AMERICAN ACADEMY OF OPHTHALMOLOGY®

Prevalence of Transcription Factor 4 Gene Triplet Repeat Expansion Associated with Fuchs' Endothelial Corneal Dystrophy in the United States and Global Populations

Xunzhi Zhang,[1,*] *Ashwani Kumar, MS,*[1,*] *Xin Gong, MD,*[2] *Chao Xing, PhD,*[1,3,4] *V. Vinod Mootha, MD*[1,2]

Objective: An intronic cytosine-thymine-guanine (CTG) triplet repeat expansion in the transcription factor 4 gene (*TCF4*) gene (CTG18.1) confers significant risk for the development of Fuchs' endothelial corneal dystrophy (FECD). The objective of this study was to conduct an unbiased survey of the CTG18.1 repeat expansion allele frequencies in a multiethnic population-based cohort from the United States and in global populations.

Design: Cross-sectional study.

Subjects: Dallas Heart Study (DHS) cohort including 1599 African Americans (AAs), 1028 European Americans (EAs), and 458 Latinos; 2500 individuals from the 1000 Genomes Project (1KGP) sampled from 26 populations across 5 continents.

Methods: We genotyped the CTG18.1 short tandem repeat (STR) in DHS using targeted polymerase chain reaction amplification followed by fragment analysis. We also inferred the CTG18.1 repeat genotype based on short-read whole-genome sequencing in 1KGP using the computational tool ExpansionHunter.

Main Outcome Measures: The prevalence of an expanded CTG18.1 allele with $\geq$ 40 repeats was determined in United States and global populations.

Results: The carrier rates of the expanded allele were 3.1%, 8.1%, and 3.3% in AAs, EAs, and Latinos, respectively, in the DHS, and 2.7%, 9.5%, 5.2%, 7.2%, and 5.2% in the African (AFR), European (EUR), East Asian, South Asian, and admixed American continental populations, respectively, in the 1KGP. The distributions of the CTG18.1 repeat in DHS and in 1KGP are similar. The median repeat length was ~17 with the interquartile range between 12 and 23 in the DHS populations. The median repeat length was ~19 in all the 1KGP populations with the interquartile range between 13 and 26. The highest prevalence of the expanded allele carriers ranging from 12.1% to 12.5% was observed in some EUR and admixed American subpopulations. The frequency of expanded alleles carriers was absent or low (0%–1.9%) in subpopulations of West Africa but was present at 6.2% in a Kenyan subpopulation in East Africa.

Conclusions: The *TCF4* repeat expansion is most prevalent in people of EUR ancestry and least in AFR ancestry, which is consistent with FECD prevalence. The expanded *TCF4* CTG18.1 allele is the most common disease-causing STR in humans with worldwide implications for corneal disease.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100611* © *2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Fuchs' endothelial corneal dystrophy (FECD) is an age-related degenerative disease of the cornea that can lead to significant vision loss. Late-onset FECD typically presents during the fourth decade of life or later. In the United States, FECD disease prevalence has been estimated to be >4% of the population of European ancestry (EUR) over the age of 40 and increases with age.[1,2] Fuchs' endothelial corneal dystrophy is characterized by progressive loss of the normal morphology and cell density of the corneal endothelium accompanied by diffuse thickening of its underlying basement membrane, Descemet's membrane, with focal excrescences called guttae. Fuchs' endothelial corneal dystrophy can progress to corneal edema and scarring and represents the leading indication for corneal transplantation in the United States and other developed countries.[3,4] It will be of clinical importance and public health interest to develop diagnostic and prognostic markers of FECD.

Genome-wide association studies of FECD highlighted the association of single nucleotide polymorphisms across

https://doi.org/10.1016/j.xops.2024.100611
ISSN 2666-9145/24

the transcription factor 4 gene (*TCF4*),[5–7] which encodes a conserved class I basic helix-loop-helix transcription factor.[8] Expansions of a cytosine-thymine-guanine (CTG) trinucleotide repeat polymorphism (CTG18.1) in the intron of *TCF4* were reported to be strongly associated with FECD.[9] Familial studies examining the cosegregation of the repeat expansion with FECD[10] as well as the replication of the disease association in many cohorts of different ethnicities[10–21] have established that expanded alleles confer significant risk for the development of disease. Expanded cytosine-uracil-guanine repeat RNA transcripts from the CTG18.1 locus accumulate in the corneal endothelium and appear to bind and functionally sequester the muscleblind-like family of splicing factors to result in missplicing of muscleblind-like sensitive exons in this corneal tissue layer[22–25] (for a review of disease mechanisms see reference[26]).

The CTG18.1 short tandem repeat (STR) was originally discovered without knowledge of its associated phenotype using the repeat expansion detection assay on peripheral blood genomic DNA of individuals from 15 Centre d'Etude du Polymorphisme Humain pedigrees.[27] Alleles of >37 CTG repeats were found be unstable in parent-child transmissions. A frequency of 3% of the expanded allele was estimated based on the studies on the Centre d'Etude du Polymorphisme Humain pedigrees and 48 other families with bipolar disease without mention of their ethnicity. The only other estimates of the frequencies of the expanded allele were based on recent association studies comparing FECD cases and controls. The expanded alleles, the cutoff criterion varying from 40 to 50 repeats, have been estimated to be in ~20% to 80% of FECD cases and ~0% to 10% of controls.[26] The carrier rates were lower in FECD subjects of East Asian (EAS) ancestry and African (AFR) ancestry than in people of EUR ancestry.

To date, unbiased estimates of the prevalence of the *TCF4* CTG18.1 repeat expansion in the major ethnic groups of the United States and in global populations are lacking. In this study, we determined the prevalence of the repeat expansion in the Dallas Heart Study (DHS), a population-based cohort comprised of African Americans (AAs), Whites or European Americans (EAs), and Latinos, as well as in the 1000 Genomes Project (1KGP) which sampled participants from 26 populations across 5 continental regions of the world.[28] In addition, we used this opportunity to compare the concordance between the 3 different approaches used to detect expanded CTG18.1 alleles, namely targeted polymerase chain amplification (PCR) followed by fragment analysis, short-read sequencing followed by statistical inference, and long-read sequencing.

