Article

Population-scale repeat expansions elucidate disease risk and brain atrophy

https://doi.org/10.1038/s41586-026-10345-6

Received: 16 May 2025

Accepted: 2 March 2026

Published online: 08 April 2026

Open access

Check for updates

Vijay Kumar Pounraja[1], Jae Hoon Sul[1], Joseph Herman[1], Sean O'Keeffe[1], Veera Rajagopal[1], Xiaodong Bai[1], Michael D. Kessler[1], Neelroop Parikshak[1], Karl Landheer[1], Xingmin Zhang[1], Sean Yu[1], Lance Zhang[1], Michelle G. LeBlanc[1], Jennifer Rico-Varela[1], Frederic Grau[2], Sarah Wolf[1], Sriramkumar Sundaramoorthy[2], Farshid Sepehrband[2], Eli A. Stahl[1], Yuda Huo[2], Mohsin Ahmed[2], Susan Croll[2], GHS-RGC DiscovEHR Collaboration*, Mayo-RGC Project Generation*, Penn Medicine BioBank*, William Salerno[1], John D. Overton[1], Jonathan Marchini[1], Jeffrey Reid[1], Luca A. Lotta[1], Aris Baras[1], Regeneron Genetics Center*, Goncalo R. Abecasis[1], Giovanni Coppola[1,9] & Sahar Gelfman[1,9] ✉

Pathogenic expansions of short tandem repeats (STRs) cause over 70 neurological diseases[1–3]. Here we performed a population-scale survey of pathogenic repeat expansions by analysing repeat length in 37 disease-associated STR loci in a diverse set of 1,020,833 samples using short-read sequencing whole-exome and whole-genome data. Consistent with previous findings, we found that the frequency of pathogenic repeats is higher than the prevalence of corresponding diseases for most loci[4,5]. Associations of repeat length with 7,671 binary traits captured known locus–trait associations, including *HTT* and Huntington's disease, *DMPK* and myotonic disorders and *C9orf72* and motor neuron disease, among others. Finally, we found that, even before disease diagnosis, repeat expansions in several loci strongly associate with increased levels of neurofilament light chain (NfL) and a loss of brain volume in specific disease-associated regions. For example, carriers of *HTT* expansions exhibited a 22.1% loss of putamen volume, and carriers of *CACNA1A* expansions showed a 24.6% loss of cerebellar volume. These observations suggest that both decreased brain volumes and increased NfL levels occur earlier than disease diagnosis. This study demonstrates the use of characterizing repeat expansions from short-read sequencing data in diverse population-scale cohorts and its application to epidemiology and clinical biomarker development.

STRs are an important class of genetic variants that can lead to many neurodegenerative and neuromuscular diseases, including Huntington's disease (HD), motor neuron disease (MND), spinocerebellar ataxias (SCAs), myotonic dystrophies 1 and 2 (DM1 and DM2) and spinal and bulbar muscular atrophy (SBMA)[6–9]. The increase in expansion size that occurs during transmission across generations or somatically, along with changes in repeat motif composition, have been shown to affect disease risk and penetrance, severity, progression and age of onset[10–13]. Since the discovery of the first association between pathogenic repeats in *FMR1* and fragile-X syndrome in 1991, more than 70 neurological diseases have now been associated with repeat expansions[1,2,14]. Repeat-associated diseases can manifest through various mechanisms including toxicity through accumulation of RNA (DM1), misfolded proteins (SCA2), post-translational modifications (SCA1) and transcriptional repression due to hypermethylation (fragile-X syndrome)[15,16]. Although each of these diseases is rare, their collective societal impact is disproportionately higher due to the high cost associated with treatment, caregiving and loss of income, highlighting the need to understand them better to enable and support drug discovery[17].

Although prevalence estimates among specific countries and populations are available for some repeat-expansion diseases (for example, HD and SCAs)[18,19], the frequency and penetrance estimates of most STRs suffer from biases stemming from ascertainment, rarity, awareness of disease, access to robust healthcare and the willingness of patients to participate in genetic testing[4,5,17]. In fact, two of the largest population studies to use PCR to estimate repeat lengths in *DMPK* (among over 50,000 individuals) and *HTT* (among over 7,000 individuals) estimated the frequency of repeat expansion carriers to be higher than previously reported[4,20]. These observations illustrate the need for population-scale studies that take a genotype-first approach to examine the frequency of expanded repeats, risk associated with increased repeat length and the variability of the motifs in the repeat loci. Until recently, studies focusing on repeat-expansion-associated diseases have been mostly disease specific or locus specific, with sample sizes constrained by the rarity of disease and the prohibitive cost of assays to estimate repeat length[21–23]. However, population-scale analyses that can call repeats at multiple loci simultaneously and provide insights into the prevalence of potentially pathogenic expansions are now emerging due to the

[1]Regeneron Genetics Center, Tarrytown, NY, USA. [2]Regeneron Pharmaceuticals, Tarrytown, NY, USA. [9]These authors jointly supervised this work: Giovanni Coppola, Sahar Gelfman. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: sahar.gelfman@regeneron.com

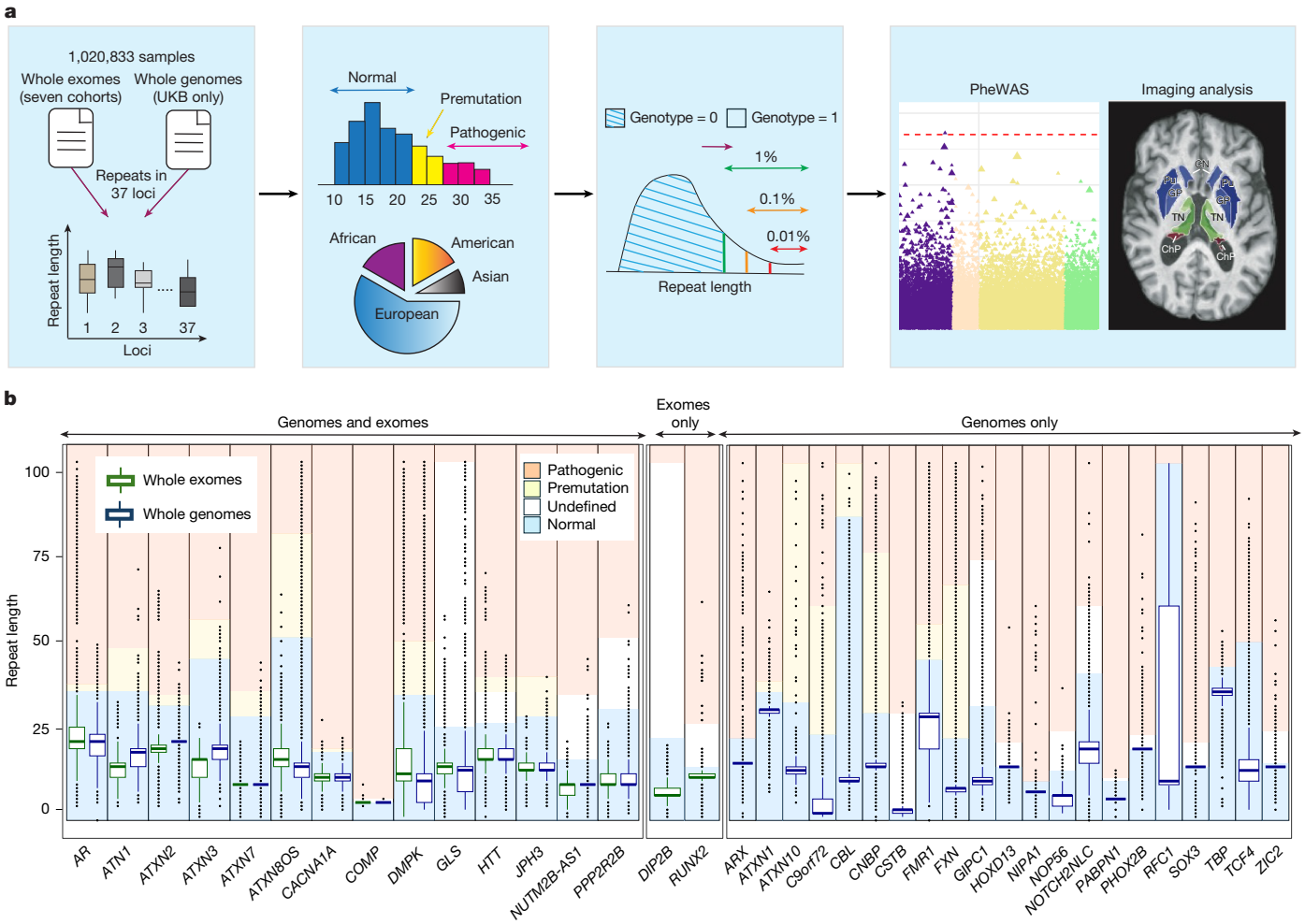


Fig. 1 | Overview of the STR analysis pipeline and the distributions of repeat length in 37 loci. The distribution of repeat length is represented as box and whisker plots with thick whiskers extending from either side of the box representing data variability and individual points representing outliers. **a**, Analysis of repeats identified from 1,020,833 whole exomes and 465,021 whole genomes. For each locus, the frequency of repeat carriers in the premutation and pathogenic ranges was estimated (left) and compared across ancestries (middle left). Next, repeat-length thresholds at the top 1, 0.1 and 0.01 percentiles of the distribution (middle right), along with the premutation and pathogenic length thresholds obtained from the literature, were used for association analyses between 37 loci and 7,671 binary traits and 1,201 brain imaging traits (right). **b**, The repeat-length distribution in 37 loci from 1,020,833 whole-exome (green) and 465,021 whole-genome (blue) samples. Locus names are shown (*x* axis). Background colours represent normal (bright blue), premutation (yellow) and pathogenic (light red) ranges. A white background represents an undefined range for loci that only have a normal range and a pathogenic cut-off in literature. Left, 14 loci for which both exome and genome results were available. Middle, two loci exclusive to whole exomes. Right, 21 loci exclusive to whole genomes. The box plots show the median (centre line), 25% quantile (lower hinge), 75% quantile (upper hinge), the smallest observation greater than or equal to lower hinge − 1.5 × interquartile range (lower whisker) and the largest observation less than or equal to upper hinge + 1.5 × interquartile range (upper whisker).

wide availability of biobank-scale whole-exome sequencing (WES) and whole-genome sequencing (WGS) data as well as dedicated computational methods[3,5,24–27].

Here we describe a population-scale survey of repeat expansions in 1,020,833 samples from 7 diverse cohorts. We estimated repeat size in 37 disease-associated loci, calculated carrier frequencies of pathogenic repeats and compared them across ancestry groups and sequencing platforms. We show that the frequencies of pathogenic expansions in many loci exceed the reported prevalence of the corresponding diseases, vary among ancestry groups and correspond to differences in disease prevalence between subpopulations reported in epidemiological studies. We further evaluated associations between expanded repeats and disease phenotypes, brain imaging traits and plasma protein levels. We replicated expected repeat–trait associations, identified a gradual increase in disease risk and penetrance associated with increased repeat length for a range of repeat associated diseases, and identified a significant loss of brain volume in carriers of pathogenic expansions before to disease diagnosis. This study brings forth the advantages of diverse sets of population-scale WES and WGS datasets in improving our understanding of repeat expansions and the diseases they are associated with.

Repeat expansions in disease-associated loci

We designed a workflow to call and analyse STRs from large electronic health record (EHR) cohorts (Fig. 1a). First, we called repeats in 42 disease-associated loci in 42 distinct genes (here, 'loci' and 'genes' are used interchangeably) from 1,020,833 WES samples using GangSTR[28] from 7 cohorts: the UK Biobank (UKB), the Geisinger Health System (GHS) MyCode cohort, the Mexico City Prospective Study (MCPS) cohort, the Mayo-Clinic Biobank, the Center for Non-Communicable Diseases cohort (CNCD), the University of Pennsylvania Penn Medicine Biobank and the Mount Sinai BioMe BioBank. We supplemented the WES calls with repeats called independently in 38 loci from WGS

samples using ExpansionHunter[29] for 465,021 UKB samples that have both WGS and WES data. This resulted in 42 unique loci, 24 of which were called from both WES and WGS. Although WGS is considered a reliable source for calling repeats accurately using ExpansionHunter[29,30] due to its longer 150 bp paired-end reads, using WES data with 75 bp reads made it challenging for GangSTR to reliably call repeats longer than the average insert size and in less well-covered regions such as UTRs. Furthermore, in contrast to ExpansionHunter, GangSTR could not model complex repeat motifs such as 'GCN' (Supplementary Information 2). To assess the accuracy and consensus among the data sources, we first devised a strict quality control (QC) criteria for repeats called from WES and WGS data using both coverage and quality measurements, and next compared, correlated and validated both data sources for several loci using in silico methods and manual read-stack reviews (see 'Repeat expansion genotyping and QC' section of the Methods; Supplementary Information 3–6). As part of the QC, we excluded loci from WES analysis when they either involved complex repeat motifs or correlated poorly with WGS calls, including *ARX*, *HOXD13*, *NOP56*, *PABPN1*, *SOX3*, *RUNX2* and *ZIC2*. Overall, we identified 37 total unique loci that passed QC in either data source, while 14 loci passed QC in both (Fig. 1a (left) and Table 1).

We next calculated the frequency of individuals that carry repeats above the reported premutation (also referred to in the literature as reduced penetrance or pre-risk alleles) and pathogenic thresholds in each of these loci in the full cohort, as well as within five groups defined by genetic ancestry (Fig. 1a (middle left)). Next, we defined repeat expansion thresholds for each locus independently based on (1) premutation and (2) pathogenic cut-offs curated from an extensive literature search including sources curated in ref. 3 (STRipy) and ref. 31 (GeneReviews), as well as (3) the top 1%, (4) 0.1% and (5) 0.01% of repeat lengths derived empirically from their distributions within each cohort. We then performed a phenome-wide association study (PheWAS) between these repeat length genotypes and 7,671 ICD10 based binary traits (Fig. 1a (middle right and right)). Finally, we compared 1,201 brain-imaging-derived quantitative volumetric phenotypes and 2,939 protein expression levels between carriers and non-carriers of pathogenic repeat expansions (Fig. 1a (right)).

