nature medicine

Article

https://doi.org/10.1038/s41591-024-03190-5

Increased frequency of repeat expansion mutations across different populations

Received: 22 June 2023

Accepted: 11 July 2024

Published online: 1 October 2024

Check for updates

Kristina Ibañez[1], Bharati Jadhav[2], Matteo Zanovello[3], Delia Gagliardi[1,3], Christopher Clarkson[1], Stefano Facchini[3,4], Paras Garg[2], Alejandro Martin-Trujillo[2], Scott J. Gies[2], Valentina Galassi Deforie[3], Anupriya Dalmia[5], Davina J. Hensman Moss[6,7], Jana Vandrovcova[3], Clarissa Rocca[3], Loukas Moutsianas[8], Chiara Marini-Bettolo[9], Helen Walker[9], Chris Turner[10], Maryam Shoai[7], Jeffrey D. Long[11], Pietro Fratta[3], Douglas R. Langbehn[11], Sarah J. Tabrizi[5,7,12], Mark J. Caulfield[1], Andrea Cortese[3,4], Valentina Escott-Price[13,14], John Hardy[7], Henry Houlden[7,15], Andrew J. Sharp[2] & Arianna Tucci[1,3]

Repeat expansion disorders (REDs) are a devastating group of predominantly neurological diseases. Together they are common, affecting 1 in 3,000 people worldwide with population-specific differences. However, prevalence estimates of REDs are hampered by heterogeneous clinical presentation, variable geographic distributions and technological limitations leading to underascertainment. Here, leveraging whole-genome sequencing data from 82,176 individuals from different populations, we found an overall disease allele frequency of REDs of 1 in 283 individuals. Modeling disease prevalence using genetic data, age at onset and survival, we show that the expected number of people with REDs would be two to three times higher than currently reported figures, indicating underdiagnosis and/or incomplete penetrance. While some REDs are population specific, for example, Huntington disease-like 2 in Africans, most REDs are represented in all broad genetic ancestries (that is, Europeans, Africans, Americans, East Asians and South Asians), challenging the notion that some REDs are found only in specific populations. These results have worldwide implications for local and global health communities in the diagnosis and counseling of REDs.

Repeat expansion disorders (REDs) are a heterogeneous group of conditions that mainly affect the nervous system and include fragile X syndrome, the most common inherited form of amyotrophic lateral sclerosis and frontotemporal dementia (*C9orf72*-ALS/FTD)[1] and inherited ataxias (Friedreich ataxia (FA), *RFC1*-CANVAS (cerebellar ataxia, neuropathy, vestibular areflexia syndrome)[2]). REDs are caused by the same underlying mechanism: the expansion of short repetitive DNA sequences (1–6 bp) within their respective genes. The mutational process is gradual; normal alleles are usually passed stably from parent to child with rare changes in repeat size, and intermediate-size alleles are more likely to expand into the disease range, giving rise to pathogenic repeat lengths in the next generation. Repeat lengths are classified in ascending order as normal, intermediate, premutation, reduced penetrance or full mutations, though this classification is not universal and not all RED loci have well-defined ranges for intermediate or reduced penetrance range.

REDs are clinically heterogeneous. For example, *C9orf72* expansions can present as either FTD or ALS even within the same family;

A full list of affiliations appears at the end of the paper. e-mail: a.tucci@qmul.ac.uk

and one in three patients carrying the repeat expansion in *C9orf72* shows an atypical presentation at onset such as Alzheimer's and Huntington disease (HD) among others[3,4]. For many REDs, the variability in repeat lengths underlines the substantial clinical heterogeneity[5]; longer repeats cause more severe disease and earlier symptom onset[6].

Previous studies have estimated that REDs affect 1 in 3,000 people[7]. Despite their broad distribution in human populations, few global epidemiological studies have been performed. In these studies, prevalence estimates are either population based, in which affected individuals are identified on the basis of clinical presentation, or genetically tested on the basis of the presence of a relative with a RED. Given that one of the most striking features of REDs is that they can present with markedly diverse phenotypes, REDs can remain unrecognized, leading to underestimation of the disease prevalence[8].

While many of the epidemiological studies so far have been conducted in cohorts of European origin, studies in other ancestries have highlighted population differences at specific RED loci[9–11]. Among the most common REDs, myotonic dystrophy type 1 (DM1) affects 1 in 8,000 people worldwide[12], ranging from 1 in 10,000 in Iceland to 1 in 100,000 in Japan[13]. Similarly, HD prevalence ranges from 0.1 in 100,000 in Asian and African countries[14,15] to 10 in 100,000 in Europeans[16]. In Europeans, it is estimated that the prevalence of *C9orf72*-FTD is 0.04–134 in 100,000, and *C9orf72*-ALS is 0.5–1.2 in 100,000 (ref. 4). The spinocerebellar ataxias (SCAs) are a group of rare neurodegenerative disorders mainly affecting the cerebellum. They are individually rare worldwide, with largely variable frequencies among populations[17], mainly due to founder effects. Overall, the worldwide prevalence of SCAs is 2.7–47 cases per 100,000 (ref. 10), with SCA3 being the most common form worldwide, followed by SCA2, SCA6 and SCA1[18].

With the advent of disease-modifying therapies for REDs, it is becoming necessary to determine comprehensively the number of patients and type of RED expected in different populations so that targeted approaches can be developed accordingly. Large-scale genetic analyses of REDs have been limited by repeat expansion profiling techniques, which historically have relied on polymerase chain reaction (PCR)-based assays or Southern blots, which by nature are targeted assays and can be difficult to scale. So far, the largest population study of the genetic frequency REDs involved the PCR-based analysis of 14,196 individuals of European ancestry[19].

In the past few years, bioinformatic tools have been developed to profile DNA repeats from short-read whole exome[20] and whole-genome sequencing (WGS) data[21]. We have recently shown that disease-causing repeat expansions can be detected from WGS with high sensitivity and specificity, making large-scale WGS datasets an invaluable resource for the analysis of the frequency and distribution of REDs[7]. Our group has previously applied this pipeline to a large WGS cohort to assess the distribution of repeat expansions in the *AR* gene, which cause spinal and bulbar muscular atrophy (SBMA), and found an unexpectedly high frequency of pathogenic alleles, suggesting underdiagnosis or incomplete penetrance of this RED[22]. However, a comprehensive study of REDs in the general population and across different ancestries using WGS has never been performed.

Here, we used large-scale genomic databases to address two main questions: (1) What is the frequency of RED mutations in the general population? (2) How does the frequency and distribution of REDs vary across populations?

Results

Cohort description

We analyzed RED loci from two large-scale medical genomics cohorts with high-coverage WGS and rich phenotypic data: the 100,000 Genomes Project (100K GP) and Trans-Omics for Precision Medicine (TOPMed). The 100K GP is a program to deliver genome sequencing of people with rare diseases and cancer within the National Health Service (NHS) in the United Kingdom[23,24]. TOPMed is a clinical and genomic program focused on elucidating the genetic architecture and risk factors of heart, lung, blood and sleep disorders from the National Institutes of Health (NIH)[25].

First, we selected WGS data generated using PCR-free protocols and sequenced with paired-end 150 bp reads (Methods and Supplementary Table 1). To avoid overestimating the frequency of REDs, we excluded individuals with neurological diseases, as their recruitment was driven by the fact that they had a neurological disease potentially caused by a repeat expansion. We then performed relatedness and principal component (PC) analyses to identify a set of genetically unrelated individuals and predict broad genetic ancestries based on 1000 Genomes Project phase 3 (1K GP3) superpopulations[26]. The resulting dataset comprised a cross-sectional cohort of 82,176 genomes from unrelated individuals (median age 61 years, Q1 (first quartile)–Q3 (third quartile): 49–70, 58.5% females, 41.5% males; Supplementary Table 2 and Extended Data Fig. 1), genetically predicted to be of European (n = 59,568), African (n = 12,786), American (n = 5,674), South Asian (n = 2,882) and East Asian (n = 1,266) descent (Methods and Extended Data Fig. 2).

RED mutation frequency

To estimate the number of individuals carrying premutation or full-mutation alleles (Fig. 1b), we selected repeats in RED genes[27] for which WGS can accurately discriminate between normal and pathogenic alleles[7], based on either or both of the following conditions: the threshold between premutation and full mutation is shorter than the sequencing read length (and therefore WGS can accurately distinguish between premutation and full mutation), or WGS was validated against the current gold-standard PCR test (Extended Data Fig. 3 and Supplementary Table 3). For the latter, PCR tests were obtained from a cohort of individuals recruited to 100K GP who had RED testing as part of their standard diagnostic pathway (Methods and Supplementary Table 4). Within this dataset, we show the following: (1) WGS accurately classifies alleles in the normal, premutation and full-mutation range in all loci assessed except *FMR1* (which causes fragile X syndrome) (Extended Data Fig. 3a and Supplementary Table 4); (2) the accuracy of repeat sizing by WGS is not affected by genetic ancestry by comparing genotypes generated by WGS with those generated by PCR from different populations (Extended Data Fig. 3b), but it might underestimate the size of large expansions in *FMR1*, *DMPK*, *FXN* and *C9orf72*, as previously described[7].

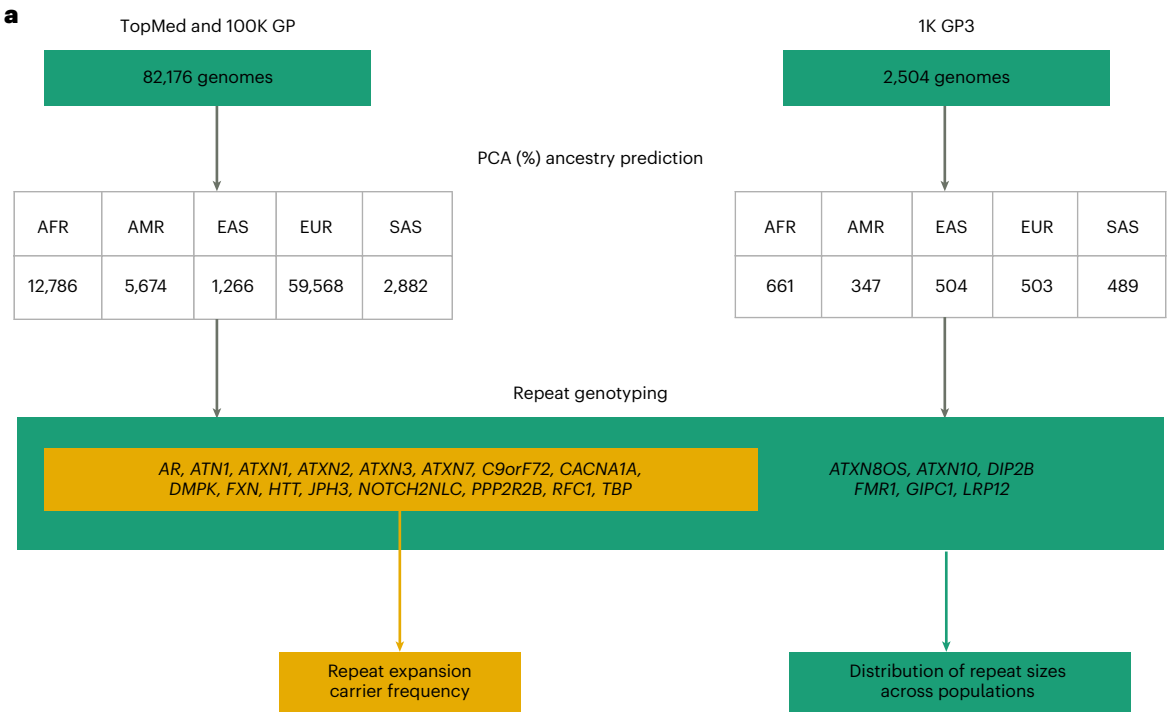
Furthermore, as we previously developed and validated a dedicated WGS analytical workflow for the repeat expansion in *RFC1* that causes CANVAS[28], this repeat was included in our analysis.

Overall, 16 RED loci pass our criteria for accurately estimating premutation and full-mutation carrier frequencies, representing a broad spectrum of REDs and different modes of inheritance: (1) autosomal dominant: HD, Huntington disease-like 2 (HDL2), DM1, *C9orf72*-ALS/FTD, the SCAs (SCA1, SCA2, SCA3, SCA6, SCA7, SCA12 and SCA17), dentatorubral–pallidoluysian atrophy (DRPLA) and *NOTCH2NLC*, which causes a spectrum of neurological disorders, especially neuronal intranuclear inclusion disease and oculopharyngodistal myopathy; (2) autosomal recessive: FA and CANVAS; and (3) X-linked SBMA (Fig. 1b).

Our analysis workflow (Fig. 1) included profiling each RED locus, followed by quality control (QC) of all alleles (employing Expansion Hunter classifier (https://github.com/bharatij/ExpansionHunter_Classifier) and visual inspection of pileup plots as previously described[29]) predicted to be larger than the premutation threshold (Methods and Supplementary Table 5). We also retrospectively analyzed factors potentially leading to overestimating disease allele frequency, such as checking that there was no selection bias for patients with DM1, which can cause cardiac abnormalities (Supplementary Table 6).

In total, for autosomal dominant and X-linked REDs, there were 290 individuals carrying one fully expanded repeat and 1,279 individuals carrying one repeat in the premutation range, meaning that

https://doi.org/10.1038/s41591-024-03190-5



b

Disease	Locus	Premutation/reduced penetrance	Full mutation repeat threshold
Spinal and bulbar muscular atrophy	*AR*	35–37	38
Dentatorubral–pallidoluysian atrophy	*ATN1*	35–47	48
Spinocerebellar ataxia 1	*ATXN1*	39–43	44
Spinocerebellar ataxia 2	*ATXN2*	33–34	35
Spinocerebellar ataxia 3	*ATXN3*	45–59	60
Spinocerebellar ataxia 7	*ATXN7*	34–36	37
C9orf72-related frontotemporal dementia and/or amyotrophic lateral sclerosis 1	*C9orf72*	–	31
Spinocerebellar ataxia 6	*CACNA1A*	19	20
Myotonic dystrophy 1	*DMPK*	35–49	50
Friedreich ataxia	*FXN*	45–65	66
Huntington disease	*HTT*	36–39	40
Huntington disease-like 2	*JPH3*	29–39	40
Oculopharyngodistal myopathy 3; neuronal intranuclear inclusion disease (NIID)	*NOTCH2NLC*	41–55	56
Spinocerebellar ataxia 12	*PPP2R2B*	32–65	66
Cerebellar ataxia, neuropathy and vestibular areflexia syndrome	*RFC1*	–	1,042 (median number of repeats)[56] IQR = 844–1,306 range = 249–3,885
Spinocerebellar ataxia 17	*TBP*	41–48	49

Fig. 1 | Overview of the study. a, Technical flowchart. Whole-genome sequences from the 100K GP and TOPMed datasets were first selected by excluding those associated with neurological diseases. WGS data from 1K GP3 were also selected by having the same technical specifications (Methods). After inferring ancestry prediction, repeat sizes for all 22 REDs were computed by using EH v3.2.2. On one hand, for 16 REDs overall carrier frequency, disease modeling and correlation distribution of long normal alleles were computed in the 100K GP and TOPMed projects (yellow box). On the other hand, the distribution of repeat sizes across different populations was analyzed in 100K GP and TOPMed combined, and in the 1K GP3 cohorts. AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian. **b**, A list of the RED loci included in the study, including repeat-size thresholds for reduced penetrance and full mutations.

the frequency of individuals carrying full-expansion and premutation alleles among this large cohort is 1 in 283 and 1 in 64, respectively (Supplementary Table 7).

