation of this. Telomeres are lengthened during development by the enzyme telomerase, but this is usually switched off in adults and so unusually short telomeres in a child may result in health defects arising from the lack of a proper buffer. Shortened telomeres are one of the triggers for cells to enter senescence, which may reduce the liklihood of cancer occurrence by limiting the number of cell divisons that can occur (i.e. the number of DNA replications and therefore the probability of an error). Reactivation of telomerase in adults may therefore increase the incidence of cancer so its inactivation in normal adults is not a bad thing (4)⁠. In fact, around 85% of cancers may escape apoptosis/senescence through reactivation of telomerase (5)⁠.

Figure 1: The final primer is synthesised around 70-100 bases from the end of the chromosome, anything further along is left out and lost.

  
Figure 2: The final primer is synthesised around 70-100 bases from the end of the chromosome, anything further along is left out and lost. This is because DNA replication on the lagging strand must begin

As a result of the processes outlined above, telomere length generally correlates negatively with age and premature shortening of telomeres may bring about age-associated, as well as more unusual, negative health effects (6)⁠. Telomere shortening may occur more quickly in times of environmental stress (7)⁠. Shortened telomeres are associated with some cancers as a result of causing genomic instability, often leading to sudden loss of telomere length and resectioning of the chromosome itself. Telomere length abnormalities are also associated with certain rare diseases such as primary immunodeficiency. Table 1 contains a list of monogenic variant information associated with telomere phenotypes. Several of these genes aencode components of the telomerase pathway, for example TERC and TERT encode the two major moities of the telomerase enzyme, allowing both the extension of the 3’ tail described previously as well as RNA binding to the chromosome. As described later, estimation of telomere lengths in patients displaying these variants may be helpful in determining pathogenicity.

Table 1 – Monogenic variants associated with telomere phenotype. Phenotypes listed in column two are simplified for brevity, see OMIM references for full information.

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | HGNC ID | Condition/Phenotype (OMIM) | Reference |
| TERC | 11727 | Aplastic anemia  Pulmonary fibrosis  Dyskeratosis congenita | (11)⁠ |
| TERT | 11730 | Dyskeratosis congenita  Leukemia  Melanoma  Pulmonary fibrosis | (12)⁠ |
| NHP2 | 14377 | Dyskeratosis congenita | (13)⁠ |
| TINF2 | 11824 | Dyskeratosis congenita  Revesz syndrome | (14)⁠ |
| NOP10 | 14378 | Dyskeratosis congenita | (15)⁠ |
| PARN | 8609 | Dyskeratosis congenita  Pulmonary fibrosis | (16)⁠ |
| ACD | 25070 | Dyskeratosis congenita | (17)⁠ |
| WRAP53 | 25522 | Dyskeratosis congenita | (18)⁠ |
| CTC1 | 26169 | Cerebroretinal microangiopathy | ⁠(8)⁠ |
| RTEL1 | 15888 | Dyskeratosis congenita  Pulmonary fibrosis | (9)⁠ |
| DKC1 | 2890 | Dyskeratosis congenita | (10)⁠ |

**Current methods of telomere length measurement**

Wet lab methods for measuring telomere length include Southern blot analysis of terminal restriction fragments (mTRF), FISH-based techniques (fluorescence in situ hybridisation), and quantitative PCR-based assays. The mTRF method measures telomere length in specific cell samples through electrophoresis and chemiluminescence, though this requires relatively high DNA yields to be successful (6,19).There are a number of FISH-based assays used in the measurement of telomere length. These measure telomere length in individual cells. As such, the requirement for intact cells makes the technique potentially expensive and difficult, especially with respect to fixation techniques (6,20,21)⁠. qPCR assays are considerably quicker and cheaper than Southern blot analysis, but may be less accurate, particularly when considering inter-assay consistency (6,22)⁠ (up to 20% between protocols (2)⁠). In addition to the techniques described above, various methods exist that may be used to measure either the total chromosome abundance or the length of a specific chromosomal section (namely long read sequencing in the latter case).

Computational NGS-based methods also exist that can piggyback on routine NGS sequencing, using off-target sequencing reads (in the case of targetted capture sequencing) that match the telomeric repeat pattern, and therefore avoid the time and expense required to order extra samples. Of course if necessary, NGS can be performed purely for the sake of this measurement. Currently, within the CUH Genomics Laboratory, NGS (Illumina) data is generated via several capture assays, including Solid Cancer, Myeloid cancer, Whole Exome, and some smaller targetted panels. The most pertinent of these to the subject of this project is the Whole Exome Sequencing (WES) as this is the larger of the capture sets being used and therefore best placed to generate informative reads for computational telomere analysis. These methods rely on estimations of telomeric content from unaligned read pairs, since their repetitive nature makes alignment to a reference genome (and therefore precise length of the repeat region) either impossible or uninformative. Of course, this does mean that estimation using these methods can only be performed in cases where there are sufficient telomeric reads, i.e. not where capture efficiency is so high that very few off-target reads are generated. In most cases these techniques have been developed with WGS data in mind, given that the aim there is for even coverage across the genome rather than just select parts. Telseq and Telomerecat are examples of computational telomere length estimation methods and are perhaps the most advanced at this time.

Telseq (1)⁠ estimates telomere length by categorising each read in a sample as either non-telomeric or telomeric and counting the latter. The total telomeric read length is then presented as a proportion of total read length. The tool categorises read pairs based on a threshold repeat number, i.e. if the threshold is n, a read is telomeric if it contains at least n TTAGGG repeats. This was validated against mTRF measurments of the same samples (see earlier Southern blot explanation) and seems to consistently report shorter measurements than that method, presumably due to the difficulty in distinguishing between telomeric and subtelomeric regions (the mTRF method in particular is known to overestimate telomere length (21)⁠). The principal limitation of this method is that the number of chromosomes must be known as it acts as a constant within the calculations. Therefore this method is not suitable for applications where aneuploidy is suspected (primarily somatic applications).