Prevalence of pathogenic repeat expansions

We generated the distribution of repeat length for each of the 37 loci with reliable calls from either WES or WGS data or both (Table 1 and Supplementary Table 1) and calculated the frequency of repeat carriers in premutation and pathogenic ranges (Fig. 1b and Supplementary Figs. 1–37). Before estimating frequencies, we assessed the reliability of the WES caller (GangSTR)[28] by validating its predictions for *HTT* repeats in 59 samples from the GHS cohort using the AmplideX PCR/CE HTT kit. GangSTR correctly predicted 19 out of 25 PCR-validated expanded cases to carry at least 36 repeats, while incorrectly predicting 6 cases as having lower than 36 repeats (false negatives). GangSTR also correctly identified all normal length controls (<27 repeats). Additional PCR validation of 222 samples showed that GangSTR overestimates length above normal by a median of three repeats (Supplementary Information 5 and Supplementary Figs. 38 and 39). We estimated the population frequency of pathogenic expansions in *HTT* before and after PCR-informed length adjustment to be 0.053% (53:100,000) and 0.016% (16:100,000), respectively (Supplementary Fig. 40). Both frequencies were close to the 0.03% (30:100,000) estimated from WGS samples and higher than the reported prevalence of HD (3–7:100,000)[18]. The enrichment of pathogenic carriers compared with disease prevalence was also observed for *CACNA1A*, *C9orf72* and *DMPK*, among others (Fig. 2a). Specifically, the frequency of pathogenic repeat carriers in *CACNA1A* (WES and WGS: 0.02% or 20:100,000), *C9orf72* (WGS: 0.15% or 150:100,000) and *DMPK* (WES: 0.18% or 180:100,000; WGS: 0.07% or 70:100,000) were all higher than the prevalences of their associated diseases: SCA6 (<1:100,000)[32], MND (3.4:100,000)[33] and DM1 (9:100,000)[34], respectively. Homozygous carrier frequencies calculated for loci such as *FXN* and *ARX* that are associated with recessive disorders exhibited similar enrichment (Fig. 2a and Supplementary Table 2). These results are consistent with recent studies that estimate the overall burden of expanded alleles to be higher than expected[4,5]. Owing to the adult-onset nature of some of the STR-associated diseases, we also examined whether stratifying the cohort to individuals older than 65 will affect these results. As the average age in the overall cohort was 57.9 years (median, 58 years), we observed only a slight trend towards increased prevalences for most loci (Supplementary Fig. 41). While these results suggest that expanded repeats might be less penetrant than previously suspected, additional factors can partially explain them as well, including false-positive calls driven by variations in sequence composition, misdiagnosis or a later onset—which may also include the effects of DNA mismatch repair polymorphism on the age at onset[4,5] and tissue-specific somatic expansions[35], ascertainment bias and environmental factors (Discussion).

In addition to the above, sequence interruptions in *HTT* and *ATXN1* were previously reported as risk modifiers[36,37]. Loss of CAA interruption (LOI) in *HTT* was previously reported to decrease the age at onset of HD, while duplication of interruption (DOI) was shown to delay HD onset[38]. We therefore called LOI and DOI data for *HTT* and *ATXN1* in both WES and WGS data (UKB) using ExpansionHunter, which has the ability to call complex motifs. The readouts for *HTT* CAA interruptions and *ATXN1* $(CAGCAT)_2$ interruptions were validated using REViewer—a dedicated read stack viewer for STR regions[39] (Supplementary Information 7 and 8).

For *HTT*, we found 198 *HTT* LOI heterozygous carriers (allele frequency = 0.05%) and no homozygous LOIs. *HTT* DOIs were more frequent in the general population with 18,944 carriers (allele frequency = 4.5%) and 293 homozygous carriers (allele frequency = 0.07%). Among the 198 LOI carriers, none were diagnosed with HD and 1 out of 198 carried expansions in the pathogenic range (≥40). Among the DOI carriers, 10 out of 18,944 carried *HTT* repeats in the pathogenic range and one of them was diagnosed with HD. Excluding the ten DOI carriers from the carriers of pathogenic *HTT* expansions in the UKB had only a slight effect on the pathogenic expansion frequency (0.054% versus 0.052%). We conclude that neither LOI nor DOI has a significant effect on the population frequency of *HTT* pathogenic expansions.

For *ATXN1*, only ten samples carrying normal-length alleles were found to have a complete LOI. Within the 1,353 carriers of *ATXN1* repeat expansions (≥39 repeats), 84% carried the expected $(CAGCAT)_2$ interruption motif on both alleles (that is, same as the reference genome), 6.6% carried $(CAGCAT)_2$ in one allele and $(CAGCAT)_1$ in the other allele, and 5% carried $(CAGCAT)_2$ in one allele and $(CAGCAT)_3$ in the other allele. None carried complete LOIs. A previous report suggested that 39–44 repeats in *ATXN1* is pathogenic only if uninterrupted, yet we did not find any such cases[40]. Thus, in addition to the frequency of *ATXN1* pathogenic expansions (>39, 0.29%; Supplementary Table 1), we also report the stringent frequency of *ATXN1* pathogenic expansions only beyond 44 repeats (105 carriers) to be 0.023%. Further work is required to assess the effect of sequence interruptions on the pathogenicity of repeat expansions in this locus.

Ancestry-specific differences in prevalence

To examine variations in premutation and pathogenic expansions between continental ancestry groups, we used a set of carefully chosen single-nucleotide polymorphisms (SNPs) to infer genetic architecture and subsequently assign each sample to one of five groups: European ancestry (EUR), African ancestry (AFR), south Asian ancestry (SAS), admixed Hispanic or Latin American ancestry (AMR) and east Asian ancestry (EAS) (see 'Ancestry assignment' section of the Methods). We next calculated carrier frequencies within each group and performed

Table 1 | List of 42 disease-associated repeat loci genotyped from WES or WGS samples

Locus	Type	Length cut-offs		Disease	Inheritance	Called from		QC status	
		Premutation	Pathogenic			WES (*n*=28)	WGS (*n*=38)	WES (pass=16)	WGS (pass=35)
AR	Coding	36	38	SBMA of Kennedy	XR	Yes	Yes	Pass	Pass
ATN1	Coding	36	48	Dentatorubral–pallidoluysian atrophy	AD	Yes	Yes	Pass	Pass
ATXN2	Coding	32	35	SCA2	AD	Yes	Yes	Pass	Pass
ATXN3	Coding	45	56	SCA3	AD	Yes	Yes	Pass	Pass
ATXN7	Coding	29	36	SCA7	AD	Yes	Yes	Pass	Pass
CACNA1A	Coding	19	20	SCA6	AD	Yes	Yes	Pass	Pass
COMP	Coding	–	6	Pseudoachondroplasia and multiple epiphyseal dysplasia	AD	Yes	Yes	Pass	Pass
HTT	Coding	36	40	HD	AD	Yes	Yes	Pass	Pass
ATXN8OS	3′ UTR	51	80	SCA8	AD	Yes	Yes	Pass	Pass
DMPK	3′ UTR	35	50	DM1	AD	Yes	Yes	Pass	Pass
JPH3	3′ UTR	29	40	Huntington disease-like 2	AD	Yes	Yes	Pass	Pass
GLS	5′ UTR	–	680	Glutaminase deficiency	AR	Yes	Yes	Pass	Pass
PPP2R2B	5′ UTR	–	51	SCA12	AD	Yes	Yes	Pass	Pass
NUTM2B-AS1	Non-coding	–	35	Neuronal intracellular inclusion disease	AD	Yes	Yes	Pass	Pass
ATXN1	Coding	36	39	SCA1	AD	Yes	Yes	Fail	Pass
PHOX2B	Coding	–	24	Central hypoventilation syndrome	AD	Yes	Yes	Fail	Pass
TBP	Coding	–	43	SCA17	AD	Yes	Yes	Fail	Pass
FMR1	5′ UTR	45	55–199	Fragile-X tremor ataxia syndrome	XD	Yes	Yes	Fail	Pass
			200	Fragile-X syndrome					
ATXN10	Intron	33	800	DM10	AD	Yes	Yes	Fail	Pass
CNBP	Intron	30	75	DM2	AD	Yes	Yes	Fail	Pass
FXN	Intron	23	66	Friedreich ataxia	AR	Yes	Yes	Fail	Pass
AFF2/FMR2	5′ UTR	–	200	Mental retardation, FRAXE type	XR	Yes	Yes	Fail	Pass
DIP2B	5′ UTR	–	273	Neurocognitive problems associated with *FRA12A*	AD	Yes	Yes	Pass	Fail
RUNX2	Coding	–	27	Cleidocranial dysplasia	AD	Yes	Yes	Pass	Fail
HOXA13	Coding	–	22	Hand-foot-uterus syndrome	AD	Yes		Fail	
FOXL2	Coding	–	19	Blepharophimosis	AD	Yes		Fail	
XYLT1	5′ UTR	–	110	Baratela–Scott syndrome	AR	Yes		Fail	
C11orf80	5′ UTR	–	500	Folate sensitive fragile chromosome site	Not inherited	Yes		Fail	
ARX	Coding	–	23	Epileptic encephalopathy, early infantile, 1	XR		Yes		Pass
HOXD13	Coding	–	22	Syndactyly	AD		Yes		Pass
PABPN1	Coding	–	7	Oculopharyngeal muscular dystrophy	AD or AR		Yes		Pass
SOX3	Coding	–	22	Panhypopituitarism	XR		Yes		Pass
			26	Mental retardation, X-linked, with isolated growth hormone deficiency					
ZIC2	Coding	–	25	Holoprosencephaly-5	AD		Yes		Pass
CBL	5′ UTR	85	100	11q-deletion syndrome	Not inherited		Yes		Pass
CSTB	5′ UTR	–	30	Epilepsy, progressive myoclonic 1A/Unverricht–Lundborg disease	AR		Yes		Pass
GIPC1	5′ UTR	–	73	Oculopharyngodistal myopathy 2	AD		Yes		Pass
NIPA1	5′ UTR	–	11	Amyotrophic lateral sclerosis	AD		Yes		Pass
NOTCH2NLC	5′ UTR	–	60	Neuronal intranuclear inclusion disease	AD		Yes		Pass

Continued

Locus	Type	Length cut-offs		Disease	Inheritance	Called from		QC status	
		Premutation	Pathogenic			WES (*n*=28)	WGS (*n*=38)	WES (pass=16)	WGS (pass=35)
C9orf72	Intron	24	60	Amyotrophic lateral sclerosis	AD		Yes		Pass
NOP56	Intron	–	25	SCA36	AD		Yes		Pass
RFC1	Intron	–	400	Cerebellar ataxia, neuropathy and vestibular areflexia syndrome	AR		Yes		Pass
TCF4	Intron	–	50	Fuchs endothelial corneal dystrophy 3	AD		Yes		Pass

For each locus, the location of repeats, premutation and pathogenic cut-offs for repeat length curated from the literature, disorders that they are associated with, the mode of inheritance and QC status are provided. The list includes repeats called from either WES or WGS samples or both. Among the 37 loci, 14 are covered reliably by both WES and WGS, 7 are covered reliably by WGS but not WES, 2 were covered reliably by WES but not WGS, 4 are covered exclusively by WES, and the remaining 14 are covered exclusively by WGS.

pairwise comparisons (Table 2, Supplementary Tables 3 and 4 and Supplementary Figs. 42–78). Comparisons involving AMR samples were adequately powered only within the diverse WES cohort (non-EUR = 25%; AMR = 14.6%), compared with the more homogenous UKB WGS cohort (non-EUR = 5%; AMR = 0.18%). These comparisons identified several known and novel ancestry-specific enrichments. First, we found that the frequency of pathogenic expansions in two loci, *AR* and *ATXN2*, is significantly enriched in AMR (2.2–2.5% and 0.95%) compared with the other ancestries ($P < 1 \times 10^{-10}$; Fig. 2b and Supplementary Table 5). Similarly, pathogenic expansions of both *AR* (2.51%) and *CACNA1A* (0.11%) were enriched in EAS, especially when compared with the low frequencies found in AFR (*AR*: 0.46%, $P = 1.6 \times 10^{-54}$; *CACNA1A*: 0.01%, $P = 0.048$). These enrichments are consistent with documented variations in disease prevalence among the corresponding ancestries and geographies, such as the lower prevalence of SBMA (*AR*) in AFR[41] and the higher prevalence of SCA6 (*CACNA1A*) in Japan and South Korea[19]. While the higher prevalence of SCA2 (*ATXN2*) in AMR has previously been reported[18], our WES calls for *ATXN2* are not highly correlated with WGS calls ($r = 0.51$), warranting further examination of this result in large AMR WGS cohorts. Finally, we found *DMPK* (WES), as well as *C9orf72*, *FXN* and *TCF4* (WGS) expansions to be enriched in EUR compared with in the other ancestries (Fig. 2b and Supplementary Tables 5 and 6). These are also consistent with reported higher prevalences of ALS, Fuchs endothelial corneal dystrophy and Friedreich's ataxia in EUR[42–44]. Beyond replicating reported disease enrichments, we find novel enrichments of *CNBP* and *JPH3* premutation expansions in AFR compared with in EUR (*CNBP*: WGS, 0.98% versus 0.12%; $P = 1.17 \times 10^{-38}$; *JPH3*: WGS, 0.98% versus 0%; $P = 5.66 \times 10^{-46}$). These enrichments and depletions observed within specific ancestry groups could help to establish baseline expectation for disease risk in specific subpopulations.

Disease risk increases with repeat length

Repeat expansions are often assessed through family-based studies with specific emphasis on diseases they are associated with[45–47]. Here, in addition to identifying known locus–trait associations, we intended to assess disease risk in the general population at several repeat-length thresholds for each locus and examine additional traits that associate with expanded repeats in the loci tested. For this purpose, we performed a cohort-wise PheWAS to test associations between the 37 loci and 7,671 binary traits using 759,585 samples from three cohorts (UKB, GHS DiscovEHR, Mayo-Clinic RGC Project Generation[48]). To control for possible batch effects across cohorts arising from data origin (general population versus a medical cohort), source and calling schema (WGS versus WES), as well as ancestry composition, we calculated and used cohort-specific thresholds for each cohort independently, followed by a meta-analysis (see 'Statistical analysis' section of the Methods). First, we defined five repeat expansion thresholds for each locus: two based on repeat-length thresholds curated from literature for premutation and pathogenic expansions, and three based on repeat-length percentiles derived from repeat-length distributions: top 1 percentile, 0.1 percentile and 0.01 percentile (Fig. 1a (middle right) and Supplementary Table 7). Next, for each locus, we associated each threshold with each trait using a logistic regression model that included age, sex and top 10 genetic principal components (PCs) as covariates. We applied multiple testing correction using a study-wide significance cut-off of 1.76×10^{-7} to account for all the traits and loci tested, and identified 168 significant associations for seven unique loci: *AR*, *ATXN2*, *CACNA1A*, *C9orf72*, *DMPK*, *HTT* and *TCF4*, of which 75 are unique locus–trait associations with multiple significant repeat thresholds per locus (Supplementary Table 8).

The top associations were identified between pathogenic (≥50 repeats) expansions in *DMPK* and myotonic disorders, the strongest of which presented a significant causal association with an odds ratio (OR) of 600.6 ($P = 8.8 \times 10^{-149}$; Fig. 3 (top blue triangle)). Pathogenic (≥40) expansions in *HTT* also presented a very high risk for HD with OR = 1,396 ($P = 3 \times 10^{-122}$). Among the top associations, we also observed well-established associations between pathogenic (≥60) repeat length in *C9orf72* and MND (OR = 98.6, $P = 3.7 \times 10^{-111}$), as well as *TCF4* (≥50) and hereditary corneal dystrophies (OR = 7.7, $P = 7.4 \times 10^{-65}$), both called only from WGS data. For each locus, we observed multiple significant associations with related traits. For example, cataract, cardiac pacemaker and implants were strongly associated with *DMPK* expansions, consistent with the wide and expected clinical spectrum of DM1[49]. Similarly, we found significant associations between expanded repeats in *C9orf72* and dementia phenotypes, *CACNA1A* with a lack of coordination and similar locus–trait associations that are consistent with the literature on related disorders (Supplementary Table 8).