The most common expansions (in the full-mutation range; Fig. 1b) were those in *C9orf72* (*C9orf72*-ALS/FTD) and *DMPK* (DM1) with a frequency of 1 in 839 and 1 in 1,786, respectively, followed by expansions in *AR* (SBMA: 1 in 2,561 males) and *HTT* (HD: 1 in 4,109). Surprisingly, many individuals were found to carry expansions in the SCA genes: 1 in 5,136 in *ATXN2* (SCA2), 1 in 5,136 in *CACNA1A* (SCA6), and 1 in 6,321 in *ATXN1* (SCA1). By contrast, expansions in *ATXN7* (SCA7) and *TBP* (SCA17) were present in only two individuals at each locus (1 in 41,077), and expansions in *JPH3* (HDL2) and *ATN1* (DRPLA) were very rare, with

https://doi.org/10.1038/s41591-024-03190-5

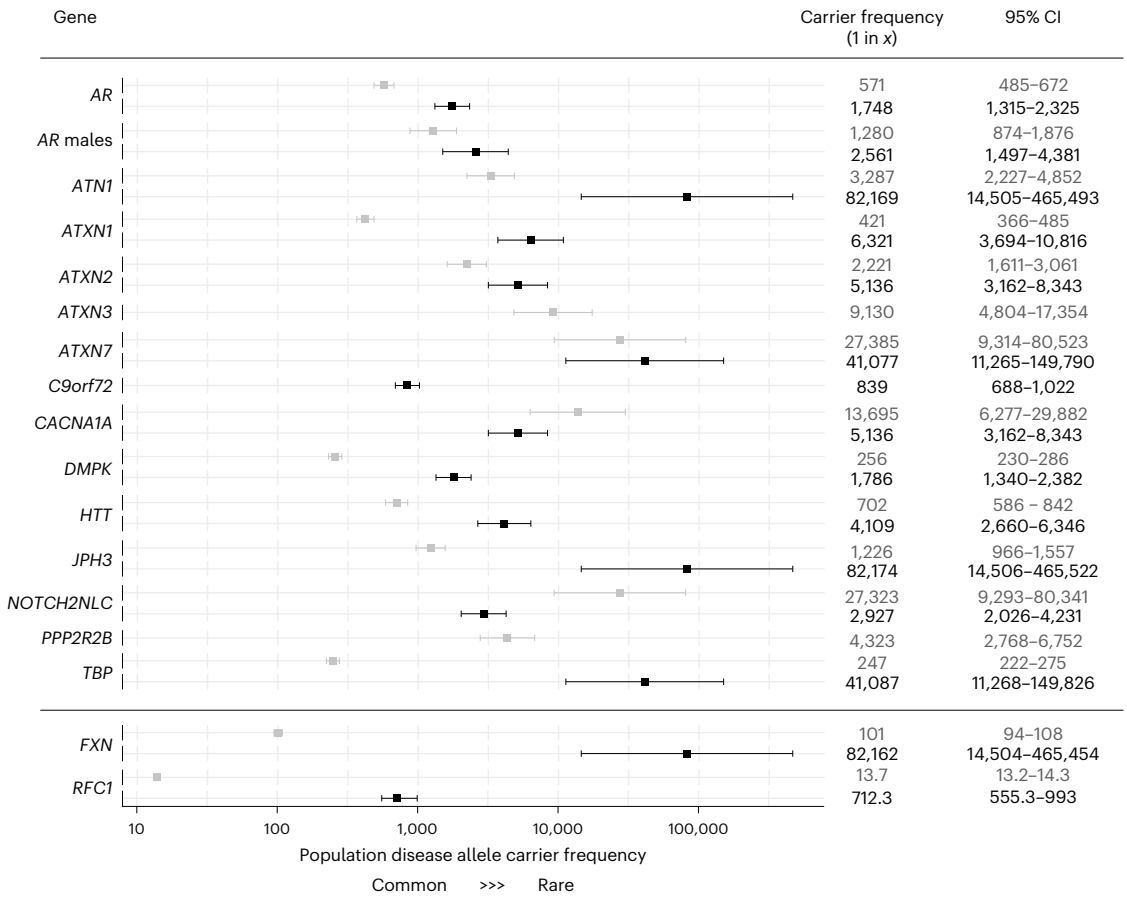


Fig. 2 | Forest plot with combined overall disease allele carrier frequency in the combined 100K GP and TOPMed datasets *N* = 82,176 (*N* individuals may vary slightly between loci owing to data quality and filtering; Supplementary Table 7). The squares show the estimated disease allele carrier frequency, and the bars show the 95% confidence interval (CI) values. Details of the statistical models are described in Methods. For autosomal dominant loci (*AR*, *ATN1*, *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *C9orf72*, *CACNA1A*, *DMPK*, *HTT*, *JPH3*, *NOTCH2NLC*, *PPP2R2B* and *TBP*), the gray and black boxes show premutation/reduced penetrance and full-mutation allele carrier frequencies. For recessive loci (*FXN* and *RFC1*) the gray and black boxes show mono- and biallelic carrier frequencies, respectively.

only a single individual at each locus identified with a repeat allele in the pathogenic full-mutation range. No pathogenic full-mutation expansions were identified in *ATXN3* (SCA3), *PPP2R2B* (SCA12) and *NOTCH2NLC* (Fig. 2 and Supplementary Table 7).

For autosomal recessive REDs, we found a carrier frequency (that is, people who carry one expanded allele) of 1 in 101 for *FXN* (FA) and 1 in 14 for *RFC1*, and a frequency of biallelic expansions of 1 in 82,176 for *FXN* (FA) and 1 in 712 for *RFC1* CANVAS (Fig. 2 and Supplementary Table 8). Demographic data available on all individuals carrying a pathogenic full-mutation repeat are listed in Supplementary Table 9. The distribution of repeat sizes overall in this cohort is represented in Extended Data Fig. 4.

Modeling the expected number of people affected by REDs

REDs have variable age at onset, disease duration and penetrance[1]. Therefore, the mutation frequency cannot be directly translated into disease frequency (that is, prevalence). To estimate the expected number of people affected by REDs, we took the mutation frequency of the most common REDs (*C9orf72*-ALS/FTD, DM1, HD, SCA1, SCA2 and SCA6) and modeled the distribution by age of those expected to be affected by REDs in the UK population. For this analysis, we used the data from the Office of National Statistics[30], and age of onset, penetrance and impact on survival of each RED based on either cohort studies or disease-specific registries (Methods).

We estimated on average a two- to threefold increase in the predicted number of people with REDs, compared with currently reported figures based on clinical observation, depending on the RED (Fig. 3). Since *C9orf72* expansions cause both ALS and FTD, we modeled both diseases separately, providing for *C9orf72*-ALS an expected number of people affected over two times higher than previous estimates (Supplementary Table 10: 1.8 per 100,000 versus 0.5–1.2 in 100,000 (refs. 4,31)) and for *C9orf72*-FTD 6.5 per 100,000 (Supplementary Table 11) within the wide reported range[32,33]. For DM1, we estimated that 15.9 per 100,000 people would be affected by the condition (Supplementary Table 12), 1.3 times higher than the estimated prevalence from clinical data (12.25 in 100,000 (ref. 34)). For HD, the majority of individuals with a pathogenic expansion in our cohort carry alleles with 40 repeats (12 out of 20 people; Supplementary Table 9). Given the well-established relationship between *HTT* repeat length and age at onset, we modeled the expected number of people with HD based on the observed frequency of the expansion, taking into account age at onset distribution and penetrance data for repeat length equal to 40 units[6]. We found that 2.3 per 100,000 people are estimated to have HD caused by 40 CAG repeats (Supplementary Table 13), over 3 times higher than the reported number of affected patients with 40 CAG repeats (0.72 per 100,000; Methods and personal communication, D.R.L. and D.H.M.). For SCA2 and SCA1, our model indicates an over threefold increase in the number of people expected with the disease compared with the reported prevalence (3 and 3.7 per 100,000, respectively, based on our estimate in Supplementary Tables 14 and 15 versus the currently reported prevalence of 1 per 100,000)[35,36]. Strikingly, we found that the expected number of people with SCA6 would be nine times higher than the reported prevalence: 9 in 100,000 versus 1 in 100,000 individuals (Methods and Supplementary Table 16). Overall,

https://doi.org/10.1038/s41591-024-03190-5

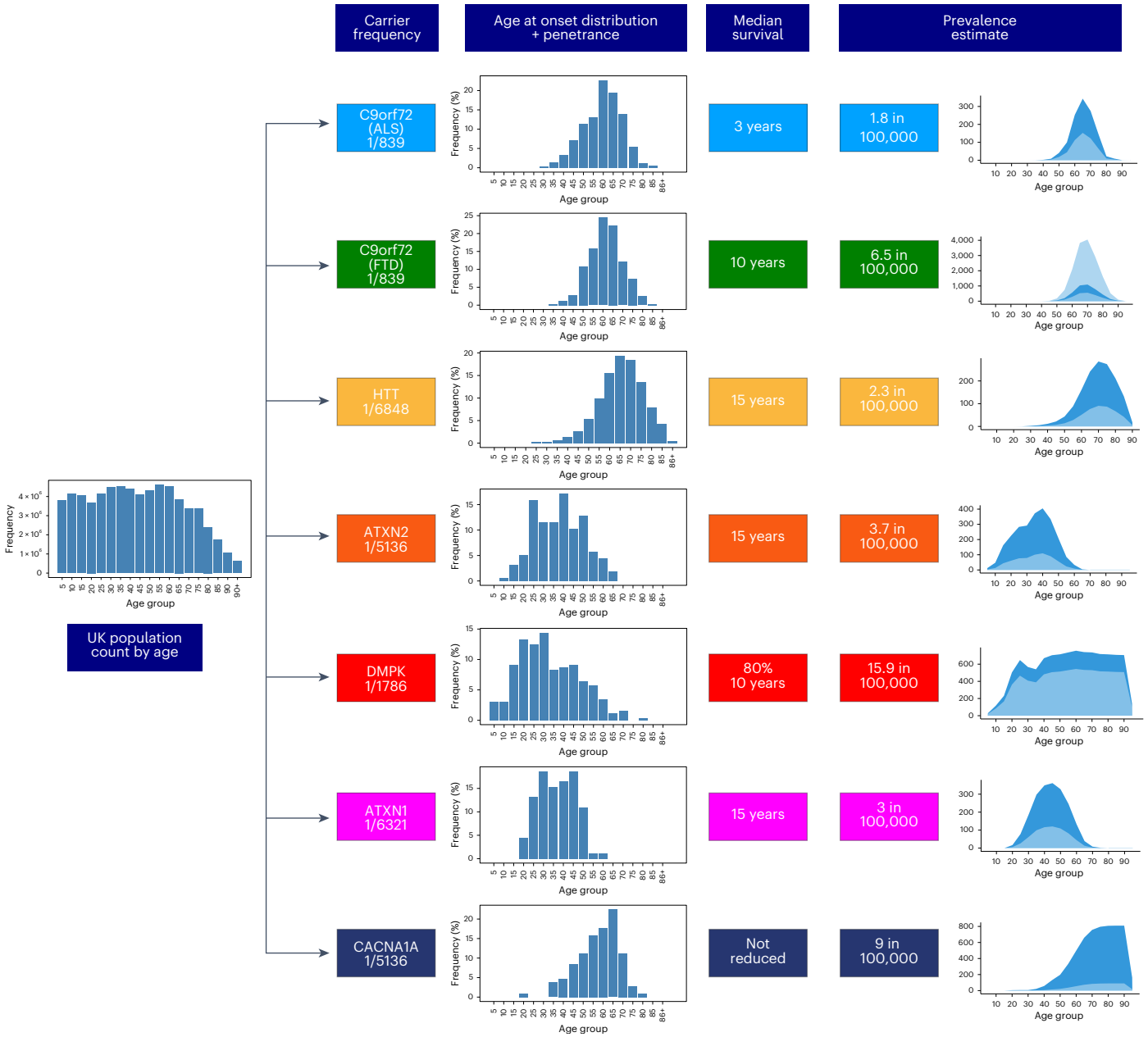


Fig. 3 | Flowchart showing the modeling of disease prevalence by age for *C9orf72*-ALS, *C9orf72*-FTD, HD in 40 CAG repeat carriers, SCA2, DM1, SCA1 and SCA6. The UK population count by age is multiplied by the disease allele frequency of each genetic defect and the age of onset distribution of each corresponding disease, and corrected for median survival. Penetrance is also taken into account for *C9orf72*-ALS and *C9orf72*-FTD. The estimated number of people affected by REDs (dark-blue area) is compared with the reported prevalence from the literature (light-blue area). x-axis: The age bins are 5 years each; y-axis: estimated number of affected individuals. For *C9orf72*-FTD, given the wide range of the reported disease prevalence[32,33], both lower and upper limits are plotted in light blue.

these data indicate either that REDs are underdiagnosed or that not all individuals who carry a repeat larger than the established full-mutation cutoff develop the condition (that is, incomplete penetrance).

RED mutation frequency in different populations

The prevalence of individual REDs varies considerably on the basis of geographic location. Hence, we set out to analyze whether these differences are reflected in the broad genetic ancestries of our cohort. First, we visualized the individuals carrying an expansion in a RED gene on the PC analysis plot (Extended Data Fig. 5). We then computed the proportion of pathogenic allele carriers (premutation and full-mutation allele carriers) in each population (Fig. 4a and Supplementary Table 17). In agreement with current known epidemiological studies, we observed that pathogenic alleles in *FXN* (FA), *C9orf72* and *DMPK* (DM1) are more common in Europeans; those in *ATN1* (DRPLA), *TBP* (SCA17) and in *NOTCH2NLC* are more common in East Asians; and those in *JPH3* (HDL2) are more common in Africans. Conversely, pathogenic alleles in *ATXN2* are more equally distributed across different populations, and those in *RFC1* are less prevalent in Africans. Moreover, pathogenic expansions within *C9orf72* and *HTT* were identified in Africans and South Asians, which so far have only been reported in smaller clinical studies[11,37–40]. Given that the initial ancestry assignments for our cohort were based on genome-wide data, we performed local ancestry analysis to check for admixture in these individuals, confirming that the expanded repeat alleles segregated on haplotypes of African and South Asian ancestry (Methods).

https://doi.org/10.1038/s41591-024-03190-5

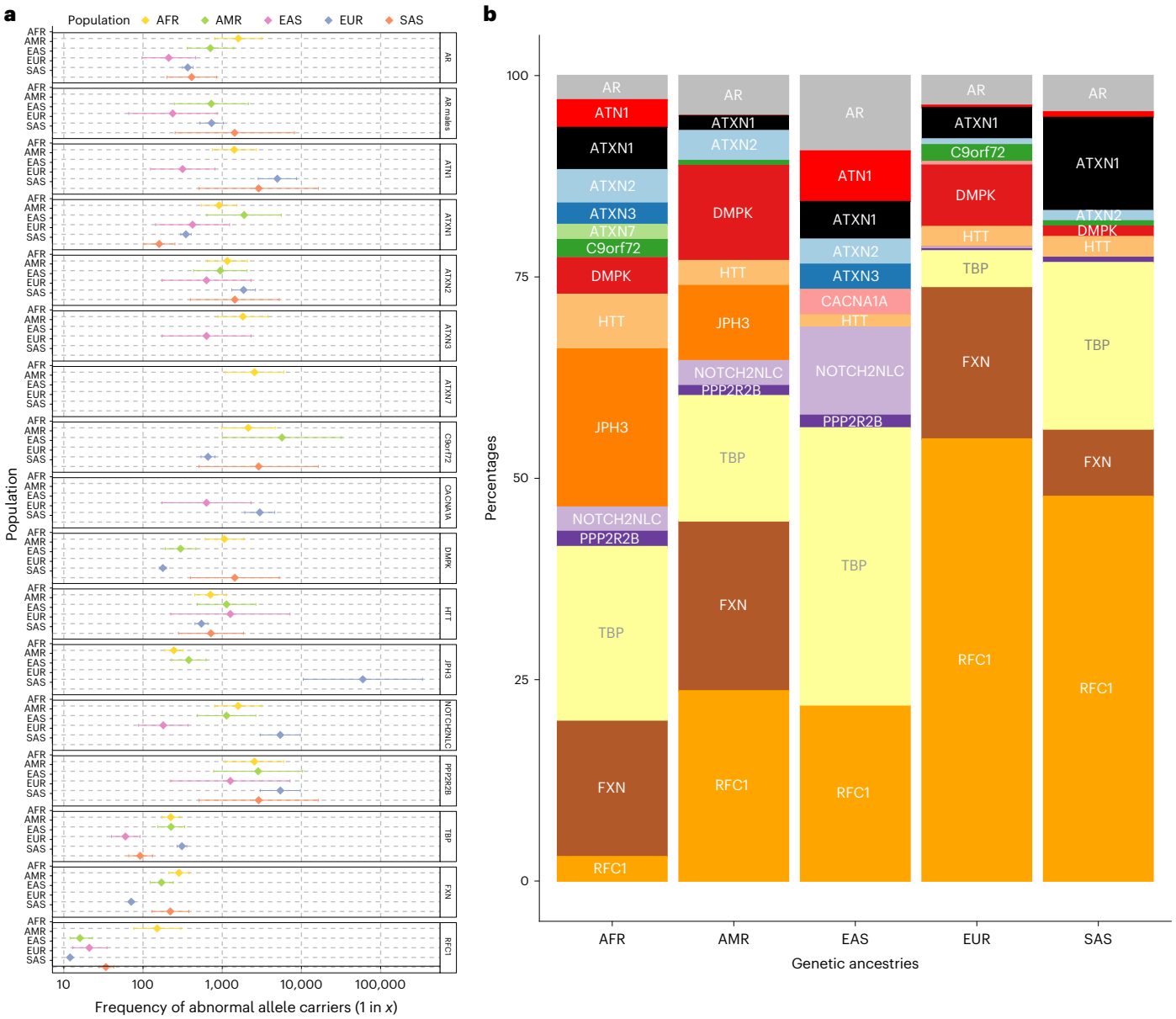


Fig. 4 | Pathogenic RED frequencies in different populations (African 12,786, American 5,674, East Asian 1,266, European 59,568, South Asian 2,882). **a**, Forest plot of pathogenic allele carrier frequency divided by population. Pathogenic alleles are defined as those larger than the premutation cutoff (Fig. 1b). The data are presented as squares showing the estimated pathogenic allele carrier frequency and bars showing the 95% confidence interval values. **b**, Bar chart showing the proportion of pathogenic allele carrier frequency repeats by ancestry. Both plots have been generated by combining data from 100K GP and TOPMed from a total of $N = 82{,}176$ unrelated genomes. N individuals may vary slightly between loci due to data quality and filtering (Supplementary Tables 17 and 18). Predicted ancestries are abbreviated as follows: AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian.

