POPULATION FREQUENCY OF REPEAT EXPANSIONS INDICATES INCREASED DISEASE PREVALENCE ESTIMATES ACROSS DIFFERENT POPULATIONS

Kristina Ibañez PhD[1], Bharati Jadhav[2], Stefano Facchini[3,4], Paras Garg[2], Matteo Zanovello[3] , Alejandro Martin-Trujillo PhD[2], Scott J Gies[2], Valentina Galassi Deforie[3], Delia Gagliardi[1,3], Davina Hensman MD PhD[5,8], Loukas Moutsianas PhD[6] , Maryam Shoai[8], Genomics England Research Consortium[7], EUROSCA network[9], Mark J Caulfield MD PhD[1], Andrea Cortese PhD[3], Valentina Escott-Price[10,11], John Hardy[8], Henry Houlden MD PhD[8,12], Andrew J Sharp PhD[2], Arianna Tucci MD PhD[1,3*]

[1]William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK, [2]Department of Genetics and Genomic Sciences and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, [3]Department of Neuromuscular Diseases, Institute of Neurology, University College London, London, UK, [4]IRCCS Mondino Foundation, Pavia, Italy, [5]St George's, University of London, London, SW17 0RE, UK, [6]Genomics England Queen Mary University of London, UK, [7]Genomics England Research Consortium, [8]Department of Neurodegenerative Disorders, Institute of Neurology, UCL, London, UK, [9]EUROSCA network (http://www.eurosca.org), [10]Department of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, UK,

[11]Dementia Research Institute, Cardiff University, UK, [12]Neurogenetics Unit, National Hospital for Neurology and Neurosurgery, London, UK

[*] correspondence to:

Arianna Tucci: William Harvey Research Institute, Queen Mary University of London, EC1M 6BQ, London, UK, a.tucci@qmul.ac.uk

Abstract

Repeat expansion disorders (REDs) are a devastating group of predominantly neurological diseases. Together they are common, affecting 1 in 3,000 people worldwide with population-specific differences. However, prevalence estimates of REDs are hampered by heterogeneous clinical presentation, variable geographic distributions, and technological limitations leading to under-ascertainment. Here, leveraging whole genome sequencing data from 82,176 individuals from different populations we found an overall carrier frequency of REDs of 1 in 340 individuals. Modelling disease prevalence using genetic data, age at onset and survival, we show that REDs are up to 3-fold more prevalent than currently reported figures. While some REDs are population-specific, e.g. Huntington's disease type 2, most REDs are represented in all broad genetic ancestries, including Africans and Asians, challenging the notion that some REDs are found only in European populations. These results have worldwide implications for local and global health communities in the diagnosis and management of REDs both at local and global levels.

Funding: Medical Research Council, Department of Health and Social Care, National Health Service England, National Institute for Health Research

MAIN

Repeat expansion disorders (REDs) are a heterogeneous group of conditions which mainly affect the nervous system, and include myotonic dystrophy (DM), Huntington's disease (HD) , and the commonest inherited form of amyotrophic lateral sclerosis and frontotemporal dementia (*C9orf72*-ALS/FTD)[1]. REDs are caused by the same underlying mechanism: the expansion of simple repetitive DNA sequences within their respective genes. The mutational process is gradual: normal alleles are usually passed stably from parent to child with rare changes in repeat size; intermediate-size (called "premutation" alleles in some cases) are more likely to expand into the disease range, both in somatic and germline cells. This is relevant for genetic counselling because the offspring of an intermediate allele carrier can be affected by REDs.

Previous studies have estimated that REDs affect 1 in 3,000 people[2], with population differences at specific RED loci. Among the most common REDs are myotonic dystrophy type 1 (DM1) and HD. DM1 affects 1 in 8,000 people worldwide, ranging from 1 in 10,000 in Iceland to 1 in 100,000 in Japan[3]. Similarly, the frequency of HD is 13.7 in 100,000 in the general population, varying between 17.2 in 100,000 in the Caucasian population, and 0.1-2 in 100,000 in Asians and Africans[3]. A repeat expansion at *C9orf72* causes both frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS): in Europeans it is estimated that the prevalence of *C9orf72*-FTD is 0.04-134 in 100,000, and *C9orf72*-ALS is 0.5-1.2 in 100,000[4]. The spinocerebellar ataxias (SCAs) are a group of rare neurodegenerative disorders, with a worldwide prevalence of SCA of 2.7-47 cases per 100,000[5] and wide regional variations mainly due to founder effects, with SCA3 being the most common form worldwide, followed by SCA2 and SCA6, SCA1, and SCA[6].

Despite their broad distribution in human populations, the few global epidemiological studies that have been performed on these disorders have focused on European populations. In these studies, prevalence estimates are either population-based, in

which affected individuals are identified based on clinical presentation, or assessed based on the presence of a relative with a RED who are then genetically tested. Given that one of the most striking features of REDs is that they can present with markedly diverse phenotypes, REDs can remain unrecognised, leading to underestimation of the disease prevalence[7].

Large scale analyses of REDs have been limited by repeat expansion profiling techniques, which historically have relied on polymerase chain reaction-based (PCR) assays or Southern blots, which by nature are targeted assays by nature and can be difficult to scale. So far, the largest population study of the genetic frequency REDs involved the analysis of 14,196 individuals of European ancestry[8].

In the last few years, bioinformatic tools have been developed to profile DNA repeats from short-read whole genome sequencing data (WGS). We have recently shown that disease-causing repeat expansions can be detected from WGS with high sensitivity and specificity, making large-scale WGS datasets an invaluable resource to analyse the frequency and distribution of disease-causing repeat expansions[2].

We here analyse disease-causing short tandem repeat (STR) loci in 82,176 individuals from two large-scale medical genomics cohorts with high-coverage WGS and rich phenotypic data: the 100,000 Genomes Project (100K GP) and Trans-Omics for Precision Medicine (TOPMed). The 100K GP is a programme to deliver genome sequencing of people with rare diseases and cancer within the National Health Service (NHS) in the United Kingdom. TOPMed is a clinical and genomic programme focused on elucidating the genetic architecture and risk factors of heart, lung, blood, and sleep disorders from the National Institute of Health (NIH). First, we selected WGS data generated using PCR-free protocols and sequenced with paired-end 150 bp reads (**Table S1**). To avoid overestimating the carrier frequency of REDs, we excluded individuals with neurological diseases, as their recruitment was driven by the fact that they had a neurological disease potentially caused by an STR expansion. We then performed relatedness and principal component analyses to identify a set of genetically unrelated individuals and predict

broad genetic ancestries based on 1,000 Genomes Project super-populations[9]. The resulting dataset comprised a cross-sectional cohort of 82,176 genomes from unrelated individuals (median age 61, Q1-Q3: 49-70, **Table S2, Fig. S1**), genetically predicted to be of European (n=59,568), African (n=12,786), American (n=5,674), South Asian (n=2,882), and East Asian (n=1,266) descent (**Fig. S2**).

We analysed repeats in RED genes for which WGS has been shown to be able to accurately discriminate between normal and pathogenic alleles[2], representing a broad spectrum of the most common REDs (**Table S3**, **Table S4**). Our analysis workflow (**Fig. 1**) included profiling each STR locus, followed by quality control of all alleles (visual inspection of pileup plots as previously described[2]) predicted to be larger than the premutation threshold (**Table S5**). Clinical and demographic data available on all individuals carrying a pathogenic repeat are listed in **Table S6**. As our cohort comprises data from different genetically predicted populations, we also compared genotypes generated by WGS and PCR for 1,006 alleles and showed that the accuracy for repeat sizing by WGS was not affected by genetic ancestry (online materials**, Table S7**, **Fig. S3**).

In total, there were 242 (0.29%) individuals carrying a fully-expanded repeat, and 798 (0.97%) individuals carrying a repeat in the premutation range (**Table S5**), meaning that frequency of individuals carrying full-expansion and premutation alleles among this large cohort is 1 in 340 people and 1 in 103 people respectively.

The most common pathogenic expansions in this cohort were those in *C9orf72* (1 in 1,126) that cause ALS-FTD, followed by expansions in *DMPK* causing DM1 (1 in 1,786). Surprisingly, we found expansions in the spinocerebellar ataxia 2 gene *ATXN2* to be almost as common as those in *DMPK*, followed by expansions in *AR* that cause spinal and bulbar muscular atrophy and in *HTT* that cause Huntington disease (**Fig. 2A**). By contrast, expansions in *JPH3* that cause Huntington disease-like 2 and in *ATN1* that cause dentatorubral-pallidoluysian atrophy were very rare, with only a single individual at each locus identified with a repeat allele in the pathogenic range. No individuals were identified with pathogenic expansions in *ATXN3* (SCA3).