As a general observation, locus–trait risks often strengthened with increased repeat-length thresholds. For example, the strongest HD risk was observed at the top 0.01% (1:10,000) *HTT* repeat-length threshold (44 and 41 repeats for WES and WGS, respectively), with an OR of 2,570 ($P = 8.4 \times 10^{-90}$; Fig. 4a (top)). However, a high risk was still observed at a 1% repeat threshold (1:100) of 32 CAG repeats (WES and WGS, OR = 62, $P = 4 \times 10^{-89}$) and 0.1% (1:1,000, 38 and 37 for WES and WGS, respectively, OR = 410.2, $P = 5.9 \times 10^{-125}$), suggesting lower penetrance for the shorter repeats that gradually increases with repeat length. As carriers of longer repeats included within the lenient frequency cut-offs (such as ≥1%) could explain the observed effect, we also estimated effects within specific frequency ranges (1% to 0.1%, 0.1% to 0.01% and above 0.01%) by considering only repeat carriers within each range and excluding carriers above that range. For example, to test risk of repeats between top 1% and 0.1%, we considered individuals carrying repeats within that range as carriers, and below that range as non-carriers (Fig. 4a–c (bottom)). Repeat ranges still exhibit a clear gradual increase in risk for *HTT* with increased repeat lengths (Fig. 4a (bottom)).

We also observed a gradual increase in risk for *DMPK* and *C9orf72* repeats: DM1 risk starts for *DMPK* repeats at the 0.1% to 0.01% frequency range (OR = 235.2, $P = 2.19 \times 10^{-63}$; Fig. 4b) and increases to reach an extremely high risk of OR = 5,878 ($P = 3.05 \times 10^{-95}$) above the 0.01%

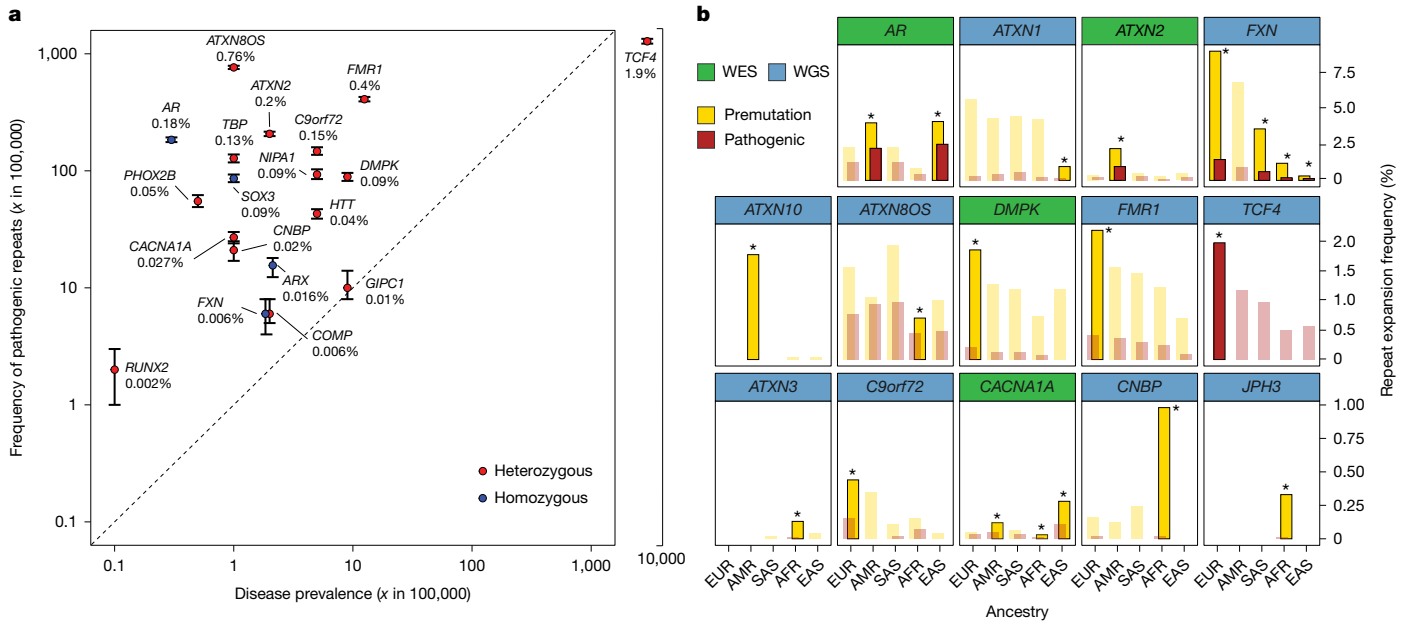


Fig. 2 | The frequency of carriers of repeat expansions within the general population and across five ancestries. a, The prevalence of pathogenic repeat carriers (*x* in 100,000; *y* axis) for 18 loci with at least ten carriers compared with the prevalence of the associated diseases (*x* in 100,000; *x* axis). Pathogenic carriers from 1,012,239 WES and 467,875 WGS samples were pooled together to estimate the average frequency (circles) and 95% confidence intervals (black lines). The black dotted line represents the $x = y$ line. **b**, Ancestry differences (*x* axis) of premutation (yellow) and pathogenic (red) repeat carrier frequencies (*y* axis) for 14 loci with statistically significant enrichments/depletions (dark colours and asterisks) supported by at least five samples and a $P < 1 \times 10^{-6}$ from two-tailed Fisher's exact tests (*P* values were not adjusted for multiple testing). Exact *P* values from pairwise comparisons across ancestries are provided in Supplementary Table 5 (WES) and Supplementary Table 6 (WGS). AFR, AMR, EAS, EUR and SAS are categorical ancestry groups; further details are provided in the 'Ancestry assignment' section of the Methods.

threshold that represents very close to full penetrance. In the case of *C9orf72*, MND risk starts at the 1% to 0.1% frequency range (OR = 3.8, $P = 5.47 \times 10^{-8}$; Fig. 4c), and peaks at the 0.1% to 0.01% frequency range (OR = 105.2, $P = 9.61 \times 10^{-78}$). That said, these frequency ranges overlap premutation thresholds, and therefore do not capture the risk within the premutation range. For example, the 1% to 0.1% range for *HTT* corresponding to repeat lengths 32/32 and 38/37 (for GHS/UKB) overlaps the premutation threshold of 36 and therefore cannot capture risk between 36 and 39 repeats. We therefore tested the premutation ranges specifically for all three loci that exhibited a gradual risk increase: *HTT* (32–35 repeats), *C9orf72* (24–60) and *DMPK* (35–50). However, we observed only one case for each locus at that range, and therefore could not confidently assess the risk at the premutation range. The pattern of gradual risk increase was not observed for the other three loci with significant associations, where both *CACNA1A* and *ATXN2* showed disease risk only above the 0.01% threshold, and *TCF4* had a similar disease risk for all three ranges (Supplementary Fig. 79).

We further examined the effect of age on the penetrance of repeats at five length cut-offs (premutation, pathogenic, 1%, 0.1% and 0.01%) for all the three loci (*HTT*, *DMPK* and *C9orf72*) that exhibited repeat-dependent increase in disease risk (see 'Penetrance calculation' section of the Methods; Fig. 4d–f). We observed two distinct patterns: first, penetrance increased gradually with repeat length as expected by the gradual increase in disease risk. Second, penetrance increased gradually with age and was more pronounced in carriers of longer pathogenic repeats.

Brain volumes associate with expansions

The loss of brain volumes and elevated levels of NfL are expected in neurodegenerative diseases and have been observed previously[50,51]. We examined whether such effects could be observed from brain images and NfL readings of carriers of repeat expansions who are not diagnosed with the corresponding disease. To do this, we selected genes leading to neurodegenerative diseases for which we had both undiagnosed carriers of pathogenic repeat expansions and corresponding MRI ($n = 66{,}258$) or NfL ($n = 54{,}572$) data from the UKB. We then measured the effect of repeat lengths estimated from WGS UKB data on (1) 1,201 brain imaging derived volumetric phenotypes derived from MRI data, and (2) plasma protein levels of 2,939 proteins available from UKB using a linear model. This analysis identified four loci to be significantly associated with volumetric phenotypes: *HTT*, *CACNA1A*, *C9orf72* and *TCF4*. Notably, for the loci causing neurodegenerative diseases, we found that the regions most affected in carriers of expansion at each locus were the same regions affected during the early stages of the corresponding diseases (Fig. 5a,b). For example, by examining brain images of carriers of *HTT* repeat expansions, we observed the strongest effects in the volumes of putamen ($P = 1.4 \times 10^{-18}$) and caudate ($P = 3.2 \times 10^{-11}$; Fig. 5c,d (top)). We next compared raw volumes of both brain regions between undiagnosed (HD) carriers of *HTT* repeats in premutation and pathogenic ranges (36–39 repeats, $n = 122$; ≥40 repeats, $n = 9$) against control carriers of normal repeats (≤26, $n = 58{,}792$) and found the average volumes to be significantly lower in carriers of expansions. This result was also consistent between the sexes (Fig. 5e (top) and Extended Data Fig. 1). While both putamen and caudate volumes were reduced by more than 4% in premutation carriers, volume loss was significantly more pronounced in pathogenic carriers at −22.1% and −20.6% for putamen and caudate, respectively ($P < 1.0 \times 10^{-4}$). In addition to the caudate and putamen, we also observed a strong reduction in the volume of the accumbens (−22.16%, $P = 8.0 \times 10^{-9}$) and pallidum (−17.03%, $P = 1.09 \times 10^{-10}$) of carriers of pathogenic expansions, mirroring the loss expected in these regions as well.

While HD cases were excluded from this analysis, we acknowledge that a few undiagnosed individuals probably experienced early symptoms. The imaging analysis specifically used UKB participants, for whom disease diagnosis was neither reported at the time of imaging nor at any phenotypic update thereafter, supporting the notion that disease-specific loss of brain volumes occurred indeed before diagnosis. Specifically, 84% of the carriers of pathogenic expansions

Table 2 | Frequency of carriers estimated from WES samples for 16 loci across 5 ancestries

Locus	Type	Threshold	Carriers (%)				
			EUR	AMR	SAS	AFR	EAS
AR	Premutation	36	2.32	4.00	2.34	0.83	4.09
	Pathogenic	38	1.29	2.23	1.28	0.46	2.51
ATN1	Premutation	36	0	0	0	0	0
	Pathogenic	48	0	0	0	0	0
ATXN2	Premutation	32	0.36	2.21	0.51	0.31	0.48
	Pathogenic	35	0.20	0.96	0.28	0.12	0.21
ATXN3	Premutation	45	0	0	0	0	0
	Pathogenic	56	0	0	0	0	0
ATXN7	Premutation	29	0	0	0	0	0
	Pathogenic	36	0	0	0	0	0
ATXN8OS	Premutation	51	0.0018	0	0	0	0
	Pathogenic	80	0	0	0	0	0
CACNA1A	Premutation	19	0.0511	0.1160	0.0577	0.0314	0.2823
	Pathogenic	20	0.0286	0.0483	0.0315	0.0121	0.1129
COMP	Premutation	6	0.0079	0.0067	0.0105	0.0217	0.0188
	Pathogenic	6	0.0079	0.0067	0.0105	0.0217	0.0188
DIP2B	Premutation	24	0	0	0	0	0
	Pathogenic	273	0	0	0	0	0
DMPK	Premutation	35	1.85	1.28	1.19	0.74	1.19
	Pathogenic	50	0.2122	0.1218	0.1259	0.0669	0
GLS	Premutation	27	0.3129	0.2121	0.1858	0.3621	0.1345
	Pathogenic	680	0	0	0	0	0
HTT	Premutation	36	0.24	0.25	0.18	0.19	0.06
	Pathogenic	40	0.05	0.05	0.04	0.05	0.00
JPH3	Premutation	29	0.0001	0	0	0	0
	Pathogenic	40	0	0	0	0	0
NUTM2B-AS1	Premutation	17	0.05	0.13	0.07	0.88	0.04
	Pathogenic	35	0	0	0	0	0
PPP2R2B	Premutation	32	0	0	0	0	0
	Pathogenic	51	0	0	0	0	0
RUNX2	Premutation	18	0.0049	0.0007	0.0038	0.0029	0
	Pathogenic	27	0.0022	0	0	0	0

The frequency of premutation and pathogenic repeat carriers in 5 ancestry groups, estimated from around 1 million WES samples for 16 loci that met QC criteria is shown. Repeat-length cut-offs for premutation and pathogenicity for each locus are also provided. For five loci (*ATN1*, *ATXN3*, *ATXN7*, *DIP2B* and *PPP2R2B*), there were no carriers within both premutation and pathogenic ranges.

in *HTT*, *C9orf72* and *CACNA1A* who participated in the imaging study had a follow-up encounter on average 4 years (ranging from 3 months to 7.8 years) after imaging, yet the corresponding diseases were not reported during that timeframe. Given that samples of individuals with a HD diagnosis were excluded from this analysis, these results strongly support previous reports showing significant brain volume loss that occurs before HD diagnosis[52]. Similarly, the strongest effect of *CACNA1A* expansions was with cerebellar grey matter volume ($P = 3.3 \times 10^{-13}$; Fig. 5c,d (middle)) that showed a 24.6% volume decrease in carriers of pathogenic repeats that were not diagnosed with hereditary ataxia compared with control carriers of normal repeats ($P < 1 \times 10^{-4}$; Fig. 5e (middle) and Extended Data Fig. 1). Along the same lines, we observed the strongest effect of *C9orf72* expansions on the volume of the thalamus ($P = 1.6 \times 10^{-43}$; Fig. 5c,d (bottom)), exhibiting a volume reduction of 9% in carriers of pathogenic repeats who were not yet diagnosed with MND compared with control carriers of normal repeats ($P < 1 \times 10^{-4}$, Fig. 5e (bottom) and Extended Data Fig. 1). While this analysis used the more reliable WGS calls for all three loci, the trend of volume decrease was also significant using WES data for both *CACNA1A* and *HTT* but was not as robust (data not shown).

Finally, we tested the effect of *HTT*, *CACNA1A* and *C9orf72* repeats on the expression of 2,939 proteins in plasma obtained from the UKB ($n = 49{,}778$). The most significant result suggested a strong association between *HTT* repeats and increased NfL levels ($P = 5.6 \times 10^{-13}$). This correlation translates to a 1.9-fold ($P = 2.1 \times 10^{-12}$) increase in NfL levels in carriers of pathogenic repeats compared with carriers of normal repeats of *HTT*. While *C9orf72* expansions were not significantly associated with NfL levels in the linear model, they did show a significant increase in NfL levels in carriers of pathogenic repeats compared with carriers of normal repeats (1.14 fold-change, $P = 4.5 \times 10^{-3}$). We did not observe such effects for carriers of *CACNA1A* expansions (Fig. 5f,g). These observations suggest that longer repeats in undiagnosed individuals strongly correlate with brain volumes and, to some extent, with NfL levels earlier than disease diagnosis. While both brain volume loss and increased NfL have been reported in families with early HD and SCAs[53,54], here we show that they can be identified from undiagnosed individuals with high accuracy through sequencing studies in the general population. The disease-specific brain volume loss that we identify in these individuals shows the ability to identify and perhaps begin to treat these diseases before clinical symptoms manifest.