Telomerecat (23)⁠ is another tool designed to estimate telomere length from NGS data, differing from other such tools by the lack of a requirement to know the number of chromosomes prior to analysis. This has important implications where cancers are involved (whether this is the reason for analysis or not) as aneuploidy may occur in such cases and cause errors in telomere length estimations where the total telomere length is divided by an erroneous chromosome number. The authors reference Castle et al, Parker et al, & Ding et al (1,5,24)⁠ as being the first to attempt computational estimation of telomere content from NGS data, but build on this further with the lesser reliance on prior knowledge. Telomerecat avoids reliance on prior knowledge of chromosome/telomere number by assessing the ratios of telomeric versus subtelomeric sequence. The amount of sequencing data from each can be assumed to increase in proportion with one another, therefore the ratio assessment should be robust to changes in chromosome number. The tool also corrects for ‘interstitial telomeric repeats’ (occurences of the TTAGGG motif within the genome proper). A more detailed explanation of the Telomerecat tool is included in the Methods section of this thesis.

**Telomeres measurement in healthcare**

Reliably assessing telomere length is one potential way of contributing to variant interpretation, among other diagnostic activities. Previous research has shown that patients with pathogenic mutations related to short telomere syndromes are frequently within the lowest deciles (of telomere length) of measured cohorts (2)⁠. Given this association, estimation of telomere length through NGS may help to classify variants in relevant genes as pathogenic (or not, if the measurement is normal). It is worth noting that telomeric length at birth varies to some degree (5)⁠, and therefore it is important to ascertain what counts as ‘normal’, or at least normalise to the population results so that patients can be categorised according to telomere length (and therefore show whether theirs are unusually short). In this way, computational telomere length measurement becomes clinically informative as it offers the unusual opportunity to infer phenotypic information directly from NGS data, as well as correlating phenotype with variant status.

Telomerecat has been validated against the mTRF and TelSeq methods and shown to give comparable results at a population level (see Figure 1 of Farmery et al (23)⁠). However, it is unclear whether this method can give clinically actionable results for the individual patient. In order for this objective to be met, individual results will need to be normalised according to age. It would be instructive to see whether a finding of excessive telomere shortening can contribute to evidence that a VUS is pathogenic, or alternatively if a patient’s telomere length is found to be well within normal ranges, that the variant is likely benign. Previous studies have already shown that short telomere individuals tend to develop symptoms from associated mutations much earlier in life than those with less telomere attrition (2)⁠.

In addition to extending the confirmatory variant interpretation work undertaken within genomics laboratories, telomere length measurement has the potential to provide diagnostic services with independent clinical outcomes. Estimates of affected individuals are in the 10s of thousands, with such individuals being prone to the variety of premature ageing disorders already discussed in addition to being susceptible to DNA-damaging therapies that would not normally pose a serious risk (e.g. high intensity marrow transplant regimes) (2)⁠. Identification of such individuals would allow healthcare professionals to better tailor their treatment regimes to the telomeric environment of the patient. Consequently, genomic counselling of patients with a finding of shortened telomeres and relevant genomic variants would also be better informed as testing would be able to show a tangible link between phenotype and genomics. Availability of a telomere length estimate (derived from flowFISH assays rather than NGS) has already shown to be clinically useful in management of bone marrow failure patients (2)⁠, so it stands to reason that such a measurement may also be useful in other clinical scenarios, especially if easily acquired through routine NGS testing.