REDs have variable age at onset, disease duration, and penetrance. Therefore, the carrier frequency cannot be directly translated into disease frequency (i.e. prevalence). To estimate the prevalence of REDs using genetic data, we modelled the distribution by age of the most common REDs (*C9orf72*-ALS/FTD, DM1, HD, and SCA2) in the UK population using data from the Office of National Statistics, taking into account the different age of onset, penetrance, and impact on survival of each RED[10]. We found an up to three-fold increase in the predicted disease prevalence compared to currently reported figures based on clinical observation, depending on the RED (**Fig. 2B**). We estimated a prevalence for DM1 of 20 per 100,000, 1.6 times higher than the estimated prevalence from clinical data[7]. Similarly, prevalence estimates for SCA2 were 3.7 per 100,000, 3.7 times higher than known SCA2 prevalence (1 per 100,000)[11]. For HD, we estimated an overall prevalence of 6.7 per 100,000. While this figure seems lower than currently reported for HD, the majority of individuals with a pathogenic expansion in *HTT* in our cohort (12 out of 20) carry alleles with 40 repeats. Given the well-established relationship between *HTT* repeat length and age at onset, we modelled the HD prevalence taking into account repeat-length and found that 1.3 per 100,000 with 40 repeats are estimated to develop HD (online methods). This is 1.8 times higher than the reported number of affected patients among these carriers (0.72 per 100,000) (personal communication, DHM). Since *C9orf72* expansions cause both ALS and FTD, we modelled the prevalence of both diseases separately, providing an estimated disease prevalence of 1.84 per 100,000 for *C9orf72*-ALS - over two times higher than previous estimates[12,13] , and 8.15 per 100,000 for *C9orf72*-FTD, within the wide reported range of *C9orf72*-FTD (online methods**, Fig. 2B**).

The prevalence of individual REDs varies considerably based on geographic location (**Fig. 3A,B**). Hence, we then set out to analyse whether these differences are reflected in the broad genetic ancestries of our cohort (**Table S8**), given the broad representation of different populations in this cohort. For this analysis, we looked at the proportion of abnormal alleles in each population after local ancestry assignment at each RED locus. In agreement with current known epidemiological studies, we

observed that the most common abnormal alleles in Europeans are those in *DMPK*, *HTT,* and *C9orf72,* in *JPH3* in Africans, *TBP, ATN1,* and *CACNA1A* in East Asians, *ATXN1* and *AR* in South Asians. Some expansions like those in *ATXN2*, *ATXN1,* and *AR* are more widely represented across all populations. Surprisingly, we identified pathogenic expansions within *C9orf72* and *HTT* in Africans and South Asians, both of which were previously thought to be found mostly in European populations. Given that the initial ancestry assignments were based on genome-wide data, we performed local ancestry analysis to check for admixture in these individuals, and confirmed that the expanded repeat alleles were carried on haplotypes of African and South Asian ancestry (**Fig. 3C, Table S9**)

We then analysed the complete distribution of repeat sizes in each population in the 100K GP and TOPMed datasets. Here we included WGS data from the 1,000 Genomes Project (1K GP3)[9] to reproduce the findings as this cohort includes a broad representation of genetic ancestries. **Fig. S4 and S5** show the distribution of repeat sizes across 13 genes among the three genomic datasets. While the overall distributions of repeat sizes are similar when comparing across ancestries for most loci, by contrast, for others such as *AR, ATN1, HTT*, and *TBP*, repeat lengths in genomes of African origin are significantly shorter ($p<10^{-16}$) than other ancestries (**Table S10, Fig. S5**). This pattern is consistent across all three datasets.

In summary, this study provides the first unbiased, worldwide population-based estimates of carrier frequency and expected disease prevalence of REDs. It shows that: i) the carrier frequency of REDs is approximately ten times higher than the previous estimates based on clinical observations, and that, based on population modelling, we estimate that REDs are predicted to affect up to three times more individuals than are currently recognized clinically; ii) while some REDs are population-specific like *JPH3*, the majority are observed in all ancestral populations, challenging the notion that some REDs (e.g. *C9orf72*) are associated with population specific founder effects; iii) an appreciable proportion of the population (1 in 103) carry alleles in the premutation range, and are therefore at risk of having children with REDs.

Different factors are likely to contribute to the increased prevalence estimate in the current study compared to others. First, our study estimates disease prevalence based on carrier frequency in large admixed cohorts, as opposed to studies based on the identification of clinically affected individuals in smaller populations. As REDs have variable clinical presentation and age at onset, it is likely that many individuals with REDs in a given population remain undiagnosed and undetected in clinically based prevalence studies. Second, even in symptomatic patients, there is a significant delay in diagnosis in many individuals with REDs[2]. The clinical data available on people carrying a pathogenic repeat expansion in this study is not suggestive of the corresponding RED. This might be explained by reduced or age-related penetrance of the mutation, which may go on to present at a later age. This is confirmed by the fact that we observed a large number of individuals carrying repeats in the lower end of the pathogenic range (e.g. *HTT* and *ATXN2*, and *DMPK* **Table S5)**, indicating that they will likely develop milder disease later in life. For example, it is well documented that carriers of small *DMPK* expansions (50–100 repeats) have milder disease with clinical features that may go unnoticed, especially early in their disease course[14]. Finally, these individuals may carry genetic modifiers of REDs.

One limitation of this study is that WGS cannot accurately size repeats larger than the sequencing read length (150 bp). For this reason, we did not assess some REDs like Fragile X syndrome, as WGS cannot distinguish between premutation and pathogenic full mutation alleles[2]. Another limitation of this study is the potential for recruitment bias, especially within the TOPMed cohort: individuals with an overt REDs have a reduced likelihood of being recruited to such studies because of the severity of their disease. For example, we note the absence of expansions in *ATXN3*, the most commonly reported SCA in patients affected by spinocerebellar ataxia.

Both 100K GP and TOPMed datasets are Euro-centric, comprising over 62% of European samples. TOPMed is more diverse, with 24% and 17% of African and

American genomes respectively, which are only present at 3.2% and 2.1% frequency in the 100K GP. East and South Asian backgrounds are underrepresented in both datasets, limiting the ability to detect rarer repeat expansions.

Further analyses on more heterogeneous and diverse large scale WGS datasets are necessary not only to confirm our findings, but also to shed light into additional ancestries. With regards to this, there are multiple ongoing projects with Asian populations[15,16]. Countries including China, Japan, Qatar, Saudi Arabia, India, Nigeria, and Turkey have all launched their own genomics projects during the last decade[17]. Analysing genomes from these coming genomic programmes will yield more detail on the prevalence of REDs around the world.

Despite efforts to estimate the frequency globally and locally, there is uncertainty surrounding the true prevalence of REDs, limiting the knowledge of the burden of disease required to secure dedicated resources to support health services, such as the estimation of the numbers of individuals profiting from drug development and novel therapies, or participating in clinical trials.

The finding that REDs are more prevalent than was previously thought has important implications. Clinicians should have a higher index of suspicion when a patient presents with symptoms compatible with a RED, and clinical diagnostic pathways should facilitate genetic testing for REDs. The presence of expansions within *HTT* and *C9orf72* in African and Asian populations supports diagnostic testing for them in people presenting with features of Huntington's disease and ALS-FTD whatever their ethnicity. There are currently no disease modifying treatments for REDs, however both disease specific treatments, and drugs which target repeat expansions more generally are in development. We have established that the numbers of people who may benefit from such treatments are greater than previously thought.

REFERENCES

1. Paulson, H. Repeat expansion diseases. *Handb. Clin. Neurol.* **147**, 105–123 (2018).
2. Ibañez, K. *et al.* Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* **21**, 234–245 (2022).
3. Bird, T. D. Myotonic dystrophy type 1. *GeneReviews®[Internet]* (2019).
4. Gossye, H., Engelborghs, S., Van Broeckhoven, C. & van der Zee, J. *C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis*. (University of Washington, Seattle, 2020).
5. Teive, H. A. G., Meira, A. T., Camargo, C. H. F. & Munhoz, R. P. The Geographic Diversity of Spinocerebellar Ataxias (SCAs) in the Americas: A Systematic Review. *Mov Disord Clin Pract* **6**, 531–540 (2019).
6. Schöls, L., Bauer, P., Schmidt, T., Schulte, T. & Riess, O. Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. *Lancet Neurol.* **3**, 291–304 (2004).
7. 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).
8. Gardiner, S. L. *et al.* Prevalence of Carriers of Intermediate and Pathological Polyglutamine Disease-Associated Alleles Among Large Population-Based Cohorts. *JAMA Neurol.* **76**, 650–656 (2019).
9. 1000 Genomes Project Consortium, {fname} *et al.* A global reference for human

genetic variation. *Nature* **526**, 68–74 (2015).