Methods

Study Participants

The study protocol had the approval of the institutional review board of the University of Texas Southwestern Medical Center and was in compliance with the tenets of the Declaration of Helsinki. Study participants were enrolled after written informed consent.

The DHS is a multiethnic population-based cohort in Dallas County[29] mainly consisting of AAs, EAs, and Latinos. Race was self-identified in questionnaires at the time of recruitment.

CTG18.1 in the DHS

Genomic DNA was extracted from leukocytes of peripheral blood samples of 3085 DHS participants (1599 AAs, 1028 EAs, and 458 Latinos). The *TCF4* CTG18.1 triplet repeat expansion was genotyped using a combination of STR analysis and triplet repeat primed PCR (TP-PCR) as previously described.[10] For the STR assay, primers flanking the repeat region were utilized for PCR amplification with 1 primer labeled with fluorescein amidite on 5′ end. Triplet repeat primed polymerase chain amplification assay was performed using the same flanking 5' fluorescein amidite labeled primer paired with repeat sequence targeted primers for PCR amplification. Polymerase chain amplification amplicons were subsequently loaded on an ABI 3730XL DNA analyzer (Applied Biosystems) and the results were analyzed using ABI GeneMapper 4.0 (Applied Biosystems).

We analyzed the STR tracings to detect 2 CTG18.1 alleles with repeat lengths up to ~95 CTG repeats in 2522 of the 3085 DHS samples. To resolve the zygosity of the 558 samples with only 1 CTG18.1 allele detected by the STR analysis, we reviewed the corresponding TP-PCR electropherogram tracings for the presence of the characteristic continuation ladder pattern of an expanded allele. In this manner, we were able to ascertain that 112 of these samples harbored a large CTG18.1 allele beyond the detection limits of the STR analysis and that the other 446 samples were homozygous for the allele detected by STR analysis. In the remaining 5 samples with no alleles detected by the STR analysis, the TP-PCR tracings detected 2 expanded alleles with the presence of the characteristic ladder pattern. Note that the TP-PCR assay can detect the presence of a large expanded CTG18.1 allele but cannot measure its exact repeat length.

We dichotomized the CTG18.1 trinucleotide repeat alleles at number 40, as we did in previous reports,[10,11] and defined those with $\geq$ 40 repeats as expanded alleles. The expanded and nonexpanded alleles were coded as "*L*" and "*S*," respectively. The genotype of rs613872, the leading single nucleotide polymorphism in the original genome-wide association studies of FECD,[5] based on the Illumina HumanExome BeadChip, was also extracted, and the degree of linkage disequilibrium between the 2 loci was calculated.

CTG18.1 in the 1000 Genomes Project

There were a total of 3202 samples that underwent PCR-free high-coverage short reads whole genome sequencing (WGS) in the 1KGP.[30] Libraries were prepared using the TruSeq DNA PCR-Free High Throughput Library Prep Kit and sequenced on an Illumina NovaSeq 6000 system using 2x150bp cycles. Read alignment to the Genome Reference Consortium Human Build 38 using BWA-MEM (v0.7.15), duplicate marking using Picard MarkDuplicates (v2.4.1), and base quality score recalibration using GATK BaseRecalibrator (v3.5) were performed according to the functional equivalence pipeline standard developed for the Centers for Common Disease Genomics project.[31] The average coverage was 34X with a range of 27X to 71X. The compressed reference-oriented alignment map files are available at the International Genome Sample Resource https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. To survey the frequencies of CTG18.1 repeat expansions, we focused on the 2504 unrelated samples from phase III of 1KGP.[32] The samples were from 26 populations across 5 continental ancestry groups: AFR, EUR, EAS, South Asian, and admixed American. There were 4 samples (HG03745, HG03874, NA19428, and

NA19437) without sequence coverage at the *TCF4* CTG18.1 locus. In the end, there were a total of 2500 samples—n = 659, 503, 504, 487, and 347 in AFR, EUR, EAS, South Asian, and admixed American, respectively—with the repeat number estimated.

The CTG18.1 repeat length was inferred using ExpansionHunter (v5.0.0) and visualized using REViewer.[33,34] ExpansionHunter estimates sizes of repeats by performing a targeted search through a binary alignment map/compressed reference-oriented alignment map file for reads that span, flank, and are fully contained in the repeats. A repeat fragment shorter than the read length can be measured exactly, and a repeat fragment longer than the read length is estimated by modeling the read length and counts. In the human reference genome Genome Reference Consortium Human Build 38, the *TCF4* CTG18.1 repeats structure can be described as either cytosine-adenine-guanine $(CAG)_{24}$ spanning chr18:55,586,155-55,586,227 or (adenine-guanine-cytosine $(AGC)_{25}$ spanning chr18:55,586,154-55,586,229. In the current study we used the structure $(AGC)_n$ to be consistent with the repeats track by RepeatMasker v4.0.7 Dfam 2.0 in the University of California, Santa Cruz genome browser.

Correlation between CTG18.1 Genotyping Approaches

To examine the accuracy of ExpansionHunter in estimating the number of CTG18.1 repeats, we also called the repeats number of 4 samples (HG00731, HG00732, NA19238, and NA19239) that underwent long-read HiFi sequencing by Pacific Biosciences. Note that there was one more sample—HG00513—documented[35] to have HiFi sequencing reads, which, however, was absent from the European Nucleotide Archive sequence read archive file transfer protocol site (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/HGSVC2_pacbio.index). HiFi reads were aligned to human hg38 genome using pbmm2 (v1.12.0; https://github.com/PacificBiosciences/pbmm2). The CTG18.1 repeat size was called using Tandem Repeat Genotyper (v0.4.0) and visualized with Tandem Repeat Visualization (v0.4.0).[36]

Genomic DNA of peripheral blood from 2 cornea clinic patients from UT Southwestern (VVM20 and VVM127) underwent Illumina short-read (2x150bp) WGS sequencing with >30X coverage. We genotyped their CTG18.1 repeat expansion by both ExpansionHunter and STR analysis to assess concordance between these 2 genotyping approaches.