We then analyzed the relative frequency of pathogenic allele carriers within each population (Fig. 4b and Supplementary Table 18), highlighting differences in the proportion of REDs within and among populations. Pathogenic allele carriers were observed in most REDs across all populations (for example, those in *AR*, *ATXN1*, *ATXN2*, *HTT* and *TBP*), though in variable proportions and with some notable exceptions. Pathogenic alleles in *RFC1* are by far the most widely represented (followed by those in *AR* and *TBP*). The African population is the most diverse, with pathogenic expansions present across all RED loci, except *CACNA1A*. The East Asian population is the one with the more striking differences in the relative frequency of REDs and, notably, the absence of pathogenic alleles in *FXN* and *DMPK* and the large proportion of *TBP* (signal driven mainly by reduced penetrance alleles; Supplementary Table 7 shows the carrier frequency of reduced penetrance alleles).

Distribution of repeat lengths in different populations

REDs are thought to arise from large normal polymorphic repeats (large normal or 'intermediate' range repeats), as they have an increased propensity to further expand upon transmission from parent to progeny, moving into the pathogenic range. The uneven RED prevalence across major populations has been associated with the variable frequency of intermediate alleles[41,42].

Therefore, we analyzed intermediate allele frequencies for those genes where WGS can accurately size intermediate alleles (Methods) across populations and confirmed that (1) the overall distribution of repeat lengths varies across populations (Fig. 5a, Extended Data Fig. 6 and Supplementary Table 19); for example, the median repeat size of *PPP2R2B* is higher in East and South Asians compared to Europeans (13 versus 10 repeats)) and (2) overall, the frequency of intermediate alleles varies in each population and correlates with the frequency

of pathogenic alleles; ($R = 0.65$; $P = 3.1 \times 10^{-7}$, Spearman correlation) (Fig. 5b, Extended Data Fig. 7 and Supplementary Table 20). These data suggest that different distributions of repeat lengths underlie differences in the epidemiology of REDs.

In fact, for HD[43,44], specific structures have been proposed as *cis*-acting modifiers of HD[39]. In a typical *HTT* allele, the pure CAG tract (Q1) is followed by an interrupting CAACAG sequence (Q2). These are followed by a polyproline region encoded by a CCGCCA sequence (P1), then by stretches of CCG repeats (P2) and, lastly, by CCT repeats (P3) (Fig. 6a). Variations in this sequence have been described, including duplication or loss of Q2 or loss of P1, and variation in the number of the downstream P2 and P3 repeats. To analyze population differences in the repeat structures of *HTT*, we developed an analytical workflow to determine accurately phased Q1, Q2 and P1 elements of the *HTT* structure from WGS (Methods). We confirm that the typical structure $Q1_n$–$Q2_1$–$P1_1$ (also known as 'canonical') is the most common across populations and that other structures are present in different proportions in different populations: P1 loss alleles are more prevalent in Africans and East Asians, while overall noncanonical alleles are more frequent in African ($P < 0.0001$ two-tailed chi-square test; Fig. 6b). Moreover, variable Q1 repeat lengths are associated with different structures (linear model, $P < 2.2 \times 10^{-16}$): shorter Q1 lengths are found on chromosomes with Q2 duplication, and larger Q1 lengths are found in those with Q2 loss and Q2–P1 loss (Fig. 6c).

Population distribution of other REDs

WGS cannot accurately size repeats larger than the read length (here, 150 bp) and is therefore unable to distinguish between premutation and full mutations of some REDs (for example, *FMR1*). However, the technology can be used to determine the distribution of repeat lengths within each population, with the largest percentiles (99.9th percentiles) reflecting the variable presence of expanded alleles in each population. For example, the 99.9th percentile of repeat sizes of *DMPK* is 39 in Europeans and 30 in Africans (Fig. 5a and Supplementary Table 19).

Hence, we set out to extend our analysis of population differences to other RED loci. For this analysis, we used 1K GP3 data and selected RED genes that are caused by expansion of the reference sequence: *FMR1* (fragile X syndrome), *DIP2B* (intellectual disability FRA12A type), *ATXN8* (SCA8), *ATXN10* (SCA10), *LRP12* and *GIPC1* (oculopharyngodistal myopathy 1 and 2, respectively[45]). In line with epidemiological studies, we found that for *FMR1*, *DIP2B* and *ATXN8* the largest percentiles are those found in Europeans, for *ATXN10* are those in Americans and for *LRP12* are those found in East Asians. Surprisingly, we found that Africans have larger repeats in *GIPC1* compared with other ancestries (Extended Data Fig. 8 and Supplementary Tables 21 and 22). The different distributions of REDs reported in this analysis reflect the relatively smaller proportion of large normal/intermediate alleles among populations, which may provide some explanation for the different frequencies of REDs in different populations.

Discussion

By analyzing a cross-sectional cohort of 82,176 people, this study provides the largest population-based estimate of disease allele carrier frequency and RED distribution in different populations. We show that (1) the disease allele carrier frequency of REDs is approximately ten times higher than the previous estimates based on clinical observations and that, based on population modeling, REDs would be predicted to affect, on average, two to three times more individuals than are currently recognized clinically; (2) while some REDs are population specific like *JPH3* (HDL2), the majority are observed in all ancestral populations, challenging the notion that some REDs are associated with population-specific founder effects (for example, *C9orf72*); (3) the different distribution of repeat lengths between population broadly reflects the known differences in disease epidemiology; (4) an appreciable proportion of the population carry alleles in the premutation range and are, therefore, at risk of having children with REDs.

As the data for the cohort in which we carried out this study were collected for medical sequencing purposes, we controlled for factors potentially leading to overestimating disease allele carrier frequency, such as excluding people with neurological disorders and checking that there was no selection bias for patients with DM1 (Supplementary Table 6), which can cause cardiac abnormalities. Our estimates for DM1 and SCAs match those previously reported using PCR-based approaches to determine the genetic prevalence of REDs[8,19,46], confirming the accuracy of our results.

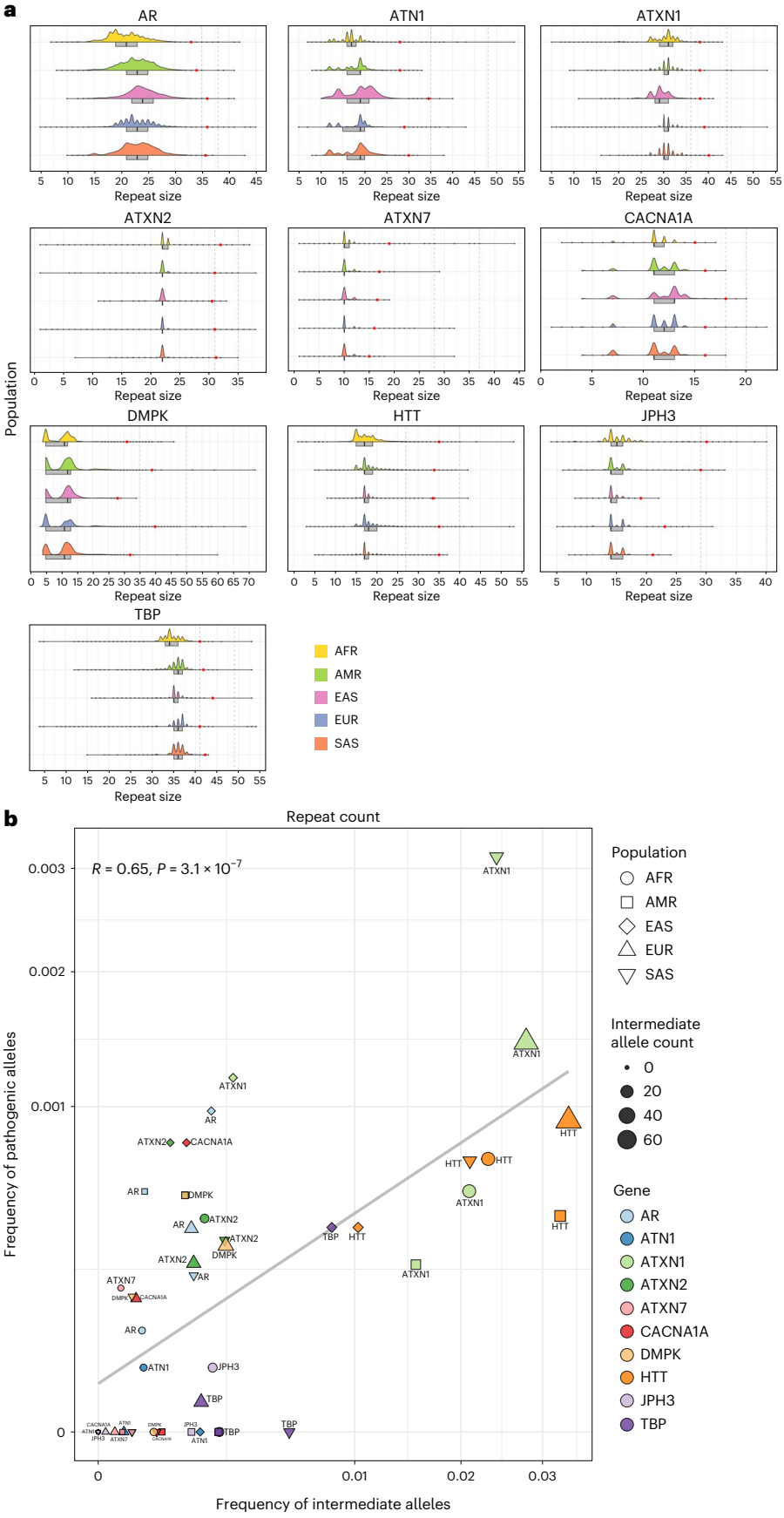
Different factors might explain the discrepancy between the increased number of people carrying disease alleles in our cohort compared with known RED epidemiology. First, our estimates are based on large admixed cohorts, as opposed to epidemiological studies based on clinically affected individuals in smaller populations. As REDs have variable clinical presentation and age at onset, individuals with REDs may remain undiagnosed in studies in which estimates of disease frequency rely on clinical ascertainment of patients. Notably, the first descriptions of the clinical phenotype of RED were based on families collected for linkage studies with an intrinsic ascertainment bias for more severe disease manifestations, resulting in a lack of very mild cases in the phenotypic spectrum. Because of the wide spectrum of milder phenotypic presentations of REDs, the prevalence of these diseases might have been underestimated. This might be true particularly for milder forms of the disease spectrum, such as DM1; it is well documented that carriers of small *DMPK* expansions (50–100 repeats) have a milder disease with clinical features that may go unnoticed, especially early in their disease course[47]. In fact, we observed a large number of individuals carrying repeats in the lower end of the pathogenic range (for example *HTT*, *ATXN2* and *DMPK*; Supplementary Table 9).

Moreover, prevalence studies are only based on individuals with manifest disease, leading to a potential bias in the disease penetrance from those who have not developed the illness. It is believed that the penetrance of REDs is characterized by a threshold effect, with people carrying an allele above a particular repeat length certainly developing the disease, as opposed to those carrying shorter repeats. Given the relationship between the size of the repeat expansion and the disease onset and progression, it is possible that individuals carrying alleles currently classified as fully penetrant (for example, ≥40 CAG repeats in *HTT*) may sometimes remain asymptomatic. In this regard, previously published studies on HD[6] and SBMA[48] have suggested incomplete penetrance of repeats in the lower end of the pathogenic range.

Finally, these individuals may carry genetic modifiers of REDs, such as interspersions. We visually inspected all alleles in the pathogenic range and did not identify atypical sequence structures within the expanded repeat. Accordingly, in *HTT*, where we performed a dedicated structural analysis (Fig. 6), all individuals carrying a fully expanded repeat show a typical canonical structure.

Fig. 5 | The distribution of repeat lengths in different populations. **a**, Half-violin plots showing the distribution of alleles in different populations (African 12,786, American 5,674, East Asian 1,266, European 59,568, South Asian 2,882) for 10 loci (Methods) from the combined 100K GP and TOPMed cohorts. The box plots highlight the interquartile range and median, and the black dots show values outside 1.5 times the interquartile range. The red dots mark the 99.9th percentile for each population and locus. The vertical bars indicate the intermediate and pathogenic allele thresholds (Supplementary Table 20). Predicted ancestries are abbreviated as follows: AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian. **b**, A scatter plot showing the frequency of intermediate allele carriers against the frequency of pathogenic allele carriers. The data points are divided by population ($n = 5$) and gene ($n = 10$), and the size represents the total number of intermediate alleles. Correlations were computed using the Spearman method and two-tailed *P* values.