10. Zanovello, M. *et al.* Unexpected frequency of the pathogenic AR CAG repeat expansion in the general population. *Brain* (2023) doi:10.1093/brain/awad050.
11. Bhandari, J., Thada, P. K. & Samanta, D. *Spinocerebellar Ataxia*. (StatPearls Publishing, 2022).
12. Van Mossevelde, S., Engelborghs, S., van der Zee, J. & Van Broeckhoven, C. Genotype-phenotype links in frontotemporal lobar degeneration. *Nat. Rev. Neurol.* **14**, 363–378 (2018).
13. Hogan, D. B. *et al.* The Prevalence and Incidence of Frontotemporal Dementia: a Systematic Review. *Can. J. Neurol. Sci.* **43 Suppl 1**, S96–S109 (2016).
14. Thornton, C. A. Myotonic dystrophy. *Neurol. Clin.* **32**, 705–19, viii (2014).
15. Wu, D. *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736–749.e15 (2019).
16. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
17. Kumar, R. & Dhanda, S. K. Current Status on Population Genome Catalogues in different Countries. *Bioinformation* **16**, 297–300 (2020).

FIGURES

Figure 1. Technical flowchart

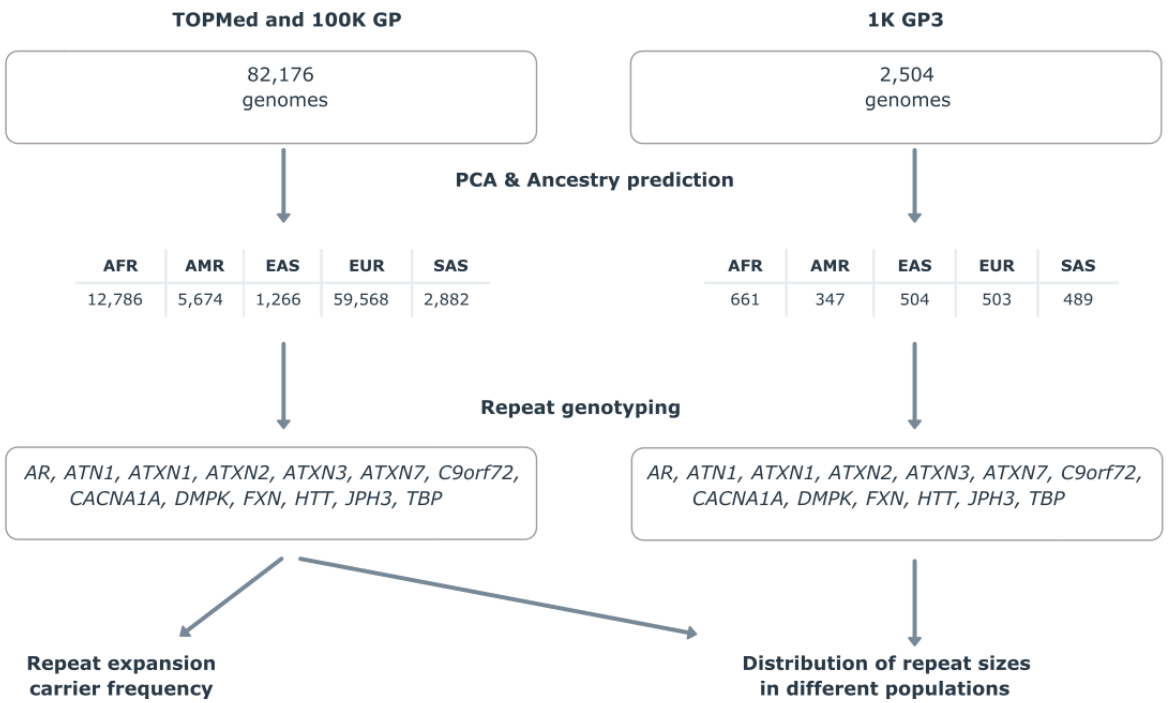


Figure 2. A) Forest plot with combined overall carrier frequency and 95% CI values in the combined 100K GP and TOPMed datasets together. Grey and black boxes show premutation and full-mutation overall carrier frequencies for each locus, respectively. B) Modelling of disease prevalence by age of DM1, SCA2, HD, *C9orf72*-ALS, and *C9orf72*-FTD. Age bins are 5-years each. Estimated prevalence (dark blue area) is compared to the reported prevalence from the literature (light blue area). For *C9orf72*-FTD, given the wide range of the reported disease prevalence[12,13],both lower and upper limits are plotted in light blue.