Results

The distributions of the CTG18.1 repeat length and the expanded allele *L* in the 3 ethnic groups in the DHS measured by targeted PCR amplification followed by fragment analysis are summarized in Table 1 and Figure 1. The median repeat length was ~17 with the interquartile range between 12 and 23. The frequencies of the *L* allele were highest in EAs (0.042) and the lowest in AAs (0.016). The carrier rates of the expanded allele *L* were 3.1%, 8.1%, and 3.3% in AAs, EAs, and Latinos, respectively, in the DHS (Table 1). There were 1, 3, and 1 homozygous carriers in AAs, EAs, and Latinos, respectively.

The distributions of the CTG18.1 repeat length and the expanded allele *L* in the 5 continental ancestry groups measured by ExpansionHunter are summarized and compared with the DHS in Table 1 and Figure 1. The distribution of the CTG18.1 repeat measured by ExpansionHunter in the 1KGP is similar to that measured by STR/TP-PCR in the DHS. The median repeat length was ~19 in all the 1KGP populations with the interquartile range between 13 and 26. The frequencies of the *L* allele were $\leq$ 0.050 with the highest in EUR (0.050) and the lowest in AFR (0.014). Accordingly, the carrier rates were the highest in EUR (9.5%) and the lowest in AFR (2.7%). There were a total of 5 homozygous carriers observed (HG00264, HG02657, NA12003, NA19786, and NA20812). There was no statistically significant difference in the distribution of the expanded allele *L* between males and females in each subpopulation of both studies (Table S2, available at www.ophthalmologyscience.org).

Table 1. Distribution of *TCF4* CTG18.1 Repeat Expansion* and Its Correlation with rs613872 in Different Populations of the Dallas Heart Study and 1000 Genomes Project

Population	N	Median (IQR)	CTG18.1 "*L*" Allele Carrier Counts (%)	CTG18.1 "*L*" Allele Frequency	rs613872 "G" Allele Frequency	*D'*	r^2
Dallas Heart Study							
AAs	1599	18 (15, 23)	49 (3.1)	0.016	0.026	0.371	0.082
EAs	1028	16 (12, 18)	83 (8.1)	0.042	0.157	0.874	0.180
Latinos	458	16.5 (12, 18)	15 (3.3)	0.018	0.074	0.712	0.113
1000 Genomes Project							
AFR	659	19 (17, 24)	18 (2.7)	0.014	0.005	0.267	0.027
EUR	503	17 (13, 19)	48 (9.5)	0.050	0.157	0.753	0.159
EAS	504	19 (13, 26)	26 (5.2)	0.026	0.004	1.000	<0.001
SAS	487	19 (13, 26)	35 (7.2)	0.037	0.098	0.057	0.001
AMR	347	18 (13, 19)	18 (5.2)	0.027	0.079	0.450	0.066

AAs = African Americans; AFR = African; AMR = admixed American; EAs = European Americans; EAS = East Asian; EUR = European; IQR = interquartile range; SAS = South Asian; *TCF4* = transcription factor 4 gene.
*The *TCF4* CTG18.1 repeat expansion genotype was determined by a combination of short-tandem repeat analysis and triplet repeat primed polymerase chain reaction in the Dallas Heart Study and inferred by ExpansionHunter based on the Illumina whole genome short reads sequences in the 1000 Genomes Project. The *TCF4* CTG18.1 trinucleotide alleles ≥40 repeats were coded as "*L*." The 1000 Genomes Project individuals were classified into 5 continental ancestry groups: African, European, East Asian, South Asian, and admixed American.

