Polygenic modifiers impact penetrance and expressivity in telomere biology disorders

Michael Poeschla[1,2,3,4,5], Uma P. Arora[1,2,3,4], Amanda Walne[6], Lisa J. McReynolds[7], Marena R. Niewisch[7,8], Neelam Giri[7], Logan Zeigler[7,9], Alexander Gusev[4,10], Mitchell J. Machiela[7], Hemanth Tummala[6], Sharon A. Savage[7], Vijay G. Sankaran[1,2,3,4,11]

[1]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA.

[2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA.

[3]Howard Hughes Medical Institute, Boston, MA 02115, USA.

[4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

[5]Harvard/MIT MD-PhD Program, Program in Biomedical Informatics, Boston, MA 02115, USA.

[6]Genomics and Child Health, Blizard Institute, Queen Mary University of London, Newark Street, London E1 2AT, UK.

[7]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20850, USA.

[8]Department of Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany

[9]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, NCI, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Bethesda, MD 20892, USA.

[10]Division of Population Sciences, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA.

[11]Corresponding author

Address correspondence to: Vijay G. Sankaran, Division of Hematology/Oncology, Boston Children's Hospital, 1 Blackfan St, Boston, USA, 02115. Phone: 617-355-6000. Email: sankaran@broadinstitute.org

Conflict-of-interest statement

V.G.S. is an advisor to Ensoma, Cellarity, and Beam Therapeutics, unrelated to this work. There are no other relevant declarations to report.

Abstract

BACKGROUND. Telomere biology disorders (TBDs) exhibit incomplete penetrance and variable expressivity, even among individuals harboring the same pathogenic variant. We assessed whether common genetic variants associated with telomere length combine with large-effect variants to impact penetrance and expressivity in TBDs.

METHODS. We constructed polygenic scores (PGS) for telomere length in the UK Biobank to quantify common variant burden, and assessed the PGS distribution across patient cohorts and biobanks to determine whether individuals with severe TBD presentations have increased polygenic burden causing short telomeres. We also characterized rare TBD variant carriers in the UK Biobank.

RESULTS. Individuals with TBDs in cohorts enriched for severe pediatric presentations have polygenic scores predictive of short telomeres. In the UK Biobank, we identify carriers of pathogenic TBD variants who are enriched for adult-onset manifestations of TBDs. Unlike individuals in disease cohorts, the PGS of adult carriers do not show a common variant burden for shorter telomeres, consistent with the absence of childhood-onset disease. Notably, TBD variant carriers are enriched for idiopathic pulmonary fibrosis diagnoses, and telomere length PGS stratifies pulmonary fibrosis risk. Finally, common variants affecting telomere length were enriched in enhancers regulating known TBD genes.

CONCLUSION. Common genetic variants combine with large-effect causal variants to impact clinical manifestations in rare TBDs. These findings offer a framework for understanding phenotypic variability in other presumed monogenic disorders.

FUNDING. This work was supported by National Institutes of Health grants R01DK103794, R01HL146500, R01CA265726, R01CA292941, and the Howard Hughes Medical Institute.

Introduction

Variable expressivity and incomplete penetrance – related phenomena where individuals with identical genetic variants display variable symptoms or no clinical symptoms at all – remain poorly understood. Deeper insights into the mechanisms underlying expressivity and penetrance are critical for accurate genetic prognostication in clinical medicine (1, 2). While a variety of factors could contribute, one proposed mechanism is common genetic variation modifying the effect of high-impact rare alleles. Supporting this, in complex diseases, polygenic variation has been shown to modify the phenotypic expression of high-impact variants (3–11). Additionally, in presumed monogenic disease, individual loci have been identified which modify risk (12–14). However, in extremely rare monogenic diseases, where pathogenic variants might be assumed to overwhelm any impact of common variation (15), the extent to which genome-wide polygenic variation affects penetrance and expressivity has not been characterized. This in part stems from a lack of quantifiable phenotypic variability in rare disease cohorts, as well as from a paucity of individuals harboring these rare and deleterious alleles in population biobanks, making it incredibly challenging to study the role of common variation in such conditions.

To assess the potential impact of polygenic variation on penetrance and expressivity in rare monogenic disease, we focused on dyskeratosis congenita and related telomere biology disorders (TBDs), which exhibit remarkable clinical heterogeneity despite their presumed monogenic nature. The TBDs are extremely rare disorders - dyskeratosis congenita, the archetypal TBD, has an estimated prevalence of approximately one in one million in the general population - driven by pathogenic germline variants in genes that regulate telomere length, maintenance, structure, and function (16–63). Variants in these genes display striking variable expressivity and incomplete penetrance related to symptom severity, age of onset, and organ involvement. For example, some carriers of known pathogenic variants in TBD-associated genes present with severe childhood-onset bone marrow failure, while others present in adulthood with idiopathic pulmonary fibrosis or liver disease, and others may have no associated clinical presentation at all (51, 64–67). Genetic anticipation, where phenotypes manifest earlier or become more severe in successive generations, has been suggested to play a role in phenotypic variability in families with TBDs, as has the affected gene and mode of inheritance, but these factors cannot explain all the heterogeneity present (16, 68–70).

As impaired regulation of telomere maintenance is causal for the TBDs and the underlying biological mechanisms are relatively well understood, this group of diseases presents a unique opportunity to test the idea of polygenic modification affecting penetrance and expressivity (16, 17, 71, 72). Rare variants in genes that cause TBDs have been identified, and conversely, common variants that collectively underlie a considerable fraction of population-level variation in telomere length have been identified in diverse cohorts (73–76). This makes it possible to use telomere length measured in large population cohorts as a common analogous trait to the biology underlying TBDs, providing a unique opportunity to study the impact of common variants in the extremely rare TBDs. We hypothesized that common genome-wide genetic polymorphisms, which contribute to the inter-individual variation in telomere length in the general population, may combine with large-effect monogenic TBD-causal genetic variants to impact variable expressivity and incomplete penetrance in TBDs.

Here, we show that common genetic variation across the entire genome contributes to penetrance and expressivity in telomere biology disorders, both in TBD cohorts enriched for individuals presenting with severe childhood-onset bone marrow failure or other hematologic manifestations, as well as in adult biobank cohorts (**Figure 1**). We further demonstrate that polygenic variation can contribute to variable disease manifestations within a single family with a shared high-impact variant, and that common and rare variation converge on the same biological mechanisms implicated in telomere maintenance. These results provide a framework for exploring the effects of polygenic variation on penetrance and expressivity in rare diseases.

Results

Common genetic variation associated with telomere length contributes to TBD variant penetrance and expressivity in inherited bone marrow failure syndrome cohorts

To assess whether common genetic variation impacts penetrance and expressivity in TBDs, we developed polygenic scores (PGS) using genome-wide common single-nucleotide polymorphisms associated with telomere length (73, 74, 77). For a given individual, these scores provide an estimate of the combined effect of common genetic variants across the genome that increase or decrease telomere length (78, 79). While these common polymorphisms underlie subtle variation in telomere length in the healthy population, we reasoned that the PGS might have a more profound impact in the context of high-impact monogenic alleles that are causal for the TBDs. We therefore applied the most predictive PGS as determined using PRCise-2 (Table S1) to the National Cancer Institute (NCI) longitudinal cohort of individuals with inherited bone marrow failure syndromes, including those with TBDs (ClinicalTrials.gov Identifier: NCT00027274). The NCI cohort included 92 patients with dyskeratosis congenita and related TBDs and is significantly enriched for individuals presenting with bone marrow failure or other severe phenotypes (80).

We reasoned that if both monogenic germline mutations and polygenic predisposition to short telomeres contribute to the clinical severity of TBDs and the likelihood of early-onset manifestations, the distribution of genetically predicted telomere length in this clinically ascertained cohort would be shifted towards shorter telomeres compared to the population average (Figure 2B). Consistent with this, individuals with TBDs enriched for early-onset bone marrow failure phenotypes (n = 92) have a median polygenic score 0.44 standard deviations (SDs) shorter than the UK Biobank (p=1.04E-4), and 0.37 SDs shorter than the external All of Us cohort (p=5.82E-4) (Figure 2C & S1A). To estimate the effect size of the polygenic contribution, we binned the PGS distribution into quintiles based on the UK Biobank population distribution and calculated odds ratios with the number of TBD patients in each quintile representing cases and with UK Biobank participants as controls. Individuals in the lowest quintile of genetically predicted telomere length had approximately three-fold odds of being a TBD case compared to those in the highest quintile (Figure 2D). We validated these findings in a separate cohort with 190 TBD patients from the Dyskeratosis Congenita Registry (DCR) at Queen Mary University of London (81). The DCR cohort has a broader referral base and is less enriched for reported severe phenotypes compared to the NCI cohort (80, 81); consistent with this, we observe a slightly attenuated, but consistent effect compared to our original NCI discovery cohort (median difference -0.20 SDs, p = 0.009) (Figure S1B). A combined analysis across both TBD cohorts further demonstrated a consistent and strongly significant association between an individual's PGS and their odds of having a TBD (median difference -0.28 SDs, p = 1.18E-5) (Figure 2E and S1C).

Interestingly, in both the NCI and DCR cohorts, a subset of patients harbor no known causal high-impact TBD mutation. These patients may have lower-effect-size mutations that past TBD gene discovery efforts have been unable to detect. Under a simple liability-threshold model in which rare large-effect variants, common small-effect variants, and environmental effects combine to drive disease risk, patients with no identified large-effect variant are expected to have a more significant polygenic contribution to their disease risk on average (82, 83). To test this, we separated patients with and without a known TBD variant. While both patients with and without a known variant have a significantly shifted PGS predictive of short telomeres compared to the population average, those with no known variant had an increased PGS burden for shorter telomeres compared to individuals with a known variant (Figure S1D), though the difference was not statistically significant. While underpowered, this analysis hints at a model in which polygenic variation might contribute to TBD risk and variant penetrance and expressivity.