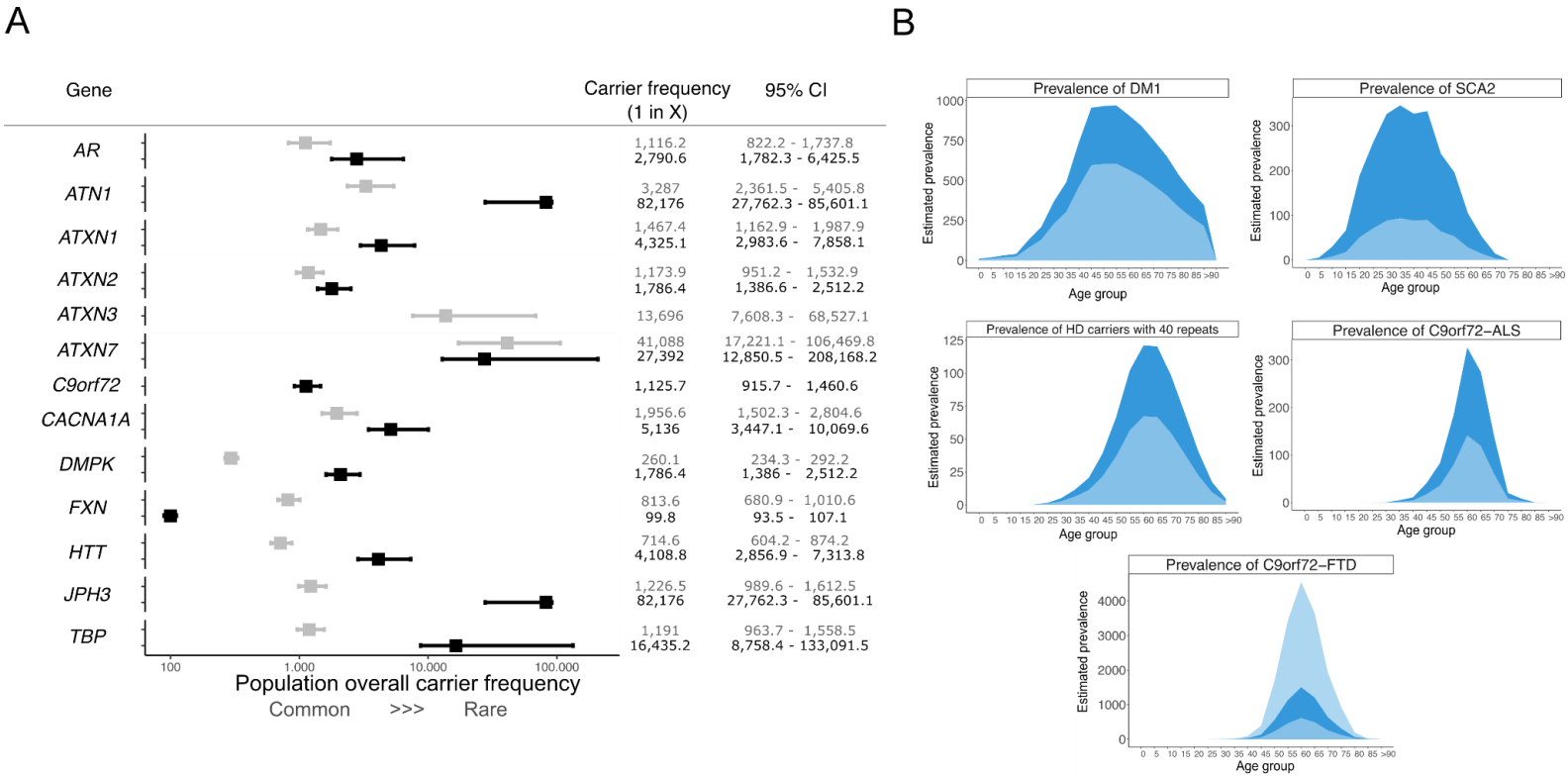
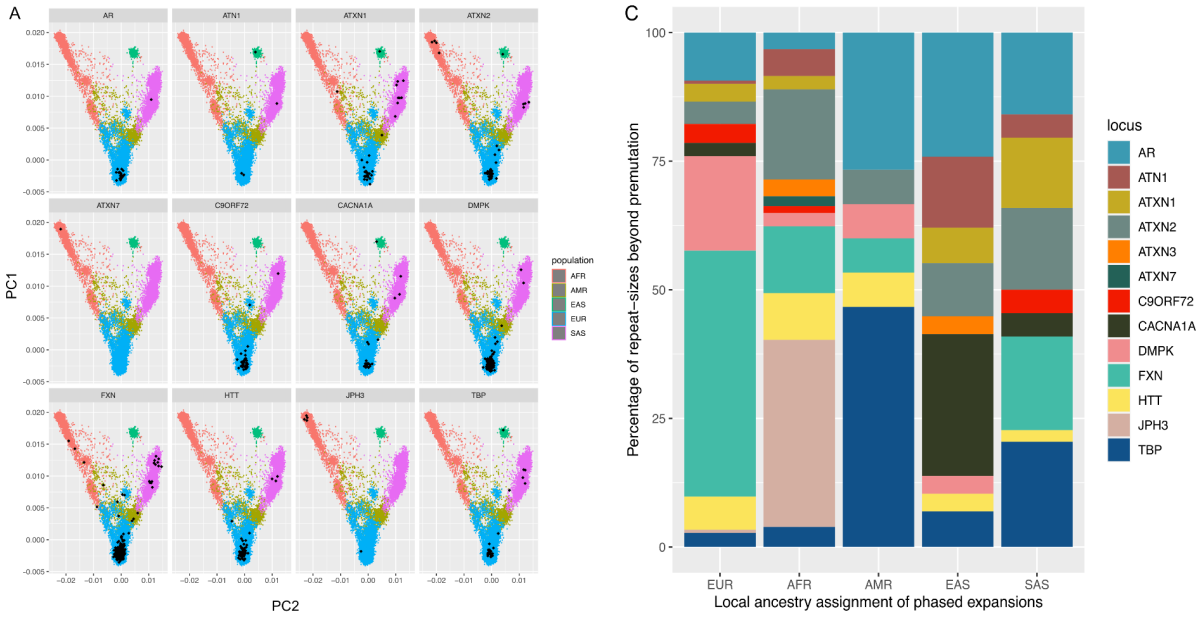
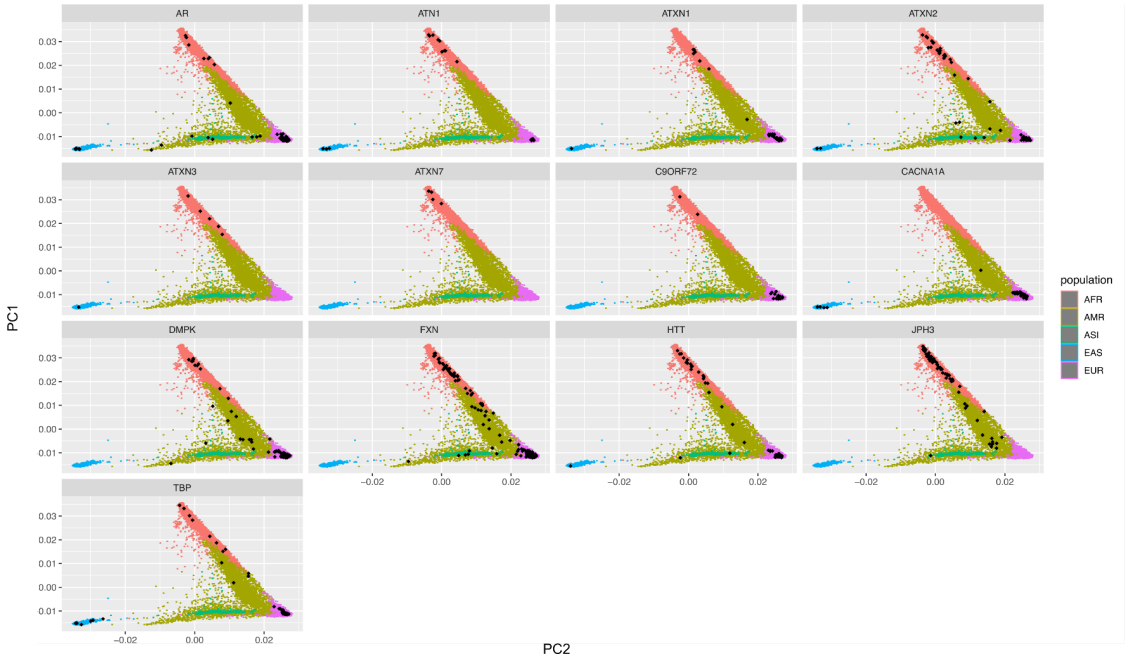


Figure 3. Principal component (PC) values on all genomes within (A) the 100K GP and B) TOPmed cohorts. Black dots represent genomes having a repeat-size beyond premutation and full-mutation range, split by gene. C) Local ancestry bar plot showing the repeats beyond the premutation threshold by super population within 100K GP and TOPMed cohorts.





SUPPLEMENTARY FIGURES

Figure S1: Pyramid demographics in (A) the 100K GP and (B) TOPMed cohorts.

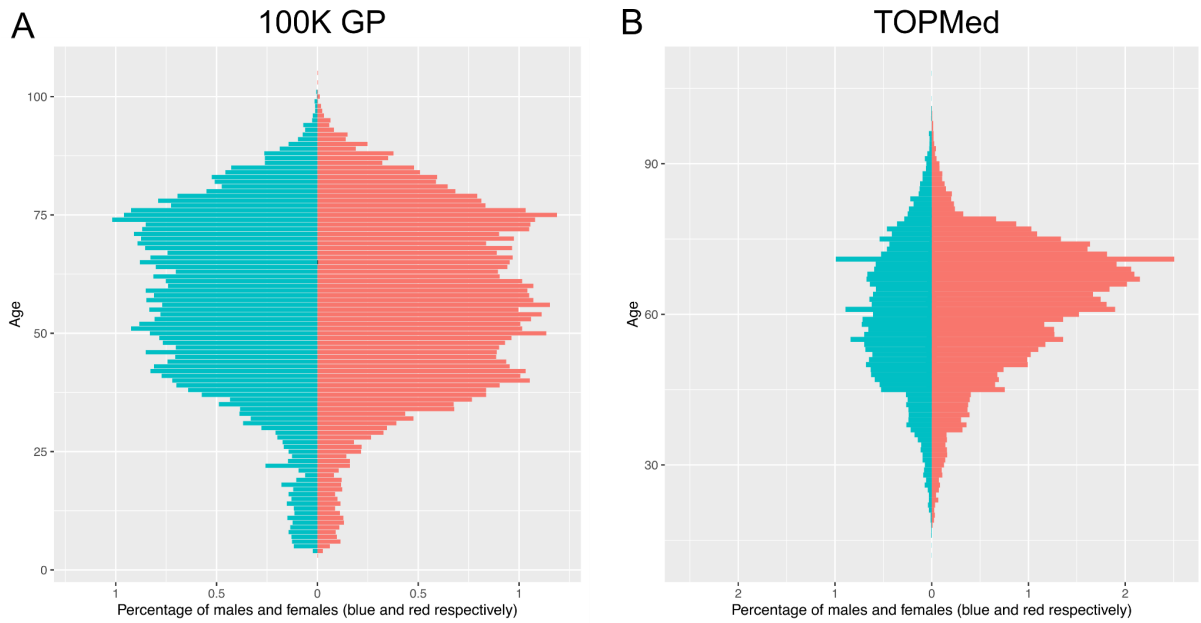


Figure S2. First 2 principal components (PCs) derived from PCA on A) the 100K GP and B) TOPMed samples (see **online materials** for more information).

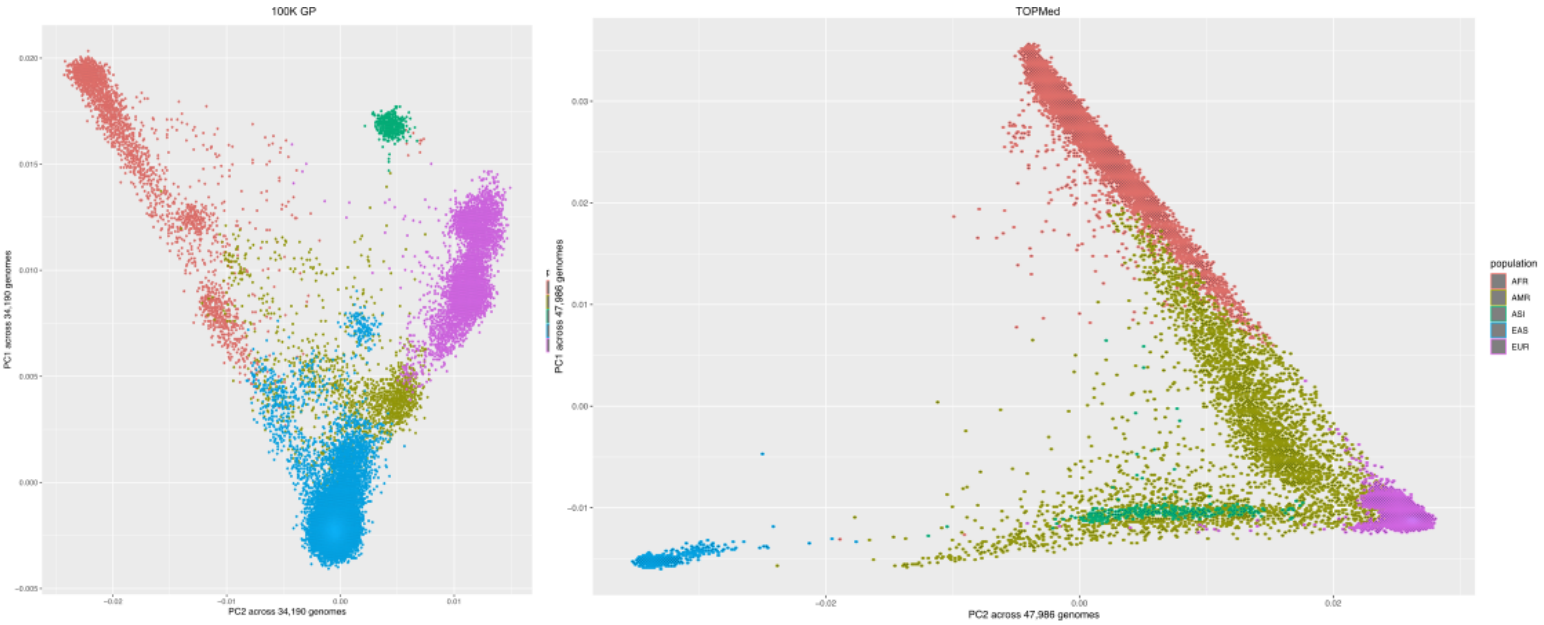


Figure S3. Experimental estimations of RE sizes using PCR vs genotypes generated by EH v3.2.2, split by super-population (samples of EAS ancestry were not tested). Points indicate the RE size estimated by both PCR and EH v3.2.2. We show the R correlation coefficient calculated using Pearson's equation.

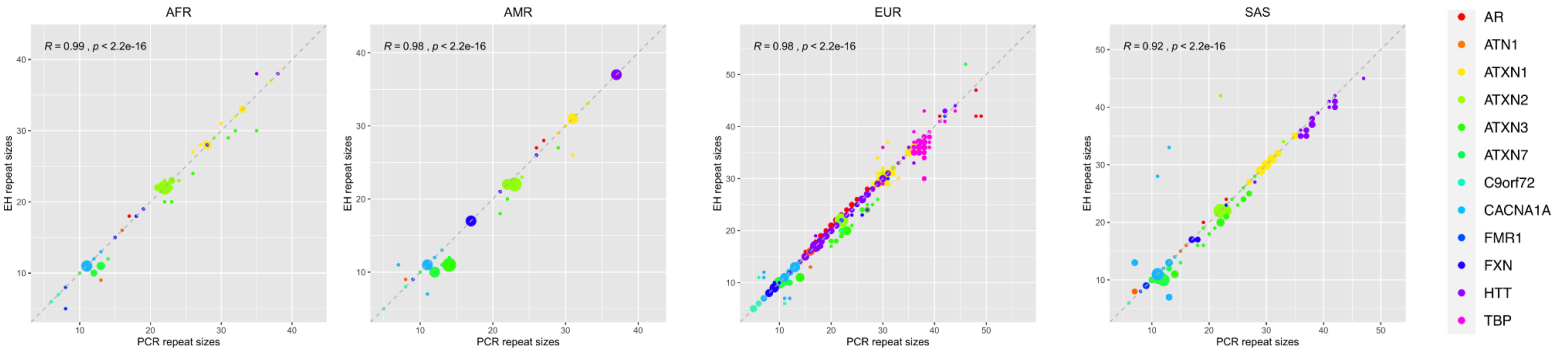


Figure S4. Distribution of disease RE sizes for genes merged within the 100K GP and TOPMed (before quality control). Bar plots showing the allele frequency percentage predicted by ExpansionHunter (before quality control) in both the 100K GP and TOPMed cohorts. The regions are shaded to indicate non-expanded (blue), premutation (yellow), and full-mutation expanded (red) ranges for each gene, as indicated in **Table S3**.

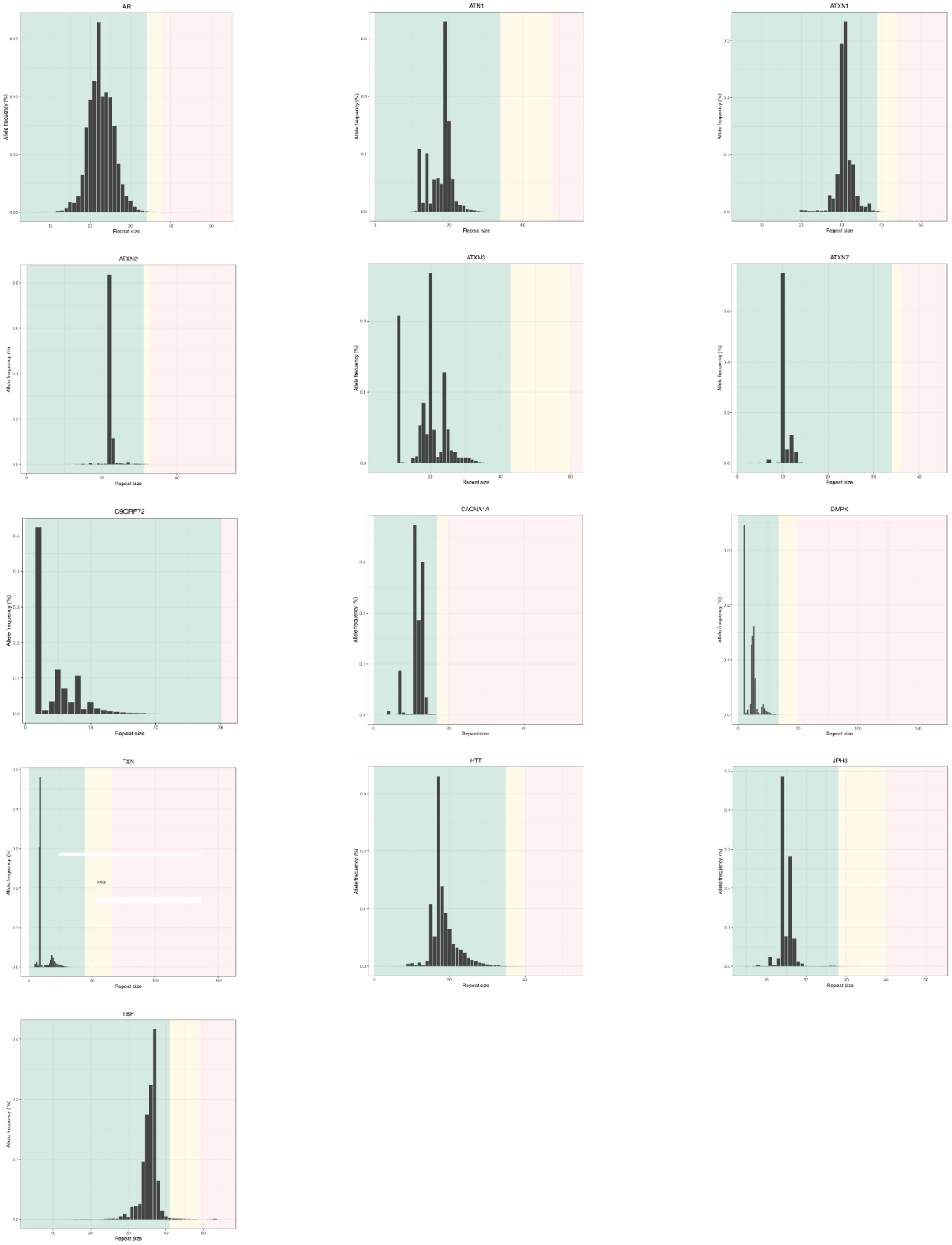
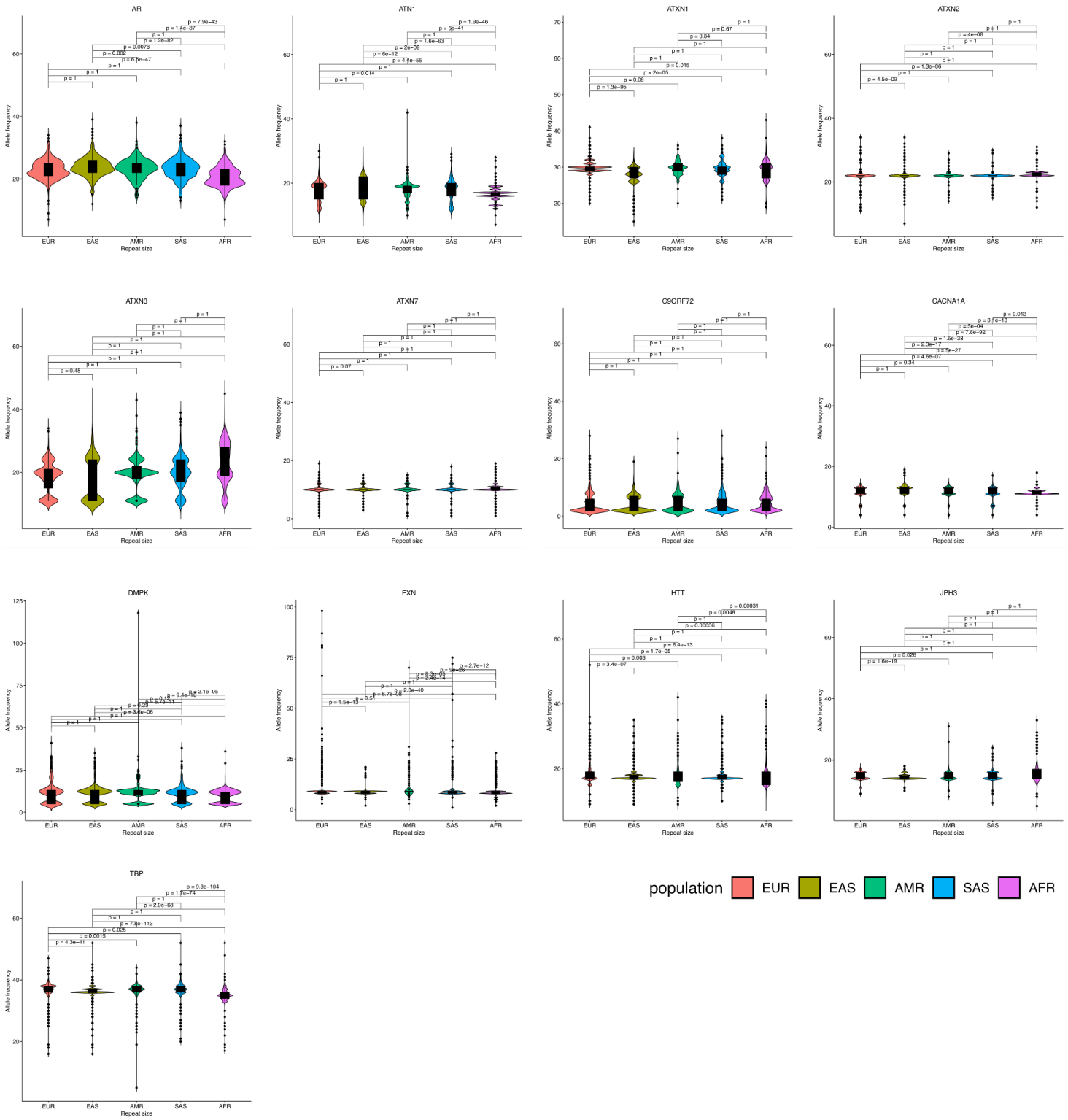