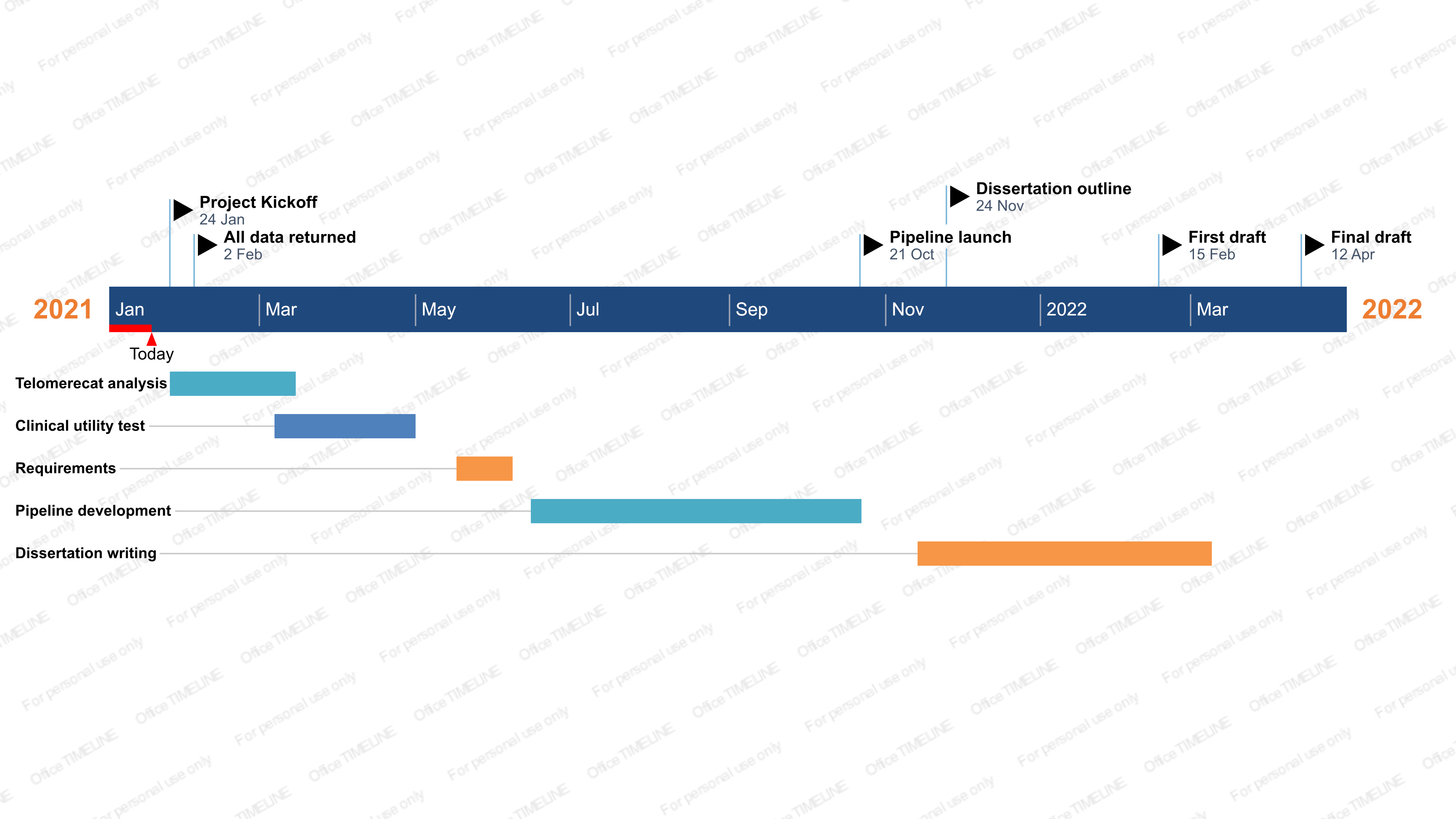
Most importantly, within the context of this project, telomere length estimations may be made without collecting additional patient samples as they will form an additional service based upon the NGS analysis already being undertaken within the diagnostic laboratory. Perhaps intuitively, the computational NGS based methods above are more effective with greater sequencing depth. This presents its own challenges in a healthcare setting, especially where the aim is to use existing data, because the various capture methods used in NGS are not usually aiming to capture telomeric sequence. For example, clinical exome sequencing and whole exome sequencing (WES) aim to be as efficient as possible in capturing only those sequences that match exonic sequence (i.e. by design these assays will generate relatively minimal quantities of intronic, intragenic, & telomeric sequences). The arrival of whole genome sequencing (WGS) in NHS genomic laboratories will of course mitigate this issue as the read pairs generated will no longer be limited to pre-determined regions of the genome. Routine administration of WGS in medical diagnostic laboratories is some time away still but the direction of travel suggests that data generated by such labs will only become more suitable for applications such as telomere length measurement. In any case, mediocre capture efficiency in the data generated in the Addenbrookes lab prior to 2021 means that, while the exome sequencing data gives fewer informative read pairs than would be obtained through WGS, the dataset available to this project may be sufficiently enriched with telomeric read pairs to enable an informative evaluation of the telomeric length estimation technique.

**Aims and Objectives**

* To assess the performance of Telomerecat at an individual level using data from clinical exome and whole exome panels.
  + Test whether patients with pathogenic variants (in TL related genes) are in the lowest deciles of telomere length.
* To develop an automated telomere length pipeline that can be deployed within the department.
  + Develop app that runs the tool within DNA-Nexus cloud workflow & produces report sent to clinical scientists.

**Project Plan**

The clinical exome data became available in early 2021 (referred to as TruSight One Extended, or TSOE, in most of this thesis after the assay kit used) andis made up of the final runs before replacement of the service with WES. Whole Exome sequencing (WES) entered routine testing around May/June 2021 and has been running approximately 24-48 samples per week since then. Both datasets will be used to assess the preformance of the Telomerecat tool on targetted capture sequencing data, as opposed to whole genome data. Below is a Gantt chart showing the main milestones for the project, along with approximate timescales for each of the component sections, as planned at the start of the project. Dates are approximate & subject to change. There is some contingency included at each stage to account for delays.



**Ethics**

Guidance regarding ethical approval has been sought from the Research Governance department at Addenbrookes Hospital (Cambridge University Hospitals). The advice was that the project falls into the category of service evaluation/improvement and as such does not require ethics approval. No special access to data is required as the data is that routinely collected within the genomics laboratory and all the patient data to be used has been consented at collection for the purposes of service improvement. The data is stored within the laboratory’s local databases and cloud-based storage assets. The cloud storage and processing is approved for clinical use and any data uploaded there is pseudonymised through sample specific identifiers that can only be matched to patients via a separate database on site.

**Risk Assessment**

The principle risk identified during earlier discussions was that of the intended dataset not materialising as its collection is out of my control. Although this has since been reduced by the fact that some of the data has been returned and is ready for analysis, the risk may be further mitigated by the availability of GEL data. This data is another source that would be appropriate for use in this study. The methods to be tested may be tested on any clinically relevant exome data so publicly available data may serve as a backup for the purposes of proving the concept. The ideal course of action would be to use our own routinely generated data as this can then act as a past cohort from which to derive allele & phenotype frequencies that inform our assessment of patient TL rankings (including in routine practice if this leads to a production application).

**Methods**

**Telomerecat v3.4.0 – Farmery et al**

This is an explanation of the methods behind the Telomerecat tool, to aid the reader’s understanding of the project. It is not intended to imply any ownership of the tool.

A WES or clinical exome (TSOE) dataset is made up of sequencing reads which are primarily mapped to the relevant targetted regions of the genome (the exons from the capture targets). However. this capture technology is imperfect and thus the dataset also contains a variable number of so-called “off-target” reads. A capture-based sequencing assay generally aims to keep the number of off-target reads to a minimum for efficiency purposes but some will always remain. At least some of these off-target reads will, by chance, originate from the telomeres and it is the ratio of these reads compared to subtelomeric reads (defined later) that allows Telomerecat to estimate the total length of the telomere.