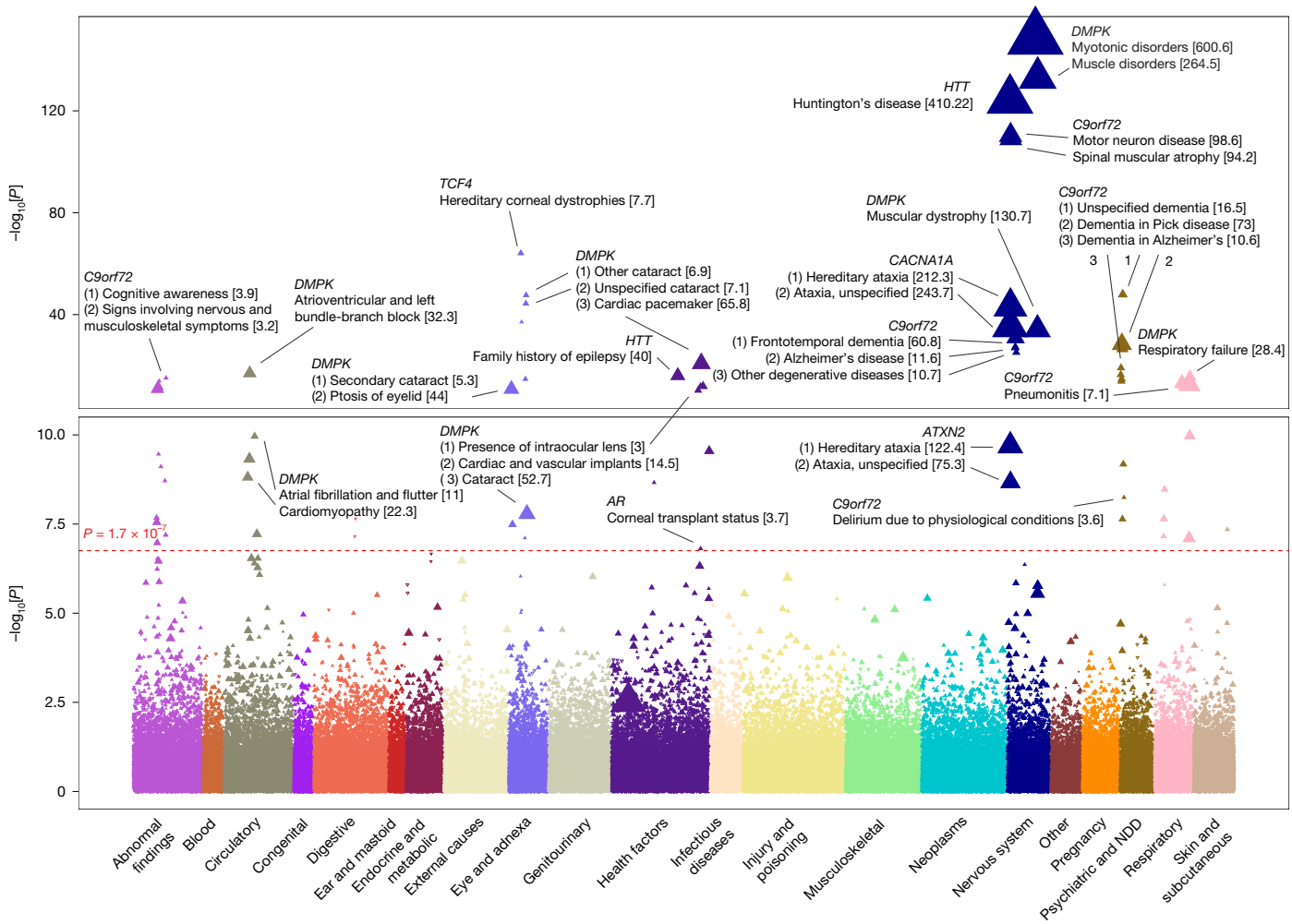


Fig. 3 | Manhattan-style plot with results from phenome-wide association analysis of 7,671 binary traits. The plot illustrates the $-\log_{10}[P]$ for all associations (*y* axis) between repeat-length thresholds and traits. Association models were run with age, age^2, sex, age × sex, and ten ancestry-informed PCs as covariates. *P* values are uncorrected and were calculated using two-sided approximate Firth logistic regression. Association results (triangles) are grouped and coloured on the basis of ICD-10 categories (*x* axis). The triangle size corresponds to the size of the effect; effect sizes are provided in brackets. Only associations supported by at least five case carriers are shown. The study-wide significance threshold is indicated by the dashed red line. The top association for each grouped ICD-10 category is labelled. NDD, neurodevelopmental disorders.

Discussion

Here we performed a very large investigation of repeat expansions in the general population, providing a comprehensive overview of repeat expansions in 37 disease-associated loci in over a million individuals. Our approach to leverage repeats identified independently from WES and WGS data enabled analyses that benefited from the strengths of each data source. The diverse WES cohort that included samples from five ancestry groups enabled the comparisons of pathogenic repeat frequencies across these groups, while the WGS cohort provided estimates of pathogenic repeats in loci not supported by WES. Calling repeats reliably from short-read sequencing data continues to remain a formidable challenge that methods such as GangSTR[28] and ExpansionHunter[29] attempt to address. We encountered several limitations using these tools due to the inherent challenges in (1) estimating repeats longer than the average insert size of short reads, (2) modelling complex repeat motifs, (3) calling repeats reliably in intronic and untranslated regions (UTRs) that are not well covered by WES and (4) identifying interruptions reliably from WGS data that have lower coverage. We tried to address these issues by using PCR validations, manually reviewing read stacks, software and data source comparisons, and stringent coverage-based QC criteria (Supplementary Information 4–6). In our efforts to identify QC metrics suitable for calling *HTT* based on PCR validation, we found that we could not identify QC criteria that fit all loci due to factors such as location of repeats within the gene, coverage, repeat length and complexity of the motif (Supplementary Table 9). We therefore decided to apply uniform QC to all loci. A customized approach tailored to each locus might potentially yield better results.

Large-scale population results show the frequency of pathogenic expansions in loci with highly reliable calls such as *HTT* and *CACNA1A* to be higher than the prevalences of corresponding diseases. After accounting for possible overestimation of *HTT* WES repeat calls based on PCR validation, we still observed a prevalence of 16–53:100,000, almost an order of magnitude higher than the estimated HD prevalence of around 3–7:100,000[18]. These results are consistent with previous estimates from population-scale sequencing studies[4,5], as well as ones derived from PCR validations[4]. Higher prevalences of expansions were also observed for other loci (Fig. 2) and have been previously attributed to heterogeneity in geography, methodologies and ascertainment, among others. To augment this point, our results suggest that longer repeats in genes such as *HTT*, *C9orf72* and *DMPK* might not be fully penetrant. Multiple recent reports discuss the reasons for population-based penetrance being lower than family-based penetrance estimates[55,56]. Higher penetrance in family-based clinical cohorts can manifest from earlier disease onset, more severe disease

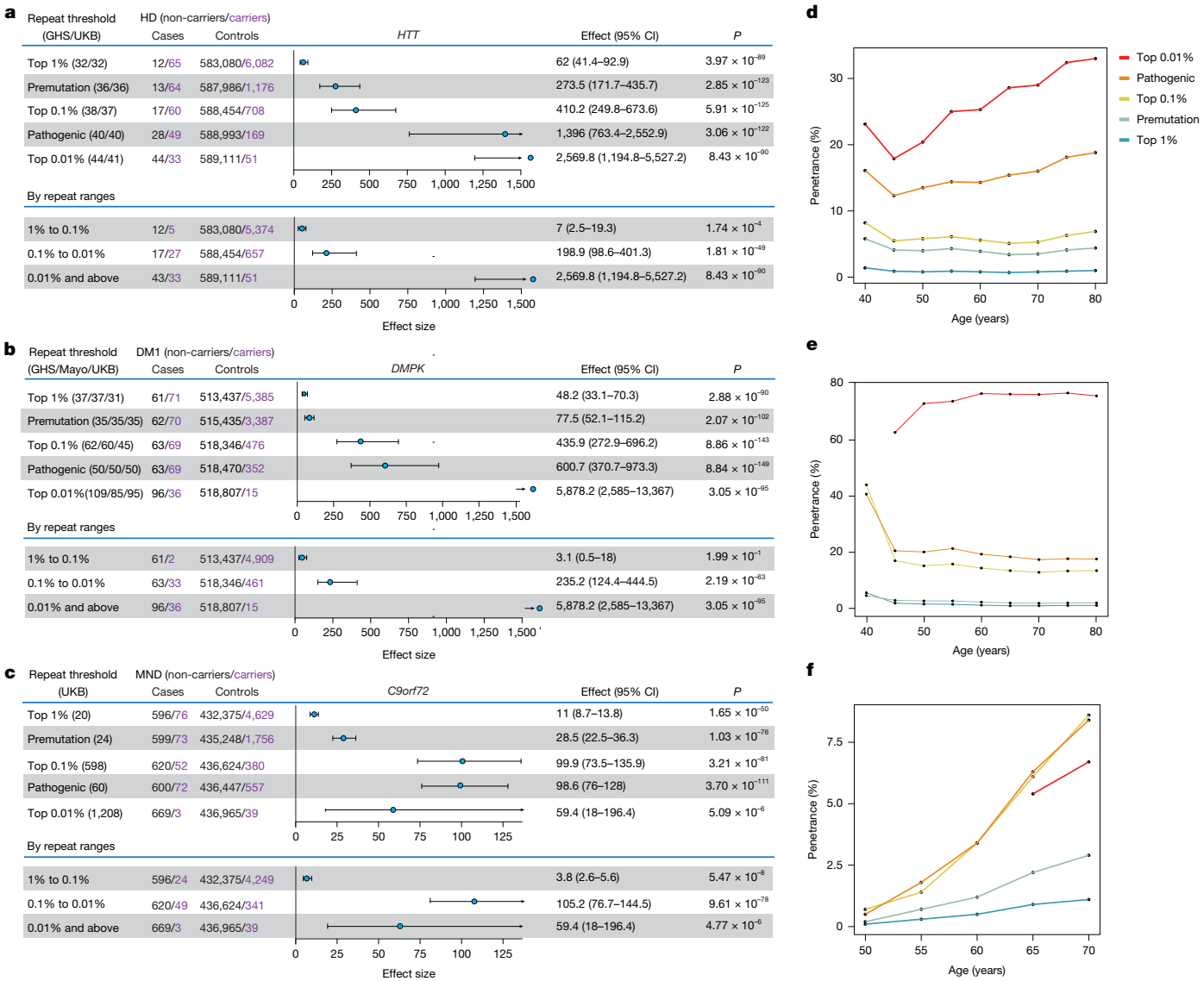


Fig. 4 | Disease risk and penetrance estimated at multiple repeat-length thresholds. a–**c**, The associations between the carrier status of repeats in *HTT* (**a**), *DMPK* (**b**) and *C9orf72* (**c**) versus diagnosis of HD, myotonic disorders and MND, respectively. Association models were run with age, age^2, sex, age × sex, and ten ancestry-informed PCs as covariates. *P* values are uncorrected and were calculated using two-sided approximate Firth logistic regression. The blue circles within the forest plot represent the estimated effect (mean); the black error bars around them represent the 95% confidence intervals. The top portions of each plot show disease risk emerging at five length thresholds (top 1%, top 0.1%, top 0.01%, premutation, pathogenic), and the bottom portion shows the risk within three specific ranges (1% to 0.1%, 0.1% to 0.01%, 0.01% and above). **d**–**f**, The penetrance (*y* axis) of repeats in *HTT* (**d**), *DMPK* (**e**) and *C9orf72* (**f**) by age bins (*x* axis) at five length thresholds (lines) as in **a**.

and the proportion of affected individuals being higher due to ascertainment bias. By contrast, a healthy or a general population cohort has a different ascertainment, and may have a less severe disease that manifests later in life, leading to a lower calculated penetrance. That said, the average age in the overall cohort reported in this work is 57.9 years, limiting our ability to examine younger population below the age at onset, where the penetrance of expansions might be even lower.

The excess of carriers of pathogenic expansions that we observe in this study could therefore be due to several factors including, but not limited to (1) false positives: in silico methods are error-prone and tend to predict longer repeats with wider confidence intervals, even when the accuracy observed through PCR validation for both ExpansionHunter[29] and GangSTR[28] is high (this and other studies)[57]; (2) carriers might not have developed symptoms yet due to later disease onset, and onset might extend beyond life expectancy; (3) misdiagnosis due to inaccurate assessment of early symptoms and underdiagnosis due to various cultural and socioeconomic factors; (4) genetic modifiers: accumulating evidence suggests both age-at-onset and progression of HD are strongly affected by proteins involved in the DNA mismatch repair mechanism, which might explain why many carriers of pathogenic repeats did not yet develop the disease[21,58,59]; (5) changes in sequence composition, such as loss or gain of CAA interruptions in *HTT*, can affect the penetrance of repeats and make motifs harder to identify using available methods, as we show in this study. As advances in long-read sequencing technologies become more accessible and enable accurate genotyping of longer repeat structures and interruptions, they may prove to be a strong validation strategy. Alternatively, SNPs tagging the repeats could serve as effective markers in detecting expansions that could not otherwise be detected from short-read sequencing data. Even with these considerations, repeats called in this study accurately identified expected ancestry-specific enrichments of pathogenic repeats that match the reported disease prevalences within subpopulations, suggesting that the identified expansions