https://doi.org/10.1038/s41591-024-03190-5



https://doi.org/10.1038/s41591-024-03190-5

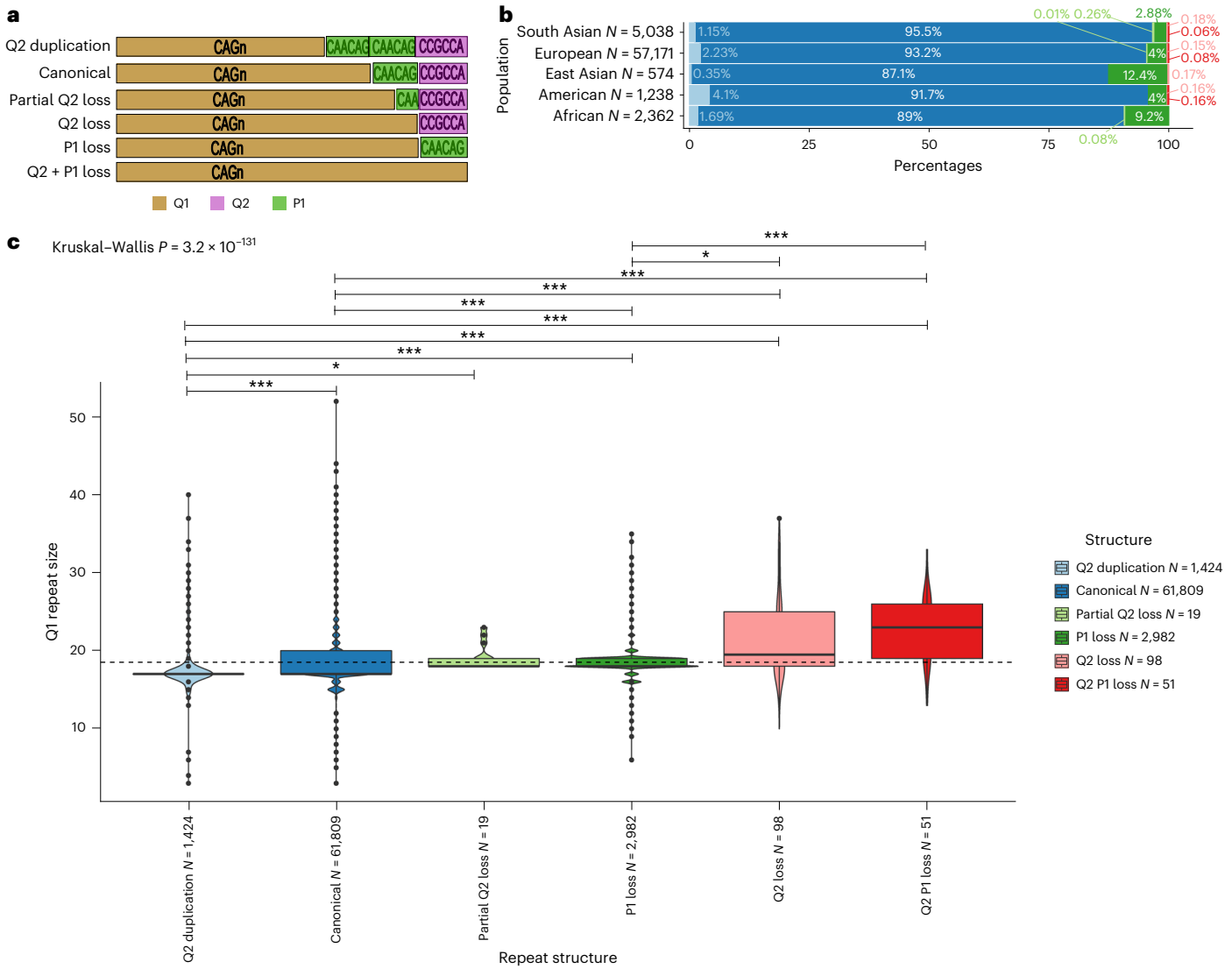


Fig. 6 | *HTT* repeat structures show varied prevalence across genetic ancestries and are associated with CAG repeat size. a, Allele structures observed within exon 1 of *HTT*. The CAG repeat is denoted as 'Q1' and marked in gold. The CAACAG unit is referred to as 'Q2' and is marked in green. The first proline-encoding 'CCGCCA' repeat element is referred to as 'P1' and is marked in purple. **b**, The prevalence of the allele structures is plotted across the studied genetic ancestries in bar plots. The ancestries are defined on the *y* axis. The number of alleles in each of the genetic ancestries is denoted as '*N* = …' at each of the *y*-axis ticks. **c**, Box plots displaying the distribution of CAG repeat sizes across different repeat structures. The box plots highlight the median (horizontal lines in the center of each box plot) and interquartile range (bounds), and the black dots show values outside 1.5 times the interquartile range. The number of alleles with different repeat structures is denoted as '*N* = …' on the *x* axis. A linear model was used to compare the repeat size distribution of the canonical alleles versus that of all atypical structures. Kruskal–Wallis tests with Dunn's correction for multiple comparisons *P* value; *P* values resulting from pairwise tests are displayed above each structure (***$P < 0.001$; *$P < 0.05$). Q2 versus canonical ($P = 6.4 \times 10^{-32}$), Q2 versus partial Q2 loss ($P = 3.5 \times 10^{-2}$), Q2 duplication versus P1 loss ($P = 5.9 \times 10^{-98}$), Q2 duplication versus Q2 loss ($P = 8.5 \times 10^{-16}$); Q2 duplication versus Q2–P1 loss ($P = 6.2 \times 10^{-20}$), canonical versus P1 loss ($P = 2.4 \times 10^{-80}$), canonical versus Q2 loss ($P = 2.8 \times 10^{-8}$), canonical versus Q2–P1 loss ($P = 1.2 \times 10^{12}$), P1 loss versus Q2 loss ($P = 2.8 \times 10^{-2}$), P1 loss versus versus Q2–P1 loss ($P = 5.6 \times 10^{-6}$).

The finding that a much larger number of people in the general population carry pathogenic alleles of REDs has important implications both for diagnosis and genetic counseling of RED. For diagnosis, when a patient presents with symptoms compatible with a RED, clinicians should have a higher index of suspicion of these diseases, and clinical diagnostic pathways should facilitate genetic testing for REDs. Currently, genetic testing for REDs tends to be a PCR-based targeted assay, with clinicians suspecting a RED ordering a test for a specific gene. As REDs are clinically and genetically heterogeneous with a tendency to have overlapping features, REDs may remain undiagnosed. The wider use of WGS and the advent of genetic technologies such as long-read sequencing can potentially address this by simultaneously interrogating an entire panel of RED loci[49]. The broader availability of these diagnostic tools would increase the diagnostic rate for REDs, thus closing the gap between disease incidence rates and estimates based on population genetic sequencing. As for genetic counseling, when a RED expansion is identified incidentally in an individual clinically unaffected, it would be important to address the potentially incomplete penetrance of the repeat, especially for small expansions. Further studies both in clinically affected individuals and in large clinical and genomic datasets from the general population are needed to address the full clinical spectrum and the penetrance by repeat sizes.

Our results are concordant with current epidemiological studies about the relative frequency of REDs, with the most common being DM1 and *C9orf72*-ALS/FTD (autosomal dominant) and CANVAS and sensory neuropathy (recessive). One exception is *ATXN3*, the most commonly

reported SCA locus in patients affected by SCA, which was absent in our cohort. This might be due either to a recruitment bias, with individuals with overt REDs having a reduced likelihood of being recruited to such studies because of the severity of their disease (TOPMed cohort), or to the fact that there are no or very few premutation alleles (0 in 59,568 Europeans), indicating that the expansion mutation is linked to few rare ancestral haplotypes[50] rather than being a gradual process arising from common large normal alleles like other REDs.

The presence of rare expansions of REDs previously thought to occur only in Europeans (for example, *C9orf72)* in African and Asian populations supports diagnostic testing for them in people presenting with features of ALS–FTD independently of their ethnicity. The lowest observed rates of some REDs in some populations (for example, *FXN* and *DMPK* in East Asians), consistent with known epidemiological studies, might be due to reduced mutation rates. Further research is needed to study the potential role of population-specific *cis* and *trans* genetic modifiers of repeat expansion mutations that underlie the marked global differences in prevalence found in the present study.

One limitation of this study is that WGS cannot accurately size repeats larger than the sequencing read length, and it is therefore not possible to accurately estimate the disease allele frequency of all RED loci. Of the 46 RED loci that have been linked to human disease[27], we included all loci where it is technically possible to address our questions: (1) to accurately estimate the disease allele carrier frequency, 16 REDs were selected; (2) to analyze the distribution of repeat lengths in different populations, 6 further REDs were selected, covering a total of 22 RED loci and providing the basis for the different prevalence of REDs in different populations. RED loci that are not included are those caused by an insertion of a nonreference sequence (currently there is no validated pipeline that can accurately size and sequence large repeat expansions in such loci, except *RFC1*) and those caused by nonpure sequences, such as 'GCN' motifs (caused by a different mutational mechanism, namely unequal allelic homologous recombination)[51]. We note that many newly discovered REDs are caused by large expansions[52]; only a broader availability of long-read sequencing technologies will facilitate addressing important questions about the frequency of these mutations.

Both 100K GP and TOPMed datasets are Eurocentric, comprising over 62% of European samples. TOPMed is more diverse, with 24% and 17% of African and American genomes, respectively, which are only present at 3.2% and 2.1% frequency in 100K GP. East and South Asian backgrounds are underrepresented in both datasets, limiting the ability to detect rarer repeat expansions in these populations. Further analyses on more heterogeneous and diverse large-scale WGS datasets are necessary not only to confirm our findings but also to shed light on additional ancestries. With regard to this, there are multiple ongoing projects with Asian populations[53]. Countries including China, Japan, Qatar, Saudi Arabia, India, Nigeria and Turkey have launched their own genomics projects during the past decade[54]. Analyzing genomes from these programs will yield more detail on the prevalence of REDs around the world.

Despite efforts to estimate the frequency of REDs globally and locally, there is uncertainty surrounding their true prevalence, limiting the knowledge of the burden of disease required to secure dedicated resources to support health services, such as the estimation of the numbers of individuals profiting from drug development and novel therapies, or participating in clinical trials. There are currently no disease-modifying treatments for REDs; however, both disease-specific treatments and drugs that target the mechanisms leading to repeat expansions are in development. We have established that the number of people who may benefit from such treatments is greater than previously thought.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03190-5.

References

1. Paulson, H. Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
2. Cortese, A. et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat. Genet.* **51**, 649–658 (2019).
3. Moore, K. M. et al. Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study. *Lancet Neurol.* **19**, 145–156 (2020).
4. Gossye, H., Engelborghs, S., Van Broeckhoven, C. & van der Zee, J. *C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis* (Univ. Washington, 2020).
5. van der Ende, E. L. et al. Unravelling the clinical spectrum and the role of repeat length in C9ORF72 repeat expansions. *J. Neurol. Neurosurg. Psychiatry* **92**, 502–509 (2021).
6. Langbehn, D. R. et al. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.* **65**, 267–277 (2004).
7. Ibañez, K. et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* **21**, 234–245 (2022).
8. Johnson, N. E. et al. Population-based prevalence of myotonic dystrophy type 1 using genetic analysis of statewide blood screening program. *Neurology* **96**, e1045–e1053 (2021).
9. de Castilhos, R. M. et al. Spinocerebellar ataxias in Brazil—frequencies and modulating effects of related genes. *Cerebellum* **13**, 17–28 (2014).
10. Teive, H. A. G., Meira, A. T., Camargo, C. H. F. & Munhoz, R. P. The geographic diversity of spinocerebellar ataxias (SCAs) in the Americas: a systematic review. *Mov. Disord. Clin. Pract.* **6**, 531–540 (2019).
11. Baine, F. K., Peerbhai, N. & Krause, A. A study of Huntington disease-like syndromes in black South African patients reveals a single SCA2 mutation and a unique distribution of normal alleles across five repeat loci. *J. Neurol. Sci.* **390**, 200–204 (2018).
12. Bird, T. D. *Myotonic Dystrophy Type 1* (Univ. Washington, 2024).
13. Theadom, A. et al. Prevalence of muscular dystrophies: a systematic literature review. *Neuroepidemiology* **43**, 259–268 (2014).
14. Pringsheim, T. et al. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
15. Hayden, M. R., MacGregor, J. M. & Beighton, P. H. The prevalence of Huntington's chorea in South Africa. *S. Afr. Med. J.* **58**, 193–196 (1980).
16. Rawlins, M. D. et al. The prevalence of Huntington's disease. *Neuroepidemiology* **46**, 144–153 (2016).
17. Sequeiros, J., Martins, S. & Silveira, I. Epidemiology and population genetics of degenerative ataxias. *Handb. Clin. Neurol.* **103**, 227–251 (2012).
18. Schöls, L., Bauer, P., Schmidt, T., Schulte, T. & Riess, O. Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. *Lancet Neurol.* **3**, 291–304 (2004).
19. Gardiner, S. L. et al. Prevalence of carriers of intermediate and pathological polyglutamine disease-associated alleles among large population-based cohorts. *JAMA Neurol.* **76**, 650–656 (2019).
20. van der Sanden, B. P. G. H. et al. Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genet. Med.* **23**, 1569–1573 (2021).

21. Tanudisastro, H. A., Deveson, I. W., Dashnow, H. & MacArthur, D. G. Sequencing and characterizing short tandem repeats in the human genome. *Nat. Rev. Genet.* **25**, 460–475 (2024).
22. Zanovello, M. et al. Unexpected frequency of the pathogenic AR CAG repeat expansion in the general population. *Brain* **146**, 2723–2729 (2023).
23. 100,000 Genomes Project. *Genomics England* https://www.genomicsengland.co.uk/initiatives/100000-genomes-project (2022).
24. 100,000 Genomes Project Pilot Investigators. et al. 100,000 Genomes pilot on rare-disease diagnosis in health care—preliminary report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
25. About TOPMed. *NIH* https://topmed.nhlbi.nih.gov/ (2024).
26. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
27. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
28. Sullivan, R. et al. *RFC1* repeat expansion analysis from whole genome sequencing data simplifies screening and increases diagnostic rates. Preprint at *medRxiv* https://doi.org/10.1101/2024.02.28.24303510 (2024).
29. Dolzhenko, E. et al. REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* **14**, 84 (2022).
30. Home. *Office for National Statistics* https://www.ons.gov.uk/ (2024).
31. Zampatti, S. et al. C9orf72-related neurodegenerative diseases: from clinical diagnosis to therapeutic strategies. *Front. Aging Neurosci.* **14**, 907122 (2022).
32. Van Mossevelde, S., Engelborghs, S., van der Zee, J. & Van Broeckhoven, C. Genotype–phenotype links in frontotemporal lobar degeneration. *Nat. Rev. Neurol.* **14**, 363–378 (2018).
33. Hogan, D. B. et al. The prevalence and incidence of frontotemporal dementia: a systematic review. *Can. J. Neurol. Sci.* **43**, S96–S109 (2016).
34. Liao, Q., Zhang, Y., He, J. & Huang, K. Global prevalence of myotonic dystrophy: an updated systematic review and meta-analysis. *Neuroepidemiology* **56**, 163–173 (2022).
35. De Mattei, F. et al. Epidemiology of spinocerebellar ataxias in Europe. *Cerebellum* **23**, 1176–1183 (2023).
36. Opal, P. & Ashizawa, T. *Spinocerebellar Ataxia Type 1* (Univ. Washington, 2023).
37. Nel, M. et al. C9orf72 repeat expansions in South Africans with amyotrophic lateral sclerosis. *J. Neurol. Sci.* **401**, 51–54 (2019).
38. Tan, Y. J. et al. C9orf72 expansions are the most common cause of genetic frontotemporal dementia in a Southeast Asian cohort. *Ann. Clin. Transl. Neurol.* **10**, 568–578 (2023).
39. Dawson, J. et al. A probable *cis*-acting genetic modifier of Huntington disease frequent in individuals with African ancestry. *HGG Adv.* **3**, 100130 (2022).
40. Muthinja, M. J. et al. An exploration of the genetics of the mutant Huntingtin (mHtt) gene in a cohort of patients with chorea from different ethnic groups in sub-Saharan Africa. *Ann. Hum. Genet.* (2024).
41. Kay, C. et al. The molecular epidemiology of Huntington disease is related to intermediate allele frequency and haplotype in the general population. *Am. J. Med. Genet. B* **177**, 346–357 (2018).
42. Takano, H. et al. Close associations between prevalences of dominantly inherited spinocerebellar ataxias with CAG-repeat expansions and frequencies of large normal CAG alleles in Japanese and Caucasian populations. *Am. J. Hum. Genet.* **63**, 1060–1066 (1998).
43. Monckton, D. G. & Caskey, C. T. Unstable triplet repeat diseases. *Circulation* **91**, 513–520 (1995).
44. Ciosi, M. 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* **48**, 568–580 (2019).
45. Ishiura, H. et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat. Genet.* **51**, 1222–1232 (2019).
46. Kay, C. et al. Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
47. Thornton, C. A. Myotonic dystrophy. *Neurol. Clin.* **32**, 705–719 (2014).
48. Laskaratos, A., Breza, M., Karadima, G. & Koutsis, G. Wide range of reduced penetrance alleles in spinal and bulbar muscular atrophy: a model-based approach. *J. Med. Genet.* **58**, 385–391 (2021).
49. Miyatake, S. et al. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *NPJ Genom. Med.* **7**, 62 (2022).
50. Gaspar, C. et al. Ancestral origins of the Machado–Joseph disease mutation: a worldwide haplotype study. *Am. J. Hum. Genet.* **68**, 523–528 (2001).
51. Amiel, J., Trochet, D., Clément-Ziza, M., Munnich, A. & Lyonnet, S. Polyalanine expansions in human. *Hum. Mol. Genet.* **13**, R235–R243 (2004).
52. Vegezzi, E. et al. Neurological disorders caused by novel non-coding repeat expansions: clinical features and differential diagnosis. *Lancet Neurol.* **23**, 725–739 (2024).
53. Wu, D. et al. Large-scale whole-genome sequencing of three diverse asian populations in Singapore. *Cell* **179**, 736–749.e15 (2019).
54. Kumar, R. & Dhanda, S. K. Current status on population genome catalogues in different countries. *Bioinformation* **16**, 297–300 (2020).