We considered the possibility that population stratification and other demographic factors contributing to differences in the PGS across populations could underlie our observations (84–87). We used Kinship-based INference for GWAS (KING) and 1000 Genomes reference data to infer ancestry in our patient cohorts and the UK Biobank (88, 89). The vast majority of individuals in each of the NCI, DCR, and UK Biobank cohorts were of predominantly European ancestry, followed by Admixed American ancestry (Figure S1E). In conducting the GWAS and constructing the polygenic scores, we restricted our analyses to the European ancestry subset in the UK Biobank to minimize the effects of population stratification and maximize predictive accuracy in the TBD patients (Methods). We closely examined our results to determine if ancestry had a significant effect. We found that the PGS distribution did not differ between European and non-European individuals in the patient cohorts (p = 0.87) (Figure S1F). Furthermore, the PGS in the cohorts was not associated with the ancestry principal components inferred by KING (Figure S1G) (84–87). As a negative control, we also assessed cases of non-TBD inherited bone marrow failure syndrome cases that included individuals diagnosed with Diamond-Blackfan Anemia, Fanconi Anemia, and Shwachman-Diamond Syndrome, and that came from the same NCI bone marrow failure syndrome cohort as the TBD patients. For these conditions, telomere length does not drive the disease process, and non-telomere related gene mutations are implicated (71, 72, 90–92). As expected, the polygenically predicted telomere length for these patients is indistinguishable from the UK Biobank population average (p = 0.51) (Figure 2F), and the telomere length PGS for the TBD patients was 0.49 standard deviations shifted towards shorter predicted telomeres compared to non-TBD inherited bone marrow failure syndrome patients (p=3.99E-4), indicating that the observed phenomenon is telomere disease-specific and is unlikely to be due to population stratification (Figure S1H).

Polygenic variation associated with telomere length impacts TBD penetrance and expressivity in population biobanks

Taken together, our analyses of the NCI and DCR cohorts indicate that common genetic variants associated with short telomere length contribute to ascertainment as a TBD case in disease cohorts enriched for individuals with childhood-onset bone marrow failure. We reasoned that the reverse should also be true: TBD causal variants should be present in adult population biobanks, and adults with a pathogenic variant who avoided the severe childhood-onset manifestations of TBDs should not have polygenically predicted short telomere length (Figure 2A). To test this, we examined the UK Biobank for carriers of variants in the genes known to cause TBDs (Methods).

To be maximally comprehensive while maintaining stringency, we defined multiple variant sets, given that each variant annotation approach has limitations for predicting true pathogenicity (93). First, we included variants annotated in ClinVar as causing dyskeratosis congenita or a related TBD with high confidence, hereafter referred to as “ClinVar Pathogenic.” Applying quality-control, dominance, and ancestry filters, we identified 213 variant carriers (Figure 1B). Next, we defined a more restrictive subset of ClinVar variants including only carriers of variants specifically annotated to cause childhood-onset TBDs in a dominant manner and male carriers of *DKC1*, resulting in 22 variant carriers, referred to as “ClinVar Dominant-Acting”. Finally, to be maximally inclusive of potentially pathogenic variants which may not have been annotated in ClinVar, we defined a set of predicted pathogenic rare coding variants in TBD genes using a consensus of Ensembl Variant Effect Predictor, LOFTEE, and AlphaMissense annotations, resulting in 1666 carriers in the UK Biobank (“Consensus Predicted Pathogenic”) (Methods) (94–96). Genes which harbor TBD-causal mutations are often classified based on mode of inheritance (16, 17, 97). For all of these analyses, we excluded genes which cause TBDs in an exclusively autosomal recessive manner, given the challenges present in determining the phase of mutations (Methods). Supporting the pathogenicity of the variants in the associated sets, individuals in each group in the UK Biobank had shorter measured TL compared to non-carriers. Furthermore, the magnitude of effect was concordant with the expected order of pathogenicity, with the most inclusive Consensus Predicted Pathogenic cohort associated with the smallest average decrease in TL (0.31 SDs), the more restrictive ClinVar Pathogenic cohort associated with a 0.85 SD decrease in mean TL, and the most restrictive ClinVar Dominant-Acting cohort showing the largest average decrease in TL (1.15 SDs) (Figure 3A).

Despite having short telomeres, all three variant carrier cohorts had a population-normal polygenic contribution to telomere length on average compared to non-carriers of TBD variants in the UK Biobank (Figure 3B), and significantly longer predicted telomere length than the TBD cohorts (Figure 3C). Interestingly, the 22 carriers of the most severe mutations ("ClinVar Dominant-Acting") appeared to have a right-shifted polygenic score distribution relative to non-carriers, suggestive of a protective effect that could help explain why carriers of these large-effect variants escaped early-life manifestations of disease, but this difference was not statistically significant given the small sample size (Figure 3B). These findings complement our results in patient cohorts, demonstrating that individuals with TBD mutations who do not have early clinical presentations have a relatively decreased common variant burden for short telomeres in comparison with childhood-onset disease cohorts. Together, these findings support the idea that the risk of childhood-onset severe TBD manifestations due to large-effect causal TBD gene variants can be modified by common variants that impact telomere length.

We then assessed whether these pathogenic variant carriers were enriched for childhood or adult-onset TBD phenotypes relative to non-carriers in the UK Biobank. Importantly, examining childhood-onset TBD phenotypes, we found no evidence for increased risk of bone marrow failure or altered blood counts in these TBD variant carriers (Figure 3D and 3E). We then examined idiopathic pulmonary fibrosis, an adult TBD manifestation (97, 98). We found that being a carrier of a pathogenic variant was associated with greatly increased odds of presenting with idiopathic pulmonary fibrosis (Figure 3D). We wondered whether common variation also affects the penetrance of these adult-onset manifestations of TBDs. Strikingly, we found that both within variant carriers and in non-carriers, telomere length PGS stratifies risk of idiopathic pulmonary fibrosis (Figure 3F and S2D, combined VEP and ClinVar carrier analysis and separated, respectively).

We sought to quantify the effects of PGS in both rare TBD variant carriers and non-carriers and also asked whether there was evidence for a non-additive interaction between polygenic effects on telomere length and pathogenic variants. We regressed idiopathic pulmonary fibrosis disease status on PGS, variant carrier status, and an interaction term between the two, while controlling for age, sex, and the first 4 ancestry principal components (Table S13). Including both carriers and non-carriers, a one standard deviation decrease in PGS (predicting shorter telomere length) was associated with an odds ratio of 1.22 for idiopathic pulmonary fibrosis (p = 9.04E-27). The included interaction term between rare variant status and PGS was not statistically significant and did not affect the coefficient estimates (Table S14). Restricting to only rare variant carriers, this association was consistent, but not statistically significant (though possibly limited in statistical power due to small sample size), with a one-unit decrease in PGS associated with an odds ratio of 1.31 for having idiopathic pulmonary fibrosis (p = 0.108) (Table S15). We confirmed through a mediation analysis that the effect of telomere length PGS on pulmonary fibrosis risk is mediated through telomere length (Table S16 and Supplemental Note) (99). These results support a model in which small-effect polygenic variants and pathogenic large-effect variants independently contribute to adult TBD manifestations by impacting telomere length and maintenance, although this analysis may be limited by power (Methods).

In summary, we find that carriers of TBD-causing variants in the UK Biobank, in contrast to TBD cohorts, have a population-normal PGS and no enrichment for bone marrow failure, but do have increased risk of idiopathic pulmonary fibrosis, a common adult manifestation of TBDs. While the UK Biobank variant carriers are not enriched for PGS associated with short telomere length overall, common genetic influences captured by the PGS do impact the penetrance of TBD mutations associated with idiopathic pulmonary fibrosis. Collectively, these findings demonstrate that both pathogenic mutations and common genetic variation associated with telomere length combine to impact penetrance and expressivity.

Within-family polygenic effects on disease risk

Having observed a significant effect of common polygenic variation associated with telomere length on clinical manifestations of TBDs in large disease cohorts and population biobanks, we wondered whether polygenic variation also affects penetrance and expressivity within a single family with a shared causal variant. To explore this question, we analyzed a large kindred with multiple dyskeratosis congenita cases and a heterozygous pathogenic variant in the *TERT* gene (ClinicalTrials.gov Identifier: NCT00027274). Of the 22 family members for whom genotype data was available, there were 12 carriers of the pathogenic *TERT* variant, three of whom had clinically diagnosed TBDs (Figure 4A). We restricted all analyses to the 12 *TERT* mutation carriers, comparing the three clinically affected family members to the 9 unaffected family members.

We reasoned that the strict clumping and pruning approach that was optimal to construct polygenic scores in the UK Biobank may not be best-suited to detect relatively subtle within-family common genetic variation, given significant shared variation among the family members. Therefore, alongside the PGS used throughout the study, we constructed multiple other polygenic predictors including more variants, reasoning that this would be more likely to pick up any subtle differences that exist within a family. As an orthogonal approach, we also constructed a polygenic score including all conditionally genome-wide significant SNP signals (100), with the idea that this would enable detection of multiple independent effects on different haplotypes which could be segregating within this family (100). We accounted for family structure using a linear mixed model with kinship as a random effect. Remarkably, across all tested approaches, the clinically affected family members had a more negative PGS on average than the unaffected variant-carrying family members, indicating a greater burden of polygenic variation associated with shorter telomeres (Figures 4B, 4C and S3B, S3C). We speculate that the affected family members have disease not because they have a shifted PGS causing a somewhat shorter mean telomere length, but rather that the PGS hints that these patients present with disease at least in part because of a decreased general ability to protect and repair telomeres, on the background of a major perturbation caused by the rare disease-driving mutation.