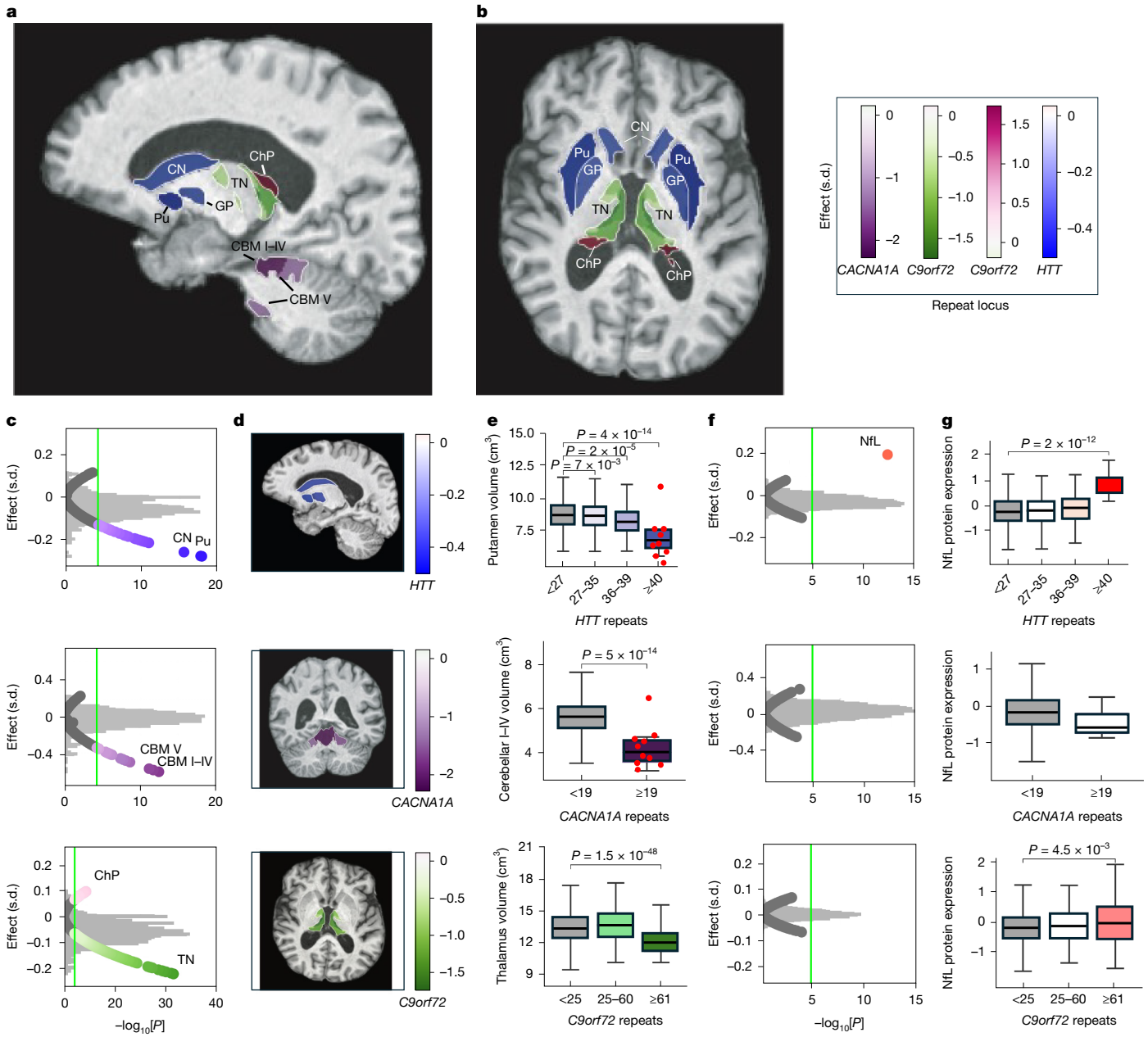


Fig. 5 | The effects of repeat expansions on brain volumes and NfL levels. **a**, Sagittal view of a brain with identified regions coloured by repeat-locus-specific effect: putamen (Pu, *HTT*, blue), caudate nucleus (CN, *HTT*, blue), globus pallidus (GP, blue), thalamic nuclei (TN, *C9orf72*, green), choroid plexus (ChP, *C9orf72*, red) and cerebellar grey matter regions I–IV/V (CBM I–IV/V, *CACNA1A*, purple). **b**, Axial view of a brain with identified regions coloured as in **a**. **c**, Volcano plots representing associations of image-derived phenotypes with pathogenic repeats in *HTT* (top), *CACNA1A* (middle) and *C9orf72* (bottom). Oversized circles represent $-\log_{10}[P]$ across all of the traits analysed, and the grey histograms represent the distribution of corresponding effect sizes. Statistically significant associations are highlighted with different colours. **d**, Brain images depicting the region and effect size for each locus as in **c**. **e**, Comparison of brain volume changes in the putamen among carriers of *HTT* repeats in normal ($n = 58,792$) versus intermediate ($n = 4,039$), premutation ($n = 122$) and pathogenic ($n = 9$) ranges (top). The red circles represent individual datapoints. Middle, comparison of the cerebellum volume between carriers of *CACNA1A* repeats in the normal ($n = 62,938$) versus pathogenic ($n = 10$) range. Bottom, comparison of the thalamus volume between carriers of *C9orf72* repeats in the normal ($n = 62,725$) versus premutation ($n = 134$) and pathogenic ($n = 67$) ranges. **f**, Associations of protein levels with pathogenic repeats in *HTT* (top), *CACNA1A* (middle) and *C9orf72* (bottom). **g**, Comparison of NfL levels among carriers of *HTT* repeats in the normal ($n = 48,404$) versus intermediate ($n = 3,247$), premutation ($n = 108$) and pathogenic ($n = 13$) ranges (top). Middle, comparison of NfL levels between carriers of *CACNA1A* repeats in the normal ($n = 51,713$) versus pathogenic ($n = 13$) range. Bottom, comparison of NfL levels between carriers of *C9orf72* repeats in the normal ($n = 51,278$) versus premutation ($n = 107$) and pathogenic ($n = 68$) ranges. *P* values are unadjusted and were computed using pairwise linear regression analysis after correcting for known confounding factors. Box plots as in Fig. 1.

strongly reflect the reported natural history of diseases. For example, the enrichment of pathogenic *CACNA1A* expansions in the EAS group corresponds to the increased prevalence of SCA6 in Japan[18,19], and a similar enrichment of *C9orf72* expansions in the EUR group corresponds to the increased prevalence of ALS in Caucasian individuals[43]. Moreover, these data offer information on expected ancestry-specific frequencies in cases in which disease prevalence information is lacking. For example, we estimate an enrichment of premutation carriers for *CNBP* and *JPH3* in AFR populations, which might serve as a baseline expectation for diseases such as DM2 and HDL2, and suggest that these diseases

might be either understudied or underdiagnosed within African subpopulations.

The availability of thousands of ICD10-based phenotypes from the EHR cohorts enabled us to perform a PheWAS that replicated known and expected repeat–trait associations, suggested gradual repeat-dependent increase in disease risk and enabled the calculation of penetrance in expansion carriers. We replicate many strong associations including *CACNA1A* and *ATXN2* expansions with hereditary ataxia, both constituting only a small fraction of SCA cases[60]. Similarly, we found very strong associations for *HTT* expansions with HD, *C9orf72* expansions with MND, and *DMPK* expansions with muscle disorders, among others. These results reiterate the reliability and sensitivity of the repeats called from WES/WGS at these loci that strongly identify associations with rare and composite traits, and provide further support to the repeat-dependent penetrance and gradual disease risks observed (Fig. 4). These observations, when taken together with recent results suggesting that *HTT* somatic expansions in striatal tissue take decades to expand from 40 to 80 repeats[35], support the idea that shorter expansions of repeats conferring lower risk and penetrance can take much longer to become pathogenic, and might not manifest during a person's lifetime. Last, we show that reliable repeat expansion calls can identify expected disease pathology in undiagnosed carriers of repeat expansions. A strong loss of volume was observed in the putamen (and caudate) of carriers of *HTT* expansions (−22%, on average), exhibiting a loss akin to pre-HD individuals sooner than 10.8 years prior to disease onset[52]. A similar loss was observed in the cerebellar grey matter of carriers of *CACNA1A* expansions (−24.6%, on average), and the thalamus of carriers of *C9orf72* expansions (−9%, on average). While the EHR-based cohorts used in this study might not capture disease diagnosis as accurately as disease-specific cohorts, the UKB-specific imaging analysis was performed with samples from individuals for whom disease diagnosis was neither reported at the time of imaging nor at any phenotypic update thereafter. These results illustrate how calling repeat expansions in the general population can help to identify individuals who are vulnerable to disease based on their carrier status during prediagnostic stages of disease and provide a wider window for therapeutic intervention.

In conclusion, this work shows the great value in calling repeat expansions from large sequenced cohorts. Even with the caveats and limitations of calling repeats from short-read sequencing, we accurately recapture several well-established ancestry and geographical differences in carrier frequencies as well as strong disease risks. Furthermore, we identify changes in brain volume and NfL levels before disease diagnosis, highlighting the predictive power of this approach. We also shed light on previously unreported enrichments of repeat expansion in specific ancestry groups, and report the risk and age-dependent penetrance arising from the extremely rare or the more common repeat expansions in the general population, offering population-level insights.

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/s41586-026-10345-6.

1. 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).
2. Hiatt, L. et al. STRchive: a dynamic resource detailing population-level and locus-specific insights at tandem repeat disease loci. *Genome Med.* **17**, 29 (2025).
3. Halman, A., Dolzhenko, E. & Oshlack, A. STRipy: a graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Hum. Mutat.* **43**, 859–868 (2022).
4. Kay, C. et al. Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
5. Ibanez, K. et al. Increased frequency of repeat expansion mutations across different populations. *Nat. Med.* **30**, 3357–3368 (2024).
6. MacDonald, M. E. et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
7. Matsuura, T. et al. Mapping of the gene for a novel spinocerebellar ataxia with pure cerebellar signs and epilepsy. *Ann. Neurol.* **45**, 407–411 (1999).
8. brook, J. D. et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **69**, 385 (1992).
9. Spada, A. R. L., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
10. Anvret, M. et al. Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum. Mol. Genet.* **2**, 1397–1400 (1993).
11. Duyao, M. et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* **4**, 387–392 (1993).
12. Igarashi, S. et al. Strong correlation between the number of CAG repeats in androgen receptor genes and the clinical onset of features of spinal and bulbar muscular atrophy. *Neurology* **42**, 2300–2302 (1992).
13. Rosenblatt, A. et al. Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov. Disord.* **27**, 272–276 (2012).
14. Verkerk, A. J. M. H. et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
15. Gatchel, J. R. & Zoghbi, H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.* **6**, 743–755 (2005).
16. Malik, I., Kelley, C. P., Wang, E. T. & Todd, P. K. Molecular mechanisms underlying nucleotide repeat expansion disorders. *Nat. Rev. Mol. Cell Biol.* **22**, 589–607 (2021).
17. Lieberman, A. P., Shakkottai, V. G. & Albin, R. L. Polyglutamine repeats in neurodegenerative diseases. *Ann. Rev. Pathol.* **14**, 1–27 (2019).
18. Medina, A., Mahjoub, Y., Shaver, L. & Pringsheim, T. Prevalence and incidence of Huntington's disease: an updated systematic review and meta-analysis. *Mov. Disord.* **37**, 2327–2335 (2022).
19. van Prooije, T., Ibrahim, N. M., Azmin, S. & van de Warrenburg, B. Spinocerebellar ataxias in Asia: prevalence, phenotypes and management. *Parkinsonism Relat. Disord.* **92**, 112–118 (2021).
20. 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).
21. Genetic Modifiers of Huntington's Disease Consortium. CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell* **178**, 887–900 (2019).
22. McAllister, B. et al. Exome sequencing of individuals with Huntington's disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. *Nat. Neurosci.* **25**, 446–457 (2022).
23. Takahashi, H. et al. A clinical and genetic study in a large cohort of patients with spinocerebellar ataxia type 6. *J. Hum. Genet.* **49**, 256–264 (2004).
24. Cui, Y. et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* **187**, 2336–2341 (2024).
25. Wendt, F. R., Pathak, G. A. & Polimanti, R. Phenome-wide association study of loci harboring de novo tandem repeat mutations in UK Biobank exomes. *Nat. Commun.* **13**, 7682 (2022).
26. Tang, H. et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).
27. Hujoel, M. L. A. et al. Insights into DNA repeat expansions among 900,000 biobank participants. *Nature* **650**, 920–929 (2026).
28. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
29. Dolzhenko, E. et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
30. Ibanez, 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).
31. Wallace, S. E. & Bean, L. J. in *GeneReviews* (eds Adam, M. P. et al.) (Univ. Washington, Seattle, 1993–2025).
32. Casey, H. L. & Gomez. C. M. in *GeneReviews* (eds Adam, M. P. et al.) (Univ. Washington, Seattle, 2019).
33. Park, J., Kim, J. E. & Song, T. J. The global burden of motor neuron disease: an analysis of the 2019 Global Burden of Disease Study. *Front. Neurol.* **13**, 864339 (2022).
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. Handsaker, R. E. et al. Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease. *Cell* **188**, 623–639.e19 (2025).
36. Dawson, J. et al. The frequency and clinical impact of synonymous HTT loss-of-interruption and duplication-of-interruption variants in a diverse HD cohort. *Genet. Med.* **26**, 101239 (2024).
37. Nethisinghe, S. et al. PolyQ tract toxicity in SCA1 is length dependent in the absence of CAG repeat interruption. *Front. Cell Neurosci.* **12**, 200 (2018).
38. Wright, G. E. B. et al. Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *Am. J. Hum. Genet.* **104**, 1116–1126 (2019).
39. Dolzhenko, E. et al. REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* **14**, 84 (2022).
40. Menon, R. P. et al. The role of interruptions in polyQ in the pathology of SCA1. *PLoS Genet.* **9**, e1003648 (2013).

41. Spada, A. L. in *GeneReviews* (eds Adam, M. P. et al.) (Univ. Washington, Seattle, 2022).
42. Vankan, P. Prevalence gradients of Friedreich's ataxia and R1b haplotype in Europe co-localize, suggesting a common Palaeolithic origin in the Franco-Cantabrian ice age refuge. *J. Neurochem.* **126**, 11–20 (2013).
43. Rechtman, L., Jordan, H., Wagner, L., Horton, D. K. & Kaye, W. Racial and ethnic differences among amyotrophic lateral sclerosis cases in the United States. *Amyotroph. Lateral Scler. Frontotemporal Degener.* **16**, 65–71 (2015).
44. Mahr, M. A., Baratz, K. H., Hodge, D. O. & Erie, J. C. Racial/ethnic differences in rates of penetrating or endothelial keratoplasty for fuchs endothelial corneal dystrophy among US Medicare beneficiaries. *JAMA Ophthalmol.* **134**, 1178–1180 (2016).
45. Renton, A. E. et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
46. Seixas, A. I. et al. A Pentanucleotide ATTTC repeat insertion in the non-coding region of DAB1, mapping to SCA37, causes spinocerebellar ataxia. *Am. J. Hum. Genet.* **101**, 87–103 (2017).
47. Corbett, M. A. et al. Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. *Nat. Commun.* **10**, 4920 (2019).
48. Olson, J. E. et al. Characteristics and utilisation of the Mayo Clinic Biobank, a clinic-based prospective collection in the USA: cohort profile. *BMJ Open* **9**, e032707 (2019).
49. De Antonio, M. et al. Unravelling the myotonic dystrophy type 1 clinical spectrum: a systematic registry-based study with implications for disease classification. *Rev. Neurol.* **172**, 572–580 (2016).
50. Schulz, J. B. et al. Visualization, quantification and correlation of brain atrophy with clinical symptoms in spinocerebellar ataxia types 1, 3 and 6. *NeuroImage* **49**, 158–168 (2010).
51. Schonecker, S. et al. Atrophy in the thalamus but not cerebellum is specific for C9orf72 FTD and ALS patients—an atlas-based volumetric MRI study. *Front. Aging Neurosci.* **10**, 45 (2018).
52. van den Bogaard, S. J. et al. Early atrophy of pallidum and accumbens nucleus in Huntington's disease. *J. Neurol.* **258**, 412–420 (2011).
53. Johnson, E. B. et al. Neurofilament light protein in blood predicts regional atrophy in Huntington disease. *Neurology* **90**, e717–e723 (2018).
54. Jung, B. C. et al. MRI shows a region-specific pattern of atrophy in spinocerebellar ataxia type 2. *Cerebellum* **11**, 272–279 (2012).
55. Kingdom, R. & Wright, C. F. Incomplete penetrance and variable expressivity: from clinical studies to population cohorts. *Front. Genet.* **13**, 920390 (2022).
56. Gelb, B. D. 2024 ASHG presidential address: Incomplete penetrance and variable expressivity: Old concepts, new urgency. *Am. J. Hum. Genet.* **112**, 461–466 (2025).
57. van der Sanden, B. et al. Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield. *Genet. Med.* **23**, 1569–1573 (2021).
58. Lee, J. M. et al. Genetic modifiers of Huntington disease differentially influence motor and cognitive domains. *Am. J. Hum. Genet.* **109**, 885–899 (2022).
59. Wexler, N. S. et al. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl Acad. Sci. USA* **101**, 3498–3503 (2004).
60. Bhandari, J., Thada, P. K. & Samanta, D. in *StatPearls* (StatPearls Publishing, 2023).

GHS-RGC DiscovEHR Collaboration

Adam Buchanan[3], David J. Carey[3], Christa L. Martin[3], Michelle Meyer[3], Kyle Retterer[3] & David Rolston[3]

[3]Geisinger Health System, Danville, PA, USA.