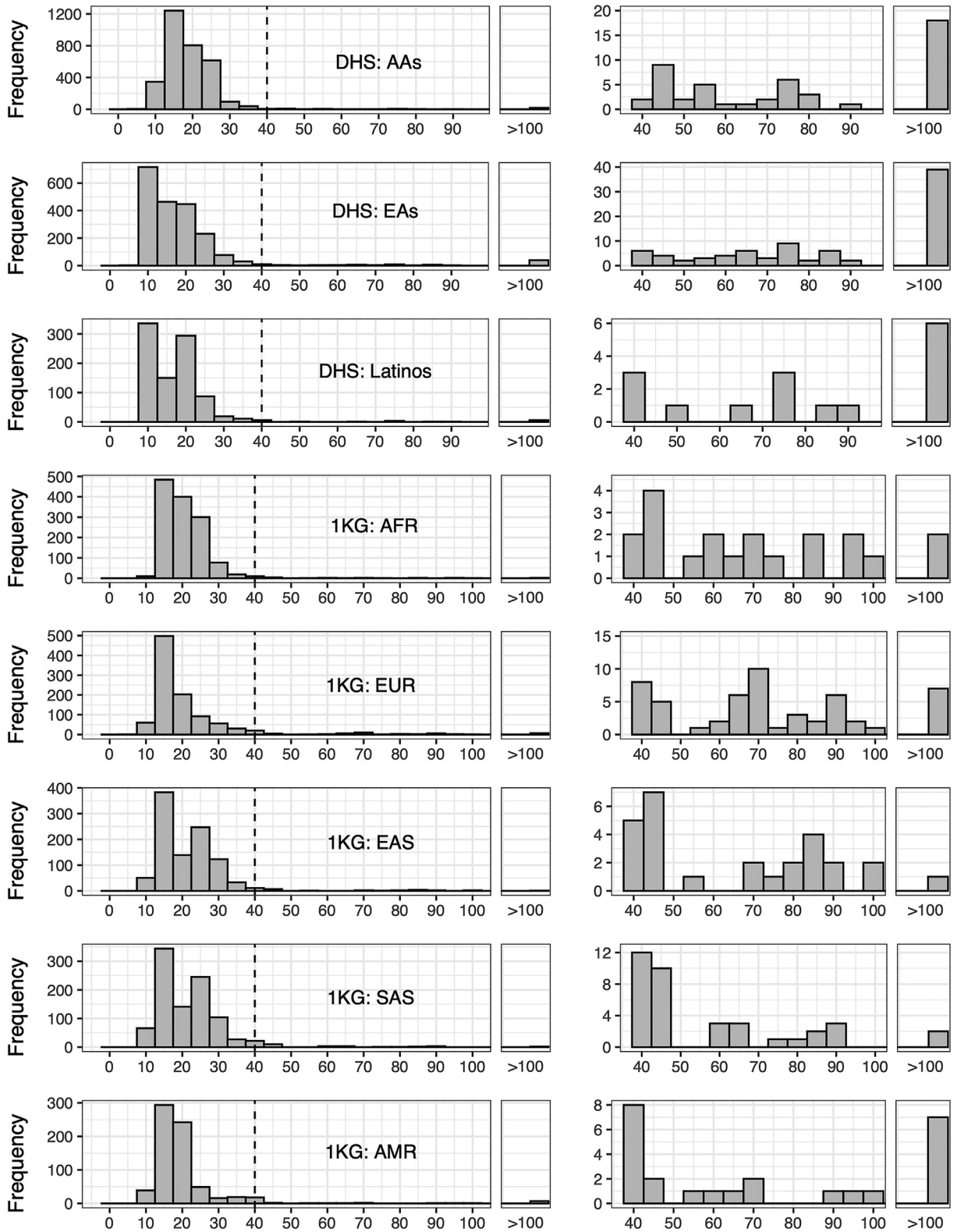

Figure 1. Distribution of *TCF4* CTG18.1 trinucleotide repeat size. For the DHS subjects—AAs (n = 1599), EAs (n = 1028), and Latinos (n = 458), the repeat length was measured by short tandem repeat analysis and triplet repeat primed polymerase chain amplification. For the 1000 Genome Project subjects—AFR (n = 659), EUR (n = 503), EAS (n = 504), SAS (n = 487), and AMR (n = 347)—the repeat length was estimated by ExpansionHunter based on Illumina short-read sequencing (2x150bp). The vertical dashed line indicates the cutoff (40 repeats) for expanded alleles. The x-axis denotes the repeat size. The short tandem repeat analysis could detect CTG18.1 alleles with repeat lengths up to ~95 CTG repeats, and the triplet repeat primed

The *L* and *S* alleles were in Hardy−Weinberg equilibrium ($P > 0.05$) in each subpopulation of both studies, and so were the rs613872 alleles. The rs613872 G allele was most prevalent in individuals of EUR ancestry—frequency equal to 0.157 in both EUR of 1KGP and EAs of DHS, and the 2 loci were in moderate linkage disequilibrium—$D' = 0.753$ and 0.874 in EUR and EA, respectively, which could explain that the FECD genome-wide association studies signal was first detected at rs613872 in EUR descendants.[5] In EAS, there were only 4 copies of the rs613872 G allele, and all of them were on the same haplotype with a CTG18.1 *S* allele, which led to an inflated D' of 1.0, but the r^2 was <0.001.

The distributions in the 26 subpopulations of the 1KGP are summarized in Table 3. Of note, the highest carrier rates of the expanded allele *L* were 12.5%, 12.1%, 12.1%, and 11.2% in subjects of Mexican ancestry from Los Angeles, Great Britain, Utah residents with Northern and Western European ancestry, and Toscani in Italy subpopulations, respectively. The expanded allele *L* was absent in the West AFR subpopulations of the Mende in Sierra Leone and Esan in Nigeria. Interestingly, however, there was a 6.2% prevalence of the expanded allele in the East AFR subpopulation, Luhya in Webuye, Kenya.

Of the 4 1KGP samples (HG00731, HG00732, NA19238, and NA19239) with both Illumina short-read and Pacific Biosciences long-read sequences, the repeat numbers are consistent between the 2 approaches but none of these 4 samples carried an expanded allele (Fig S2, available at www.ophthalmologyscience.org). Of the 2 clinical samples (VVM20 and VVM127) who were known to carry an expanded allele as previously genotyped by STR analysis and now assayed by Illumina short-read WGS, the shorter allele measures by the 2 methods were identical, and the ExpansionHunter estimates of the longer alleles were greater than that of STR analysis (Fig S3, available at www.ophthalmologyscience.org).

Discussion

The *TCF4* CTG18.1 STR can be genotyped using targeted PCR amplification followed by fragment analysis.[10] This approach combining STR analysis and TP-PCR is an effective method in resolving zygosity and detecting expanded alleles and has been used in many FECD association studies.[9,10,37] However, this approach cannot accurately determine the repeat length of very large expanded alleles without use of the Southern blot technique, which is labor- and time-intensive; therefore, it may not be practical to apply them to large scale studies. In recent years amplification-free sequencing methods have been developed that enrich targeted DNA by clustered regularly interspaced palindromic repeats (CRISPR)/CRISPR associated protein 9 and sequence by long-read single molecule sequencing platforms.[38,39] The long-read sequencing data can not only provide accurate measures of repeat length but also reveal the dynamic instability of the expanded alleles. While long-read sequencing strategies are increasing in popularity, they are still expensive for large-scale genetic studies. In the meantime, many methods to genotype STRs based on Illumina short-read sequences have been developed.[33,34,40,41] These methods can infer the length of a repeat allele that is longer than the sequence read length, and some can even detect novel STRs. There have been novel pathogenic repeats detected by this approach (for a review see, e.g., reference[42]). In this study, we compared the STR analysis with Illumina short-read sequencing with use of ExpansionHunter and established concordance between these genotyping approaches to discriminate between normal and pathogenic expanded *TCF4* CTG18.1 alleles. ExpansionHunter can measure the exact size of repeats shorter than the read length but can only estimate the size of repeats longer than the read length by statistical modeling. In the 1KGP the read length was 150 bp, thus the CTG18.1 repeats with size <50 were accurately measured. As the criterion of ≥ 40 repeats was used to define an expanded allele, it is valid to use ExpansionHunter to estimate the prevalence of the CTG18.1 repeat expanded allele *L*, even if the estimates of the repeat number of an expanded allele may not be exact. ExpansionHunter is a targeted tool that requires an STR to be specified by its reference coordinates and repeat motif. Therefore, without preknowledge it cannot automatically discover repeats interruption, a mechanism that influences the age of onset of Huntington's disease.[43−45] Therefore, we did visual inspection of the ExpansionHunter aligned reads of 19 alleles with the estimated size ≥ 100 as well as 20 random samples using REViewer and did not find credible insertions that are supported by >2 reads. It will be of interest to examine whether the same mechanism applies to FECD by long-read sequencing.

Our results indicate that the expanded *TCF* CTG18.1 allele is represented in all broad genetic ancestries. In the DHS, prevalence of the expanded allele *L* was highest in EAs and lowest in AAs. One caveat on the frequency estimates of the expanded allele in the DHS is that they were based on individuals classified by their self-identified race and ethnicity. On one hand, the genetic make-up of the United States population is shaped by migration from distant continents and admixture of migrants and Native Americans; on the other hand, the

polymerase chain amplification assay could detect the presence of a large expanded CTG18.1 allele but cannot measure its exact repeat length. Therefore, we broke the x-axis at the repeat size of 100 and stacked all the alleles >100 together. To make the style consistent, we broke the x-axis at 100 for the 1KGP data, too, though the exact repeat size measured by ExpansionHunter is up to 50. The distributions of all alleles were shown on the left column and the distributions of expanded alleles were amplified on the right column. Note that in the current study we used the structure $(AGC)_n$ to be consistent with the repeats track by RepeatMasker v4.0.7 Dfam 2.0 in the UCSC genome browser—$(AGC)_{25}$ spanning chr18:55,586,154-55,586,229 in the human reference genome GRCh38. 1KGP = 1000 Genomes Project; AAs = African Americans; AFR = African; AGC = adenine-guanine-cytosine; AMR = admixed American; CTG = cytosine-thymine-guanine; DHS = Dallas Heart Study; EAs = European Americans; EAS = East Asian; EUR = European; GRCh38 = Genome Reference Consortium Human Build 38; SAS = South Asian; UCSC = University of California, Santa Cruz.

Table 3. Distribution of *TCF4* CTG18.1 Repeat Expansion* in the 26 Populations of 1000 Genomes Project

Superpopulation	Subpopulation	N	Median (IQR)	CTG18.1 "*L*" Allele Carrier Counts (%)	CTG18.1 "*L*" Allele Frequency
SAS	BEB	86	19 (13, 26)	5 (5.8)	0.029
	PJL	96	17 (13, 26)	8 (8.3)	0.047
	ITU	101	19.5 (13, 27)	9 (8.9)	0.045
	STU	101	20.5 (13, 26)	5 (5.0)	0.025
	GIH	103	20 (13, 26)	8 (7.8)	0.039
EUR	GBR	91	16.5 (13, 19)	11 (12.1)	0.066
	CEU	99	17 (13, 24)	12 (12.1)	0.066
	FIN	99	16 (13, 19)	5 (5.1)	0.025
	IBS	107	17 (13, 19)	8 (7.5)	0.037
	TSI	107	17 (13, 26)	12 (11.2)	0.061
EAS	CDX	93	21.5 (13, 27.8)	8 (8.6)	0.043
	KHV	99	19 (13, 26)	4 (4.0)	0.020
	CHB	103	19 (13, 26.75)	4 (3.9)	0.019
	JPT	104	19 (13, 26)	4 (3.8)	0.019
	CHS	105	19 (13, 27)	6 (5.7)	0.029
AMR	MXL	64	18 (13, 19)	8 (12.5)	0.070
	PEL	85	18 (13, 19)	3 (3.5)	0.018
	CLM	94	19 (13, 21)	5 (5.3)	0.027
	PUR	104	16.5 (13, 19)	2 (1.9)	0.010
AFR	ASW	61	19 (17, 23)	4 (6.6)	0.033
	MSL	85	18 (16, 23.8)	0 (0.0)	0
	ACB	96	20 (16.8, 24)	4 (4.2)	0.021
	LWK	97	20 (17, 23)	6 (6.2)	0.031
	ESN	99	20 (17, 24)	0 (0.0)	0
	YRI	108	19 (17, 23)	2 (1.9)	0.009
	GWD	113	19 (17, 25)	2 (1.8)	0.009

ACB = African Caribbean in Barbados; AFR = African; AMR = admixed American; ASW = African ancestry in Southwest United States; BEB = Bengali in Bangladesh; CDX = Chinese Dai in Xishuangbanna, China; CEU = Utah residents (CEPH) with Northern and Western European ancestry; CHB = Han Chinese in Beijing, China; CHS = Han Chinese South; CLM = Colombian in Medellin, Colombia; EAS = East Asian; ESN = Esan in Nigeria; EUR = European; FIN = Finnish in Finland; GBR = British in England and Scotland; GIH = Gujarati Indians in Houston, Texas; GWD = Gambian in Western Division, The Gambia − Mandinka; IBS = Iberian populations in Spain; IQR = interquartile range; ITU = Indian Telugu in the United Kingdom; JPT = Japanese in Tokyo, Japan; KHV = Kinh in Ho Chi Minh City, Vietnam; LWK = Luhya in Webuye, Kenya; MSL = Mende in Sierra Leone; MXL = Mexican ancestry in Los Angeles, California; PEL = Peruvian in Lima, Peru; PJL = Punjabi in Lahore, Pakistan; PUR = Puerto Rican in Puerto Rico; SAS = South Asian; STU = Sri Lankan Tamil in the United Kingdom; *TCF4* = transcription factor 4 gene; TSI = Toscani in Italy; YRI = Yoruba in Ibadan, Nigeria.

The 1000 Genomes Project individuals were classified into 5 continental ancestry groups: AFR, European, EAS, SAS, and AMR from 26 populations—ACB, ASW, YRI, GWD, MSL, ESN, LWK, PUR, CLM, PEL, MXL, CHS, KHV, JPT, CHB, CDX, FIF, CEU, IBS, TSI, GBR, BEB, PJL, GIH, STU, and ITU.

*The *TCF4* CTG18.1 repeat expansion genotype was inferred by ExpansionHunter based on the Illumina whole genome short reads sequences in the 1000 Genomes Project. The *TCF4* CTG18.1 trinucleotide alleles ≥40 repeats were coded as "*L*."

genetic ancestry of self-described groups varies across geographic regions. Therefore, the allele frequency estimates have large variance and can be quite different from estimates from populations of different locations.[46,47] In the 1KGP, the prevalence of the expanded allele *L* was highest in subjects from the EUR continent, especially in northwestern EUR subpopulations, lowest in subjects from the AFR continent, and totally absent in some subpopulations of Western Africa. Interestingly, however, there was 6.2% prevalence of the expanded allele in the LWD subpopulation of East Africa. Based on prevalence of the 1KGP subpopulations, we hypothesize that the expanded CTG18.1 repeat allele may have originated in East Africa, where some argue our species evolved and migrated out of Africa.[48]

Few studies have surveyed the prevalence of FECD on a large population-based scale. The prevalence of FECD in the United States and Europe is generally thought to be in the 4% to 5% range in individuals over the age of 40 years.[1,2,49,50] A study of central corneal guttae in 1016 people conducted by Lorenzetti et al in 1967 is often cited for the prevalence of FECD in the United States.[50] By reanalysis of the primary data from this classic paper, we estimate the prevalence of FECD (as defined as ≥1−2 mm of central confluent guttae) to be 2.5% and 6.6% in AAs and EAs, respectively, of age >40 years. A

population-based study in Iceland found a prevalence of central guttae in 11% of females and 7% of males over the age of 55 years[51] while another study conducted in Japan reported central guttae in 4.1% of individuals over the age of 40 years.[52] A comparative study found the incidence of corneal guttae was significantly higher in Singaporeans than in the Japanese.[53] Including studies in various clinical settings, it is consistently observed that the FECD prevalence is higher in people of EUR ancestry than in other ancestral groups; in the United States, the prevalence is higher in EAs than in AAs and Latinos (for a review, see, e.g., references[26,49]). Notably, it is in line with the frequencies of CTG18.1 repeat expansion alleles across populations.

We hypothesize that the prevalence of the expanded CTG18.1 alleles in different populations may be greater than the actual incidence of FECD in these groups due to incomplete penetrance. The focus of studies to date has been on determining the proportion of the expanded CTG18.1 allele carriers in FECD patients. All the association studies between CTG18.1 and FECD (including ours) have been based on the recruitment of patients with FECD and unaffected controls presenting to tertiary care eye clinics and thus, may be subject to ascertainment bias.[26] About 3% to 11% of control subjects in these association studies conducted on population of EUR ancestry had the expanded CTG18.1 allele without any findings of FECD. Population-based studies of carriers of the expanded CTG18.1 allele in different ethnic groups are required to determine unbiased estimates of the penetrance of the expanded allele and to study the impact of other factors such as gender and environmental factors.

Not only is the CTG18.1 repeat expansion causally associated with FECD, but it also shows high specificity—>90% in nearly all studies of various ethnicities, which makes it a potential biomarker for prediction, diagnosis, and prognosis of FECD. Besides the ethnic disparity due to genetic background, FECD is more prevalent in females.[54,55] However, we found no sex bias for the prevalence of the CTG18.1 expansion in the general populations. It is been hypothesized that disease pathogenesis may be influenced by environmental factors such as smoking and exposure to ultraviolet light.[51,56] Future studies to include the CTG18.1 repeat expansion dosage as well as other factors to construct an FECD prediction model are warranted.

DNA repeat expansion disorders (REDs) are a heterogenous group of diseases that result in neurodegenerative disease including myotonic dystrophy, Huntington's disease, and the common form of amyotrophic lateral sclerosis and frontotemporal dementia. Previous studies of neurodegenerative disorders mediated by expansion of repetitive DNA sequences within their respective genes estimated that REDs affect 1 in 3000 humans with population based differences at specific RED loci.[57] In a survey of EUR-based cohorts of 9 REDs including Huntington's disease, 1.3% of participants had expanded alleles in the disease causing genes.[58] In comparison, the high prevalence of the expanded *TCF4* CTG18.1 alleles in all broad ancestries indicates that it is the most common disease-causing STR in humans.

Availability of Data and Materials

The 1KG Illumina short read WGS data can be accessed through https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

The Pacbio HiFi sequencing data for the 4 samples (HG00731, HG00732, NA19238, and NA19239) can be accessed through https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/HGSVC2_pacbio.index.

Acknowledgments

The authors acknowledge the Texas Advanced Computing Center (https://www.tacc.utexas.edu) at The University of Texas at Austin for providing high performance computing resources that have contributed to the research results reported within this paper.

Footnotes and Disclosures

Originally received: March 15, 2024.
Final revision: August 20, 2024.
Accepted: August 22, 2024.
Available online: August 30, 2024. Manuscript no. XOPS-D-24-00087.

[1] Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas.

[2] Department of Ophthalmology, University of Texas Southwestern Medical Center, Dallas, Texas.

[3] Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas.

[4] O'Donnell School of Public Health, University of Texas Southwestern Medical Center, Dallas, Texas.

*X.Z. and A.K. contributed equally to this work.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

V.V.M.: Patents planned, issued or pending — Co-inventor on patent application related to treatments and biomarkers for Fuchs' dystrophy.

This study was supported by grants R01EY022161 and P30 EY030413 (V.V.M.) from the National Institutes of Health, Bethesda, MD and a Challenge Grant from Research to Prevent Blindness, New York.

HUMAN SUBJECTS: Human subjects were included in this study. The study protocol had the approval of the institutional review board of the University of Texas Southwestern Medical Center and was in compliance with the tenets of the Declaration of Helsinki. Study participants were enrolled after written informed consent.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Xing, Mootha

Data collection: Zhang, Kumar, Gong, Xing, Mootha

Analysis and interpretation: Zhang, Kumar, Gong, Xing, Mootha

Obtained funding: Mootha

Overall responsibility: Zhang, Kumar, Gong, Xing, Mootha

Abbreviations and Acronyms:
1KGP = 1000 Genomes Project; **AA** = African American; **AFR** = African; **CTG** = cytosine-thymine-guanine; **DHS** = Dallas Heart Study; **EA** = European American; **EAS** = East Asian; **EUR** = European; **FECD** = Fuchs' endothelial corneal dystrophy; **PCR** = polymerase chain amplification; **RED** = repeat expansion disorder; **STR** = short tandem repeat; ***TCF4*** = transcription factor 4 gene; **TP-PCR** = triplet repeat primed polymerase chain amplification; **WGS** = whole genome sequencing.

Key words:
CTG18.1 trinucleotide repeat expansion, Fuchs' endothelial corneal dystrophy, Repeat expansion disorder, Short tandem repeat, *TCF4*.

Correspondence:
Chao Xing, PhD, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, MC 8591, Dallas, TX 75390. E-mail: chao.xing@utsouthwestern.edu; and V. Vinod Mootha, MD, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, MC 9057, Dallas, TX 75390. E-mail: vinod.mootha@utsouthwestern.edu.

References

1. Krachmer JH, Purcell Jr JJ, Young CW, Bucher KD. Corneal endothelial dystrophy. A study of 64 families. *Arch Ophthalmol*. 1978;96:2036–2039.
2. Aiello F, Gallo Afflitto G, Ceccarelli F, et al. Global prevalence of fuchs endothelial corneal dystrophy (FECD) in adult population: a systematic review and meta-analysis. *J Ophthalmol*. 2022;2022:3091695.
3. Gain P, Jullienne R, He Z, et al. Global survey of corneal transplantation and eye banking. *JAMA Ophthalmol*. 2016;134:167–173.
4. Eye Bank Association of America. *2019 Eye Banking Statistical Report*. Eye Bank Association of America; 2020. https://restoresight.org/wp-content/uploads/2020/04/2019-EBAA-Stat-Report-FINAL.pdf. Accessed September 23, 2024.
5. Baratz KH, Tosakulwong N, Ryu E, et al. E2-2 protein and Fuchs's corneal dystrophy. *N Engl J Med*. 2010;363:1016–1024.
6. Afshari NA, Igo RP, Morris NJ, et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat Commun*. 2017;8:14898.
7. Gorman BR, Francis M, Nealon CL, et al. A multi-ancestry GWAS of Fuchs corneal dystrophy highlights the contributions of laminins, collagen, and endothelial cell regulation. *Commun Biol*. 2024;7:418.
8. Murre C, Bain G, van Dijk MA, et al. Structure and function of helix-loop-helix proteins. *Biochim Biophys Acta Gene Struct Expr*. 1994;1218:129–135.
9. Wieben ED, Aleff RA, Tosakulwong N, et al. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts fuchs corneal dystrophy. *PLoS One*. 2012;7:e49083.
10. Mootha VV, Gong X, Ku HC, Xing C. Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in fuchs' endothelial corneal dystrophy. *Invest Ophthalmol Vis Sci*. 2014;55:33–42.
11. Xing C, Gong X, Hussain I, et al. Transethnic replication of association of CTG18.1 repeat expansion of TCF4 gene with fuchs' corneal dystrophy in Chinese implies common causal variant. *Invest Ophthalmol Vis Sci*. 2014;55:7073–7078.
12. Nanda GG, Padhy B, Samal S, et al. Genetic association of TCF4 intronic polymorphisms, CTG18.1 and rs17089887, with fuchs' endothelial corneal dystrophy in an Indian population. *Invest Ophthalmol Vis Sci*. 2014;55:7674–7680.
13. Eghrari AO, Vahedi S, Afshari NA, et al. CTG18.1 expansion in TCF4 among african Americans with fuchs' corneal dystrophy. *Invest Ophthalmol Vis Sci*. 2017;58:6046–6049.
14. Rao BS, Tharigopala A, Rachapalli SR, et al. Association of polymorphisms in the intron of TCF4 gene to late-onset Fuchs endothelial corneal dystrophy: an Indian cohort study. *Indian J Ophthalmol*. 2017;65:931–935.
15. Okumura N, Hayashi R, Nakano M, et al. Association of rs613872 and trinucleotide repeat expansion in the TCF4 gene of German patients with fuchs endothelial corneal dystrophy. *Cornea*. 2019;38:799–805.
16. Okumura N, Puangsricharern V, Jindasak R, et al. Trinucleotide repeat expansion in the transcription factor 4 (TCF4) gene in Thai patients with Fuchs endothelial corneal dystrophy. *Eye*. 2020;34:880–885.
17. Kuot A, Hewitt AW, Snibson GR, et al. TGC repeat expansion in the TCF4 gene increases the risk of Fuchs' endothelial corneal dystrophy in Australian cases. *PLoS One*. 2017;12:e0183719.
18. Skorodumova LO, Belodedova AV, Antonova OP, et al. CTG18.1 expansion is the best classifier of late-onset fuchs' corneal dystrophy among 10 biomarkers in a cohort from the European part of Russia. *Invest Ophthalmol Vis Sci*. 2018;59:4748–4754.
19. Foja S, Luther M, Hoffmann K, et al. CTG18.1 repeat expansion may reduce TCF4 gene expression in corneal endothelial cells of German patients with Fuchs' dystrophy. *Graefes Arch Clin Exp Ophthalmol*. 2017;255:1621–1631.
20. Vasanth S, Eghrari AO, Gapsis BC, et al. Expansion of CTG18.1 trinucleotide repeat in TCF4 is a potent driver of fuchs' corneal dystrophy. *Invest Ophthalmol Vis Sci*. 2015;56:4531–4536.
21. Zarouchlioti C, Sanchez-Pintado B, Hafford Tear NJ, et al. Antisense therapy for a common corneal dystrophy ameliorates TCF4 repeat expansion-mediated toxicity. *Am J Hum Genet*. 2018;102:528–539.
22. Mootha VV, Hussain I, Cunnusamy K, et al. TCF4 triplet repeat expansion and nuclear RNA foci in fuchs' endothelial corneal dystrophy. *Invest Ophthalmol Vis Sci*. 2015;56:2003–2011.
23. Du J, Aleff RA, Soragni E, et al. RNA toxicity and missplicing in the common eye disease fuchs endothelial corneal dystrophy. *J Biol Chem*. 2015;290:5979–5990.
24. Chu Y, Hu J, Liang H, et al. Analyzing pre-symptomatic tissue to gain insights into the molecular and mechanistic origins of late-onset degenerative trinucleotide repeat disease. *Nucleic Acids Res*. 2020;48:6740–6758.
25. Zhang X, Kumar A, Sathe AA, et al. Transcriptomic meta-analysis reveals ERRalpha-mediated oxidative phosphorylation is downregulated in Fuchs' endothelial corneal dystrophy. *PLoS One*. 2023;18:e0295542.
26. Fautsch MP, Wieben ED, Baratz KH, et al. TCF4-mediated Fuchs endothelial corneal dystrophy: insights into a common trinucleotide repeat-associated disease. *Prog Retin Eye Res*. 2021;81:100883.
27. Breschel TS, McInnis MG, Margolis RL, et al. A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene

on chromosome 18q21.1. *Hum Mol Genet*. 1997;6:1855−1863.
28. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061−1073.
29. Victor RG, Haley RW, Willett DL, et al. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol*. 2004;93:1473−1480.
30. Byrska-Bishop M, Evani US, Zhao X, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022;185:3426−3440.e19.
31. Regier AA, Farjoun Y, Larson DE, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun*. 2018;9:4038.
32. The 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526:68−74.
33. Dolzhenko E, van Vugt JJFA, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res*. 2017;27:1895−1903.
34. Dolzhenko E, Deshpande V, Schlesinger F, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*. 2019;35:4754−4756.
35. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372:eabf7117.
36. Dolzhenko E, English A, Dashnow H, et al. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol*. 2024;42:1606−1614.
37. Warner JP, Barron LH, Goudie D, et al. A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J Med Genet*. 1996;33:1022−1026.
38. Wieben ED, Aleff RA, Basu S, et al. Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. *PLoS One*. 2019;14:e0219446.
39. Hafford-Tear NJ, Tsai YC, Sadan AN, et al. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy−associated TCF4 triplet repeat. *Genet Med*. 2019;21:2092−2102.
40. Dolzhenko E, Bennett MF, Richmond PA, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol*. 2020;21:102.
41. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res*. 2019;47:e90.
42. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Res*. 2022;32:1−27.
43. Ciosi M, Maxwell A, Cumming SA, et al. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine*. 2019;48:568−580.
44. Lee J-M, Correia K, Loupe J, et al. CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell*. 2019;178:887−900.e14.
45. Wright GEB, Collins JA, Kay C, et al. Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *Am J Hum Genet*. 2019;104:1116−1126.
46. Bryc K, Durand Eric Y, Macpherson JM, et al. The genetic ancestry of african Americans, Latinos, and European Americans across the United States. *Am J Hum Genet*. 2015;96:37−53.
47. Dai CL, Vazifeh MM, Yeang CH, et al. Population histories of the United States revealed through fine-scale migration and haplotype analysis. *Am J Hum Genet*. 2020;106:371−388.
48. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. *Science*. 2009;324:1035−1044.
49. Eghrari AO, Riazuddin SA, Gottsch JD. Chapter seven - Fuchs corneal dystrophy. In: Hejtmancik JF, Nickerson JM, eds. *Progress in Molecular Biology and Translational Science*. 134. Amsterdam: Elsevier; 2015:79−97.
50. Lorenzetti DWC, Uotila MH, Parikh N, Kaufman HE. Central cornea guttata: incidence in the general population. *Am J Ophthalmol*. 1967;64:1155−1158.
51. Zoega GM, Fujisawa A, Sasaki H, et al. Prevalence and risk factors for cornea guttata in the reykjavik eye study. *Ophthalmology*. 2006;113:565−569.
52. Higa A, Sakai H, Sawaguchi S, et al. Corneal endothelial cell density and associated factors in a population-based study in Japan: the kumejima study. *Am J Ophthalmol*. 2010;149:794−799.
53. Kitagawa K, Kojima M, Sasaki H, et al. Prevalence of primary cornea guttata and morphology of corneal endothelium in aging Japanese and Singaporean subjects. *Ophthalmic Res*. 2002;34:135−138.
54. Yeh P, Colby K. Corneal endothelial dystrophies. In: Foster C, Azar D, DH D, eds. *Smolin and Thoft's The Cornea: Scientific Foundations and Clinical Practice*. Philadelphia, PA: Lippincott Williams & Wilkins; 2015:849−873.
55. Wilson SE, Bourne WM. Fuchs' dystrophy. *Cornea*. 1988;7:2−18.
56. Liu C, Miyajima T, Melangath G, et al. Ultraviolet A light induces DNA damage and estrogen-DNA adducts in Fuchs endothelial corneal dystrophy causing females to be more affected. *Proc Natl Acad Sci U S A*. 2020;117:573−583.
57. Paulson H. Repeat expansion diseases. *Handb Clin Neurol*. 2018;147:105−123.
58. Gardiner SL, Boogaard MW, Trompet S, et al. Prevalence of carriers of intermediate and pathological polyglutamine disease-associated alleles among large population-based cohorts. *JAMA Neurol*. 2019;76:650−656.