Thus, even in the relatively controlled setting of a family with a shared causal variant and overall similar genetic background, our results suggest that random segregation of common variants that alter disease biology might contribute to variable expressivity and penetrance. These results provide a framework for larger studies of trios and families to further elucidate how common genetic variation may help explain why some family members with a disease-causing variant have severe symptoms, while others remain clinically unaffected.

Convergence of common and rare genetic variation in telomere biology disorders

A key question regarding polygenic modifiers of disease is whether common and rare variation converges on the same genes and biological pathways (101). In autism spectrum disorder, examples of convergence of common and rare variation at the same loci have been observed (102). Similarly, in Hirschsprung's disease and craniosynostosis, shared signaling and regulatory pathways appear impacted by rare and common risk variants (6, 9, 12). In contrast, in sickle cell disease and beta-thalassemia, common genetic variation largely impacts disease expressivity through modulation of fetal hemoglobin gene transcriptional regulation, a mechanism distinct from the primary disease-causing mutations that alter adult hemoglobin (10, 103). Having shown that common polygenic variation affects penetrance and expressivity in telomere biology disorders, we asked whether these variants act upon the same or different genes as the causal high-impact variants underlying TBDs.

In the TBDs, causal variants affect genes regulating telomere length, maintenance, and function (17, 97). Using multiple gene prioritization approaches (Methods), we found that common variation associated with telomere length and TBD expressivity implicates genes that strongly overlap the set of genes implicated as high-impact monogenic variants in TBDs (p = 5.41E-17) (Figure 5, S4A). The polygenic variants are primarily noncoding, with >97% of credible set variants mapping to introns or intergenic regions (Figure S4B). These variants show a striking enrichment for enhancers in CD34$^+$ hematopoietic stem and progenitor cells, possibly explaining the association we observe in bone marrow failure and likely underlying impacts in these progenitors for all blood and immune cells (Figure S4C). Collectively, these findings suggest a model in which common, noncoding variation converges upon the same genes implicated by high-effect Mendelian coding mutations.

Discussion

Variable expressivity and incomplete penetrance are important challenges for rare, monogenic disorders (1). Because telomere length is a measurable population-level trait, the TBDs provide a unique opportunity to test the contribution of common genome-wide variation to phenotype expressivity in a rare, presumed monogenic disease. By utilizing both disease cohorts enriched for severe, childhood-onset TBD phenotypes as well as population biobanks that are depleted of individuals with severe pediatric disease (104), we show that even in a rare-disease setting where pathogenic variants are expected to be highly penetrant, common variation plays an important role in determining phenotype expression.

First, we show that common variation affecting telomere length influences the severity and clinical presentation of telomere biology disorders. We find that individuals in disease cohorts enriched for cases with severe childhood-onset TBD manifestations have a left-shifted PGS compared to the general population, indicating that polygenic variants associated with telomere length contribute to disease liability. Importantly, there is phenotypic heterogeneity in these cohorts, and we might underestimate the impact of polygenic variation as a result. Next, we show that carriers of these same variants in the UK Biobank, a population depleted of severe childhood disease, have a population-normal PGS. Consistent with this, variant carriers in the UK Biobank do not have enrichment for severe childhood-onset TBD manifestations, but are enriched for the adult-onset TBD presentation of idiopathic pulmonary fibrosis. We also show that telomere length PGSs and rare variants independently contribute to the likelihood of having idiopathic pulmonary fibrosis in the UK Biobank. We further suggest that within a single family sharing a rare disease-causing variant, common genetic variation might affect an individual's likelihood of developing the disease. Finally, we show that common, noncoding variants and rare, highly penetrant coding variants converge on genes that regulate telomere length and maintain telomere homeostasis.

Our work builds upon previous studies which have shown that polygenic variation interacts with high-impact variants in relatively more common disease settings (3, 4, 7, 8). Here, we have identified an example of polygenic modifiers impacting an extremely rare disease by affecting the same genes and biological mechanisms as high-impact rare variants. Future work will likely uncover other examples, as exemplified through studies of disease modifiers such as fetal hemoglobin, where modifiers affect phenotypes through distinct biological mechanisms (10, 103). Increasingly large population sequencing studies have identified many individual rare and common variant associations with traits (93). However, for many phenotypes, it is still challenging to explain why some variant carriers will present with severe disease, while others will have more moderate presentations, or not have any clinical manifestations at all (1). Our findings lay the groundwork for future work to precisely quantify the polygenic contribution to penetrance and expressivity across a range of common and rare diseases. For the TBDs specifically, penetrance and expressivity have been discussed as important clinical challenges, but most prior work has focused on the importance of high-impact rare coding variants (17, 97). Our work establishes common variation as a contributor to this phenomenon and lays the groundwork for detailed clinical studies delineating the prognostic and therapeutic value of common modifiers of telomere length for TBDs.

Limitations

Our analyses are based on average telomere length in white blood cells, which is only a correlated proxy for the length of the shortest telomere, which is likely the more biologically meaningful measurement in telomere homeostasis and the TBDs (105–107). Furthermore, telomere length is correlated yet variable across cell and tissue types, and TBDs are a multisystem disease. While mean leukocyte TL is the best available metric to assess common genome-wide variation impacting telomere homeostasis, it provides an imperfect estimate of the common genetic variant effects which contribute to penetrance and expressivity in the TBDs, and better telomere measurements across multiple tissues will enable more precise future population-level analyses (108, 109) Moreover, both disease cohorts and population biobanks have strengths and weaknesses for assessing variant penetrance and expressivity. Disease cohorts may overestimate variant pathogenicity, and adult biobanks are likely to underestimate variant impact (110). Different adult biobanks have variable ascertainment characteristics, which can affect estimates of variant expressivity (104, 110–112). For this reason, in this study we utilize both disease cohorts and include both the UK Biobank and All of Us cohorts, which have different recruitment mechanisms and demographic characteristics, enabling complementary insights.

Finally, TBDs are heterogeneous; mode of inheritance and effects of specific causal variants have been shown to be associated with variation in outcome (16). In this study, we grouped different genotypes together to power analyses of the effects of common genetic variation, likely obscuring some effect heterogeneity. Other factors may contribute to variable expressivity, including genetic anticipation, environmental influences, and somatic genetic rescue (17). Future, larger studies will be necessary to tease out the effects of variant- and gene-specific interactions and other mechanisms underlying variable expressivity and incomplete penetrance.

METHODS

Further information is available in Supplemental Methods.

Sex as a biological variable

For all analyses, individuals of both male and female sex were included. For genome-wide association studies and PGS construction, sex was included as a covariate. For statistical testing, sex was included as a covariate to assess the significance of biological sex for effects studied, as detailed in the relevant Methods sections below.

Study participant details

UK Biobank

The UK Biobank is a large prospective cohort with extensive phenotype and molecular data available, including whole-genome sequencing (113, 114). The UK Biobank was accessed under application number 31063. Cohort analysis details are under METHOD DETAILS - *UK Biobank*.

All of Us

All of Us is a longitudinal cohort that currently contains short-read whole genome sequencing data and phenotypic data from 245,394 participants. Whole genome sequencing was performed on participants to a mean depth of 30x. Informed consent for all participants is conducted in person or through an eConsent platform that includes primary consent, HIPAA Authorization for Research use of EHRs and other external health data, and Consent for Return of Genomic Results. The protocol was reviewed by the Institutional Review Board (IRB) of the All of Us Research Program. The All of Us IRB follows the regulations and guidance of the NIH Office for Human Research Protections for all studies, ensuring that the rights and welfare of research participants are overseen and protected uniformly. Cohort analysis details are under METHOD DETAILS - *All of Us*.

National Cancer Institute Inherited Bone Marrow Failure Syndrome samples

Processed genotype array files were obtained as part of the NCI Inherited Bone Marrow Failure Study (ClinicalTrials.gov Identifier: NCT00027274). Genotype data was available for 92 dyskeratosis congenita cases, 49 Diamond-Blackfan Anemia cases, 45 Fanconi Anemia cases, and 19 Shwachman-Diamond Syndrome cases. Genotype data was also available for 22 members of a family with a shared *TERT* mutation that included 4 dyskeratosis congenita cases. Genotype data was also available for 120 unaffected family members of dyskeratosis congenita cases in the cohort. Deleterious variants were originally identified in these patients using TBD gene sequencing panels.

Dyskeratosis Congenita Registry Queen Mary University of London samples

Two sample sources were used as part of the DCR cohort. The first included 49 samples of unknown mutation that had previously undergone sequencing. For the second, DNA from 132 cases of known genotype was sequenced using low-pass whole-genome sequencing (mean coverage 0.5X) by Azenta Life Sciences.

See Supplemental Methods for processing details for each cohort.

Method details

Genome-wide association analysis for telomere length

Phenotype and sample construction

For the telomere phenotype, we used the qPCR-based leukocyte telomere length measurements generated, batch-corrected and quality controlled in 474,074 UKB participants (UKB Data Field 22191) (74). The phenotype was log-normalized and z-scored as in Codd et al (73). Of the remaining individuals, we then removed the individuals with the highest and lowest 0.5% of telomere length values to minimize the impact of outliers.

The set of filtered, phenotyped individuals was randomly split into two non-overlapping subsets. Two-thirds of the individuals (n = 259,142) were used in the GWAS analysis of telomere length, and the remaining one-third of samples (n = 129,547) were used for subsequent polygenic risk score construction.