Mayo-RGC Project Generation
PG Leadership Team
James R. Cerhan[4], Fergus J. Couch[4] & Janet E. Olson[4]

Statistical Genetics, Bioinformatics
Nicholas B. Larson[4] & Zachary S. Fredericksen[4]

Laboratory Operations
Mine Cicek[4]

[4]Mayo-Clinic, Rochester, MN, USA

Registry Principal Investigators
Joanna M. Biernacka[4], Victor M. Karpyak[4], Prashanthi Vemuri[4], Vijay K. Ramanan[4], Owen A. Ross[4], Mark A. Frye[4], Jeanette E. Eckel Passow[4], Robert R. Jenkins[4], Daniel H. Lachance[4], Kristen L. Drucker[4], Paul A. Decker[4], Matthew L. Kosel[4], Sarah A. McLaughlin[4], Janet E. Olson[4], Fergus J. Couch[4], Kathryn J. Ruddy[4], Nicholas J. Boddicker[4], Wenan Chen[4], Suzette J. Bielinski[4], John C. Lieske[4], W. Michael Hooten[4], Lisa A. Boardman[4], Richard B. Kennedy[4], James R. Cerhan[4], Andrew D. Badley[4], Sean C. Dowdy[4], Shariska Harrington[4], Gretchen E. Glaser[4], Ping Yang[4], Celine M. Vachon[4], Stacey Winham[4], Angela Dispenzieri[4], Samuel O. Antwi[4], Ann L. Oberg[4], Kari G. Rabe[4], Scott H. Kaufmann[4], Ellen L. Goode[4], William A. Cliby[4], Jamie Bakkum-Gamez[4], Sun-Hee Lee[4], J. Eric Ahlskog[4], James H. Bower[4], Peter C. Harris[4], Naveen L. Pereira[4], Mine Cicek[4], Nadia N. Laack[4], Daniel J. Ma[4] & Robert W. Mutter[4]

Management
Jonathan J. Harrington[4]

Mexico city Prospective Study
Jason Torres[5], Jonathan R. Emberson[5], Rory Collins[5], Jaime Berumen[6], Jesús Alegre-Díaz[6], Roberto Tapia-Conyer[6] & Pablo Kuri-Morales[7]

[5]University of Oxford, Oxford, UK. [6]Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico. [7]Tecnológico de Monterrey, Monterrey, Mexico.

Penn Medicine BioBank
PMBB Leadership Team
Daniel J. Rader[8] & Marylyn D. Ritchie[8]

Patient Recruitment, Regulatory Oversight
JoEllen Weaver[8], Nawar Naseer[8], Giorgio Sirugo[8], Afiya Poindexter[8], Yi-An Ko[8], Kyle P. Nerz[8], Jenna Dever[8], Aidan Harvey[8] & Sydney Linn[8]

Lab Operations
JoEllen Weaver[8], Meghan Livingstone[8], Fred Vadivieso[8], Stephanie DerOhannessian[8], Teo Tran[8], Julia Stephanowski[8], Salma Santos[8], Ned Haubein[8] & Joseph Dunn[8]

Clinical Informatics
Anurag Verma[8], Colleen Morse Kripke[8], Marjorie Risman[8], Renae Judy[8] & Colin Wollack[8]

Genome Informatics
Anurag Verma[8], Shefali S. Verma[8], Scott Damrauer[8], Yuki Bradford[8], Scott Dudek[8] & Theodore Drivas[8]

[8]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Regeneron Genetics Center
RGC Management, Leadership Team
Aris Baras[1], Gonçalo Abecasis[1], Adolfo Ferrando[1], Giovanni Coppola[1,9], Andrew Deubler[1], Luca A. Lotta[1], John D. Overton[1], Jeffrey G. Reid[1], Alan Shuldiner[1], Katherine Siminovitch[1], Jason Portnoy[1], Marcus B. Jones[1], Lyndon Mitnaul[1], Alison Fenney[1], Jonathan Marchini[1], Manuel Allen Revez Ferreira[1], Maya Ghoussaini[1], Mona Nafde[1], William Salerno[1], Cristen Willer[1], Lourdes Crane[1], Niek Verweij[1], Eric Jorgenson[1] & Joseph Pickrell[1]

Sequencing, Lab Operations
John D. Overton[1], Christina Beechert[1], Erin Fuller[1], Laura M. Cremona[1], Eugene Kalyuskin[1], Hang Du[1], Caitlin Forsythe[1], Zhenhua Gu[1], Kristy Guevara[1], Michael Lattari[1], Alexander Lopez[1], Kia Manoochehri[1], Prathyusha Challa[1], Manasi Pradhan[1], Raymond Reynoso[1], Ricardo Schiavo[1], Maria Sotiropoulos Padilla[1], Chenggu Wang[1], Sarah E. Wolf[1], Hang Du[1] & Kristy Guevara[1]

Genome Informatics, Data Engineering
Jeffrey G. Reid[1], Mona Nafde[1], Manan Goyal[1], George Mitra[1], Rouel Lanche[1], Vrushali Mahajan[1], Sai Lakshmi Vasireddy[1], Gisu Eom[1], Krishna Pawan Punuru[1], Sujit Gokhale[1], Shehroze Aamer[1], Pooja Mule[1], Mudasar Sarwar[1], Muhammad Aqeel[1], Xiaodong Bai[1], Lance Zhang[1], Sean O'Keeffe[1], Razvan Panea[1], Evan Edelstein[1], Devika Torvi[1], Ayesha Rasool[1], William Salerno[1], Evan K. Maxwell[1], Boris Boutkov[1], Alexander Gorovits[1], Ju Guan[1], Alicia Hawes[1], Olga Krasheninina[1], Samantha Zarate[1], Adam J. Mansfield[1], Lukas Habegger[1], Stephen Tahan[1] & Naveen Karumuri[1]

Analytical Genetics, Data Science
Gonçalo Abecasis[1], Manuel Allen Revez Ferreira[1], Joshua Backman[1], Kathryn Burch[1], Adrian Campos[1], Liron Ganel[1], Sheila Gaynor[1], Benjamin Geraghty[1], Arkopravo Ghosh[1], Christopher Gillies[1], Lauren Gurski[1], Eric Jorgenson[1], Tyler Joseph[1], Michael Kessler[1], Jack Kosmicki[1], Adam Locke[1], Priyanka Nakka[1], Jonathan Marchini[1], Karl Landheer[1], Olivier Delaneau[1], Maya Ghoussaini[1], Anthony Marcketta[1], Joelle Mbatchou[1], Jonathan Ross[1], Carlo Sidore[1], Eli Stahl[1], Timothy Thornton[1], Rujin Wang[1], Kuan-Han Wu[1], Bin Ye[1], Blair Zhang[1], Andrey Ziyatdinov[1], Yuxin Zou[1], Jingning Zhang[1], Kyoko Watanabe[1], Mira Tang[1], Frank Wendt[1], Suganthi Balasubramanian[1], Suying Bao[1], Kathie Sun[1], Chuanyi Zhang[1], Sean Yu[1], Aaron Zhang[1], David Corrigan[1], Dhruv Shidhaye[1], Chen Wang[1], Keyrun Adhikari[1], Alexander Lachmann[1], Anna Alkelai[1], Mark Weiner[1] & Julian Stamp[1]

Therapeutic Area Genetics
Adolfo Ferrando[1], Giovanni Coppola[1,9], Luca A. Lotta[1], Alan Shuldiner[1], Katherine Siminovitch[1], Brian Hobbs[1], Jon Silver[1], William Palmer[1], Rita Guerreiro[1], Amit Joshi[1], Antoine Baldassari[1], Cristen Willer[1], Sarah Graham[1], Ernst Mayerhofer[1], Erola Pairo Castineira[1], Mary Haas[1], Niek Verweij[1], George Hindy[1], Jonas Bovijn[1], Tanima De[1], Luanluan Sun[1], Olukayode Sosina[1], Arthur Gilly[1], Peter Dornbos[1], Moeen Riaz[1], Manav Kapoor[1], Gannie Tzoneva[1], Veera Rajagopal[1], Sahar Gelfman[1,9], Vijay Kumar[1], Jacqueline Otto[1], Jose Bras[1], Silvia Alvarez[1], Jessie Brown[1], Hossein Khiabanian[1], Joana Revez[1], Kimberly Skead[1], Jae Soon Sul[1], Lei Chen[1], Sam Choi[1], Amy Damask[1], Nan Lin[1], Charles Paulding[1], Sameer Malhotra[1], Joseph Herman[1], Jacob McPadden[1], David Blair[1], Joshua Motelow[1] & Julie Horowitz[1]

Research Program Management, Strategic Initiatives
Marcus B. Jones[1], Michelle G. LeBlanc[1], Nadia Rana[1], Jennifer Rico Varela[1], Jaimee Hernandez[1], Larizbeth Romero[1] & Ashley Paynter[1]

Senior Partnerships, Business Operations
Randi Schwartz[1], Lourdes Crane[1], Alison Fenney[1], Jody Hankins[1], Anna Han[1], Samuel Hart[1], Ryan Smith[1] & Sarah Murphy[1]

Business Operations, Administrative Coordinators
Ann Perez-Beals[1], Gina Solari[1], Johannie Rivera-Picart[1], Michelle Pagan[1] & Sunilbe Siceron[1]

Methods

Sample preparation and exome sequencing

Exome sequencing was performed at the Regeneron Genetics Center using a custom automated sample preparation approach (Supplementary Information 9). Samples from MAYO-CLINIC and part of the CNCD (n = 25,580) cohorts were captured with Twist Comprehensive Exome probes and the remaining cohorts were captured with IDT xGen v1 and sequenced using the Illumina HiSeq 2500-v4 or Illumina NovaSeq instrument, with 75-bp paired-end reads and two index reads. For the MAYO-CLINIC cohort sequenced with Twist, probes also included the Twist Diversity SNP panel, for which multipoint refinement was conducted using GLIMPSE before further genotype QC and imputation[61].

Study populations

This study included 1,020,833 participants from seven cohorts: 469,662 participants from the UKB, 173,585 participants from the GHS cohort, 140,996 participants from the MCPS cohort, 116,345 participants from the Mayo-Clinic Biobank, 44,970 participants from the CNCD cohort, 44,027 participants from the University of Pennsylvania Penn Medicine Biobank, 31,248 participants from the Mount Sinai cohort (Supplementary Table 10 and Supplementary Information 1).

Genetic relatedness analysis

To estimate the portion of identical-by-descent (IBD) genomic regions shared between pairs of individuals in each study, we first obtained a set of high-equality common SNPs from the exome variant set by excluding SNPs with minor allele frequency < 10% and genotype missingness >5% and all indels. Furthermore, variants with abnormal heterozygosity rates based on the expected (exp) versus observed (obs) heterozygosity calculations based on empirically determined cutoffs (obs − exp > 0.01 or exp − obs > 0.1) were excluded from further analysis. The asymmetry in the cut-off values was selected to account for the Wahlund effect. IBD estimates were calculated among individuals within the same ancestral superclass that was determined in ancestry predictions as mentioned above using PLINK with a minimum PI_HAT cut-off of 0.1875 to capture out to second-degree relationships, which generates ancestry–version IBD estimates. A separate IBD estimation was calculated among all individuals using a minimum PI_HAT cut-off of 0.3 to identify the first-degree relationships among all samples to generate first-degree family networks, which are connected components of individuals (nodes) and first-degree relationships (edges). Each first-degree family network was analysed using the prePRIMUS pipeline built into PRIMUS[62] using the default settings to produce improved IBD estimates for the relationships within each family network and capture close relationships that span more than one ancestral superclass that were not captured in the ancestry–version IBD estimates. The two versions of IBD estimates were combined in the form of a PLINK .genome file and subject to summary analysis after removing overly related samples with more than 100 close relatives (PI_HAT > 0.1875) or 25,000 relatives (PI_HAT > 0.08). All of the samples in the predicted first- and second-degree relationships were removed to generate the maximum unrelated data set for further analysis.

Ancestry assignment

We used array data released by the UKB study to determine continental ancestry super-groups (AFR, AMR, EAS, EUR and SAS) by projecting each sample onto reference PCs calculated from the HapMap3 reference panel. In brief, we merged our samples with HapMap3 samples and retained only SNPs in common between the two datasets. We further excluded SNPs with minor allele frequency < 10%, genotype missingness > 5% or Hardy–Weinberg equilibrium test $P < 1 \times 10^{-5}$. We calculated PCs for the HapMap3 samples and projected each of our samples onto those PCs. To assign a continental ancestry group to each non-HapMap3 sample, we trained a kernel density estimator (KDE) using the HapMap3 PCs and used the KDEs to calculate the likelihood of a given sample belonging to each of the five continental ancestry groups. When the likelihood for a given ancestry group was greater than 0.3, the sample was assigned to that ancestry group. When two ancestry groups had a likelihood of greater than 0.3, we arbitrarily assigned AFR over EUR, AMR over EUR, AMR over EAS, SAS over EUR, and AMR over AFR. Samples were excluded from analysis if no ancestry likelihoods were greater than 0.3, or if more than three ancestry likelihoods were greater than 0.3.

These sample analysed in this study include 763,174 participants of EUR ancestry, 149,124 participants of AMR ancestry, 57,196 participants of SAS ancestry, 41,405 participants of AFR ancestry and 5,313 participants of EAS ancestry (Supplementary Table 10). Within the UKB cohort, repeats called from WGS data were available for 465,021 samples, of which 441,379 were of EUR ancestry, 848 were of AMR ancestry, 10,085 of SAS ancestry, 9,173 of AFR ancestry and 2,287 of EAS ancestry.

Repeat expansion genotyping and QC

Repeats were called from whole exomes using GangSTR[28] (v.2.5.0) from the CRAM files obtained by aligning the exome sequencing reads to the GRCh38 human reference genome. We used the list of 832,380 (version 13) repeat loci provided by GangSTR to curate a list of 7,046 loci that either overlapped with the exome capture probes or are one of the 34 loci associated with neurological disorders. We then ran GangSTR for each sample using the --targeted and --nonuniform options and provided insert size metrics as inputs. Both methods were run using DNANexus applets built to process samples in batches of 50 using AWS instances with dual core CPUs, 8 GB memory and 75 GB of storage. We curated a high-quality WES call set by identifying calls that either had quality score (Q score reported by GangSTR) of at least 95% or that met the following conditions, (1) be supported by read(s) that enclose the expansion; (2) read depth ≥ 10; (3) Q score ≥ 5%, and ($) maximum-likelihood (ML) estimate within 95% confidence interval (CI). That is, when the point estimate for ML falls outside the 95% CI, the authors of GangSTR deem such predictions to be unreliable and recommend discarding them. For the WGS calls, we used the 'FILTER' column provided by ExpansionHunter to remove calls with the 'LowDepth' flag and considered only calls with the 'PASS' flag for subsequent analyses. Moreover, we applied two filters at the locus level on both generated GangSTR calls and the obtained ExpansionHunter WGS calls (Table 1): loci were excluded (1) if after applying QC, they had fewer than 20% remaining samples; (2) or in cases in which the repeat distribution fell outside of the expected normal range based on literature thresholds. To identify the latter, we calculated for each locus the upper whisker of the repeat length distribution (that is, third quartile + [1.5 × interquartile range]) and excluded loci where the upper whisker fell over the expected pathogenic threshold. These QC criteria identified 16 and 35 loci with reliable distributions from WES and WGS, respectively, 14 of which overlapped between the two sources (Table 1).

STRs genotyped in UKB 500k Whole Genome data using ExpansionHunter (v.4)[29] on the DRAGEN (v.3.7.8) platform were made available as part of the UK Biobank 500k Whole Genome Sequencing data release in November 2023. These STR genotypes were downloaded from the UKB Research Analysis Platform[63] and subsequently used to estimate carrier frequencies and identify phenotypic associations. Spearman correlation between the repeat lengths of the longer allele estimated from WES and WGS samples for the 14 loci was done using the base R function cor.test[64] (Supplementary Table 11).