Figure S5. Distribution of disease RE sizes for genes within the 1K GP3 split by population. Violin plots with boxplots represent the repeat size distribution of each locus across all ancestries. Repeat size median (Q1-Q3) among all ancestries across the 13 repeat loci are in **Table S10**.



SUPPLEMENTARY APPENDIX

INVESTIGATORS

The members of The Genomics England Research Consortium are:

J. C. Ambrose[1], P. Arumugam[1], E. L. Baple[1], M. Bleda[1], F. Boardman-Pretty[1,2], J. M. Boissiere[1], C. R. Boustred[1], H. Brittain[1], M. J. Caulfield[1,2], G. C. Chan[1], C. E. H. Craig[1], L. C. Daugherty[1], A. de Burca[1], A. Devereau[1], G. Elgar[1,2], R. E. Foulger[1], T. Fowler[1], P. Furió-Tarí[1], J. M. Hackett[1], D. Halai[1], A. Hamblin[1], S. Henderson[1,2], J. E. Holman[1], T. J. P. Hubbard[1], K. Ibáñez[1,2], R. Jackson[1], L. J. Jones[1,2], D. Kasperaviciute[1,2], M. Kayikci[1], L. Lahnstein[1], K. Lawson[1], S. E. A. Leigh[1], I. U. S. Leong[1], F. J. Lopez[1], F. Maleady-Crowe[1], J. Mason[1], E. M. McDonagh[1,2], L. Moutsianas[1,2], M. Mueller[1,2], N. Murugaesu[1], A. C. Need[1,2], C. A. Odhams[1], C. Patch[1,2], D. Perez-Gil[1], D. Polychronopoulos[1], J. Pullinger[1], T. Rahim[1], A. Rendon[1], P. Riesgo-Ferreiro[1], T. Rogers[1], M. Ryten[1], K. Savage[1], K. Sawant[1], R. H. Scott[1], A. Siddiq[1], A. Sieghart[1], D. Smedley[1,2], K. R. Smith[1,2], A. Sosinsky[1,2], W. Spooner[1], H. E. Stevens[1], A. Stuckey[1], R. Sultana[1], E. R. A. Thomas[1,2], S. R. Thompson[1], C. Tregidgo[1], A. Tucci[1,2], E. Walsh[1], S. A. Watters[1], M. J. Welland[1], E. Williams[1], K. Witkowska[1,2], S. M. Wood[1,2], M. Zarowiecki[1]

1. Genomics England, London, UK
2. William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK

ONLINE METHODS

Whole genome sequencing datasets

Both 100,000 Genomes Project (100K GP) and Trans-Omics for Precision Medicine (TOPMed) include whole genome sequencing (WGS) data optimal to genotype short DNA repeats: WGS libraries generated using PCR-free protocols, sequenced at 150 base-pair read-length, and with a 35x mean average coverage (**Table S1**).
For the both 100K GP and TOPMed cohorts, the following genomes were selected: i) WGS from genetically unrelated individuals (see Ancestry and relatedness inference below); ii) WGS from people not presenting with a neurological disorder - these people were excluded to avoid overestimating the frequency of a repeat expansion due to individuals recruited due to symptoms related to a RED.

The TOPmed project has generated omics data, including WGS, on over 180,000 individuals with heart, lung, blood and sleep disorders (see NHLBI Trans-Omics for Precision Medicine WGS-About TOPMed (nih.gov)). TOPMed has incorporated samples gathered from dozens of different cohorts, each of which were collected using different ascertainment criteria. The specific TOPMed cohorts included in this study are described in **Table S11**.

To analyse the distribution of repeat lengths of RED genes in different populations, we used the 1000 Genomes Project phase 3 (1K GP3) as the WGS data are more equally distributed across the continental groups (**Table S2**). Genome sequences with read lengths of ~150bp were considered, with an average minimum depth of 30x (**Table S1**).

Correlation between PCR and ExpansionHunter

Results were obtained on samples tested as part of routine clinical assessment. Repeat expansions were assessed by polymerase chain reaction (PCR) amplification and fragment analysis Southern blotting was performed for large *C9orf72* expansions as previously described[1].

A dataset was set up from the 100K GP samples comprising a total of 198 genomes with PCR-quantified lengths across 11 loci (*AR, ATN1, ATXN1, ATXN2, ATXN3, ATXN7, CACNA1A, C9orf72, FXN, HTT, TBP)*. Repeats smaller than the read-length (i.e. 150bp) were only considered since ExpansionHunter estimates these accurately. Out of 512 cases 720 had repeats smaller than any cut-off (i.e. negatives), and 23 and 14 were expansions beyond pathogenic and premutation thresholds, respectively. **Fig. S3** shows the distribution of repeat sizes quantified by PCR compared to those estimated by EH after visual inspection, split by super-population.

Ancestry and relatedness inference

For relatedness inference WGS VCFs were aggregated with Illumina's agg or gvcfgenotyper (https://github.com/Illumina/gvcfgenotyper). All genomes passed the following quality control (QC) criteria: cross-contamination <5% (VerifyBamId)[2], mapping-rate >75%, mean-sample coverage >20, and insert size > 250. No variant QC filters were applied in the aggregated dataset, but VCF filter was set to `PASS` for variants which passed GQ (genotype quality), DP (depth), missingness, allelic imbalance, and Mendelian error filters. From here, by using a set of ~65,000 high quality SNPs, a pairwise kinship matrix was generated using the PLINK2 implementation of the KING-Robust algorithm (www.cog-genomics.org/plink/2.0/)[3]. For relatedness, PLINK2 `--king-cutoff` (www.cog-genomics.org/plink/2.0/) relationship-pruning algorithm [3] was used with a threshold of 0.044. These were then partitioned into 'related' (up to, and including 3rd degree relationships) and 'unrelated' sample lists. Only unrelated samples were selected for this study.

1K GP3 data was used when inferring ancestry, by taking the unrelated samples and by calculating the first 20 PCs using GCTA2. We then projected the aggregated data (100K GP and TOPMed separately) onto 1K GP3 PC loadings, and a random forest model was trained to predict ancestries based on 1) First 8 1K GP3 PCs, 2) setting `Ntress` to 400, and 3) train and predict on 1kPG3 five broad super-populations: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS).

In total, the following WGS data were analysed: 34,190 individuals in the 100K GP; 47,986 in TOPMed; 2,504 in the 1K GP3. The demographics describing each cohort can be found in **Table S2**.

Repeat expansion genotyping and visualisation

ExpansionHunter v3.2.2 (EH) software package was used for genotyping repeats in disease-associated loci[4,5]. EH assembles sequencing reads across a predefined set of DNA repeat using both mapped and unmapped reads (with the repetitive sequence of interest) to estimate the size of both alleles from an individual.

REViewer software package was used to enable direct visualisation of haplotypes and corresponding read pileup of the EH genotypes[6]. **Table S4** includes the genomic coordinates for the loci analysed. **Table S5** lists repeats before and after visual inspection. Pileup plots are available upon request.

Computation of genetic prevalence

For each gene, the frequency of each repeat size across the 100K GP and TOPMed genomic datasets was determined. Genetic prevalence was calculated as the number of genomes with repeats exceeding the full-mutation and permutation cutoffs (**Table S3**) compared to the overall cohort (**Table S8**).

Overall unrelated and non-neurological disease genomes corresponding to both programmes were considered, breaking down by ancestry.

CARRIER FREQUENCY ESTIMATE (1 in xx)

- freq_carrier = round(total_unrel / total_exp_after_VI_locus, digits = 2)
- ci_max=round(total_unrel/(total_unrel*((total_exp_after_VI_locus/total_unrel)-1.96*sqrt((total_exp_after_VI_locus/total_unrel)*(1-total_exp_after_VI_locus/total_unrel)/total_unrel))), digits = 2)
- ci_min=round(total_unrel/(total_unrel*((total_exp_after_VI_locus/total_unrel)+1.96*sqrt((total_exp_after_VI_locus/total_unrel)*(1-total_exp_after_VI_locus/total_unrel)/total_unrel))), digits = 2)

PREVALENCE ESTIMATE (x in 100,000)

x = 100,000/freq_carrier

new_freq = 100000*(1/frequency_cohort2_df$carrier_freq)

new_low_ci = 100000*low-ci

new_high_ci = 100000*high_ci

Modelling disease prevalence using carrier frequency

To estimate the prevalence of REDs based on the carrier frequency, we modelled the distribution by age of the most common REDs (C9orf72-ALS/FTD, DM1, HD, and SCA2) in UK population in mid-2020 taken from the Office of National Statistics[7], considering:

(i) Combined carrier frequency from 100K GP and TOPMed datasets. We used the carrier frequency of 100K GP to model C9orf72-ALS/FTD and HD, being more representative of the UK population which we are using the data to compare with.

(ii) Age at onset distribution of the specific disease, available from cohort studies or international registries. For disease modelling of

C9orf72-disease, we tabulated the distribution of disease onset of 811 patients with C9orf72-ALS pure and overlap FTD, and 323 patients with C9orf72-FTD pure and overlap ALS[8]. HD onset was modelled on 246 patients from the UK's General Practice Research Database[9] and DM1 was modelled on a cohort of 395 patients[10]. Data of 157 patients with SCA2 and *ATXN* allele size equal or higher than 35 repeats from EUROSCA were used to model prevalence of SCA2[11].

(iii) Mortality from disease. Median survival length is approximately three years for C9orf72-ALS and ten years for C9orf72-FTD[12]. HD and SCA2 have a median survival of fifteen years[13,14]. Given that approximately 22% of patients with DM1 die over a period of 11 years, we estimated a survival of 80% after 10 years[15].

(iv) Other factors that affect age at disease onset: As regards *ATXN2*, it is known that 33 and 34 CAG repeats are considered reduced-penetrance alleles[16]. Hence, for disease modelling, we used a carrier frequency of 1 in 5170, considering only carriers with allele size equal or higher than 35 repeats. When modelling HD prevalence in 40-CAG repeat carriers, the estimate was corrected by the chance to be symptomatic (stage 2 or 3 according to Huntington's Disease-Integrated Staging System[17] for a 40-CAG repeat carrier.

(v) Reduced penetrance, e.g., C9orf72-carriers may not develop symptoms even after 90 years of age[8]. Thus, age-related penetrance of C9orf72-ALS/FTD was derived from the red curve in

Fig. 2B reported by Murphy et al[8], and was used to correct C9orf72-ALS and C9orf72-FTD prevalence by age.

Both general UK population and age at onset distribution of each disease were divided into age groups. To account for mortality, age group length for a given disease was equal to the median survival length for that disease. For DM1 we subtracted the 20% of the predicted affected individuals every 10 years and we computed a cumulative distribution of age at onset.

For each disease, we multiplied the distribution of the disease onset by the corresponding general population count for each age group and by carrier frequency., and by penetrance (*C9orf72*). The resulting estimated prevalence of *C9orf72*-ALS/FTD, HD, SCA2 and DM1 by age group were plotted in **Fig. 2B** (dark blue)**.** The literature reported prevalence by age for each disease was represented as a dashed line for comparison and was obtained by dividing the new estimated prevalence by age by the ratio between the two prevalences.

To compare the new estimated prevalence to the known reported disease prevalence figures for each disease:

i) *C9orf72*-FTD: the median prevalence of FTD was obtained from studies included in the systematic review by Hogan and colleagues[18] (83.5 in 100,000). Since 4-29% of FTD patients carry a *C9orf72* repeat expansion[19], we calculated *C9orf72*-FTD prevalence by multiplying this proportion range by median FTD prevalence (3.3 - 24.2 in 100,000, mean 13.78 in 100,000).

ii) C9orf72-ALS: The reported prevalence of ALS is 5-12 in 100,000 [20] and *C9orf72* repeat expansion is found in 30%-50% of individuals with familial forms and in 4%-10% of people with sporadic disease [21]. Given that ALS is familial in 10% of cases and sporadic in 90%, we estimated the prevalence of C9orf72-ALS by

calculating the [(0.4 of 0.1) + (0.07 of 0.9)] of known ALS prevalence of 0.5-1.2 in 100,000 (mean prevalence is 0.8 in 100,000);

iii) HD prevalence ranges from 0.4 in 100,000 in Asian countries[22] to 10 in 100,000 in Europeans[23], and mean prevalence is 5.2 in 100,000. 40-CAG repeat carriers represent the 7.4% of patients clinically affected by HD according to the Enroll-HD[24] version 6. Considering an average reported prevalence of 9.7 in 100,000 in Europeans, we calculated a prevalence of 0.72 in 100,000 for symptomatic 40-CAG carriers;

iv) Prevalence of SCA2 is unknown, but it represents the second most common form of SCA. Since global prevalence of SCA is 5 in 100,000 and SCA2 represents up to 18% of forms, we estimated SCA2 prevalence to be approximately 1 in 100,000[25].

Local ancestry prediction

100K GP

For each RE locus and for each sample with a pre- or a full mutation, we obtained a prediction for the local ancestry in a region of +/- 5Mb around the repeat, as follows:

1. We extracted VCF files with SNPs from the selected regions and phased them with SHAPEIT v4. As a reference haplotype set, we used non-admixed individuals from the 1kG project. Additional non-default parameters for SHAPEIT: --mcmc-iterations 10b,1p,1b,1p,1b,1p,1b,1p,10m –pbwt-depth 8.

2. The phased VCFs were merged with non-phased genotype prediction for the repeat length as provided by ExpansionHunter. These combined VCFs were then phased again using Beagle v4.0. This separate step is necessary because SHAPEIT

does not accept genotypes with more than the two possible alleles (as is the case for repeat expansions).

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

TOPMed

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

1, We extracted SNPs with maf>=0.01 that were within +/-5 Mb of the Tandem Repeats and ran beagle (version .22Jul22.46e) on these SNPs to perform phasing with parameters burnin=10 and iterations=10.

```
 SNP phasing using beagle
java -jar ./beagle.22Jul22.46e.jar \
gt=${input} \
ref=../RefVCF/hgdp.tgp.gwaspy.merged.chr${chr}.merged.cleaned.vcf.gz \
out=Topmed.SNPs.maf0.001.chr${prefix}.beagle \
chrom=$region \
burnin=10 \
iterations=10 \
map=./genetic_maps/plink.chr${chr}.GRCh38.map \
nthreads=${threads}  \
impute=false
```

2. Next, we merged the unphased Tandem Repeat genotypes with the respective phased SNP genotypes using the bcftools. We used beagle version r1399, incorporating the parameters burnin-its=10, phase-its=10, and usephase=true. This version of beagle allows multiallelic Tander Repeat to be phased with SNPs.

```
ml beagle
java -jar ./beagle.r1399.jar \
gt=${input} \
out=${prefix} \
burnin-its=10 \
phase-its=10 \
map=./genetic_maps/plink.${chr}.GRCh38.map \
```

```
nthreads=${threads}  \
usephase=true
```

3. To conduct local ancestry analysis (LAI), we used RFMIX[26] with the parameter -n 5 -e 1 -c 0.9 -s 0.9 and -G 15. We utilised phased genotypes of 1,000 genomes as a reference panel[27].

```
time rfmix \
  -f $input \
  -r ../RefVCF/hgdp.tgp.gwaspy.merged.${chr}.merged.cleaned.vcf.gz \
  -m samples_pop \
  -g genetic_map_hg38_withX_formatted.txt \
  --chromosome=$c \
  -n 5 \
  -e 1 \
  -c 0.9 \
  -s 0.9 \
  -G 15 \
  --n-threads=48 \
  -o $prefix
```

Repeat size distribution analysis

The distribution of each RE was analysed across the 100K GP and TOPMed datasets (**Fig. S4**), and reproduced afterwards on the 1K GP3. Per each gene the distribution of the repeat-size across each super-population subset was analysed using the Wilcoxon test (**Fig. S5**).

Supplementary Tables

Supplementary_tables_final.xlsx

REFERENCES

1. Ibañez, K. *et al.* Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* **21**, 234–245 (2022).
2. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
3. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
4. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
5. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
6. Dolzhenko, E. *et al.* REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* **14**, 84 (2022).
7. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. https://onsdigital.github.io/dp-filter-a-dataset-prototype/v2/pop-est-current/.
8. Murphy, N. A. *et al.* Age-related penetrance of the C9orf72 repeat expansion. *Sci. Rep.* **7**, 2116 (2017).
9. Evans, S. J. W. *et al.* Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *J. Neurol. Neurosurg. Psychiatry* **84**, 1156–1160 (2013).

10. Vanacore, N. *et al.* An Age-Standardized Prevalence Estimate and a Sex and Age Distribution of Myotonic Dystrophy Types 1 and 2 in the Rome Province, Italy. *Neuroepidemiology* **46**, 191–197 (2016).
11. EUROSCA. http://www.eurosca.org/.
12. Glasmacher, S. A., Wong, C., Pearson, I. E. & Pal, S. Survival and Prognostic Factors in C9orf72 Repeat Expansion Carriers: A Systematic Review and Meta-analysis. *JAMA Neurol.* **77**, 367–376 (2020).
13. Bates, G. P. *et al.* Huntington disease. *Nat Rev Dis Primers* **1**, 15005 (2015).
14. Diallo, A. *et al.* Survival in patients with spinocerebellar ataxia types 1, 2, 3, and 6 (EUROSCA): a longitudinal cohort study. *Lancet Neurol.* **17**, 327–334 (2018).
15. Wahbi, K. *et al.* Development and Validation of a New Scoring System to Predict Survival in Patients With Myotonic Dystrophy Type 1. *JAMA Neurol.* **75**, 573–581 (2018).
16. Pulst, S. M. *Spinocerebellar Ataxia Type 2*. (University of Washington, Seattle, 2019).
17. Tabrizi, S. J. *et al.* A biological classification of Huntington's disease: the Integrated Staging System. *Lancet Neurol.* **21**, 632–644 (2022).
18. Hogan, D. B. *et al.* The Prevalence and Incidence of Frontotemporal Dementia: a Systematic Review. *Can. J. Neurol. Sci.* **43 Suppl 1**, S96–S109 (2016).
19. Van Mossevelde, S., Engelborghs, S., van der Zee, J. & Van Broeckhoven, C. Genotype-phenotype links in frontotemporal lobar degeneration. *Nat. Rev. Neurol.* **14**, 363–378 (2018).
20. Gossye, H., Engelborghs, S., Van Broeckhoven, C. & van der Zee, J. *C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis*. (University of Washington, Seattle, 2020).

21. Zampatti, S. *et al.* C9orf72-Related Neurodegenerative Diseases: From Clinical Diagnosis to Therapeutic Strategies. *Front. Aging Neurosci.* **14**, 907122 (2022).
22. Pringsheim, T. *et al.* The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
23. Rawlins, M. D. *et al.* The Prevalence of Huntington's Disease. *Neuroepidemiology* **46**, 144–153 (2016).
24. Sathe, S. *et al.* Enroll-HD: An Integrated Clinical Research Platform and Worldwide Observational Study for Huntington's Disease. *Front. Neurol.* **12**, 667420 (2021).
25. Bhandari, J., Thada, P. K. & Samanta, D. *Spinocerebellar Ataxia*. (StatPearls Publishing, 2022).
26. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
27. 1000 Genomes Project Consortium, {fname} *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
28. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

Acknowledgements

This work was supported by and funding from the UKRI (MR/S006753/1), Barts charity (MGU0569) and a Medical Research Council Clinician Scientist award (MR/S006753/1) to A.T.. A.J.S. received support from NIH grants AG075051, NS105781, HD103782 and NS120241, and A.M.T. received support from NHLBI Biodata Catalyst fellowship 5120339.

Data used in the preparation of this publication were obtained from the Rare Disease Cures Accelerator - Data and Analytics Platform (RDCA-DAP) funded by FDA Grant U18FD005320 and administered by Critical Path Institute. The data was provided to RDCA-DAP by Universitätsklinikum Bonn and Universitätsklinikum Tübingen [Universitätsklinikum Bonn and Universitätsklinikum Tübingen of datasets used by Arianna Tucci in March 2023.

Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: The BioMe Biobank at Mount Sinai” (phs001644.v1.p1) was performed at the McDonnell Genome Institute (3UM1HG008853-01S2). Genome sequencing for “NHLBI TOPMed: Women's Health Initiative (WHI)” (phs001237.v2.p1) was performed at the Broad Institute Genomics Platform (HHSN268201500014C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center

(3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Women's Health Initiative (WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. This manuscript was not prepared in collaboration with investigators of the WHI, and does not necessarily reflect the opinions or views of the WHI investigators, or NHLBI.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institute of Health, Department of Health and Human Services, under contract numbers (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR000040, UL1-TR-001420, UL1-TR-001881, and DK063491.

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors

also wish to thank the staff and participants of the JHS.

This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

This research used data generated by the COPDGene study, which was supported by NIH grants U01 HL089856 and U01 HL089897. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board composed of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion.

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

The Mount Sinai BioMe Biobank is supported by The Andrea and Charles Bronfman Philanthropies.

Data availability

For the 100K GP, full data is available in the Genomic England Secure Research Environment. Access is controlled to protect the privacy and confidentiality of participants in the Genomics England 100,000 Genomes Project and to comply with the consent given by participants for use of their healthcare and genomic data. Access to full data is permitted to researchers after registration with a Genomics England Clinical Interpretation Partnership (GeCIP) (https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/) and by contacting the corresponding author upon reasonable request.

For TOPMed, a detailed description of the TOPMed participant consents and data access is provided in Box 1[28]. TOPMed data used in this manuscript are available through dbGaP. The dbGaP accession numbers for all TOPMed studies referenced in this paper are listed in Extended Data Tables 2 and 3[28]. A complete list of TOPMed genetic variants with summary level information used in this manuscript is available through the BRAVO variant browser (bravo.sph.umich.edu). The TOPMed imputation reference panel described in this manuscript can be used freely for imputation through the NHLBI BioData Catalyst at the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/). DNA sequence and reference placement of assembled insertions are available in VCF format (without individual genotypes) on dbGaP under the TOPMed GSR accession phs001974.

Inclusion & Ethics

The 100 000 Genomes Project is a UK programme to assess the value of whole genome sequencing in patients with unmet diagnostic needs in rare disease and cancer. Following ethical approval for the 100 000 Genomes Project by the East of England Cambridge South Research Ethics Committee (reference 14/EE/1112), including for data analysis and return of diagnostic findings to the patients, these patients were recruited by health-care professionals and researchers from 13 Genomic Medicine Centres in England, and were enrolled in the project if they or

their guardian provided written consent for their samples and data to be used in research, including this study.

For ethics statements for the contributing TOPMed studies, full details are provided in the original description of the cohorts (Supplementary material).[28]