Genome-wide association study for telomere length

GWAS was performed with genotype array data for REGENIE Step 1 and whole-genome sequencing data for REGENIE Step 2. Age, sex, and the first 10 genetic principal components were included as covariates. Additional QC, filtering parameters, and full details are provided in the Supplemental Info.

Polygenic risk score construction using telomere length summary statistics

Primary polygenic risk score construction

The summary statistics and processed whole-genome sequences variant call files were then used to construct polygenic risk scores using the 129,547 samples that were not included in the GWAS. As the goal of constructing these scores was for use as an instrument to assess the burden of common genetic variation affecting telomere length across different cohorts, we restricted analysis to SNPs that were well imputed (INFO > 0.6) across each of the different cohorts that were tested. This retained 5,279,945 SNPs genome-wide.

Polygenic scores were constructed using the common clumping and thresholding approach. In clumping and thresholding, index SNPs are selected, and SNPs within a prespecified distance window and above a prespecified correlation threshold are identified (“clumped”) and removed (“pruned”), to remove redundant effects from SNPs on the same haplotype, and this process is iteratively repeated across the genome. For the variants remaining after pruning, a range of p-value thresholds is then applied, removing all variants above the p-value threshold. The resulting variant sets are then used to compute polygenic scores defined as the sum of variant allele counts weighted by effect size estimates from the telomere length GWAS. See **Supplemental Methods** for details on score selection, which yielded the best performing polygenic score with 304 genome-wide SNPs, and a variance explained for telomere length of 0.071 (Table S1).

Comparing polygenic risk scores across cohorts

The SNPs which were included in the best PGS score computed in the UK biobank were used to directly compute scores in each different cohort. See Supplemental Info for details.

All of Us polygenic risk score construction and testing

To compare telomere length PGS between the All of Us (111) cohort and patient samples, an identical construction procedure was followed as in UK Biobank, but restricting to SNPs genotyped with high quality in All of Us. See Supplemental Info for details.

Plotting and comparing score distributions

See Supplemental Info for details.

Computing odds ratios

To assess the relationship between PGS and the outcome of dyskeratosis congenita, we defined dyskeratosis congenita affected individuals as cases and UK Biobank participants as controls. The PGS population values were binned into 5 equal-sized quintiles, and each observation for both cases and controls were placed in the appropriate quintile. We then calculated odds ratios and 95% confidence intervals for each quintile relative to the lowest risk (longest predicted telomere length) quintile. For each quintile, a 2x2 contingency table was constructed, comparing the number of outcome cases within the quintile to those outside it. Fisher's exact test was applied to estimate the odds ratio for each quintile. The standard error of the logarithm of the odds ratio was used to compute the 95% confidence intervals.

Ancestry inference

See Supplemental Info for details.

UK Biobank pathogenic variant carrier analysis

Defining variant sets

For all analysis of variant carriers in the UK Biobank, the following genes were used as a TBD gene set, defined based on genes with strong evidence of being capable of causing dyskeratosis congenita, Revesz syndrome, Hoyeraal-Hreidarsson syndrome, or Coats plus syndrome, and having a mode of inheritance reported in the literature as causing dyskeratosis congenita/other syndromic TBDs or bone marrow failure in a monoallelic or “monoallelic or biallelic” fashion, excluding genes for which only biallelic inheritance has been reported (Table S3) (17). This resulted in inclusion of the following genes: *TERC, TINF2, ZCCHC8, ACD, RTEL1, TERT,* and *DKC1.* For *DKC1*, only male carriers (X-linked recessive) were included. Noncarriers were defined as any participants passing QC that did not carry one of the pathogenic variants included in the final analyses. See Supplemental Info for ClinVar query.

The Predicted Pathogenic dataset was produced starting from all variants in the whole exome dataset that passed QC (above) with less than 1% frequency, within the start and end positions of known TBD genes (NCBI). Variants were then annotated using the Ensembl Variant Effect Predictor. The annotations were then filtered to retain all predicted “HIGH IMPACT” variants. Variants with predicted “MODERATE IMPACT” were further annotated using the AlphaMissense released pathogenicity scores, and variants predicted as “likely_pathogenic” were retained.(94, 95)

Pre-processing the UK Biobank whole-exome sequencing data

See Supplemental Methods.

Identification and phenotypic analysis of carriers of whole-exome variants

We used the UK Biobank Research Analysis Platform to extract phenotype data. We first filtered individuals using the criteria described in Supplemental Info for quality control and White British ancestry. We extracted phenotypic information on blood cell counts, clinical records, and telomere length. To assess for blood cell count differences, the Haemoglobin_concentration, Platelet_count, and White_blood_cell_count phenotype fields were utilized and z-scored. To assess for idiopathic pulmonary fibrosis, we used ICD-10 code of J84.1 (“other interstitial pulmonary diseases with fibrosis”). To assess for aplastic anemia, we used ICD-10 code of D61 (“Other aplastic anaemias,” which includes “Aplastic anaemia, unspecified,” “constitutional aplastic anaemia”, “idiopathic aplastic anaemia”, “other specified aplastic anaemias” and “aplastic anaemia, unspecified”).

Statistical analyses of UKB pathogenic variant carriers

See Supplemental Info for details on statistical analyses, interactions between PGS and rare variant status, and mediation analysis.

Analysis of expressivity within a family

Pedigree analysis polygenic risk score construction

We utilized the same set of 5,279,945 SNPs to compute polygenic risk scores in the pedigree cohort. We constructed four different polygenic scores. Three were constructed taking the best score using clumping and thresholding with various sets of hyperparameters, while one score was constructed using only COJO genome-wide conditionally independent SNPs, as an orthogonal approach. See Supplemental Info for details. Each pedigree score was validated by binning the PGS into four quartiles and computing the Spearman correlation between the quartile and measured telomere length in the UK Biobank.

Family data statistical analysis

Because of high relatedness within the pedigree, a linear mixed model with a kinship matrix as a random effect was used. The kinship matrix was generated using PLINK (-- make-rel square) using the pedigree genotypes to estimate relatedness. The LMM was performed using lmm.aireml, with the PGS and an intercept term as fixed effects and the kinship matrix as a random effect. Fixed effects estimates and standard errors were extracted from the LMM results. Wald test statistics were computed as the ratio of the fixed effects to their standard errors, and p-values were derived from the Wald test statistics.

Analysis of convergence of common and rare variation

Gene prioritization

From the telomere length GWAS summary statistics, genes were prioritized using MAGMA and a closest-gene approach based on COJO conditionally genome-wide significant SNPs. See Supplemental Info for details.

Finemapping

See Supplemental Methods for details.

Fine-mapped Variant Annotation

Fine-mapped variants were annotated using the Ensembl VEP and the Activity-by-Contact genome-wide enhancer maps (94, 126). A credible set of fine-mapped variants was defined as variants with posterior inclusion probability greater than 0.9. This resulted in a fine-mapped set of 529 variants. Credible set variants were annotated using Ensembl VEP for functional sequence type. ABC enhancer maps from CD34+ hematopoietic stem cells were overlapped with the credible set variants. Enrichment of fine-mapped variants within CD34+ enhancers was assessed by comparing the proportion of credible set variants within the ABC CD34+ enhancers to the proportion of variants with a posterior inclusion probability less than 0.1 within the ABC CD34+ enhancers, using a Chi-squared test.

Statistics

Statistical tests used for each analysis can be found in the figure legends with further details and explanations for choice of test in the relevant METHOD DETAILS sections. All analyses with multiple comparisons were corrected using the Bonferroni method. An adjusted *P* value less than 0.05 was considered statistically significant. All statistical tests were performed in R.

Study approval

All uses of patient data were approved by the Institutional Review Boards at Queen Mary University, the NCI, Boston Children's Hospital, and the Broad Institute. Written informed consent for use of all patient data was obtained.

Data availability

As per biobank policy, UK Biobank individual-level data are available upon request by application (https://www.ukbiobank.ac.uk/). TOPMed individual-level genomic data is available by application. Due to patient confidentiality, deidentified patient cohort data will be made available in restricted fashion upon reasonable request to the corresponding author. The data underlying the analysis of finemapped variants in Figure 5 and Figure S4 is available in the Supporting Data Values spreadsheet. Due to limitations on sharing individual-level data, summary data information for most other analyses is available in the Supporting Data Values spreadsheet.

Author contributions

M.P. and V.G.S. conceived and designed the study. M.P. and U.P.A. performed computational and statistical analyses. A.W., L.J.M., N.G., L.Z., H.T., and S.A.S. provided genotype and phenotype data from disease cohorts. A.W., L.J.M., M.R.N., N.G., A.G., M.J.M., H.T., and S.A.S. contributed ideas and insights on the analyses. M.P. and V.G.S. wrote the manuscript with input from all authors. V.G.S. supervised all analytical aspects of this work.

Acknowledgments

We thank Sankaran lab members and colleagues including D. Tang, M. Talukdar, T. de Lange, D. Nathan, and D. Ginsburg for valuable discussions on this work. This work was supported by National Institutes of Health grants R01DK103794, R01HL146500, R01CA265726, R01CA292941 (to V.G.S.), the New York Stem Cell Foundation (to V.G.S.), the Mathers Foundation (to V.G.S.), a gift in memory of Jan Ellen Paradise, MD to Boston Children's Hospital (to V.G.S.), and the Howard Hughes Medical Institute (to V.G.S.). This research has been conducted using the UK Biobank under application number 31063. Figure schematics were prepared with BioRender. M.P. was supported by NIH grants T32GM007753 and T32GM144273. The work of S.A.S., L.Z., M.R.N, N.G., and L.J.M was supported by the intramural research program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent official views of the NIH. V.G.S. is an Investigator of the Howard Hughes Medical Institute.

References

1. Kingdom R, Wright CF. Incomplete penetrance and variable expressivity: From clinical studies to population cohorts. *Front Genet*. 2022;13:920390.

2. Cooper DN, et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet*. 2013;132(10):1077–1130.

3. Fahed AC, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun*. 2020;11(1):3635.

4. Kingdom R, et al. Genetic modifiers of rare variants in monogenic developmental disorder loci. *Nat Genet*. 2024;56(5):861–868.

5. Milne RL, Antoniou AC. Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Ann Oncol*. 2011;22 Suppl 1(suppl 1):i11–7.

6. Chatterjee S, et al. Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. *Cell*. 2016;167(2):355–368.e10.

7. Oetjens MT, et al. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat Commun*. 2019;10(1):4897.

8. Goodrich JK, et al. Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat Commun*. 2021;12(1):3505.

9. Chatterjee S, et al. A multi-enhancer RET regulatory code is disrupted in Hirschsprung disease. *Genome Res*. 2021;31(12):2199–2208.

10. Cato LD, et al. Genetic regulation of fetal hemoglobin across global populations. *medRxiv*.

[published online ahead of print: March 28, 2023]. https://doi.org/10.1101/2023.03.24.23287659.

11. Niemi MEK, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*. 2018;562(7726):268–271.

12. Timberlake AT, et al. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. *Elife*. 2016;5. https://doi.org/10.7554/eLife.20125.

13. Albers CA, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet*. 2012;44(4):435–9, S1–2.

14. Lemmers RJLF, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet*. 2012;44(12):1370–1374.

15. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet*. 2001;17(9):502–510.

16. Niewisch MR, et al. Disease progression and clinical outcomes in telomere biology disorders. *Blood*. 2022;139(12):1807–1819.

17. Revy P, Kannengiesser C, Bertuch AA. Genetics of human telomere biology disorders. *Nat Rev Genet*. 2023;24(2):86–108.

18. Nachmani D, et al. Germline NPM1 mutations lead to altered rRNA 2'-O-methylation and cause dyskeratosis congenita. *Nat Genet*. 2019;51(10):1518–1529.

19. Yamaguchi H, et al. Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. *N Engl J Med*. 2005;352(14):1413–1424.

20. Vulliamy TJ, et al. Mutations in the reverse transcriptase component of telomerase (TERT) in patients with bone marrow failure. *Blood Cells Mol Dis*. 2005;34(3):257–263.

21. Toufektchan E, et al. Germline mutation of MDM4, a major p53 regulator, in a familial syndrome of defective telomere maintenance. *Sci Adv*. 2020;6(15):eaay3511.

22. Simon AJ, et al. Mutations in STN1 cause Coats plus syndrome and are associated with genomic and telomere defects. *J Exp Med*. 2016;213(8):1429–1440.

23. Walne AJ, et al. Mutations in the telomere capping complex in bone marrow failure and related syndromes. *Haematologica*. 2013;98(3):334–338.

24. Anderson BH, et al. Mutations in CTC1, encoding conserved telomere maintenance component 1, cause Coats plus. *Nat Genet*. 2012;44(3):338–342.

25. Kermasson L, et al. Inherited human Apollo deficiency causes severe bone marrow failure and developmental defects. *Blood*. 2022;139(16):2427–2440.

26. Wu P, Takai H, de Lange T. Telomeric 3’ overhangs derive from resection by Exo1 and Apollo and fill-in by POT1b-associated CST. *Cell*. 2012;150(1):39–52.

27. Wu P, et al. Apollo contributes to G overhang maintenance and protects leading-end telomeres. *Mol Cell*. 2010;39(4):606–617.

28. Ye J, et al. TRF2 and apollo cooperate with topoisomerase 2alpha to protect human telomeres from replicative damage. *Cell*. 2010;142(2):230–242.

29. van Overbeek M, de Lange T. Apollo, an Artemis-related nuclease, interacts with TRF2 and protects human telomeres in S phase. *Curr Biol*. 2006;16(13):1295–1302.

30. Margalef P, et al. Stabilization of reversed replication forks by telomerase drives telomere

catastrophe. *Cell*. 2018;172(3):439–453.e14.

31. Sharma R, et al. Gain-of-function mutations in RPA1 cause a syndrome with short telomeres and somatic genetic rescue. *Blood*. 2022;139(7):1039–1051.

32. Ballew BJ, et al. Germline mutations of regulator of telomere elongation helicase 1, RTEL1, in Dyskeratosis congenita. *Hum Genet*. 2013;132(4):473–480.

33. Deng Z, et al. Inherited mutations in the helicase RTEL1 cause telomere dysfunction and Hoyeraal-Hreidarsson syndrome. *Proc Natl Acad Sci U S A*. 2013;110(36):E3408–16.

34. Le Guen T, et al. Human RTEL1 deficiency causes Hoyeraal-Hreidarsson syndrome with short telomeres and genome instability. *Hum Mol Genet*. 2013;22(16):3239–3249.

35. Ballew BJ, et al. A recessive founder mutation in regulator of telomere elongation helicase 1, RTEL1, underlies severe immunodeficiency and features of Hoyeraal Hreidarsson syndrome. *PLoS Genet*. 2013;9(8):e1003695.

36. Walne AJ, et al. Constitutional mutations in RTEL1 cause severe dyskeratosis congenita. *Am J Hum Genet*. 2013;92(3):448–453.

37. Takai H, et al. A POT1 mutation implicates defective telomere end fill-in and telomere truncations in Coats plus. *Genes Dev*. 2016;30(7):812–826.

38. Tummala H, et al. Homozygous OB-fold variants in telomere protein TPP1 are associated with dyskeratosis congenita-like phenotypes. *Blood*. 2018;132(12):1349–1353.

39. Guo Y, et al. Inherited bone marrow failure associated with germline mutation of ACD, the gene encoding telomere protein TPP1. *Blood*. 2014;124(18):2767–2774.

40. Walne AJ, et al. TINF2 mutations result in very short telomeres: analysis of a large cohort of

patients with dyskeratosis congenita and related bone marrow failure syndromes. *Blood*. 2008;112(9):3594–3600.

41. Burris AM, et al. Hoyeraal-Hreidarsson syndrome due to PARN mutations: Fourteen years of follow-up. *Pediatr Neurol*. 2016;56:62–68.e1.

42. Dhanraj S, et al. Bone marrow failure and developmental delay caused by mutations in poly(A)-specific ribonuclease (PARN). *J Med Genet*. 2015;52(11):738–748.

43. Stanley SE, et al. Loss-of-function mutations in the RNA biogenesis factor NAF1 predispose to pulmonary fibrosis-emphysema. *Sci Transl Med*. 2016;8(351):351ra107.

44. Brailovski E, et al. Previously unreported WRAP53 gene variants in a patient with dyskeratosis congenita. *Ann Hematol*. 2022;101(4):907–909.

45. Bergstrand S, et al. Biallelic mutations in WRAP53 result in dysfunctional telomeres, Cajal bodies and DNA repair, thereby causing Hoyeraal-Hreidarsson syndrome. *Cell Death Dis*. 2020;11(4):238.

46. Zhong F, et al. Disruption of telomerase trafficking by TCAB1 mutation causes dyskeratosis congenita. *Genes Dev*. 2011;25(1):11–16.

47. Vulliamy T, et al. Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita. *Proc Natl Acad Sci U S A*. 2008;105(23):8073–8078.

48. Kannengiesser C, et al. First heterozygous NOP10 mutation in familial pulmonary fibrosis. *Eur Respir J*. 2020;55(6):1902465.

49. Walne AJ, et al. Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase-associated protein NOP10. *Hum Mol Genet*. 2007;16(13):1619–1629.

50. Xu J, et al. Investigation of chromosome X inactivation and clinical phenotypes in female carriers of DKC1 mutations. *Am J Hematol*. 2016;91(12):1215–1220.

51. Armanios MY, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med*. 2007;356(13):1317–1326.

52. Knight SW, et al. X-linked dyskeratosis congenita is predominantly caused by missense mutations in the DKC1 gene. *Am J Hum Genet*. 1999;65(1):50–58.

53. Shukla S, et al. Inhibition of telomerase RNA decay rescues telomerase deficiency caused by dyskerin or PARN defects. *Nat Struct Mol Biol*. 2016;23(4):286–292.

54. Heiss NS, et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. *Nat Genet*. 1998;19(1):32–38.

55. Vulliamy T, et al. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature*. 2001;413(6854):432–435.

56. Marrone A, et al. Telomerase reverse-transcriptase homozygous mutations in autosomal recessive dyskeratosis congenita and Hoyeraal-Hreidarsson syndrome. *Blood*. 2007;110(13):4198–4205.

57. Du H-Y, et al. Complex inheritance pattern of dyskeratosis congenita in two families with 2 different mutations in the telomerase reverse transcriptase gene. *Blood*. 2008;111(3):1128–1130.

58. Lim CJ, Cech TR. Shaping human telomeres: from shelterin and CST complexes to telomeric chromatin organization. *Nat Rev Mol Cell Biol*. 2021;22(4):283–298.

59. Cai SW, et al. POT1 recruits and regulates CST-Polα/primase at human telomeres. *Cell*. 2024;0(0). https://doi.org/10.1016/j.cell.2024.05.002.

60. Mannherz W, Agarwal S. Thymidine nucleotide metabolism controls human telomere length. *Nature genetics*. 2023;55(4). https://doi.org/10.1038/s41588-023-01339-5.

61. Tummala H, et al. Germline thymidylate synthase deficiency impacts nucleotide metabolism and causes dyskeratosis congenita. *Am J Hum Genet*. 2022;109(8):1472–1483.

62. Savage SA, et al. TINF2, a component of the shelterin telomere protection complex, is mutated in dyskeratosis congenita. *Am J Hum Genet*. 2008;82(2):501–509.

63. Orphanet: Dyskeratosis congenita [Internet]. https://www.orpha.net/en/disease/detail/1775. Accessed August 5, 2024.

64. Tsakiri KD, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci U S A*. 2007;104(18):7552–7557.

65. Alder JK, et al. Exome sequencing identifies mutant TINF2 in a family with pulmonary fibrosis. *Chest*. 2015;147(5):1361–1368.

66. Stuart BD, et al. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat Genet*. 2015;47(5):512–517.

67. Garcia CK, Wright WE, Shay JW. Human diseases of telomerase dysfunction: insights into tissue aging. *Nucleic Acids Res*. 2007;35(22):7406–7416.

68. Armanios M, et al. Haploinsufficiency of t *e* lomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc Natl Acad Sci USA*. 2005;102(44):15960–15964.

69. Vulliamy T, et al. Disease anticipation is associated with progressive telomere shortening in families with dyskeratosis congenita due to mutations in TERC. *Nat Genet*. 2004;36(5):447–449.

70. Niewisch MR, et al. Genotype and associated cancer risk in individuals with telomere biology disorders. *JAMA Netw Open*. 2024;7(12):e2450111.

71. Savage SA, Alter BP. The role of telomere biology in bone marrow failure and other disorders. *Mech Ageing Dev*. 2008;129(1-2):35–47.

72. Alter BP, et al. Telomere length in inherited bone marrow failure syndromes. *Haematologica*. 2015;100(1):49–54.

73. Codd V, et al. Polygenic basis and biomedical consequences of telomere length variation. *Nat Genet*. 2021;53(10):1425–1433.

74. Codd V, et al. Measurement and initial characterization of leukocyte telomere length in 474,074 participants in UK Biobank. *Nat Aging*. 2022;2(2):170–179.

75. Burren OS, et al. Genetic architecture of telomere length in 462,675 UK Biobank whole-genome sequences. *bioRxiv*. 2023. https://doi.org/10.1101/2023.09.18.23295715.

76. Taub MA, et al. Genetic determinants of telomere length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. *Cell Genom*. 2022;2(1). https://doi.org/10.1016/j.xgen.2021.100084.

77. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8(7). https://doi.org/10.1093/gigascience/giz082.

78. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1):44.

79. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15(9):2759–2772.

80. Alter BP, et al. Cancer in the National Cancer Institute inherited bone marrow failure syndrome cohort after fifteen years of follow-up. *Haematologica*. 2018;103(1):30–39.

81. Tummala H, et al. The evolving genetic landscape of telomere biology disorder dyskeratosis congenita. *EMBO Molecular Medicine*. 2024;1–23–23.

82. Zhou D, et al. Contextualizing genetic risk score for disease screening and rare variant discovery. *Nat Commun*. 2021;12(1):4418.

83. Bergen SE, et al. Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *Am J Psychiatry*. 2019;176(1):29–35.

84. Berg JJ, et al. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019;8. https://doi.org/10.7554/eLife.39725.

85. Sohail M, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019;8. https://doi.org/10.7554/eLife.39702.

86. Aw AJ, et al. Highly parameterized polygenic scores tend to overfit to population stratification via random effects. *bioRxiv*. 2024;2024.01.27.577589.

87. Martin AR, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet*. 2017;100(4):635–649.

88. Manichaikul A, et al. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010;26(22):2867–2873.

89. 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.

90. Kee Y, D'Andrea AD. Molecular pathogenesis and clinical management of Fanconi anemia. *J Clin Invest*. 2012;122(11):3799–3806.

91. Ludwig LS, et al. Altered translation of GATA1 in Diamond-Blackfan anemia. *Nat Med*. 2014;20(7):748–753.

92. Boocock GRB, et al. Mutations in SBDS are associated with Shwachman-Diamond syndrome. *Nat Genet*. 2003;33(1):97–101.

93. Claussnitzer M, et al. A brief history of human disease genetics. *Nature*. 2020;577(7789):179–189.

94. McLaren W, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.

95. Cheng J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492.

96. Karczewski KJ, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom*. 2022;2(9):100168.

97. Savage SA, Bertuch AA. The genetics and clinical manifestations of telomere biology disorders. *Genet Med*. 2010;12(12):753–764.

98. Dressen A, et al. Analysis of protein-altering variants in telomerase genes and their association with MUC5B common variant status in patients with idiopathic pulmonary fibrosis: a candidate gene sequencing study. *Lancet Respir Med*. 2018;6(8):603–614.

99. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*. 2013;18(2):137–150.

100. Yang J, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44(4):369–75, S1–3.

101. Weiner DJ, et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature*. 2023;614(7948):492–499.

102. Weiner DJ, et al. Statistical and functional convergence of common and rare genetic influences on autism at chromosome 16p. *Nat Genet*. 2022;54(11):1630–1639.

103. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med*. 2013;3(1):a011643.

104. Bycroft C, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203–209.

105. Shay JW, Wright WE. Telomeres and telomerase: three decades of progress. *Nat Rev Genet*. 2019;20(5):299–309.

106. Hemann MT, et al. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell*. 2001;107(1):67–77.

107. Raj HA, et al. The distribution and accumulation of the shortest telomeres in telomere biology disorders. *Br J Haematol*. 2023;203(5):820–828.

108. Karimian K, et al. Human telomere length is chromosome end–specific and conserved across individuals. *Science*;0(0):eado0431.

109. Schmidt TT, et al. High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nat Commun*. 2024;15(1):5149.

110. Fry A, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK

Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026–1034.

111. All of Us Research Program Investigators, et al. The “All of Us” Research Program. *N Engl J Med*. 2019;381(7):668–676.

112. Mapes BM, et al. Diversity and inclusion for the All of Us research program: A scoping review. *PLoS One*. 2020;15(7):e0234962.

113. Sudlow C, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.

114. Li S, et al. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv*. 2023;2023.12.06.23299426.

115. Taliun D, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290–299.

116. Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster genotype imputation. *Bioinformatics*. 2015;31(5):782–784.

117. Das S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284–1287.

118. Rubinacci S, et al. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat Genet*. 2023;55(7):1088–1090.

119. Mbatchou J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097–1103.

120. de Leeuw CA, et al. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput*

Biol. 2015;11(4):e1004219.

121. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.

122. Yang J, et al. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82.

123. Benner C, et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016;32(10):1493–1501.

124. Nasser J, et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature*. 2021;593(7858):238–243.

Figures

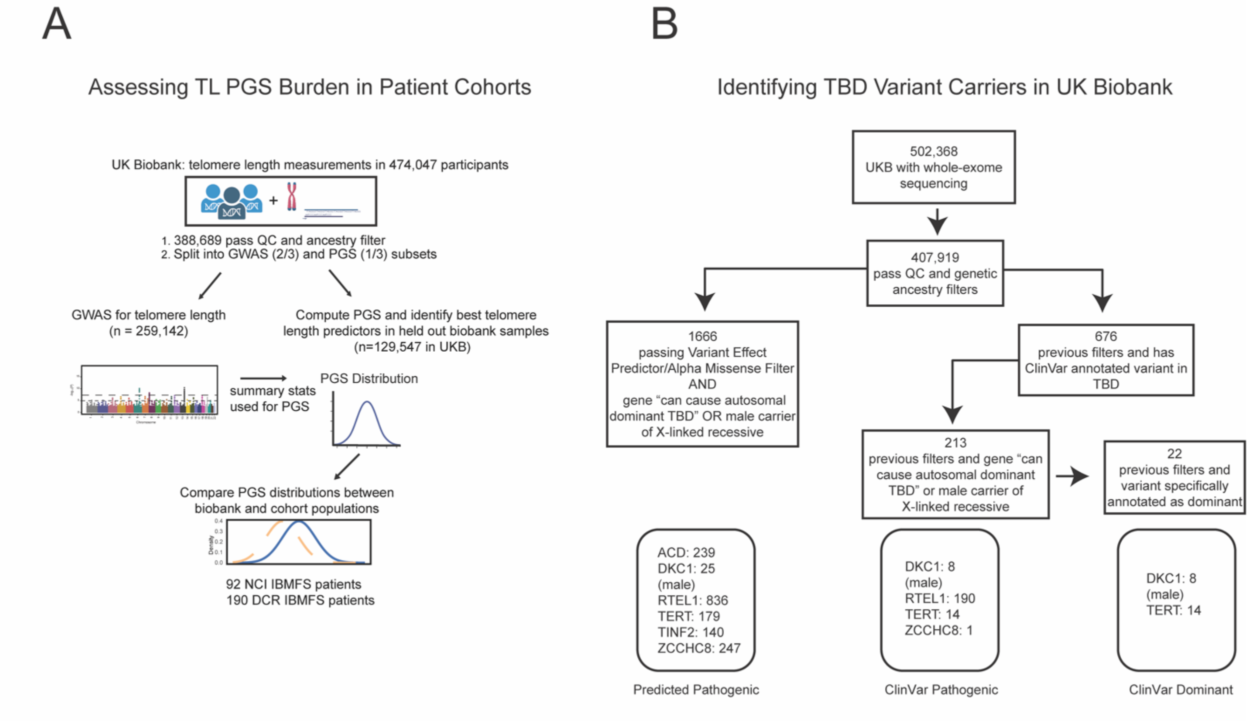

A
Assessing TL PGS Burden in Patient Cohorts
UK Biobank: telomere length measurements in 474,047 participants
1. 388,689 pass QC and ancestry filter
2. Split into GWAS (2/3) and PGS (1/3) subsets
GWAS for telomere length
(n = 259,142)
Compute PGS and identify best telomere length predictors in held out biobank samples
(n=129,547 in UKB)
summary stats used for PGS
PGS Distribution
Compare PGS distributions between biobank and cohort populations
Density
0.4
0.3
0.2
0.1
0.0
92 NCI IBMFS patients
190 DCR IBMFS patients
B
Identifying TBD Variant Carriers in UK Biobank
502,368
UKB with whole-exome sequencing
407,919
pass QC and genetic ancestry filters
1666
passing Variant Effect Predictor/Alpha Missense Filter
AND
gene "can cause autosomal dominant TBD" OR male carrier of X-linked recessive
676
previous filters and has ClinVar annotated variant in TBD
213
previous filters and gene "can cause autosomal dominant TBD" or male carrier of X-linked recessive
22
previous filters and variant specifically annotated as dominant
ACD: 239
DKC1: 25
(male)
RTEL1: 836
TERT: 179
TINF2: 140
ZCCHC8: 247
DKC1: 8
(male)
RTEL1: 190
TERT: 14
ZCCHC8: 1
DKC1: 8
(male)
TERT: 14
Predicted Pathogenic
ClinVar Pathogenic
ClinVar Dominant

Figure 1: Flow Diagram of Study Design

A) Schematic of study design to assess polygenic variation affecting telomere length across biobank and cohort populations. Individuals with measured telomere length in the UK Biobank were randomly split into GWAS and PGS subsets. A GWAS for telomere length was conducted, and then the best polygenic predictors were identified in the other subset. The resulting PGS represents the common variant burden associated with telomere length. The best PGS was compared between biobank cohorts and disease cohorts.

B) Variant carrier ascertainment strategy in UK Biobank, and summary of mutation carriers for each gene in variant carrier groups. Pathogenic variants were alternately defined based on pathogenic prediction using VEP and AlphaMissense, or using ClinVar annotated variants.

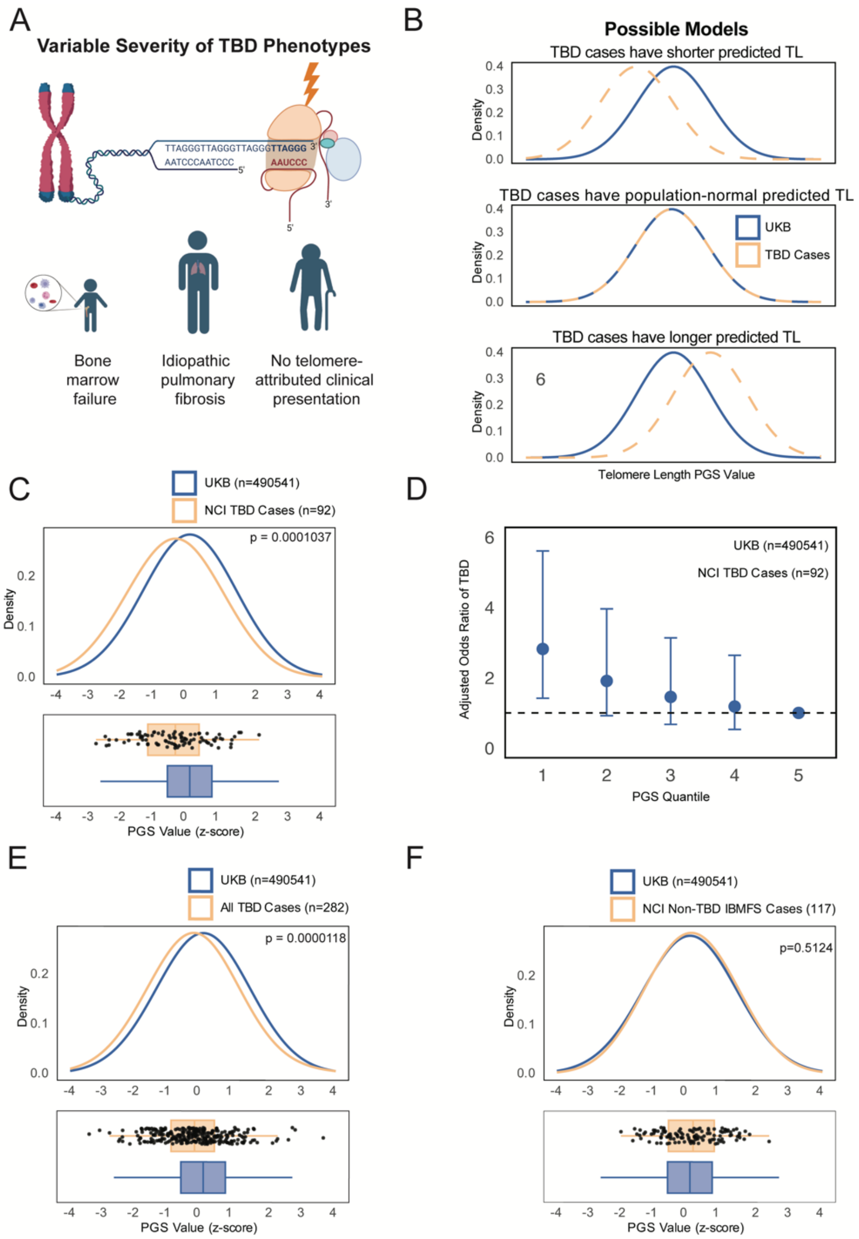


Figure 2: Polygenic modification of TBD penetrance and expressivity in disease cohorts

A) Illustration. TBD-associated germline variants affect genes involved in telomere length and integrity. Variable expressivity of TBD variants results in diverse phenotypic presentations and age of onset.

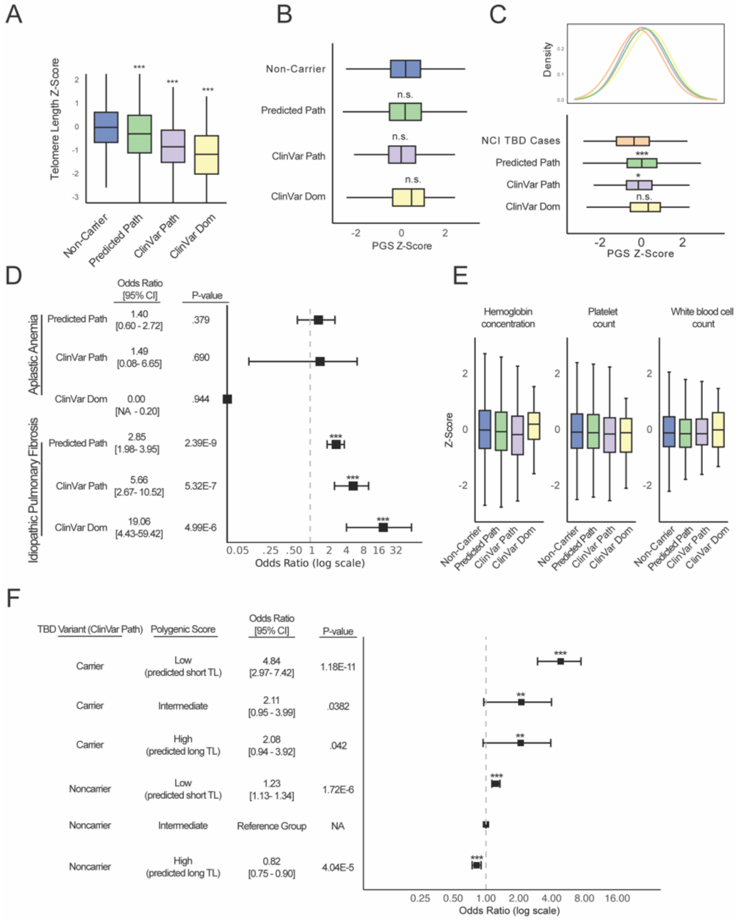
B) Schematic of distribution of TBD-case telomere length polygenic scores compared to biobanks, under different hypotheses. If common variation affecting telomere length contributes to TBD expressivity and disease cohort ascertainment, left-shifted (towards shorter TL) PGS distribution would be expected (top panel). The null hypothesis is that TBD high-impact variants overpower any effects of common variation (central panel). An alternative hypothesis is that common variation predisposing to long telomere length protects TBD variant-carriers from severe phenotypes and mortality; under this model, a right-shifted PGS could be observed (bottom panel).

C) Distribution of telomere length PGS in NCI TBD cases compared to the UK Biobank (Welch's two-tailed t-test, p = 1.037E-4).

D) Odds ratio of case-control status versus telomere length PGS quintile, NCI TBD cases.

E) Comparison of meta-analysis of telomere length PGS distribution in NCI and DCR cases vs UK Biobank (Welch's two-tailed t-test, p = 1.18E-5).

F) Comparison of NCI non-TBD IBMFS case telomere length PGS vs UK Biobank (Welch's two-tailed t-test, p = 0.5124).



		Odds Ratio [95% CI]	P-value
Aplastic Anemia	Predicted Path	1.40 [0.60 - 2.72]	.379
	ClinVar Path	1.49 [0.08- 6.65]	.690
	ClinVar Dom	0.00 [NA - 0.20]	.944
Idiopathic Pulmonary Fibrosis	Predicted Path	2.85 [1.98- 3.95]	2.39E-9
	ClinVar Path	5.66 [2.67- 10.52]	5.32E-7
	ClinVar Dom	19.06 [4.43-59.42]	4.99E-6

TBD Variant (ClinVar Path)	Polygenic Score	Odds Ratio [95% CI]	P-value
Carrier	Low (predicted short TL)	4.84 [2.97- 7.42]	1.18E-11
Carrier	Intermediate	2.11 [0.95 - 3.99]	.0382
Carrier	High (predicted long TL)	2.08 [0.94 - 3.92]	.042
Noncarrier	Low (predicted short TL)	1.23 [1.13- 1.34]	1.72E-6
Noncarrier	Intermediate	Reference Group	NA
Noncarrier	High (predicted long TL)	0.82 [0.75 - 0.90]	4.04E-5

Figure 3: Polygenic modification of TBD penetrance and expressivity in the UK Biobank
A) Measured telomere length in UK Biobank non-carriers and carriers of pathogenic TBD variants (pairwise t-tests with Bonferroni multiple testing correction, alternative = "less": non-carrier vs Predicted Pathogenic: p = 3.80E-20; non-carrier vs ClinVar Pathogenic: p = 3.46E-23; non-carrier vs ClinVar Dominant-Acting: p = 1.39E-3)
B) TL PGS in UK Biobank non-carriers and carriers of pathogenic TBD variants (pairwise t-test with Bonferroni multiple testing correction, alternative = "two-sided": non-carrier vs Predicted Pathogenic: p = 1; non-carrier vs ClinVar Pathogenic: p = .20; non-carrier vs ClinVar Dominant-Acting: p = 1).
C) TL PGS in NCI cases compared to UKB pathogenic variant carriers (pairwise t-test with Bonferroni multiple testing correction, alternative = "greater": TBD case vs Predicted Pathogenic: p = .00087; TBD case vs ClinVar Pathogenic: p = .041; TBD case vs ClinVar Dominant-Acting: p = .0896).
D) Odds ratios of aplastic anemia and idiopathic pulmonary fibrosis in carriers of pathogenic TBD variants compared to non-carriers (logistic regression adjusting for age and sex).
E) Blood cell counts in UK Biobank non-carriers and carriers of pathogenic TBD variants (pairwise t-test with Bonferroni multiple testing correction)
F) Odds ratios of idiopathic pulmonary fibrosis in UK Biobank stratified by PGS tertile and ClinVar Path or Predicted Path variant-carrier status, using non-carrier intermediate group as the control group (logistic regression adjusting for age, sex and first 4 ancestry PCs).

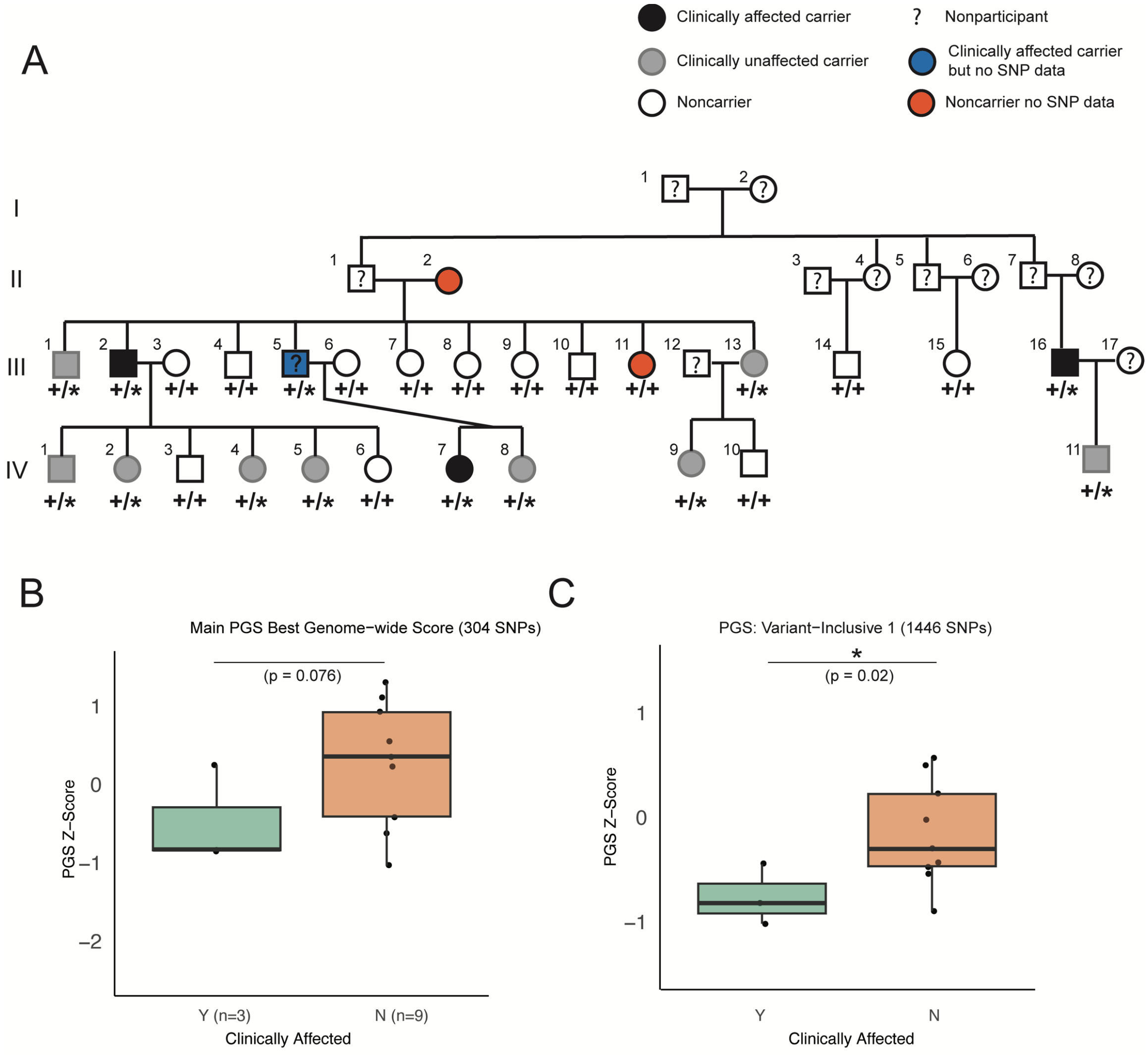


Figure 4: Polygenic modification of penetrance within a family

A) Pedigree depicting family with TERT variant. Black indicates case, gray indicates TERT non-case carrier, transparent indicates non-carrier. Square indicates male, circle indicates female. ? indicates unknown status, diamond indicates unknown sex.

B) Telomere length PGS comparing cases to non-case TERT variant carriers using same parameters as best genome-wide SNP score (linear mixed model with kinship matrix as random effect, see Methods).

C) Telomere length PGS comparing cases to non-case TERT variant carriers for Variant-inclusive PGS Score 1 (linear mixed model with kinship as random effect).

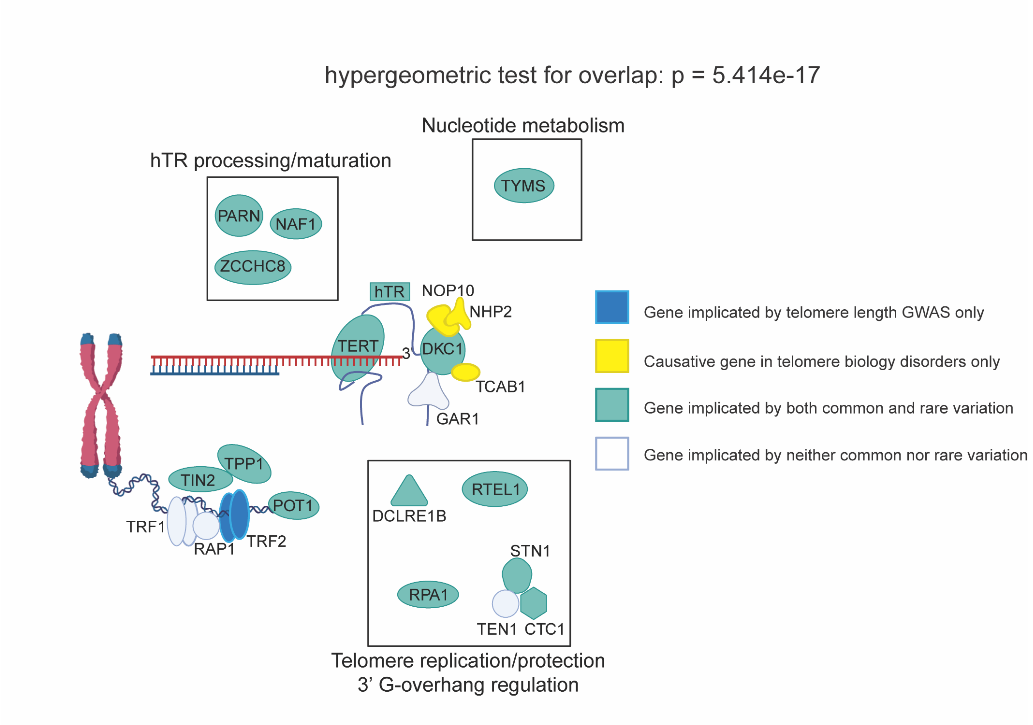

hypergeometric test for overlap: p = 5.414e-17
Nucleotide metabolism
TYMS
hTR processing/maturation
PARN
NAF1
ZCCHC8
hTR
NOP10
NHP2
TERT
3'
DKC1
TCAB1
GAR1
TPP1
TIN2
POT1
TRF1
RAP1
TRF2
DCLRE1B
RTEL1
STN1
RPA1
TEN1
CTC1
Telomere replication/protection
3' G-overhang regulation
Gene implicated by telomere length GWAS only
Causative gene in telomere biology disorders only
Gene implicated by both common and rare variation
Gene implicated by neither common nor rare variation

Figure 5: Common and rare genetic variation converges in telomere biology disorders
Genes with mutations which cause TBDs show strong overlap with genes underpinning common-variant associations with telomere length. Causal genes associated with common SNPs in telomere length PGS overlaid with genes causally implicated in telomere biology disorders.