As the repeat genotypes used in our study were generated from seven different cohorts, we assessed batch effects for three loci we report heavily in the manuscript (*CACNA1A*, *DMPK* and *HTT*). Specifically, we compared the entire repeat-length distribution across all cohorts (Supplementary Figs. 80–82) for these loci as well as the frequency of pathogenic carriers between cohorts and found evidence for batch effects likely arising from the variations in sequencing platform, capture kits

and cohort ascertainment (Supplementary Information 10 and Supplementary Table 12).

Statistical analysis

Association analyses between repeat length genotypes and binary phenotypes were performed for the three largest cohorts with ICD-10 based diagnoses (GHS, UKB and Mayo-Clinic) using REGENIE (v.3.2+), which adjusts for relatedness and population structure[65]. We used the following covariates: age, age^2, sex, age-by-sex, and age^2-by-sex, and the first ten common-variant derived genetic PCs (calculated separately for each cohort). Five genotypes for each of the 37 loci that passed QC were constructed by binarizing the repeat length of the longer allele based on (1) premutation and pathogenic repeat length cut-offs from literature; and (2) the lengths that corresponded to the 1, 0.1, and 0.01 percentiles obtained from the repeat-length distribution of each locus. The percentile cut-offs for each locus were calculated independently within each of the three cohorts (Supplementary Table 7). Repeats on the longer allele that were expanded beyond these thresholds were encoded as 1, and 0 otherwise. The encoded genotypes were then converted to the PLINK 1.9 format (.bed, .bim and .fam files) using the R package named genio (v.1.1.2)[66] and subsequently to PLINK 2.0 format (.pgen, .psam and .pvar) using the --make-pgen command via PLINK (v.2.0)[67]. For the phenome-wide association analysis, diagnoses spanning thousands of ICD-10 traits were considered. Specifically, we analysed the associations between repeat length genotypes in 37 loci and 6,140 traits from GHS, 5,657 traits from the UKB and 5,318 from Mayo-Clinic. Subsequent meta-analysis of these associations were run using the fixed-effect inverse-variance-weighted approach using METAL[68] for the 3,890 traits found in all three cohorts as well as the 1,663 traits found in two of the three cohorts. The study-wide statistical significance threshold was determined to be 1.7×10^{-7} based on the 7,671 phenotypes and 37 loci considered for the PheWAS (significance threshold = 0.05/(7,671 × 37)). In addition to this threshold, to ensure the quality of associations, we considered only the subset of associations supported by at least ten repeat carriers who are cases for the phenotype.

For the analysis of enrichments/depletions within specific ancestry groups, we considered the statistical significance threshold to be 3.3×10^{-5} based on 37 loci, two tested thresholds (premutation and pathogenic), two data sources (WES and WGS) and 10 pairwise comparisons per locus (significance threshold = 0.05/(37 × 2 × 2 × 10)). In addition to this threshold, to ensure the quality of associations, we considered only the subset of tests supported by at least five repeat carriers.

Linear regression models

Association analysis for the brain imaging traits was run using linear regression with repeat status as a categorical predictor, rank inverse normal transformed brain volume as the outcome and the following covariates: the first 10 array PCs, first 20 exome rare variant PCs, UKB sequencing batch, sex, age, age^2, age × sex, UKB assessment centre, position of brain within the scanner, and head size from SIENAX. An analogous analysis was performed with NfL levels in the plasma. To test for a linear association between repeat length and the outcome variable, we also ran the same analysis with $\min(0, L - T + 1)$ as a predictor, where L is the repeat length and T is the premutation repeat-length threshold. Quoted P values in the main text are for the coefficients on repeat status/length in these models.

Penetrance calculation

Penetrance was calculated as follows: the calculation considered the age at diagnosis for cases and the age at the last encounter for controls. In the case of a recorded death, we considered age at death as the most recent age. The samples were further divided into repeat expansion threshold groups based on their length. The calculation of the proportion of cases was cumulative, considering all cases and controls up until the tested age bin. The age-dependent penetrance was calculated as follows:

penetrance = (case carriers of repeat length>t threshold who are younger than age a)/(case and control carriers of repeat length > t threshold who are younger than age a).

For example, at a = 50 and a threshold of t = 35 repeats, the penetrance was calculated as

penetrance = (case carriers of repeat length > 35 who are younger than 50)/(case and control carriers of repeat length > 35 who are younger than 50). At the age of 80, this calculation included all samples younger than 80 years of age.

Brain imaging and plasma protein data from the UKB

T1-weighted brain images and plasma protein data were obtained from the UKB for 66,258 and 54,572 individuals, respectively. The raw structural T1-weighted brain images were processed through an internal pipeline similar to that presented previously[69]. In brief, the raw .zip files were accessed from the UKB Research Access Platform and converted to NIfTI using dcm2niix[70]. The image was then brain-extracted[71], rigidly transformed to MNI152 space using FLIRT[72] (FSL v.6.0.7.8) followed by a deformable registration to MNI152 space using FNIRT[73]. The rigid and deformable transformations are combined into a single transformation. The inverse of this transformation is then obtained, and is used transform the MNI152 atlas to the structural space, which is then used as a mask for a final brain extraction. Tissue-type segmentation was then performed using FSL's FAST[74], which also provides the intensity bias, which is used to obtain the final brain-extracted and intensity-bias-corrected T1-weighted image. This final image was then segmented using FastSurfer[75] (v.2.3.0), and cortical thicknesses and other morphological traits were obtained using FreeSurfer (v.7.3.2). Cerebellum volumes were obtained from CerebNet[76].

Samples with repeat information for *HTT* (n = 62,963) were split into four groups based on repeat length: normal (<27 repeats), intermediate (27–35; n = 4,049), premutation (36–39; n = 122) and pathogenic (≥40; n = 9). Similarly, individuals with repeat information for *CACNA1A* (n = 63,960) were split into two groups: normal (<19) and premutation + pathogenic (≥19; n = 11).

HTT PCR validations

The AmplideX PCR/CE *HTT* Kit (49657; Asuragen) was used to PCR amplify the *HTT* trinucleotide CAG fragment starting from 2 µl of purified genomic DNA. The samples were prepared for the PCR with a master mix from Asuragen containing *HTT* PCR Mix (5.0 µl), *HTT* forward and reverse primer mix (3.0 µl). Tubes of master mix were vortex-mixed, centrifuged and transferred into a thermal cycler. Thermal cycling was performed on Eppendorf Nexus Gradient Mastercycler (Eppendorf) using the following cycle: 95 °C for 5 min, 10 cycles of 97 °C for 35 s, 64 °C for 35 s, 68 °C for 4 min, 18 cycles of 97 °C for 35 s, 64 °C for 35 s, 68 °C for 4 min plus 20 s per cycle, and final extension at 72 °C for 10 min, 4 °C hold. Then, 2 µl of PCR product was mixed with 11 µl Hi-Di Formamide (Applied Biosystems) and 2 µl ROX1000TM Size Ladder (Asuragen). The samples prepared were then denatured at 95 °C for 2 min and cooled at 4 °C and held until ready for analysis by capillary electrophoresis (3730xL Genetic Analyzer, Applied Biosystems). The FAM-labelled amplicons were detected using the following fragment analysis protocol: 50 cm capillary, 2.5 kV, 20 s injection and 15 kV run for 4,800 s. Data analysis and interpretation was conducted using the fragment analysis software GeneMapper v.6 and an Excel-based analysis tool, AmplideX PCR/CE *HTT* Macro (Asuragen). The AmplideX PCR/CE *HTT* Kit Macro can determine size and mobility correction, and repeat size was determined using a linear fit adjustment of the ROX ladder size peaks to the PCR/CE control sample alleles. Alleles are reported as integer CAG repeats. The largest allele size determines genotype category: normal, intermediate, reduced penetrance or expanded. The alleles up to 200 CAG repeats are reported; the alleles with >200 repeats are identified as '>200'.

Validation of pathogenic expansions using Integrated Genome Viewer

Integrated Genome Viewer (v.2.19.4), a read stack viewer, was used to visually inspect the alignment and document evidence of expansion within the repeat region of *HTT*, *CACNA1A* and *DMPK* (Supplementary Information 6, Supplementary Table 13 and Supplementary Figs. 83–85).

Validation of interruptions using REViewer

REViewer (v.0.2.7), a tool for visualizing alignment of reads, was used to validate interruptions identified in *HTT* and *ATXN1* using Expansionhunter[39]. Interruptions were called from both WES and WGS samples from the UKB. Comparison of interruptions in 419,186 samples that passed QC in both WES and WGS UKB data showed interruptions called from WES data to be more reliable than WGS due to higher read depth in coding regions (Supplementary Information 8 and Supplementary Figs. 86–88). Specifically, validations showed that >93% of interruptions in *HTT* and *ATXN1* called by ExpansionHunter from WES data were accurate (Supplementary Information 7 and Supplementary Table 14).

Ethical compliance

Ethical approval for the UKB was previously obtained from the North West Centre for Research Ethics Committee (11/ NW/0382). The work described here was approved by UKB under application number 26041. Approval for GHS MyCode analyses was provided by the Geisinger Health System Institutional Review Board under project number 2006-0258. Informed consent was obtained for all of the study participants. Appropriate consent for the University of Pennsylvania Penn Medicine BioBank was obtained from each participant regarding storage of biological specimens, genetic sequencing and genotyping, and access to all available EHR data. This study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. All individuals participating in the Mayo-RGC Project Generation provided informed consent for use of specimens and data in genetic and health research and ethical approval for Project Generation was provided by the Mayo Clinic IRB (09-007763). All research performed in this study used de-identified data (without any Protected Health Information data) with no possibility of reidentifying any of the participants. For participants in the MCPS cohort, approval for the study was given by the Mexican Ministry of Health, the Mexican National Council of Science and Technology (0595 P-M) and the Central Oxford Research Ethics Committee (C99.260) and the Ethics and Research commissions from the Medicine Faculty at the National Autonomous University of Mexico (UNAM) (FMED/CI/SPLR/067/2015). All of the study participants provided written informed consent. The study participants were recruited from the BioMe Biobank Program of the Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center from 2007 onward. The BioMe Biobank Program (Institutional Review Board 07-0529) operates under a Mount Sinai Institutional Review Board-approved research protocol. All of the study participants provided written informed consent.

Reporting summary

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

Data availability

Individual-level sequencing data have been deposited in the UKB and will be freely available to approved researchers, as done for other genetic datasets to date. GangSTR calls will also be deposited in the UKB under the same guidelines. Individual-level phenotype data are available to approved researchers for the surveys and health-record datasets from which all our traits are derived. Instructions for access to UKB data is available online (https://www.ukbiobank.ac.uk/enable-your-research). Data from the MCPS are available to bona fide academic researchers. For more information, the study's data and sample sharing policy may be downloaded (in English or Spanish) from https://www.ctsu.ox.ac.uk/research/mcps. Available study data can be examined through the data showcase online (https://datashare.ndph.ox.ac.uk/mexico/). Further information can be found online (https://www.ctsu.ox.ac.uk/research/prospective-blood-based-study-of-150-000-individuals-in-mexico). Exome sequencing and genotyping data used for analysis from additional cohorts such as the GHS MyCode, the University of Pennsylvania Penn Medicine BioBank, the Mayo-Clinic Biobank, the Mount Sinai BioMe BioBank and CNCD can be made available to qualified, academic, non-commercial researchers on request through a data transfer agreement with the respective research institute.

Code availability

The REGENIE software for whole-genome regression that was used to perform all genetic association analysis is available at GitHub (https://github.com/rgcgithub/regenie). GangSTR v.2.5.0 used to call repeats from WES is available at GitHub (https://github.com/gymreklab/GangSTR). ExpansionHunter v.4, which was used to call repeats from WGS, is available at GitHub (https://github.com/Illumina/ExpansionHunter). Plink1.9/2.0, which was used for genotypic analysis, is available at GitHub (https://github.com/chrchang/plink-ng). R Statistical Computing 4.x and its libraries broom (v.1.0.7), data.table (v.1.17.0), dplyr (v.1.1.4), forestplot (v.3.1.3), reshape2 (v.1.4.4), genio (v.1.1.2), ggplot2 (v.3.5.1), wesanderson (v.0.3.7), sqldf (v.0.4-11), stringi (v.1.8.4), stringr (v.1.5.1) and tidyverse (v.2.0) were used for data wrangling, visualization and statistical analysis. Tools such FLIRT, FNIRT and FAST that are part of FSL and used for brain imaging analysis can be installed from https://web.mit.edu/fsl_v5.0.10/fsl/doc/wiki/FslInstallation.html. FastSurfer and CerebNet are available at GitHub (https://github.com/Deep-MI/FastSurfer/tree/dev). FreeSurfer is available at GitHub (https://github.com/freesurfer/freesurfer).

61. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
62. Staples, J. et al. PRIMUS: improving pedigree reconstruction using mitochondrial and Y haplotypes. *Bioinformatics* **32**, 596–598 (2016).
63. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
64. R Core Team. R: a language and environment for statistical computing v.4.3.2 (R Foundation for Statistical Computing, 2023).
65. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
66. Ochoa, A. genio: genetics input/output functions. R package version 1.1.6.9000; https://github.com/ochoalab/genio (2025).
67. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
68. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
69. Alfaro-Almagro, F. et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).
70. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264**, 47–56 (2016).
71. Smith, S. M. et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* **17**, 479–489 (2002).
72. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
73. Andersson, J. L. R., Jenkinson, M. & Smith, S. *Non-Linear Registration aka Spatial Normalisation. FMRIB Technical Report TR07JA2* (FMRIB Analysis Group Univ. Oxford, 2007).
74. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
75. Henschel, L. et al. FastSurfer—a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* **219**, 117012 (2020).
76. Faber, J. et al. CerebNet: a fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. *Neuroimage* **264**, 119703 (2022).

Acknowledgements We thank B. Vanacoro for her assistance with the graphics for main figures; and everyone who made this work possible, the professionals from the member institutions who contributed to and supported this work and especially all of the participants, without whom this research would not be possible. This study is funded by Regeneron Genetics Center and Regeneron Pharmaceuticals.

Author contributions Conceptualization: S.G., G.C. and V.K.P. Data generation and analytical pipeline development: S.O., X.B., V.K.P., J.H.S., J.H., K.L., X.Z., S.Y, L.Z., F.G., S.W., S.S., J.O. and J.M. Formal analysis: S.G., V.K.P., J.H.S., J.H. and K.L. Supervision: S.G., G.C., G.R.A., L.L. and A.B. Writing (original draft): S.G., V.K.P., J.H.S., J.H. and K.L. Writing (review and editing): S.G., V.K.P., J.H.S., J.H., V.R., M.D.K., N.P., K.L., F.S., E.A.S., Y.H., M.A., S.C., W.S., J.M., G.R.A., L.L. and A.B. Figures: S.G., V.K.P., J.H.S., J.H. and K.L. Funding acquisition: G.C. and A.B. Project management: S.G., G.C., M.G.L., and J.R. All of the authors contributed to study design and oversight and reviewed the final version of the manuscript.