[1]William Harvey Research Institute, Queen Mary University of London, London, UK. [2]Department of Genetics and Genomic Sciences and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Neuromuscular Diseases, Institute of Neurology, UCL, London, UK. [4]IRCCS Mondino Foundation, Pavia, Italy. [5]UK Dementia Research Institute, UCL, London, UK. [6]St George's, University of London, London, UK. [7]Department of Neurodegenerative Disorders, Queen Square Institute of Neurology, UCL, London, UK. [8]Genomics England, London, UK. [9]The John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [10]Centre for Neuromuscular Disease, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology and National Hospital for Neurology and Neurosurgery, London, UK. [11]Departments of Psychiatry and Biostatistics, The University of Iowa, Iowa City, IA, USA. [12]Huntington's Disease Centre, UCL, London, UK. [13]Department of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, Cardiff, UK. [14]Dementia Research Institute, Cardiff University, Cardiff, UK. [15]Neurogenetics Unit, National Hospital for Neurology and Neurosurgery, London, UK. ✉e-mail: a.tucci@qmul.ac.uk

Methods

Ethics statement inclusion and ethics

The 100K GP is a UK program to assess the value of WGS in patients with unmet diagnostic needs in rare disease and cancer. Following ethical approval for 100K GP by the East of England Cambridge South Research Ethics Committee (reference 14/EE/1112), including for data analysis and return of diagnostic findings to the patients, these patients were recruited by healthcare professionals and researchers from 13 genomic medicine centers in England and were enrolled in the project if they or their guardian provided written consent for their samples and data to be used in research, including this study.

For ethics statements for the contributing TOPMed studies, full details are provided in the original description of the cohorts[55].

WGS datasets

Both 100K GP and TOPMed include WGS data optimal to genotype short DNA repeats: WGS libraries generated using PCR-free protocols, sequenced at 150 base-pair read length and with a 35× mean average coverage (Supplementary Table 1).

For both the 100K GP and TOPMed cohorts, the following genomes were selected: (1) WGS from genetically unrelated individuals (see 'Ancestry and relatedness inference' section); (2) WGS from people not presenting with a neurological disorder (these people were excluded to avoid overestimating the frequency of a repeat expansion due to individuals recruited due to symptoms related to a RED).

The TOPMed project has generated omics data, including WGS, on over 180,000 individuals with heart, lung, blood and sleep disorders (https://topmed.nhlbi.nih.gov/). TOPMed has incorporated samples gathered from dozens of different cohorts, each collected using different ascertainment criteria. The specific TOPMed cohorts included in this study are described in Supplementary Table 23.

To analyze the distribution of repeat lengths in REDs in different populations, we used 1K GP3 as the WGS data are more equally distributed across the continental groups (Supplementary Table 2). Genome sequences with read lengths of ~150 bp were considered, with an average minimum depth of 30× (Supplementary Table 1).

Ancestry and relatedness inference

For relatedness inference WGS, variant call formats (VCF)s were aggregated with Illumina's agg or gvcfgenotyper (https://github.com/Illumina/gvcfgenotyper). All genomes passed the following QC criteria: cross-contamination <5% (VerifyBamId)[56], mapping rate >75%, mean-sample coverage >20 and insert size >250 bp. No variant QC filters were applied in the aggregated dataset, but the VCF filter was set to 'PASS' for variants that passed GQ (genotype quality), DP (depth), missingness, allelic imbalance and Mendelian error filters. From here, by using a set of ~65,000 high-quality single-nucleotide polymorphisms (SNPs), a pairwise kinship matrix was generated using the PLINK2 implementation of the KING-Robust algorithm (www.cog-genomics.org/plink/2.0/)[57]. For relatedness, the PLINK2 '--king-cutoff' (www.cog-genomics.org/plink/2.0/) relationship-pruning algorithm[57] was used with a threshold of 0.044. These were then partitioned into 'related' (up to, and including, third-degree relationships) and 'unrelated' sample lists. Only unrelated samples were selected for this study.

The 1K GP3 data were used to infer ancestry, by taking the unrelated samples and calculating the first 20 PCs using GCTA2. We then projected the aggregated data (100K GP and TOPMed separately) onto 1K GP3 PC loadings, and a random forest model was trained to predict ancestries on the basis of (1) first eight 1K GP3 PCs, (2) setting 'Ntrees' to 400 and (3) training and predicting on 1K GP3 five broad superpopulations: African, Admixed American, East Asian, European and South Asian.

In total, the following WGS data were analyzed: 34,190 individuals in 100K GP, 47,986 in TOPMed and 2,504 in 1K GP3. The demographics describing each cohort can be found in Supplementary Table 2.

Correlation between PCR and EH

Results were obtained on samples tested as part of routine clinical assessment from patients recruited to 100K GP. Repeat expansions were assessed by PCR amplification and fragment analysis. Southern blotting was performed for large *C9orf72* and *NOTCH2NLC* expansions as previously described[7].

A dataset was set up from the 100K GP samples comprising a total of 681 genetic tests with PCR-quantified lengths across 15 loci: *AR*, *ATN1*, *ATXN1*, *ATXN2*, *ATXN3*, *ATXN7*, *CACNA1A*, *DMPK*, *C9orf72*, *FMR1*, *FXN*, *HTT*, *NOTCH2NLC*, *PPP2R2B* and *TBP* (Supplementary Table 3).

Overall, this dataset comprised PCR and correspondent EH estimates from a total of 1,291 alleles: 1,146 normal, 44 premutation and 101 full mutation. Extended Data Fig. 3a shows the swim lane plot of EH repeat sizes after visual inspection classified as normal (blue), premutation or reduced penetrance (yellow) and full mutation (red). These data show that EH correctly classifies 28/29 premutations and 85/86 full mutations for all loci assessed, after excluding *FMR1* (Supplementary Tables 3 and 4). For this reason, this locus has not been analyzed to estimate the premutation and full-mutation alleles carrier frequency. The two alleles with a mismatch are changes of one repeat unit in *TBP* and *ATXN3*, changing the classification (Supplementary Table 3). Extended Data Fig. 3b shows the distribution of repeat sizes quantified by PCR compared with those estimated by EH after visual inspection, split by superpopulation. The Pearson correlation (R) was calculated separately for alleles larger (for Europeans, $n = 864$) and shorter ($n = 76$) than the read length (that is, 150 bp).

Repeat expansion genotyping and visualization

The EH software package was used for genotyping repeats in disease-associated loci[58,59]. EH assembles sequencing reads across a predefined set of DNA repeats using both mapped and unmapped reads (with the repetitive sequence of interest) to estimate the size of both alleles from an individual.

The REViewer software package was used to enable the direct visualization of haplotypes and corresponding read pileup of the EH genotypes[29]. Supplementary Table 24 includes the genomic coordinates for the loci analyzed. Supplementary Table 5 lists repeats before and after visual inspection. Pileup plots are available upon request.

Computation of genetic prevalence

The frequency of each repeat size across the 100K GP and TOPMed genomic datasets was determined. Genetic prevalence was calculated as the number of genomes with repeats exceeding the premutation and full-mutation cutoffs (Fig. 1b) for autosomal dominant and X-linked REDs (Supplementary Table 7); for autosomal recessive REDs, the total number of genomes with monoallelic or biallelic expansions was calculated, compared with the overall cohort (Supplementary Table 8).

Overall unrelated and nonneurological disease genomes corresponding to both programs were considered, breaking down by ancestry.

Carrier frequency estimate (1 in *x*)

- freq_carrier = round(total_unrel/total_exp_after_VI_locus, digits = 0), where
  - 'total_unrel' is the total number of unrelated genomes
  - 'total_exp_after_VI_locus' is the total number of genomes that have a repeat expansion beyond premutation or full-mutation after visual inspection (per each locus)

Confidence intervals:

- n is the total number of unrelated genomes
- p = total expansions/total number of unrelated genomes
- $q = 1 - p$
- z = 1.96
- $\text{ci\_max} = p + \frac{z^2}{2n} + z \times \frac{\sqrt{\frac{p \times q}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$
- $\text{ci\_min} = p - \frac{z^2}{2n} - z \times \frac{\sqrt{\frac{p \times q}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$

Prevalence estimate (x in 100,000)

x = 100,000/freq_carrier
new_low_ci = 100,000 × ci_max_final
new_high_ci = 100,000 × ci_min_final

Modeling disease prevalence using carrier frequency

The total number of expected people with the disease caused by the repeat expansion mutation in the population (M) was estimated aswhere M_k is the expected number of new cases at age k with the mutation and n is survival length with the disease in years.

M_k is estimated as $M_k = f \times N_k \times p_k$, where f is the frequency of the mutation, N_k is the number of people in the population at age k (according to Office of National Statistics[60]) and p_k is the proportion of people with the disease at age k, estimated at the number of the new cases at age k (according to cohort studies and international registries) divided by the total number of cases.

To estimate the expected number of new cases by age group, the age at onset distribution of the specific disease, available from cohort studies or international registries, was used. For *C9orf72* disease, we tabulated the distribution of disease onset of 811 patients with *C9orf72*-ALS pure and overlap FTD, and 323 patients with *C9orf72*-FTD pure and overlap ALS[61]. HD onset was modeled using data derived from a cohort of 2,913 individuals with HD described by Langbehn et al.[6], and DM1 was modeled on a cohort of 264 noncongenital patients derived from the UK Myotonic Dystrophy patient registry (https://www.dm-registry.org.uk/). Data from 157 patients with SCA2 and *ATXN2* allele size equal to or higher than 35 repeats from EUROSCA were used to model the prevalence of SCA2 (http://www.eurosca.org/). From the same registry, data from 91 patients with SCA1 and *ATXN1* allele sizes equal to or higher than 44 repeats and of 107 patients with SCA6 and *CACNA1A* allele sizes equal to or higher than 20 repeats were used to model disease prevalence of SCA1 and SCA6, respectively.

As some REDs have reduced age-related penetrance, for example, *C9orf72* carriers may not develop symptoms even after 90 years of age[61], age-related penetrance was obtained as follows: as regards *C9orf72*-ALS/FTD, it was derived from the red curve in Fig. 2 (data available at https://github.com/nam10/C9_Penetrance) reported by Murphy et al.[61] and was used to correct *C9orf72*-ALS and *C9orf72*-FTD prevalence by age. For HD, age-related penetrance for a 40 CAG repeat carrier was provided by D.R.L., based on his work[6].

Detailed description of the method that explains Supplementary Tables 10–16:

The general UK population and age at onset distribution were tabulated (Supplementary Tables 10–16, columns B and C). After standardization over the total number (Supplementary Tables 10–16, column D), the onset count was multiplied by the carrier frequency of the genetic defect (Supplementary Tables 10–16, column E) and then multiplied by the corresponding general population count for each age group, to obtain the estimated number of people in the UK developing each specific disease by age group (Supplementary Tables 10 and 11, column G, and Supplementary Tables 12–16, column F). This estimate was further corrected by the age-related penetrance of the genetic defect where available (for example, *C9orf72*-ALS and FTD) (Supplementary Tables 10 and 11, column F). Finally, to account for disease survival, we performed a cumulative distribution of prevalence estimates grouped by a number of years equal to the median survival length for that disease (Supplementary Tables 10 and 11, column H, and Supplementary Tables 12–16, column G). The median survival length (n) used for this analysis is 3 years for *C9orf72*-ALS[62], 10 years for *C9orf72*-FTD[62], 15 years for HD[63] (40 CAG repeat carriers) and 15 years for SCA2 and SCA1[64]. For SCA6, a normal life expectancy was assumed. For DM1, since life expectancy is partly related to the age of onset, the mean age of death was assumed to be 45 years for patients with childhood onset and 52 years for patients with early adult onset (10–30 years)[65], while no age of death was set for patients with DM1 with onset after 31 years. Since survival is approximately 80% after 10 years[66], we subtracted 20% of the predicted affected individuals after the first 10 years. Then, survival was assumed to proportionally decrease in the following years until the mean age of death for each age group was reached.

The resulting estimated prevalences of *C9orf72*-ALS/FTD, HD, SCA2, DM1, SCA1 and SCA6 by age group were plotted in Fig. 3 (dark-blue area). The literature-reported prevalence by age for each disease was obtained by dividing the new estimated prevalence by age by the ratio between the two prevalences, and is represented as a light-blue area.

To compare the new estimated prevalence with the clinical disease prevalence reported in the literature for each disease, we employed figures calculated in European populations, as they are closer to the UK population in terms of ethnic distribution: *C9orf72*-FTD: the median prevalence of FTD was obtained from studies included in the systematic review by Hogan and colleagues[33] (83.5 in 100,000). Since 4–29% of patients with FTD carry a *C9orf72* repeat expansion[32], we calculated *C9orf72*-FTD prevalence by multiplying this proportion range by median FTD prevalence (3.3–24.2 in 100,000, mean 13.78 in 100,000). (2) *C9orf72*-ALS: the reported prevalence of ALS is 5–12 in 100,000 (ref. 4), and *C9orf72* repeat expansion is found in 30–50% of individuals with familial forms and in 4–10% of people with sporadic disease[31]. Given that ALS is familial in 10% of cases and sporadic in 90%, we estimated the prevalence of *C9orf72*-ALS by calculating the ((0.4 of 0.1) + (0.07 of 0.9)) of known ALS prevalence of 0.5–1.2 in 100,000 (mean prevalence is 0.8 in 100,000). (3) HD prevalence ranges from 0.4 in 100,000 in Asian countries[14] to 10 in 100,000 in Europeans[16], and the mean prevalence is 5.2 in 100,000. The 40-CAG repeat carriers represent 7.4% of patients clinically affected by HD according to the Enroll-HD[67] version 6. Considering an average reported prevalence of 9.7 in 100,000 Europeans, we calculated a prevalence of 0.72 in 100,000 for symptomatic 40-CAG carriers. (4) DM1 is much more frequent in Europe than in other continents, with figures of 1 in 100,000 in some areas of Japan[13]. A recent meta-analysis has found an overall prevalence of 12.25 per 100,000 individuals in Europe, which we used in our analysis[34].

Given that the epidemiology of autosomal dominant ataxias varies among countries[35] and no precise prevalence figures derived from clinical observation are available in the literature, we approximated SCA2, SCA1 and SCA6 prevalence figures to be equal to 1 in 100,000.

Local ancestry prediction

100K GP. For each repeat expansion (RE) locus and for each sample with a premutation or a full mutation, we obtained a prediction for the local ancestry in a region of ±5 Mb around the repeat, as follows:

1. We extracted VCF files with SNPs from the selected regions and phased them with SHAPEIT v4. As a reference haplotype set, we used nonadmixed individuals from the 1 K GP3 project. Additional nondefault parameters for SHAPEIT include --mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10 m –pbwt-depth 8.
2. The phased VCFs were merged with nonphased genotype prediction for the repeat length, as provided by EH. These combined VCFs were then phased again using Beagle v4.0. This separate step is necessary because SHAPEIT does not accept genotypes with more than the two possible alleles (as is the case for repeat expansions that are polymorphic).

3. Finally, we attributed local ancestries to each haplotype with RFmix, using the global ancestries of the 1 kG samples as a reference. Additional parameters for RFmix include -n 5 -G 15 -c 0.9 -s 0.9 –reanalyze-reference.

TOPMed. The same method was followed for TOPMed samples, except that in this case the reference panel also included individuals from the Human Genome Diversity Project.

1. We extracted SNPs with minor allele frequency (maf) ≥0.01 that were within ±5 Mb of the tandem repeats and ran Beagle (version 5.4, beagle.22Jul22.46e) on these SNPs to perform phasing with parameters burnin = 10 and iterations = 10.
```
SNP phasing using beagle
java -jar./beagle.22Jul22.46e.jar \
gt = ${input} \
ref = ./RefVCF/hgdp.tgp.gwaspy.merged.chr${chr}.merged.
cleaned.vcf.gz \
out=Topmed.SNPs.maf0.001.chr${prefix}.beagle \
chrom = $region \
burnin = 10 \
iterations = 10 \
map = ./genetic_maps/plink.chr${chr}.GRCh38.map \
nthreads = ${threads} \
impute = false
```
2. Next, we merged the unphased tandem repeat genotypes with the respective phased SNP genotypes using the bcftools. We used Beagle version r1399, incorporating the parameters burnin-its = 10, phase-its = 10 and usephase = true. This version of Beagle allows multiallelic Tander Repeat to be phased with SNPs.
```
java -jar./beagle.r1399.jar \
gt = ${input} \
out = ${prefix} \
burnin-its = 10 \
phase-its = 10 \
map = ./genetic_maps/plink.${chr}.GRCh38.map \
nthreads = ${threads} \
usephase = true
```
3. To conduct local ancestry analysis, we used RFMIX[68] with the parameters -n 5 -e 1 -c 0.9 -s 0.9 and -G 15. We utilized phased genotypes of 1K GP as a reference panel[26].
```
time rfmix \
-f $input \
-r./RefVCF/hgdp.tgp.gwaspy.merged.${chr}.merged.cleaned.
vcf.gz \
-m samples_pop \
-g genetic_map_hg38_withX_formatted.txt \
–chromosome = $c \
-n 5 \
-e 1 \
-c 0.9 \
-s 0.9 \
-G 15 \
–n-threads=48 \
-o $prefix
```

Distribution of repeat lengths in different populations

Repeat size distribution analysis. The distribution of each of the 16 RE loci where our pipeline enabled discrimination between the premutation/reduced penetrance and the full mutation was analyzed across the 100K GP and TOPMed datasets (Fig. 5a and Extended Data Fig. 6). The distribution of larger repeat expansions was analyzed in 1K GP3 (Extended Data Fig. 8). For each gene, the distribution of the repeat size across each ancestry subset was visualized as a density plot and as a box blot; moreover, the 99.9th percentile and the threshold for intermediate and pathogenic ranges were highlighted (Supplementary Tables 19, 21 and 22).

Correlation between intermediate and pathogenic repeat frequency. The percentage of alleles in the intermediate and in the pathogenic range (premutation plus full mutation) was computed for each population (combining data from 100K GP with TOPMed) for genes with a pathogenic threshold below or equal to 150 bp. The intermediate range was defined as either the current threshold reported in the literature[36,69–72] (*ATXN1* 36, *ATXN2* 31, *ATXN7* 28, *CACNA1A* 18 and *HTT* 27) or as the reduced penetrance/premutation range according to Fig. 1b for those genes where the intermediate cutoff is not defined (*AR*, *ATN1*, *DMPK*, *JPH3* and *TBP*) (Supplementary Table 20). Genes where either the intermediate or pathogenic alleles were absent across all populations were excluded. Per population, intermediate and pathogenic allele frequencies (percentages) were displayed as a scatter plot using R and the package tidyverse, and correlation was assessed using Spearman's rank correlation coefficient with the package ggpubr and the function stat_cor (Fig. 5b and Extended Data Fig. 7).

***HTT* structural variation analysis.** We developed an in-house analysis pipeline named Repeat Crawler (RC) to ascertain the variation in repeat structure within and bordering the *HTT* locus. Briefly, RC takes the mapped BAMlet files from EH as input and outputs the size of each of the repeat elements in the order that is specified as input to the software (that is, Q1, Q2 and P1). To ensure that the reads that RC analyzes are reliable, we restrict our analysis to only utilize spanning reads. To haplotype the CAG repeat size to its corresponding repeat structure, RC utilized only spanning reads that encompassed all the repeat elements including the CAG repeat (Q1). For larger alleles that could not be captured by spanning reads, we reran RC excluding Q1. For each individual, the smaller allele can be phased to its repeat structure using the first run of RC and the larger CAG repeat is phased to the second repeat structure called by RC in the second run. RC is available at https://github.com/chrisclarkson/gel/tree/main/HTT_work.

To characterize the sequence of the *HTT* structure, we used 66,383 alleles from 100K GP genomes. These correspond to 97% of the alleles, with the remaining 3% consisting of calls where EH and RC did not agree on either the smaller or bigger allele.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

For 100K GP, full data are available in the Genomic England Secure Research Environment. Access is controlled to protect the privacy and confidentiality of participants in the Genomics England 100K GP and to comply with the consent given by participants for use of their healthcare and genomic data. Access to full data is permitted through the Research Network (https://www.genomicsengland.co.uk/research/academic/join-research-network). For TOPMed, a detailed description of the TOPMed participant consents and data access is provided in Box 1 of ref. 55. TOPMed data used in this manuscript are available through dbGaP. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Supplementary Table 23[55]. A complete list of TOPMed genetic variants with summary-level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). The TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/). DNA sequences and reference placement of assembled insertions are available in VCF format (without individual genotypes) on dbGaP under the TOPMed

GSR accession phs001974. For the 1000 Genomes Project, the WGS datasets are available from the European Nucleotide Archive under accessions PRJEB31736 (unrelated samples) and PRJEB36890 (related samples).

Code availability

The following GitHub repositories used in this work are free to access. ExpansionHunter v3.2.2 (to estimate the repeat size within defined loci): https://github.com/Illumina/ExpansionHunter. REViewer v0.2.7 (to generate pileup plots for quality check): https://github.com/Illumina/REViewer. ExpansionHunter_Classifier (March 2024 release, to automatically run quality assessment of EHv322 call): https://github.com/bharatij/ExpansionHunter_Classifier. Code to analyze repeat structure across *HTT*: https://github.com/chrisclarkson/gel/tree/main/HTT_work. gvcfgenotyper (to merge gVCF files, when inferring ancestry across genomes within the 100K GP and TOPMed datasets): https://github.com/Illumina/gvcfgenotyper. To compute survival curve analysis for *C9orf72*: https://github.com/nam10/C9_Penetrance.

References

55. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* **590**, 290–299 (2021).
56. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
57. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
58. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
59. Dolzhenko, E. et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
60. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. *Office for National Statistics* https://onsdigital.github.io/dp-filter-a-dataset-prototype/v2/pop-est-current (2024).
61. Murphy, N. A. et al. Age-related penetrance of the C9orf72 repeat expansion. *Sci. Rep.* **7**, 2116 (2017).
62. Glasmacher, S. A., Wong, C., Pearson, I. E. & Pal, S. Survival and prognostic factors in C9orf72 repeat expansion carriers: a systematic review and meta-analysis. *JAMA Neurol.* **77**, 367–376 (2020).
63. Bates, G. P. et al. Huntington disease. *Nat. Rev. Dis. Prim.* **1**, 15005 (2015).
64. Diallo, A. et al. Survival in patients with spinocerebellar ataxia types 1, 2, 3, and 6 (EUROSCA): a longitudinal cohort study. *Lancet Neurol.* **17**, 327–334 (2018).
65. Mathieu, J., Allard, P., Potvin, L., Prévost, C. & Bégin, P. A 10-year study of mortality in a cohort of patients with myotonic dystrophy. *Neurology* **52**, 1658–1662 (1999).
66. Wahbi, K. et al. Development and validation of a new scoring system to predict survival in patients with myotonic dystrophy type 1. *JAMA Neurol.* **75**, 573–581 (2018).
67. Sathe, S. et al. Enroll-HD: an integrated clinical research platform and worldwide observational study for Huntington's disease. *Front. Neurol.* **12**, 667420 (2021).
68. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
69. Pulst, S. M. *Spinocerebellar Ataxia Type 2* (Univ. Washington, 2019).
70. Casey, H. L. & Gomez, C. M. *Spinocerebellar Ataxia Type 6* (Univ. Washington, 2019).
71. Caron, N. S., Wright, G. E. B. & Hayden, M. R. *Huntington Disease* (Univ. Washington, 2020).
72. La Spada, A. in *GeneReviews®* (eds Adam, M. P. et al) https://www.ncbi.nlm.nih.gov/books/NBK1116/ (Univ. of Washington, 1999).
73. King, T., Butcher, S. & Zalewski, L. Apocrita - High Performance Computing Cluster for Queen Mary University of London. *Zenodo* https://doi.org/10.5281/zenodo.438045 (2017).

Acknowledgements

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. This work was supported by funding from Barts charity (MGU0569) and a Medical Research Council Clinician Scientist award (MR/S006753/1) to A.T. A.J.S. received support from NIH grants AG075051, NS105781, HD103782 and NS120241, and A.M.-T. received support from NHLBI Biodata Catalyst fellowship 5120339. Age at onset data used in the preparation of this publication for SCA were obtained from the Rare Disease Cures Accelerator-Data and Analytics Platform (RDCA-DAP) funded by FDA grant U18FD005320 and administered by Critical Path Institute. The data were provided to RDCA-DAP by Universitätsklinikum Bonn and Universitätsklinikum Tübingen [Universitätsklinikum Bonn and Universitätsklinikum Tübingen of datasets used by Arianna Tucci in March 2023. Data used in the preparation of this publication for the age at onset data of myotonic dystrophy were obtained by the UK DM Patient Registry, and we acknowledge the Participants and the Steering Committee. Research reported in this paper was supported by the Office of Research Infrastructure of the NIH under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program were supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for 'NHLBI TOPMed - NHGRI CCDG: The BioMe Biobank at Mount Sinai' (phs001644.v1.p1) was performed at the McDonnell Genome Institute (3UM1HG008853-01S2). Genome sequencing for 'NHLBI TOPMed: Women's Health Initiative (WHI)' (phs001237.v2.p1) was performed at the Broad Institute Genomics Platform (contract number HHSN268201500014C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract number HHSN2682018000021). Core support including phenotype harmonization, data management, sample-identity QC and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract number HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Women's Health Initiative (WHI) program is funded by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services through contract numbers HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C and HHSN268201600004C. This manuscript was not prepared in collaboration with investigators of the WHI and does not necessarily reflect the opinions or views of the WHI investigators, or NHLBI.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institute of Health, Department of Health and Human Services, under contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I. We thank the staff and participants of the ARIC study for their important contributions. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contract numbers HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR000040, UL1-TR-001420, UL1-TR-001881 and DK063491. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (contract number HHSN268201800013I), Tougaloo College (contract number HHSN268201800014I), the Mississippi State Department of Health (contract number HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (contract numbers HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). We also thank the staff and participants of the JHS. This research was supported by contract numbers HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083 and N01HC85086 and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at https://chs-nhlbi.org/. This research used data generated by the COPDGene study, which was supported by NIH grants U01 HL089856 and U01 HL089897. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board composed of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis and Sunovion. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (contract numbers N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University or NHLBI. The Mount Sinai BioMe Biobank is supported by The Andrea and Charles Bronfman Philanthropies. This research utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT[73]. Funding was provided by Medical Research Council, Department of Health and Social Care, NHS England, National Institute for Health Research.

Author contributions

K.I. and A.T. conceptualized the research project. K.I., B.J. M.Z., D.G. C.C., S.F., P.G., L.M., M.S., V.E.-P., A.T. and A.J.S. conducted data analysis, interpreted statistical findings and created visual representations of the data. K.I., B.J., M.Z., P.G., A.M.-T., S.J.G., V.G.D., D.J.H.M., C.R. and A.T. performed quality check visually inspecting pileup plots corresponding to repeat expansions. J.V. analyzed the carrier frequency for RFC1 within 100K GP. C.M.-B., H.W., C.T., S.T., J.D.L. and D.R.L. provided data from HD and DM1 patient registries. A.C. provided valuable insights into the genetics of RFC1. Funding and supervision: A.T., A.J.S., P.F., M.J.C., J.H., H.H. Writing—original draft: K.I., A.T., A.J.S., A.D., D.J.H.M., M.Z., C.R. and M.S. meticulously reviewed and edited the manuscript for clarity, accuracy and coherence. All authors carefully reviewed the manuscript, offering pertinent feedback that enhanced the study's quality, and ultimately approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

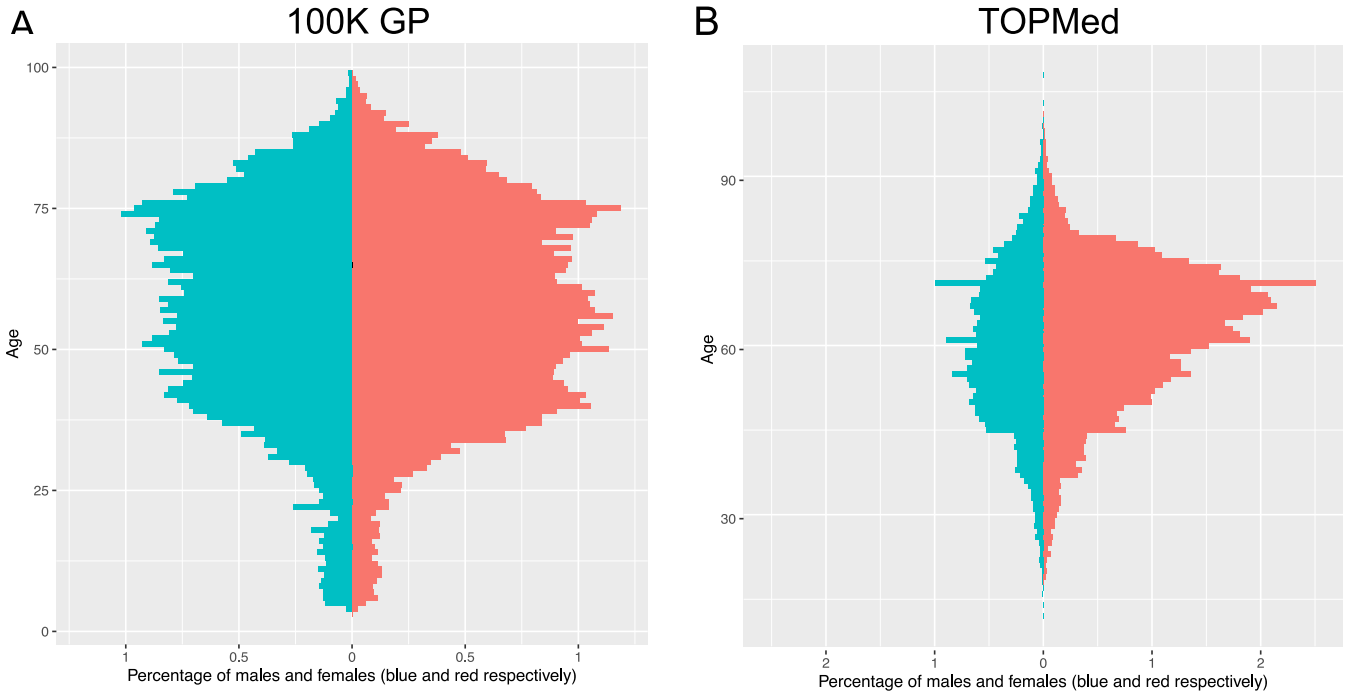
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03190-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03190-5.

Correspondence and requests for materials should be addressed to Arianna Tucci.

Peer review information *Nature Medicine* thanks Weng Khong Lim and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jerome Staal, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03190-5

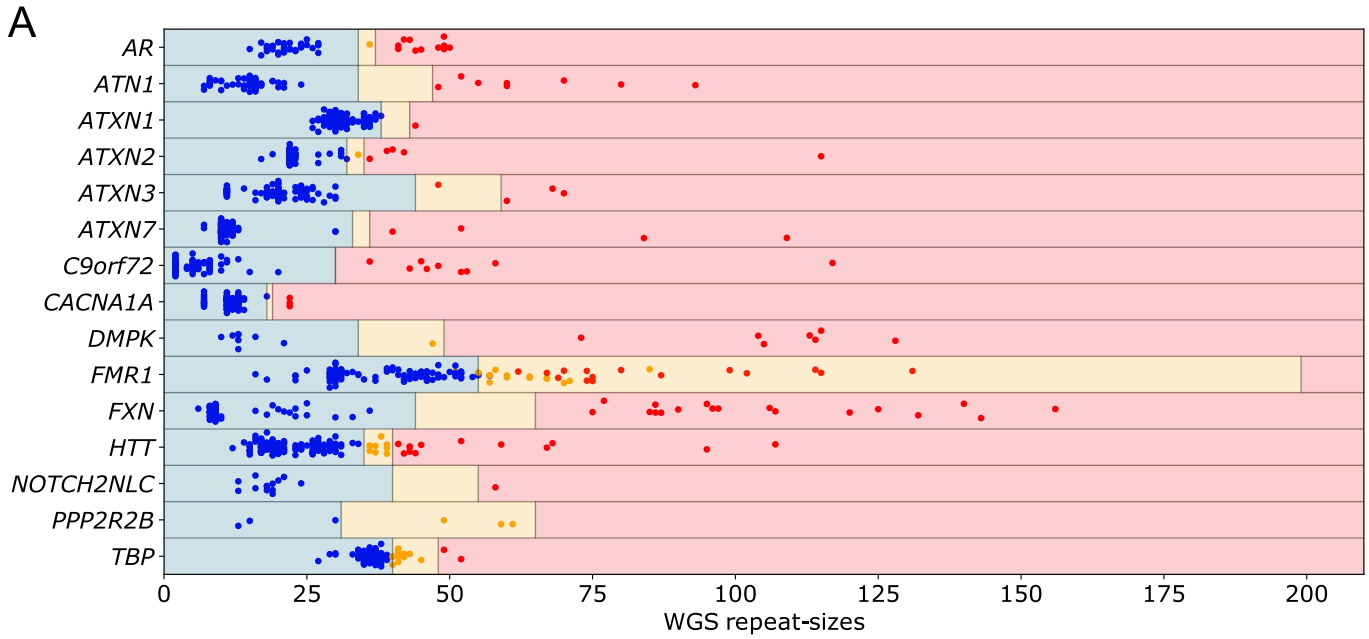


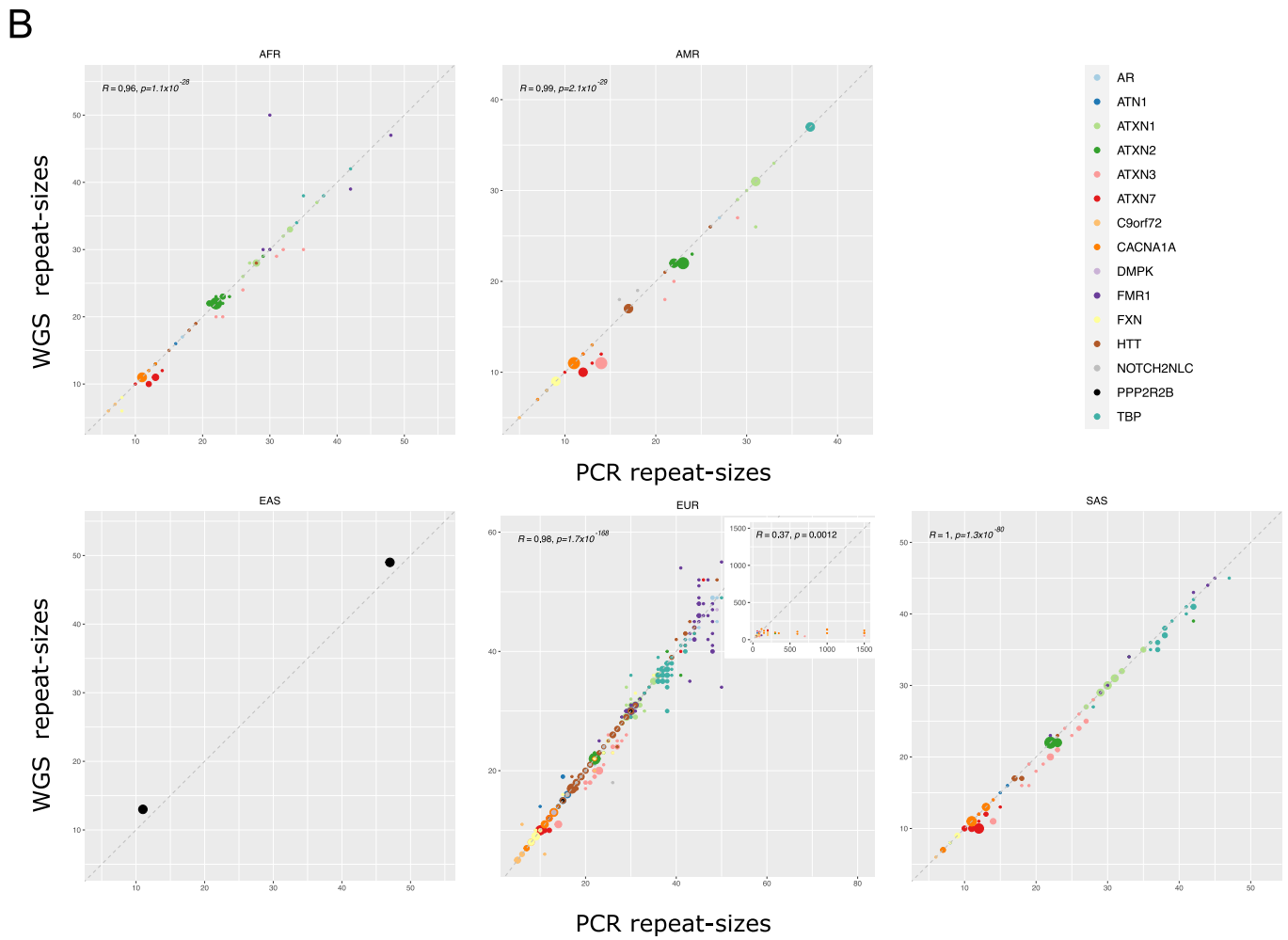
Extended Data Fig. 1 | Study cohorts by gender and age. Population pyramid of (**A**) the 100 K GP and (**B**) TOPMed cohorts.

https://doi.org/10.1038/s41591-024-03190-5

Extended Data Fig. 2 | Principal components of genetic ancestry. First two principal components derived from PCA on A) the 100 K GP and B) TOPMed samples respectively.

https://doi.org/10.1038/s41591-024-03190-5



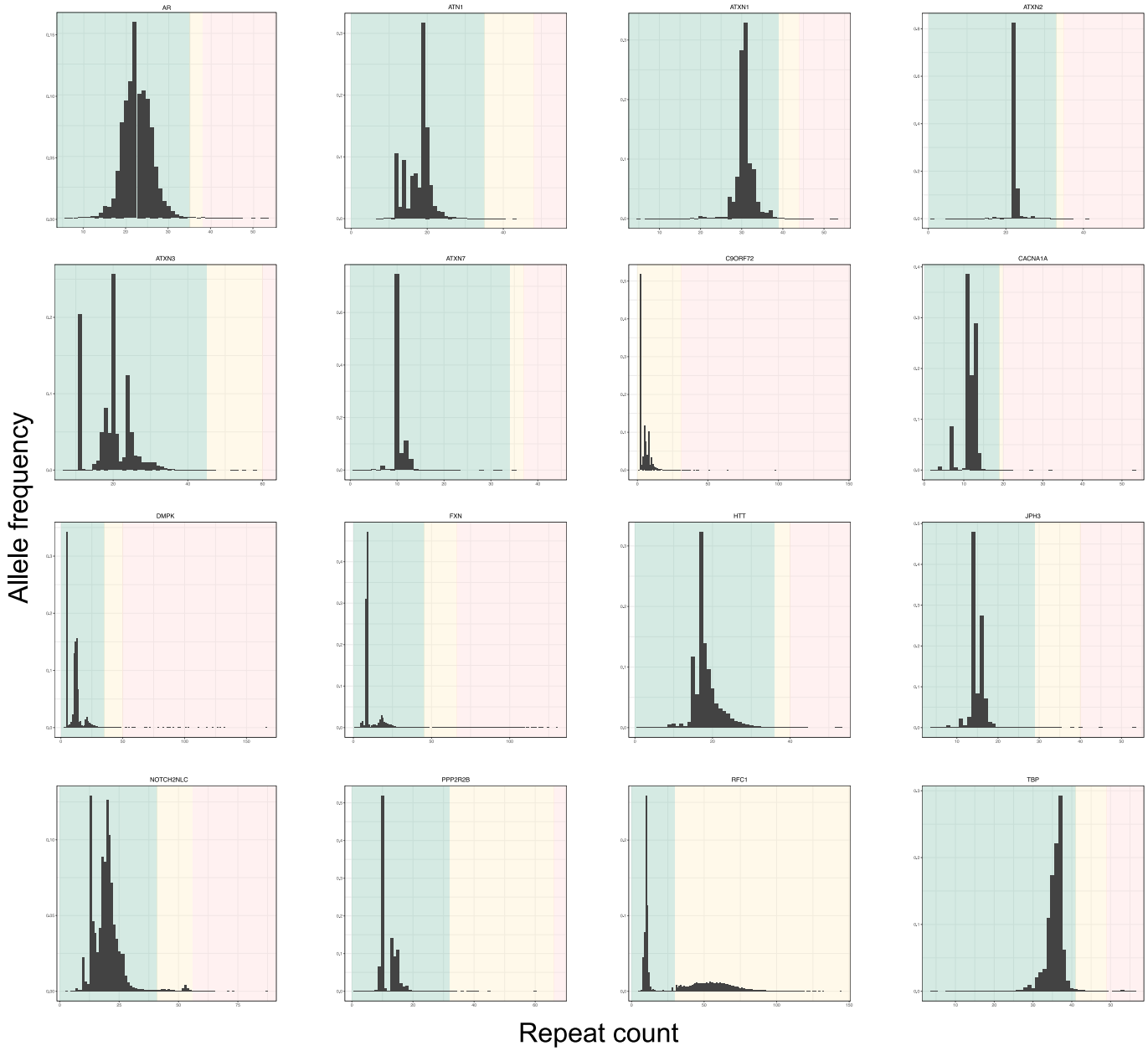


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-024-03190-5

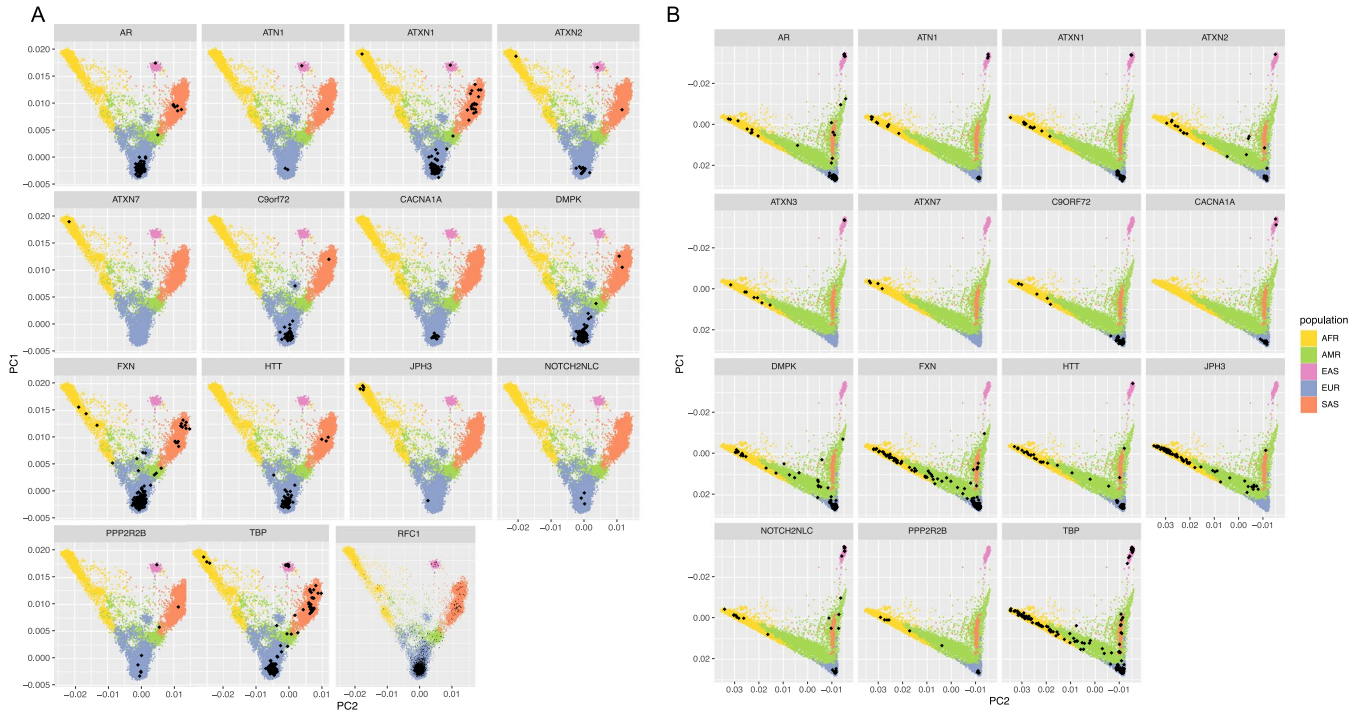
Extended Data Fig. 3 | Experimental estimations of repeat sizes using PCR versus genotypes generated by ExpansionHunter v3.2.2. a). Swim lane plot showing sizes of repeat expansions predicted by ExpansionHunter across 681 samples with expansion calls. Each genome is represented by two points, one corresponding to each allele for each locus, except for those on the X chromosome (that is *FMR1* and *AR*) in males, for which only one point is shown. Points indicate the repeat length estimated by ExpansionHunter after visual inspection and the colours indicate the repeat size as assessed by PCR (blue represents non-expanded; red represents expanded). The regions are shaded to indicate non-expanded (blue), premutation (yellow), and expanded (red) ranges for each gene, as indicated in Table 1. Blue points in yellow or red-shaded regions indicate false positives and red points in blue-shaded regions indicate false negatives. The individual calls are provided in Supplementary Table 3. **b**). Points indicate the RE size estimated by both PCR and EH v3.2.2 split by super-population. We show the R correlation coefficient calculated using Pearson's equation and two-tailed P values. Exact p-values for the regression model: AFR (1.1×10^{-28}), AMR (2.1×10^{-29}), EUR (1.7×10^{-168}), and SAS (1.3×10^{-80}).

https://doi.org/10.1038/s41591-024-03190-5



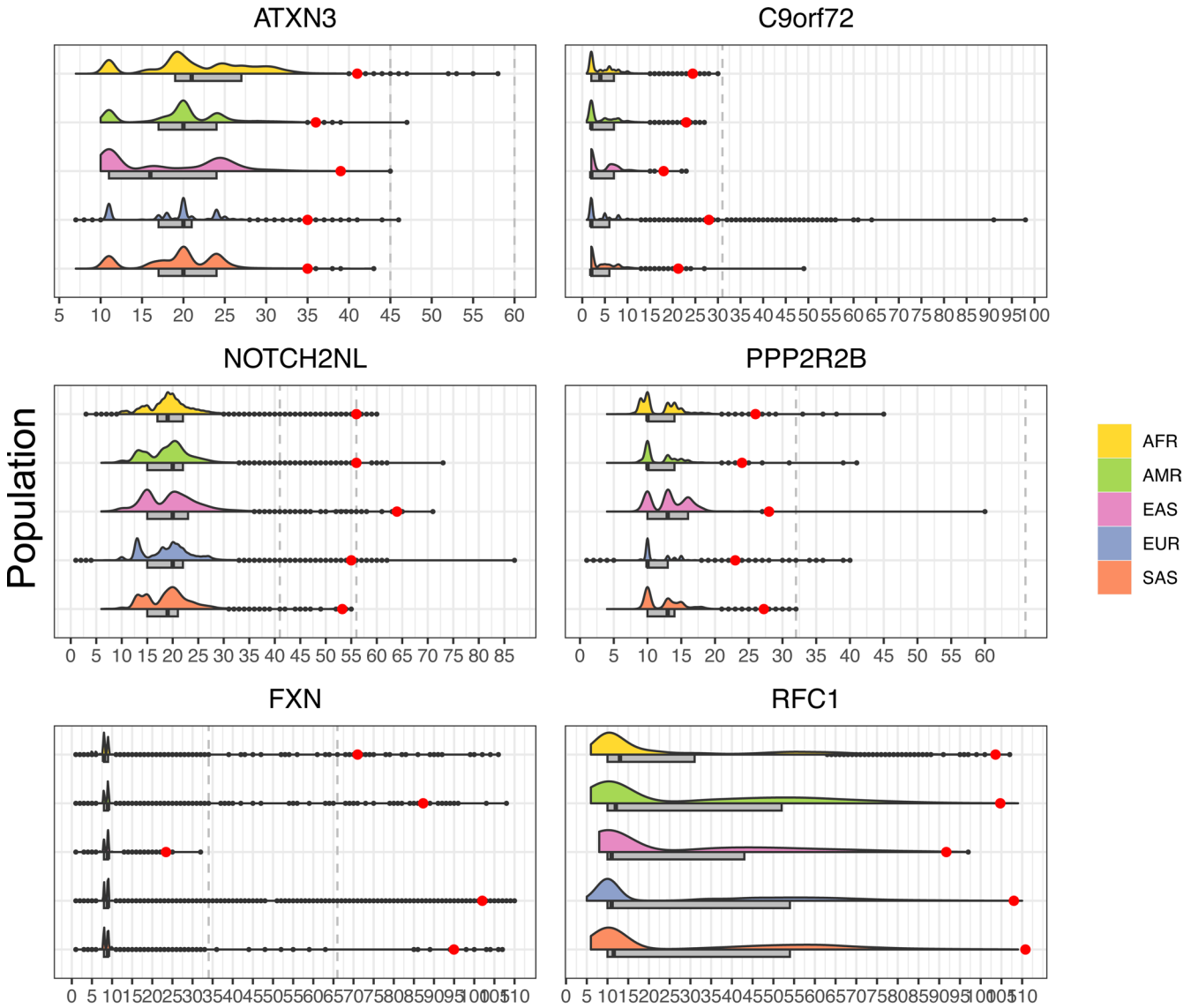
Extended Data Fig. 4 | Distribution of repeat size alleles within the combined 100 K GP and TOPMed cohort. Allele frequency (percentage) predicted by ExpansionHunter in the combined 100 K GP and TOPMed cohorts. The regions are shaded to indicate non-expanded (blue), premutation (yellow), and full mutation expanded (red) ranges for each gene, as indicated in Table 1. For *RFC1*, repeat sizes beyond 30 are shaded as repeat sizes beyond this threshold may represent expanded alleles.

https://doi.org/10.1038/s41591-024-03190-5



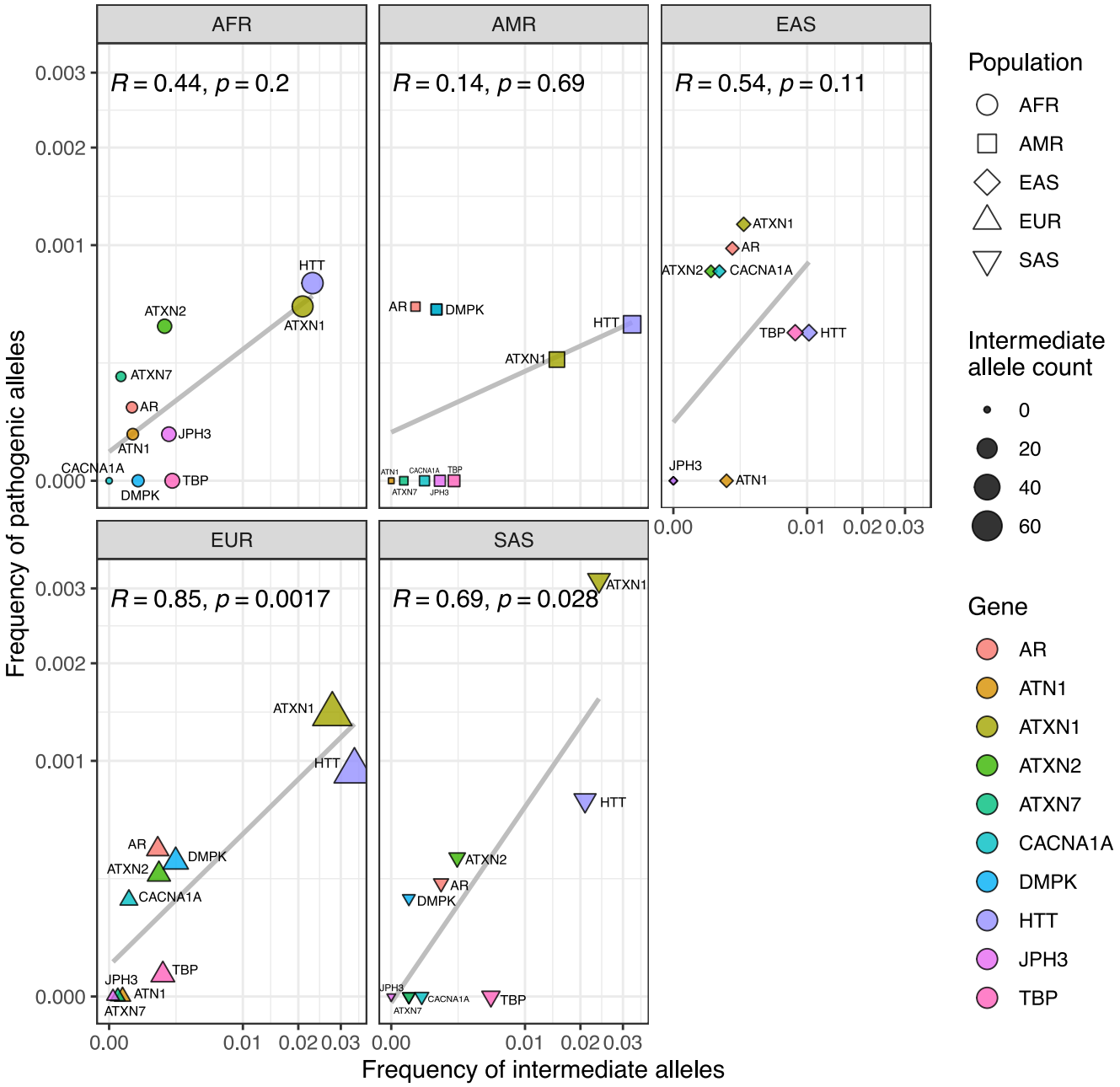
Extended Data Fig. 5 | PC values of genomes carrying normal and pathogenic alleles. Principal component (PC) values on all genomes within (**A**) the 100 K GP and (**B**) TOPMed cohorts. Black dots represent genomes having a repeat size beyond premutation and full mutation range for X-linked and autosomal dominant loci, split by locus. For recessive loci, the plot shows genomes carrying monoallelic and biallelic expansions. Note that *RFC1* has only been analysed in the 100 K GP dataset due to code availability. Note that *ATXN3* is missing from the 100 K GP panels as there are no pathogenic alleles in this cohort (Supplementary Table 7).

https://doi.org/10.1038/s41591-024-03190-5



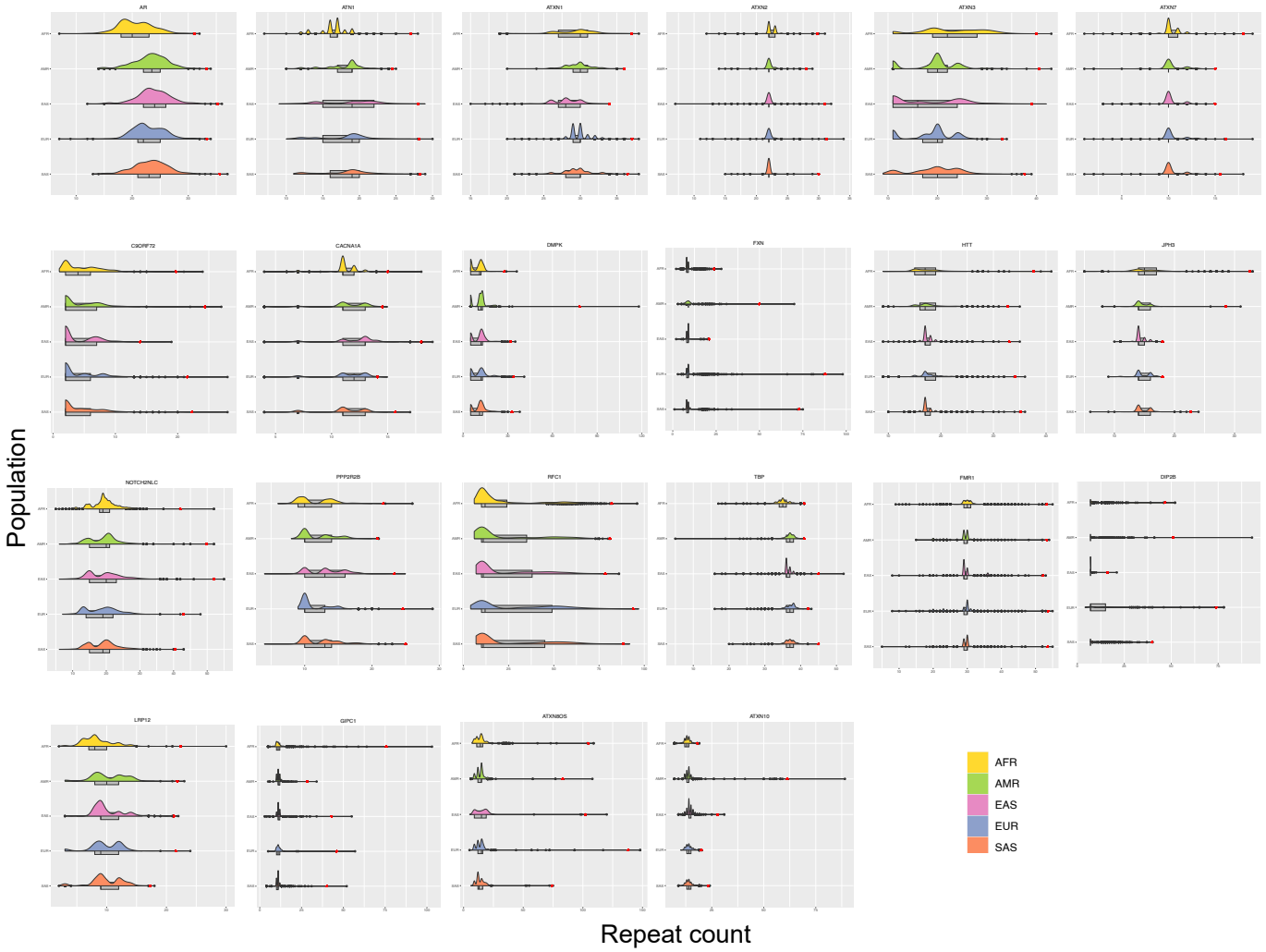
Extended Data Fig. 6 | Distribution of repeat size alleles in different populations in the combined cohort (100K GP and TOPMed). Half-violin plots showing the distribution of alleles in different populations for 6 loci excluded from the correlation analysis from the combined 100K GP and TOPMed cohort (African = 12,786; American = 5,674; East Asian = 1,266; European = 59,568; South Asian = 2,882). Boxplots highlight the interquartile range and median, and black dots show values outside 1.5 times interquartile ranges. Red dots mark the 99.9th percentile for each population and locus. Vertical bars indicate the intermediate and pathogenic allele thresholds (Supplementary Table 20).

https://doi.org/10.1038/s41591-024-03190-5



Extended Data Fig. 7 | Frequency of intermediate alleles versus frequency of pathogenic alleles by population. The scatter plots show the frequency of intermediate allele carriers (x-axis) against the frequency of pathogenic allele carriers (y-axis), based on the thresholds in Supplementary Table 20, split by population. Data points are divided by gene (n = 10), and size represents the total number of intermediate alleles. Correlations were computed using the Spearman method.

https://doi.org/10.1038/s41591-024-03190-5



Extended Data Fig. 8 | Distribution of repeat size alleles by population in the 1 K GP. Distribution of disease RE sizes for 22 genes within the 1 K GP3 split by population (African = 661; American = 347; East Asian = 504; European = 503; South Asian = 489). Half-violin plots show the distribution of alleles, while boxplots highlight the interquartile range and median, and black dots show values outside 1.5 times interquartile ranges. Red dots mark the 99.9th percentile for each population and locus. Repeat size mean;median (Q1-Q3) among all ancestries are in Supplementary Table 19.

nature portfolio

Corresponding author(s): Arianna Tucci
Last updated by author(s): 04/07/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://github.com/Illumina/ExpansionHunter (EHv322) https://github.com/Illumina/REViewer (one version available)
Data analysis	https://github.com/bharatij/ExpansionHunter_Classifier (one version available) https://odelaneau.github.io/shapeit4/ https://github.com/slowkoni/rfmix (version 2) http://faculty.washington.edu/browning/beagle/beagle.html (version 5.4) https://github.com/chrisclarkson/gel/tree/main/HTT_work https://github.com/Illumina/gvcfgenotyper https://github.com/nam10/C9_Penetrance

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

For the 100K GP, full data is available in the Genomics England Secure Research Environment. Access is controlled to protect the privacy and confidentiality of participants in the Genomics England 100,000 Genomes Project and to comply with the consent given by participants for use of their healthcare and genomic data. Access to full data is permitted through the Research Network (https://www.genomicsengland.co.uk/research/academic/join-research-network), and by contacting the corresponding author upon reasonable request.

For TOPMed, a detailed description of the TOPMed participant consents and data access is provided in Box 130. TOPMed data used in this manuscript are available through dbGaP. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Extended Data Tables 23. A complete list of TOPMed genetic variants with summary level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). The TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/). DNA sequence and reference placement of assembled insertions are available in VCF format (without individual genotypes) on dbGaP under the TOPMed GSR accession phs001974.

For the 1000 Genomes Project, the WGS datasets are available from the European Nucleotide Archive under accessions PRJEB31736 (unrelated samples) and PRJEB36890 (related samples).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex used in the paper matches with the genetically inferred sex.
Reporting on race, ethnicity, or other socially relevant groupings	The genetic ancestry used in the paper is based on the 1000 Genomes Project Consortium original work (https://www.nature.com/articles/nature15393), using a random forest model trained to predict five broad super-populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) ancestries based on principal component analysis
Population characteristics	For each participant in the 100K GP, age was calculated based on the year of birth. Clinical data was entered by health-care professionals based on eligibility criteria and rare disease model (https://files.genomicsengland.co.uk/forms/Rare-Disease-Data-Model-Conditions-Phenotypes-and-Clinical-Tests-v1.9.0.pdf). In the 100K GP 53.5% are female and 46.5% male, whereas in TOPMed 63.5% are female and 36.5% males. Median age was 50 in the 100K GP, and 62 in TopMed. The following populations were identified by genetic ancestry predictions in the 100K GP and TOPMed cohorts respectively: Africans (3.5%, 24%), Americans (1.8%, 10.5%), East Asians (0.9%, 2%), Europeans (86%, 63%), South Asians (7.5%, 0.7%).
Recruitment	For the 100K GP, participants were recruited by health-care professionals and researchers from 13 Genomic Medicine Centres in England, and were enrolled in the project if they or their guardian provided written consent for their samples and data to be used in research, including this study. Recruitment of TOPMed cohorts was performed at multiple sites within the US according to diverse inclusion and exclusion criteria, as described in the dbGAP entries for each cohort. We can provide more detailed information.
Ethics oversight	Genomics England has approval from the HRA Committee East of England – Cambridge South (REC Ref 14/EE/1112). For TOPMed, the study was approved by, and the procedures followed were in accordance with, the ethical standards of the Institutional Review Board of the Icahn School of Medicine under HS# 19-01376 and HS# 23-00469.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Any WGS data available across the 3 cohorts (100KGP, TopMed and 1KG) was included in the study if :

	-genomes sequenced following using PCR-free whole genome sequences AND mapped against Human GRCh38/hg38 assembly AND sequenced with a read-length 150bp.
Data exclusions	Genetically related genomes having up to third degree familiar relationship were excluded in all datasets. This is, only unrelated genomes were included in each cohort. Furthermore, In 100KGpand TopMed all genomes from individuals with a neurological disorder were excluded from this analysis (as these cohorts are medical sequencing studies that may have an over-representation of repeat expansions in people with a neurological disease)
Replication	Not applicable to this cross-sectional study
Randomization	Not applicable to this cross-sectional study
Blinding	Not applicable to this cross-sectional study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not relevant as this study is not an interventional trial
Study protocol	Not relevant as this study is not an interventional trial
Data collection	Data analysis was carried out from April 2020 to April 2023.
Outcomes	Frequency of repeat expansion mutation and distribution across different populations