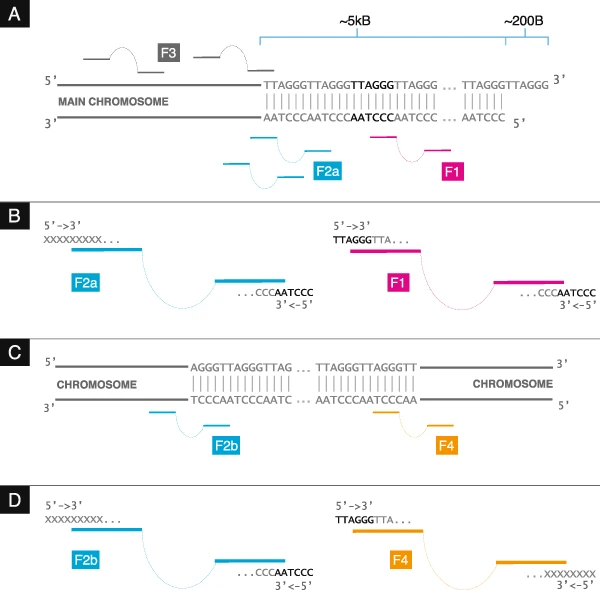
Sequencing read pairs containing the canonical telomeric repeat motif (TTAGGG) may originate from one of four locations in the genome. The first is a region fully within the telomere (i.e. the entire read is made up of the repeat sequence and its complement), the second is a region spanning the telomeric boundary (one end of the read being entirely made up of TTAGGG repeats, the other not so). The third possibility is a region within the subtelomere (a mix of TTAGGG repeats and more heterogenous sequence), and the fourth is a region containing interstitial repeats (non-telomeric TTAGGG motifs within the body of the chromosome). In this final case, the read pair will have one end made up of TTAGGG or AATCCC repeats and the other made up of heterogenous sequence. Interstitial repeats are not likely to be long enough that both ends of a read will be within the repeat, so all reads matching that pattern are put into the first (fully telomeric) category. Importantly, it is not possible to distinguish, from just the DNA sequence, between reads spanning the telomeric boundary and those spanning an interstitial repeat boundary, so an extra calculation step is performed, as described below.

The process of collecting reads containing relevant repeat sequences is error-corrected to avoid discarding true telomere read pairs that include sequencing errors. Telomere length is then estimated using an iterative model that aims to discover the true telomere length from the relative proportions of fully telomeric read pairs to subtelomeric read pairs.

The figure below shows the rationale behind the classification of read pairs. Five bins are used: read pairs with both ends containing only TTAGGG repeats are classed as F1 (fully telomeric – panel A), those with one end consisting of AATCCC (complement to TTAGGG) repeats are classed as F2 (split into F2a and F2b, see below), chromosomal read pairs (neither end fully TTAGGG or AATCCC) are F3, and pairs with one end made up of TTAGGG repeats are F4.

F4 read pairs represent pairs spanning the right hand boundary of a repeat region, as depicted in the figure (panel C). Since telomeric sequence always reads TTAGGG in a 5’ to 3’ direction away from the centromere, there is no right-hand equivalent of F4 read pairs at the other end of the chromosome. This means an F4 read must be the boundary of an interstitial repeat region.

F2 read pairs represent the pairs that span the left hand boundary and are split into F2a and F2b, the former being pairs spanning the telomeric boundary and the latter being pairs spanning the boundary of interstitial repeat regions. These are indistinguishable from sequence alone but the true F2a number can be inferred from the assumption that there must be an equal number of F2b and F4 read pairs (because an interstitial repeat region must have a start and and end, whereas a telomere only has one end connected to the chromosome).

  
Figure 3: Visual representation of the four/five types of reads assessed by Telomerecat. The key statistics relevant to telomere length estimation are the ratios between F1 reads and F2a reads. [A] Chromosome-telomere boundary. [B] Telomere boundary reads F2a and full telomere sequence F1. [C] Interstitial repeat sequence, with boundary reads either side. [D] structure of interstitial boundary reads, as distinguishable from telomeric boundary reads from panel B. Figure reproduced from the Telomerecat paper – Farmery et al 2018

Once the key statistics have been gathered, namely the F1 and F2a read counts, the tool uses an iterative simulation approach to recreate them from various possible telomere length values until the simulated values resemble the measured ones. Pseudocode for the simulation is shown below, but is explained in more detail in the paper.

Start with an arbitrary TL

Get sample insert mean and SD

Until estimated F1 and F2a match observed F1 and F2a:

Random walk simulation using arbitrary TL and insert stats

If estimate F1 is too low, increase starting TL and resimulate, & vice versa

**DNAnexus applet**

The CUH/EGLH (Cambridge University Hospitals / East Genomic Laboratory Hub) genomics laboratory currently uses cloud computing and storage infrastructure provided by DNAnexus. The service is also used by several other GLHs across the country. The DNAnexus service is a data management and analysis platform built on Amazon Web Services (AWS) and includes both a web browser interface and a command-line client for interacting with data. The platform is focussed on genomic science and provides global applications for analysis, workflow management for clinical testing, and the ability to develop custom applications. At CUH/EGLH, most NGS data is either automatically streamed directly from the sequencers to storage on DNAnexus or uploaded manually by bioinformaticians. From there the data is triaged and analysed via several assay-specific variant calling workflows, before being distilled into clinically relevant reports to be returned to scientists/clinicians. Data on DNAnexus is, from a user perspective, organised into discrete “projects”, though each data object is stored once in the platform and different projects may contain multiple references to that object. At EGLH, each project corresponds to a single sequencing run of an assay, or a single development project. There are also projects that contain only carefully managed reference data (e.g. routine production applications and reference genomes).

In order to run Telomerecat in DNAnexus, in line with current practice for NGS analysis at EGLH, it was necessary to built a DNAnexus app(let) to act as a wrapper for the application. An app(let) is made up of a wrapper shell script (bash) that contains application code and calls to external tools (including install/download commands from either local images or internet hosted repositories), as well as housekeeping code that imports input data and exports outputs to/from elsewhere on the DNAnexus platform.

A DNAnexus applet was written in order to run Telomerecat within the EGLH DNAnexus environment. The code can be seen at github.com/eastgenomics/telomerecat. The app is written to take a BAM input and output a short summary file, including a telomere length estimate. The applet takes its dependencies from docker images of the telomerecat tool and samtools. This was done in order to both maintain reproducibility and to avoid dependency on dynamic downloads. In the former case, it is an important tenet of clinical scientific practice to ensure that analysis pipelines run in exactly the same way across different sequencing runs or patient samples. This is so that the results can be shown to be produced in a comparable way and that differences in the results are due to differences in the samples not in the process used to generate the data. Any changes to analysis pipelines must normally be validated and approved before putting them into routine use, therefore it is important not to allow dynamic updates via external downloads (i.e. calling for the latest version of a tool, which may change) without deliberate review. The other reason for avoiding dynamic downloads is to ensure service continuity. The EGLH has no control over the hosting of external dependencies so it is far more robust to have a local image of a tool, which can be used within DNAnexus repeatably, than to rely on a call to an external repository that may fail. In addition, from a security perspective it is often better that app(let)s do not connect to external sites that may be compromised without our knowledge, or may allow access to the data from outside the approved web-space.

**Age correlation in exome sequencing**

*Data collection*

Five sequencing runs (a run being defined as a batch of samples that went through sequencing and analysis together) were chosen from both Twist Whole Exome (TWE) and TruSight One Expanded (TSOE) runs (10 total runs) and BAM files were transferred to a new project within the EGLH DNAnexus environment. Patient age was retrieved from the CUH Genomics LIMS (Laboratory Information Management System). No other patient data was retrieved, in accordance with data protection principles to use the minimum data necessary for the task.

Initial estimates of telomere length were generated by running the Telomerecat applet on each BAM. This gave a telomere length estimate separately for each sample. The intermediate TELBAM files were then downloaded in order to run a separate installation of Telomerecat with batch correction (error correction of F2a count via within-batch normalisation). Running this on a separate instance was necessary due to input limitations of the applet at the time, while the error correction itself was recommended practice in low coverage samples (see Telomerecat manual), which applies to this dataset due to our reliance on off-target reads. The non-corrected estimates appeared in line with the cloud-based initial runs but more consistent, so there were no concerns that this was generating markedly different results.

This generated a single file for each sequencing run, containing telomere length estimates for all samples on that run, as well as normalised informative read counts and other metadata. Estimates were generated for both the full read length (150bp) and a restricted read length (75bp). This read length trimming was performed to test whether removing the lower quality end of the reads resulted in more informative reads and therefore more reliable telomere length estimates (exploratory testing had already suggested that some samples were presenting very few informative reads and thus were not giving consistent measurements at the simulation stage). Three technical replicates were performed for each (so each run generated 6 output files) in order to check the repeatability of the measurements. These were concatenated into one excel spread pairsheet for ease of analysis. Scatter plots were generated in R-Studio, the code for which is available at github.com/chrispyatt/TelomereDissertation.

**Pathogenic variants**

A list of 136 pathogenic variants in the telomere length related genes mentioned previously was gathered from Clinvar. This list was generated by searching for all variants in those genes meeting the follwing criteria: Clinical significance = ‘Pathogenic’, Variant length = ‘Less than 51 bp’, and Method type = ‘Clinical testing’. A less stringent list (266 hits) was also generated by including non clinical testing variant reports.

This list was converted into a VCF (variant calls) by gathering the HGVS notations from column one of the Clinvar tabular output (edited manually to remove invalid characters) and then using the ensembl Variant Effect Predictor (VEP GRCh37 online version) to generate the VCF. The resulting VCF was compared to two merged VCFs of all variants found in 500 recent TWE samples, and specifically all samples used in the earlier age correlation part of this study (all sequenced as part of EGLH clinical testing) using Bedtools intersect. The process of generating these merged VCFs was conceived by Matt Garner of EGLH and is described in the file “TWE\_merge\_instructions.pdf”, available at github.com/chrispyatt/TelomereDissertation. The command used to compare the two VCFs to the pathogenic variant list is:

bedtools intersect -a TWE\_POPAF\_N500\_220202.vcf.gz -b clinvar133.vcf -header > TWE\_POPAF\_intersect.vcf

**Results**

**Age correlation of telomere estimates from exome sequencing data**

In order to check whether the telomere length estimates achieved by Telomerecat were likely to be correct and useable, they were plotted against patient birth year (as a proxy for age). Since telomere length shortens over a person’s lifespan due to replicative losses, this should yield a positive correlation if the measurement technique is working correctly (note that telomere length correlates negatively with age but we expect a positive correlation with birth year, as plotted here). This has been shown to be the case with Telomerecat estimates generated from Whole Genome sequencing data but has never been shown in other classes of NGS data.

The figures below show the correlations, first with TSOE (clinical exome) data, and then with TWE (whole exome) data. Each plot shows the mean of 3 estimates for each sample, with the variance shown by a vertical bar. In some cases an arbitrary outlier cut-off of 30,000 was imposed in order to maintain readability. Any outliers removed tended to also show a wide variance and are therefore likely to be unreliable. For each sequencing run, estimates were generated using both the whole read (150bp – shown in blue) and a limited proportion (75bp – shown in black). This was to see whether restricting the read length affected the reliability of the results. R^2 values were computed for each read length group but not separated by sex. As you can see from the figures, the estimates generated from the limited read length samples were noticeably more consistent and appear by eye to more closely follow the expected trend. It is also clear in some cases that there is a bias toward more recent samples in this dataset, presumably due to the growing use of genetic testing and the likelihood that genetic diagnoses will first become apparent during childhood.

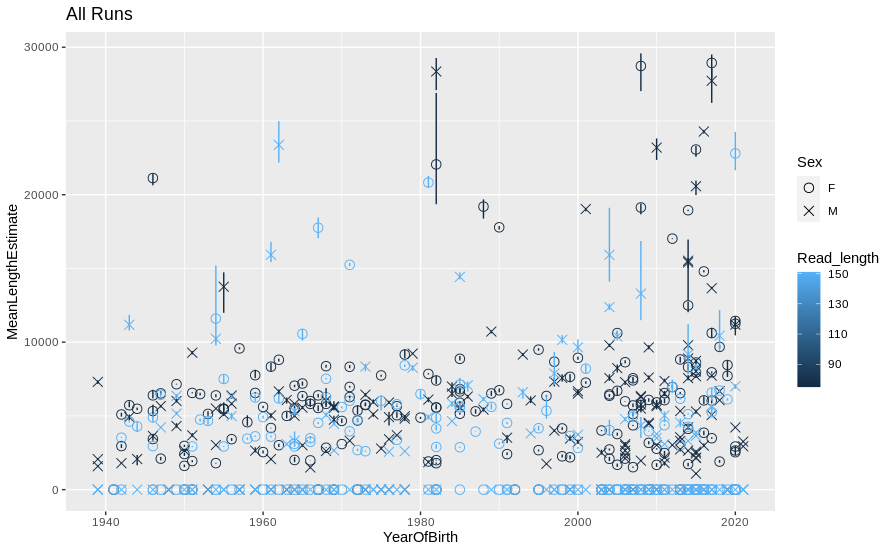
The full summary datasheet showing all information gleaned from the Telomerecat runs is contained in the spreadsheets “Summary.xlsx” and “SummarySansOutliers.xlsx”, available in the project github repository at github.com/chrispyatt/TelomereDissertation.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Moving to the TWE data, we can see that the estimates in general are more varied than in the TSOE samples. The 150bp read length group generally does not yield any estimate, shown on the plots as a measurement of zero. This is likely due to a lack of informative read pairs in what is known to be a more accurate capture dataset (remember that the capture probes for these assays are aiming for exonic sequences and so any telomeric read pairs are off-target read pairs). Telomere length estimation performance is recovered by limiting to 75bp but with markedly more variation as compared to the TSOE samples. In fact, the pattern seen in the limited estimates here resembles that of the unlimited TSOE estimates, except with more outliers. All outliers removed from the dataset were from the TWE subsets, in some cases with estimates reaching 120,000 (the expected range is 5-10,000).

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

I have also plotted all the data together (minus outliers above 30,000) in order to visualise the general trend. There may be batch effects that reduce the efficacy of this combined approach but, since the majority of estimates are within the expected range, I feel justified in concatenating the dataset here. We can again see a slight positive correlation with age, although it is not as pronounced. It is worth pointing out that this is a combined dataset (both TSOE and TWE) so there may be a mixing of the signal there (SPLIT BY ASSAY AND REMOVE ZEROS FOR CORR).



**Co-incidence of pathogenic variants in telomere related genes in patients with low telomere length estimates**

Searching 500 TWE samples for variants in the telomere length related genes mentioned in the introduction returned three samples containing the same variant (Table 2). Two of these samples were not part of the datasets used in the earlier part of this study so no telomere length estimates are available for those samples.

Searching for the same set of variants specifically in the samples used in the age correlation section (336 samples) yielded three variants, each found in a different sample (Table 3). In one case the variant is the same as the one found in the larger cohort (in the same sample, X219462, due to crossover in the sample sets). Searching for a wider set of variants (266, including non-clinically reported variants from Clinvar) did not reveal any new hits in either VCF dataset.

Table 2 – Results of searching a wider cohort of 500 TWE samples.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| #CHROM | POS | ID | REF | ALT | X218467-GM2106585-TWE-M-EGG4 | X218836-GM2111976-TWE-M-EGG4 | X219462-GM2117487-TWE-M-EGG4 |
| 17 | 8133247 | rs62637613 | A | G | 0/1:76,72:148:99:1616,0,1787 | 0/1:63,71:134:99:1473,0,1428 | 0/1:50,64:114:99:1552,0,1212 |

Table 3 – Results of searching the 336 samples (TWE and TSOE) used in the age correlation step.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| #CHROM | POS | ID | REF | ALT | X217361 | X217577 | X219462-GM2117487-TWE-M-EGG4 |
| 17 | 8132756 | rs199473680 | AG | A | 0/1:83,91:174:99:2575,0,2243 | 0/0:.:.:.:. | 0/0:.:.:.:. |
| 17 | 8133222 | . | CAG | C | 0/0:.:.:.:. | 0/1:83,85:168:99:3264,0,3308 | 0/0:.:.:.:. |
| 17 | 8133247 | rs62637613 | A | G | 0/0:.:.:.:. | 0/0:.:.:.:. | 0/1:50,64:114:99:1552,0,1212 |

Table 4 – Telomere length estimates in the three samples in which pathogenic variants were found.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Birth year | 151bp read estimate | 75bp read estimate |
| X217361 | 1956 | 6312 | 6381 |
| X217577 | 2014 | 4260 | 8307 |
| X219462 | 1997 | 0 (insufficient reads) | 3999 |

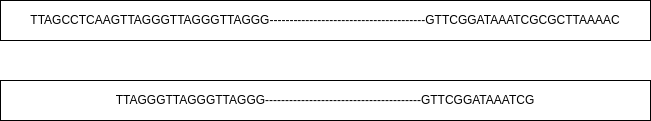
**Discussion**

The investigation followed 2 thread pairs: (1) can telomerecat produce results on our current WES data, (2) can telomerecat produce results from ANY of our data. The tool has therefore been run on a series of datasets including Whole Exome, Clinical Exome, and Whole Genome.

**Can computational telomere length estimation using Telomerecat be used with exome data?**

Here I have used the known fact that telomere length declines with age to ascertain whether the estimates generated with Telomerecat, from routinely collected NGS data within the CUH genetic laboratory, can be viewed as reliable. The results show that there is a correlation between the generated estimates and patient age, as expected, but that it is a weak one. This is likely due to a level of uncertainty in the estimates that inevitably occurs when the tool is presented with minimal data from which to judge. The simple fact is that exome based sequencing uses technology specifically designed to avoid capturing off-target sequences, and this includes telomeric sequence. In some cases the capture efficiency is low enough that sufficient informative read pairs make it through the process for an estimate of telomere length to be made. These are the cases where we can achieve success and in general, from the data collected at CUH, this seems to be the case, albeit borderline. It is clear that the more efficient recent TWE capture assay is generating fewer off-target read pairs, and thus fewer informative read pairs from a telomere measurement perspective, than previous TSOE clinical exome assays.

In both assays / datasets, it is clear that artificially limiting the read length improves telomere length estimates. In the case of the TSOE (clinical exome) data, this has the effect of tightening down the variance seen in repeated estimates of the same sample, as well as reducing the spread of estimates seen across the whole run to a range more in keeping with the expected length values (5-10,000). In the case of the TWE dataset, this has the effect of saving the results and generating estimates in cases where the unadulterated data produces such low numbers of informative read pairs that estimates are not possible. In both cases, this is because limiting the effective read length results in a greater number of read pairs being classed as subtelomeric (see diagram below) and therefore the algorithm has more data on which to base its modelling approach.

  
Figure 4: The top read will be discarded by Telomerecat as uninformative as neither end is fully telomeric. However by artificially limiting the read length, the read appears to hav eone end being entirely composed of TTAGGG repeats and can be included as an informative subtelomeric read.

The success here in rescuing the estimates for TWE is particularly good news given that this is the assay currently in use within the CUH genetics laboratory. This means that these results give a good indication that telomere length estimation could easily be incorporated into the routine reporting of the lab. In the case of both assays, the estimated telomere length is well within the range that one would expect when measuring telomere length in adults and children. This is a promising sign as it suggests that the estimates being generated are somewhere close to the real value, as opposed to just being proportionally correct within batches. This means that measurements may be compared across batches (sequencing runs) and larger scale comparisons may be possible. This goes a long way toward answering the question of whether these estimates can be used clinically, because if a large cohort database can be built up, then individual patients can be assessed against the testing population and categorised in terms of percentile (accounting for age of course). This could then contribute to variant interpretation activities as we could confidently say whether a variant in telomere length related genes correlated with shortened telomeres in that patient.

It seems obvious that, given the stated objective of the NHS to introduce Whole Genome Sequencing (WGS) into routine service in the near future, the technical feasibility of telomere length measurement from NGS data will only improve. More telomeric read pairs will improve any estimates and leave read length correction unneccessary, not to mention the fact that Telomerecat was originally demonstrated and developed on WGS data. For now though it is promising that routinely collected exome data can produce seemingly reliable technical estimates despite relatively low numbers of informative read pairs.

**Do low Telomerecat measurements correspond with pathogenic variants in telomere length related genes?**

To attempt to answer this question, a list of variants, found in any of the telomere length related genes mentioned in Table 1 (Introduction) and designated as pathogenic in Clinvar, was generated and used as a query to search 500 previous TWE samples as well as the 336 samples used in the age correlation section (with some sample crossover). This investigation revealed 3 previously seen samples that contained relevant pathogenic variants, though the results are far from conclusive. One sample, X217361, is from a male in his 60s with an estimated telomere length of around 6kb. According to Vaiserman et al, this is around the expected telomere length for the patient’s age. The second sample highlighted, X217577, is from a female patient born in 2014. Telomere length would normally be expected to be around 8-10kb at this age so the estimated length here of 4000-8000 is low. There is a relatively large difference in length estimate between the 151bp and 75bp estimates for this patient, but if we take the larger measurement (that of the 75bp) as being more believeable (due to the generally better performance of the Telomerecat algorithm at this limited read length) then this patient’s phenotype is perhaps on the low end of normal. It would be up to specialists to say whether this is likely to be clinically significant. The final patient is a TWE sample so only one estimate could be achieved due to the lower off-target read counts (the 75bp estimate). This patient is a male in his mid-twenties with an estimated telomere length of around 4kb. This is undoubtably very short for the patient’s age, given that telomeres would be expected to be at least 7-9kb at this age.

Clearly the rarity of these genomic variants makes reaching any solid conclusion in this section difficult but it would appear that at least in some cases, short telomeres do appear to coincide with the presence of pathogenic variation. Only a relatively small subsample of the WES data thus far collected at EGLH has been investigated, and routine sequencing is ongoing as part of other clinical services, so it may be useful to revisit this particular test once more patient data is available.

It is evident for the moment that telomere length estimation from NGS data is possible, even though the relative lack of informative reads in some assays results in some uncertainty. This uncertainty can be improved through artificially limiting read lengths as suggested by the Telomerecat authors, resulting in a global telomere length estimation distribution that appears in line with what is expected (i.e. the loose negative correlation with age).

**Conclusion**

Basically it kind of works but it turns out I don’t have enough data to say whether we can action anything clinically off the back of it.

The applet works and the process is technically possible. Even better when WGS turns up.

**References**

1. Ding Z, Mangino M, Aviv A, Spector T, Durbin R. Estimating telomere length from whole genome sequence data. Nucleic Acids Res [Internet]. 2014 [cited 2021 Jan 14];42(9):e75. Available from: /pmc/articles/PMC4027178/?report=abstract

2. Alder JK, Hanumanthu VS, Strong MA, DeZern AE, Stanley SE, Takemoto CM, et al. Diagnostic utility of telomere length testing in a hospital-based setting. Proc Natl Acad Sci U S A [Internet]. 2018 Mar 6 [cited 2021 Jan 16];115(10):E2358–65. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1720427115

3. Olovnikov AM. Telomeres, telomerase, and aging: Origin of the theory. Exp Gerontol. 1996 Jul 1;31(4):443–8.

4. Weinstein BS, Ciszek D. The reserve-capacity hypothesis: Evolutionary origins and modern implications of the trade-off between tumor-suppression and tissue-repair. Exp Gerontol [Internet]. 2002 [cited 2021 Jan 14];37(5):615–27. Available from: https://pubmed.ncbi.nlm.nih.gov/11909679/

5. Parker M, Chen X, Bahrami A, Dalton J, Rusch M, Wu G, et al. Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data. Genome Biol [Internet]. 2012 Dec 11 [cited 2021 Jan 15];13(12):R113. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-12-r113

6. Kimura M, Stone RC, Hunt SC, Skurnick J, Lu X, Cao X, et al. Measurement of telomere length by the southern blot analysis of terminal restriction fragment lengths. Nat Protoc [Internet]. 2010 Sep 2 [cited 2021 Jan 15];5(9):1596–607. Available from: https://pubmed.ncbi.nlm.nih.gov/21085125/

7. van Lieshout SHJ, Sparks AM, Bretman A, Newman C, Buesching CD, Burke T, et al. Estimation of environmental, genetic and parental age at conception effects on telomere length in a wild mammal. J Evol Biol [Internet]. 2020 [cited 2021 Jan 11]; Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/jeb.13728

8. OMIM Entry - \* 613129 - CONSERVED TELOMERE MAINTENANCE COMPONENT 1; CTC1 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/613129?search=ctc1&highlight=ctc1

9. OMIM Entry - \* 608833 - REGULATOR OF TELOMERE ELONGATION HELICASE 1; RTEL1 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/608833?search=rtel1&highlight=rtel1

10. OMIM Entry - \* 300126 - DYSKERIN; DKC1 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/300126?search=dkc1&highlight=dkc1

11. OMIM Entry - \* 602322 - TELOMERASE RNA COMPONENT; TERC [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/602322?search=terc&highlight=terc

12. OMIM Entry - \* 187270 - TELOMERASE REVERSE TRANSCRIPTASE; TERT [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/187270?search=tert&highlight=tert

13. OMIM Entry - \* 606470 - NUCLEOLAR PROTEIN FAMILY A, MEMBER 2; NOLA2 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/606470?search=nhp2&highlight=nhp2

14. OMIM Entry - \* 604319 - TRF1-INTERACTING NUCLEAR FACTOR 2; TINF2 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/604319?search=tinf2&highlight=tinf2

15. OMIM Entry - \* 606471 - NUCLEOLAR PROTEIN FAMILY A, MEMBER 3; NOLA3 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/606471?search=nop10&highlight=nop10

16. OMIM Entry - \* 604212 - POLYADENYLATE-SPECIFIC RIBONUCLEASE; PARN [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/604212?search=parn&highlight=parn

17. OMIM Entry - \* 609377 - ACD SHELTERIN COMPLEX SUBUNIT AND TELOMERASE RECRUITMENT FACTOR; ACD [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/609377?search=acd&highlight=acd

18. OMIM Entry - \* 612661 - WD REPEAT-CONTAINING PROTEIN ANTISENSE TO TP53; WRAP53 [Internet]. [cited 2021 Jan 16]. Available from: https://www.omim.org/entry/612661?search=wrap53&highlight=wrap53

19. Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature [Internet]. 1990 [cited 2021 Jan 15];345(6274):458–60. Available from: https://pubmed.ncbi.nlm.nih.gov/2342578/

20. Martens UM, Zijlmans JMJM, Poon SSS, Dragowska W, Yui J, Chavez EA, et al. Short telomeres on human chromosome 17p. Nat Genet [Internet]. 1998 Jan [cited 2021 Jan 15];18(1):76–80. Available from: https://www.nature.com/articles/ng0198-76

21. Baerlocher GM, Vulto I, de Jong G, Lansdorp PM. Flow cytometry and FISH to measure the average length of telomeres (flow FISH). Nat Protoc [Internet]. 2006 Dec 21 [cited 2021 Jan 15];1(5):2365–76. Available from: https://www.nature.com/articles/nprot.2006.263

22. Cawthon RM. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. Nucleic Acids Res [Internet]. 2009 [cited 2021 Jan 15];37(3). Available from: https://pubmed.ncbi.nlm.nih.gov/19129229/

23. Farmery JHR, Smith ML, Nihr &, Diseases B-R, Lynch AG. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. [cited 2021 Jan 11]; Available from: www.nature.com/scientificreports

24. Castle JC, Biery M, Bouzek H, Xie T, Chen R, Misura K, et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics [Internet]. 2010 Apr 16 [cited 2021 Jan 15];11(1). Available from: https://pubmed.ncbi.nlm.nih.gov/20398377/