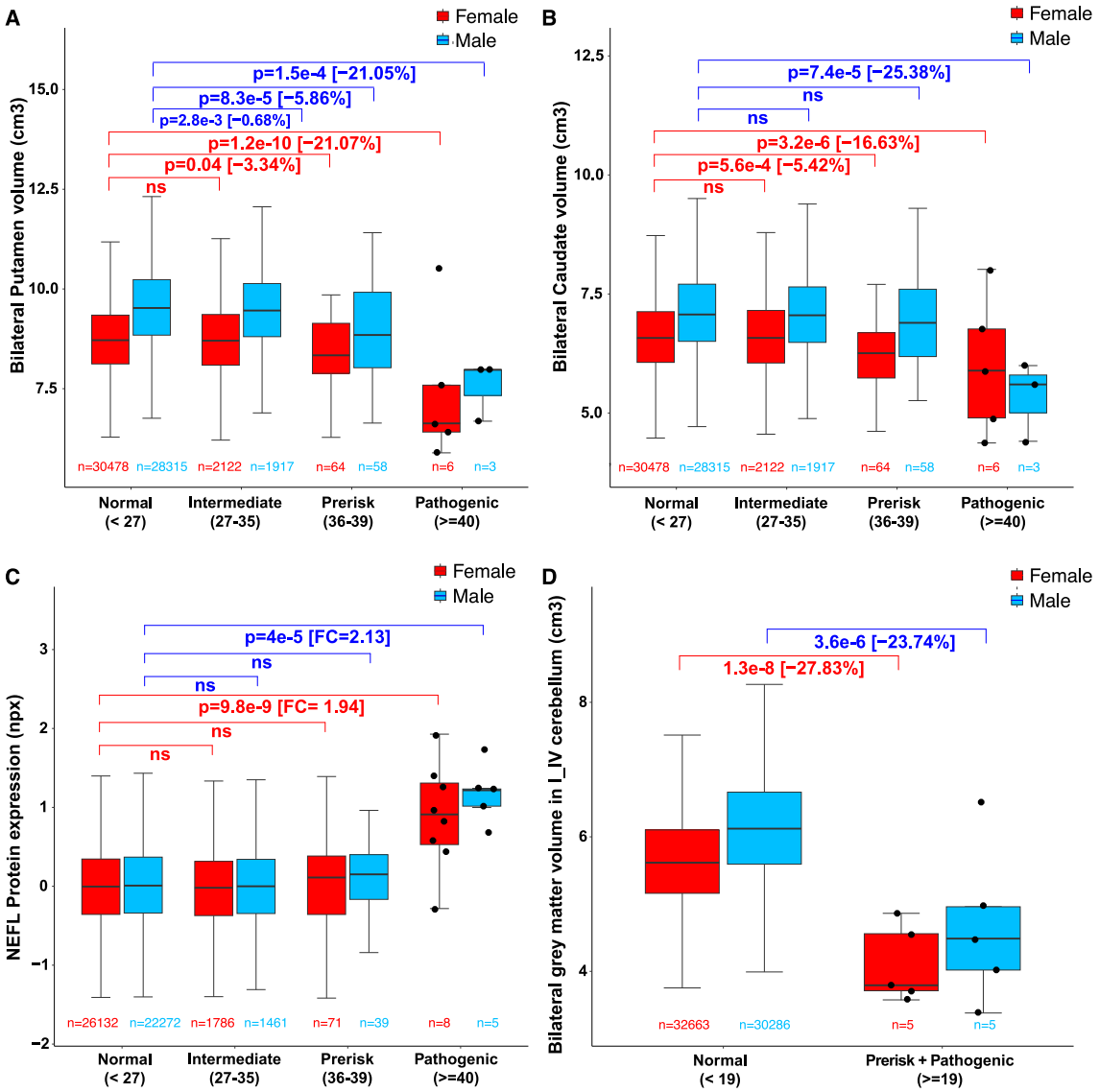
Competing interests V.K.P., J.H.S., J.H., S.O., V.R., X.B., M.D.K., K.L., X.Z., S.Y., L.Z., M.G.L., J.R.-V, F.G., S.W., S.S., F.S., E.A.S., Y.H., M.A., S.C., W.S., J.O., J.M., J.R., G.R.A., L.L., A.B., G.C. and S.G. are current employees and/or stockholders of Regeneron Genetics Center or Regeneron Pharmaceuticals. N.P. was a former employee of Regeneron Genetics Center. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10345-6.

Correspondence and requests for materials should be addressed to Sahar Gelfman.

Peer review information *Nature* thanks Matt Danzi, Robert Handsaker, Indhu-Shree Rajan-Babu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Boxplots comparing brain volumes and NfL labels in male (blue) and female (female) repeat carriers of HTT and CACNA1A. **A**) A box plot comparing brain volumes in putamen among male (blue) and female (red) *HTT* repeat carriers in the normal, intermediate, premutation and pathogenic ranges. Unadjusted p-values from linear regression analysis after correcting for known confounders are provided directly in the figure. P-values > 0.05 are not shown; *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001 **B**) A box plot comparing caudate volume between male (blue) and female (red) *HTT* repeat carriers in the normal, intermediate, premutation and pathogenic ranges. **C**) A box plot comparing plasma NfL levels among male and female *HTT* repeat carriers. **D**) A plot comparing grey matter volume in the cerebellum between male (red) and female (blue) *CACNA1A* repeat carriers in the normal, and premutation+pathogenic ranges. All box plots show, *centre:* median; *lower hinge:* 25% quantile; *upper hinge:* 75% quantile; *lower whisker:* smallest observation greater than or equal to lower hinge −1.5×interquartile range; *upper whisker:* largest observation less than or equal to upper hinge +1.5×interquartile range.

nature portfolio

Corresponding author(s): Sahar Gelfman

Last updated by author(s): Feb 18, 2026

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	No software was used
Data analysis	Association Analysis: 1) Regenie v3.2 (https://github.com/rgcgithub/regenie) 2) PLINK 1.9/2.0 (https://github.com/chrchang/plink-ng) Repeat callers: 1) GangSTR v2.5.0 (https://github.com/gymreklab/GangSTR) 2) ExpansionHunter v4 (https://github.com/Illumina/ExpansionHunter) 3) DRAGEN v3.7.8 Data wrangling and plotting: R4.x, forestplot_3.1.3, tidyverse_2.0.0, genio_1.1.2, R.utils_2.13.0, wesanderson_0.3.7, broom_1.0.7, ggplot2_3.5.1, reshape2_1.4.4, data.table_1.17.0, stringi_1.8.4, stringr_1.5.1, tidyr_1.3.1, dplyr_1.1.4, sqldf_0.4-11, fMRI/Image analysis: FSL tools (FLIRT, FNIRT & FAST) (https://web.mit.edu/fsl_v5.0.10/fsl/doc/wiki/FslInstallation.html) FastSurfer/CerebNet 2.3.0 (https://github.com/Deep-MI/FastSurfer/tree/dev) FreeSurfer 7.3.2 (https://github.com/freesurfer/freesurfer) GeneMapper v6 (https://www.thermofisher.com/order/catalog/product/A38888)

REViewer v0.2.7 (https://github.com/Illumina/REViewer)

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

Individual-level sequence data have been deposited with UK Biobank and will be freely available to approved researchers, as done with other genetic datasets to date. Individual-level phenotype data are available to approved researchers for the surveys and health-record datasets from which all our traits are derived. Quality controlled individual-level STR calls will be made available through the UKB Research Analysis Platform (RAP). Instructions for access to UK Biobank data is available at https://www.ukbiobank.ac.uk/enable-your-research. Data from the Mexico City Prospective Study are available to bona fide academic researchers. For more information, the study's Data and Sample Sharing policy may be downloaded (in English or Spanish) from https://www.ctsu.ox.ac.uk/research/mcps. Available study data can be examined through the Data Showcase at https://datashare.ndph.ox.ac.uk/mexico/. Further information can be found at https://www.ctsu.ox.ac.uk/research/prospective-blood-based-study-of-150-000-individuals-in-mexico. Exome sequencing and genotyping data used for analysis from additional cohorts such as the Geisinger Health System MyCode, the University of Pennsylvania Penn Medicine BioBank, the Mayo-Clinic Biobank, the Mount Sinai BioMe BioBank, and CNCD can be made available to qualified, academic, non-commercial researchers upon request via a Data Transfer Agreement with the respective research institute. List of disease-associated repeat expansions were curated from STRipy (https://stripy.org/database) and Genereviews (https://www.ncbi.nlm.nih.gov/books/NBK1116/). GRCh38 (hg38), the current standard human reference genome, is publicly available through the Genome Reference Consortium (GRC), NCBI (accession GCA_000001405.29), and UCSC Genome Browser.

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 was used as a covariate in association analyses. However, the findings are not meant to be sex-specific. In the analysis comparing the brain volumes of pathogenic repeat carriers vs. controls, sex-stratification was employed to make sure the results replicate for both sexes even after considering the inherent brain volume differences between sexes.
Reporting on race, ethnicity, or other socially relevant groupings	This study uses ancestry information derived from genetic data. This information was used to compare the frequencies of pathogenic repeat carriers among them to subsequently identify enrichments/depletions within subpopulations. Specifically, participants categorized into one of the five ancestries (European, African, Admixed Americans, South Asians, East Asians) and pair-wise comparison of frequencies were performed using Fisher's exact test.
Population characteristics	This study uses whole-genome and whole-exome sequencing data from seven cohorts, analyzed retrospectively without any selection criteria based on age, gender, or genotypic information. The demographics of each cohort is as follows, UK Biobank: Males = 46% & Females = 54%; 95% European; Mean age at recruitment = 56.5 GHS: Males = 39% & Females = 61%; 95% European; Mean age at recruitment = 57 MCPS: Males = 33% & Females = 67%; 99% Admixed American; Mean age at recruitment = 53 Mayo-Clinic: Males = 44% & Females = 56%; 96% European; Mean age at recruitment = 71 UPENN: Males = 50% & Females = 50%; 69% European; Mean age at recruitment = 61 SINAI:Males = 41% & Females = 59%; 44% African, 35% European, & 12% Admixed Americans; Mean age at recruitment = 55 CNCD: Males = 42% & Females = 58%; 99% South Asian; Mean age at recruitment = 57
Recruitment	No subjects were recruited specifically for this manuscript since this is a retrospective analysis.
Ethics oversight	Ethical approval for the UK Biobank was previously obtained from the North West Centre for Research Ethics Committee (11/NW/0382). The work described herein was approved by UK Biobank under application number 26041. Approval for Geisinger Health System MyCode analyses was provided by the Geisinger Health System Institutional Review Board under project number 2006-0258. Informed consent was obtained for all study participants. Appropriate consent for the University of Pennsylvania Penn Medicine BioBank was obtained from each participant regarding storage of biological specimens, genetic sequencing and genotyping, and access to all available EHR data. This study was approved by the Institutional Review Board of the University of Pennsylvania and complied with the principles set out in the Declaration of Helsinki. All subjects participating in the MAYO-RGC Project Generation provided informed consent for use of specimens and data in genetic and health research and ethical approval for Project Generation was provided by the Mayo Clinic IRB (#09-007763). All research performed in this study used de-identified data (without any Protected Health Information data) with no possibility of re-identifying any of the participants. For participants in the Mexico City Prospective Study, approval for the study was given by the Mexican Ministry of Health, the Mexican National Council of Science and Technology (0595 P-M) and the Central Oxford Research Ethics Committee (C99.260) and the Ethics and Research commissions from the Medicine Faculty at the National Autonomous University of Mexico (UNAM) (FMED/CI/SPLR/067/2015). All study participants provided written informed consent. Study participants were recruited from the BioMe Biobank Program of the Charles Bronfman Institute for Personalized Medicine at Mount Sinai Medical Center from 2007 onward. The BioMe Biobank Program (Institutional Review Board #07-0529) operates under a Mount Sinai Institutional Review Board-approved research protocol. All study participants provided written informed consent.

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	Sample size was not predetermined. Association analysis was run using all samples for which both disease diagnosis status as well as repeat length in at least one of the loci from WES/WGS data were available. All analyses were restricted to samples that passed quality control criteria. Supplementary table 10 shows the number of WES and WGS samples analyzed within each cohort. Sample sizes across our six cohorts together represent one of the largest set of exomes analyzed for repeat expansions.
Data exclusions	Repeat expansion calls that failed to meet the quality control criteria described in the 'Repeat expansion genotyping and quality control' section of 'Methods' were excluded from all analysis. Supplementary table 9 provides the summary of the number and proportion of samples removed after QC for each disease-associated gene.
Replication	This is the first paper to analyze disease-associated repeat expansions from a million human exomes, so there are no prior studies of similar scale to compare results with. We used a meta-analysis to aggregate results from phenome-wide association analysis run independently within three large cohorts. All the statistically significant associations we report replicate independently within two or more cohorts. Brain imaging and proteomics data were only available for the UK Biobank participants.
Randomization	Randomization was not required for the analyses completed in this study. To control for confounding, we performed association analyses with the following covariates included in the regression model: age, age-squared, sex, age×sex, and 10 ancestry-informed principal components.
Blinding	Blinding was not required for the analyses completed in this study. Participant recruitment and phenotype collection were obtained without prior knowledge of sample genotypes. Association analyses were performed with all available samples, without any filtering based on sample genotypes.

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

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

Magnetic resonance imaging

Experimental design

Design type	Structural MRI
Design specifications	N/A
Behavioral performance measures	N/A

Acquisition

Imaging type(s)	T1-weighted MPRAGE sequence. Other scans are available from UKBiobank (i.e., FLAIR, gradient echo, DTI, task and resting state fMRI, and ASL), but they were not used in this analysis.
Field strength	3 T (technically 2.895 T)
Sequence & imaging parameters	T1-weighted MPRAGE sequence, 1 mm3 isotropic resolution, FOV 256 mm in all three spatial dimensions, TR = 2 s, TE = 2.01 ms, TI = 880 ms, flip angle = 8 degrees. Full details available at: https://www.fmrib.ox.ac.uk/ukbiobank/protocol/
Area of acquisition	Whole brain was acquired, then specific regions were analyzed later (as explained below). In short, FreeSurfer, FastSurfer and CerbNet were used to obtain volumes of the regions of interest.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	A minimally modified version of the UK_biobank_pipeline was used to process the structural brain images, available at: https://git.fmrib.ox.ac.uk/falmagro/uk_biobank_pipeline_v_1.5. In short, the raw .zip files were accessed from the UK Biobank Research Access Platform and converted to NIfTI via dcm2niix. The image is then brain-extracted, rigidly transformed to MNI152 space via FLIRT (FSL version 6.0.7.8) followed by a deformable registration to MNI152 space via FNIRT. The rigid and deformable transformations are combined into a single transformation. The inverse of this transformation is then obtained, and is used transform the MNI152 atlas to the structural space, which is then used as a mask for a final brain extraction. Tissue-type segmentation is then performed via FSL's FAST, which also provides the intensity bias, which is used to obtain the final brain-extracted and intensity-bias-corrected T1-weighted image. This final image was then segmented via FastSurfer(version 2.3.0), and cortical thicknesses and other morphological traits were obtained via FreeSurfer (version 7.3.2). Cerebellum volumes were obtained from CerebNet.
Normalization	Normalization was not directly used in these analyses. Images were, at some point, transformed to MNI152 space in order to extract the final brain extraction (as explained above), but analysis are performed in native T1-space.
Normalization template	MNI152 space is used in preprocessing, and FreeSurfer uses several templates for parcellation (see https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation).
Noise and artifact removal	No noise or artifact removal was performed
Volume censoring	No volumes were censored

Statistical modeling & inference

Model type and settings	Association analysis was run using linear regression with repeat status as a categorical predictor, rank inverse normal transformed brain volume as the outcome, and the following covariates: the first 10 array principal components, first 20 exome rare variant principal components, UK Biobank sequencing batch, sex, age, age^2, age x sex, UK Biobank assessment centre, position of brain within the scanner, and head size from SIENAX
Effect(s) tested	Rank inverse normal transformed brain volume

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): FreeSurfer, FastSurfer and CerbNet were used to obtain volumes of the regions of interest.

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis