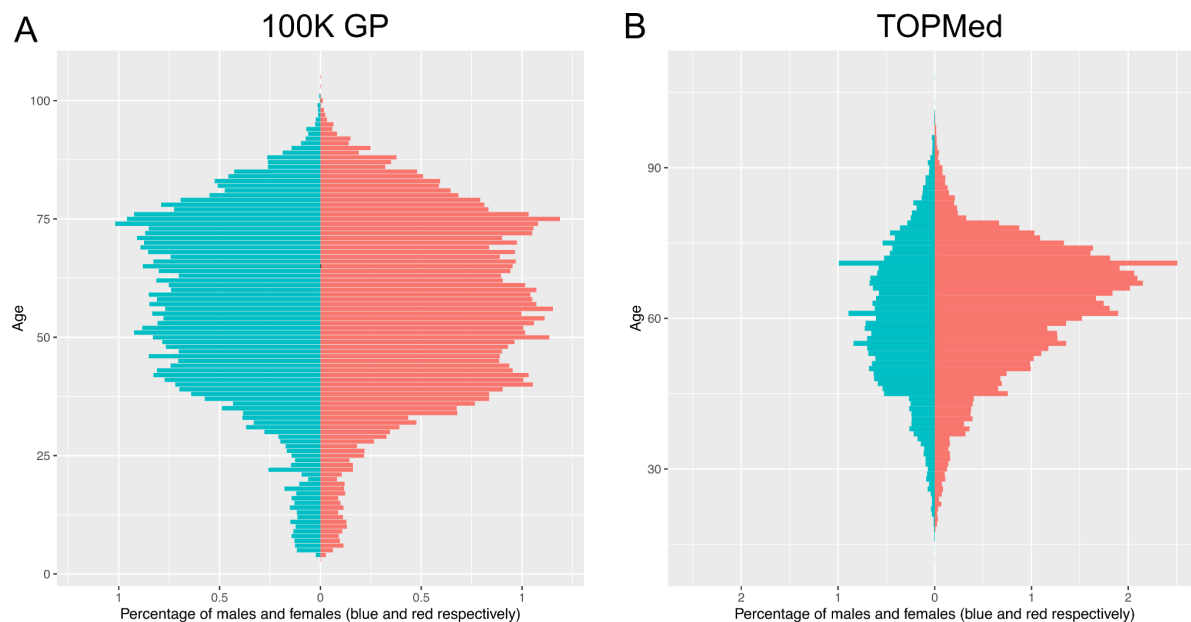
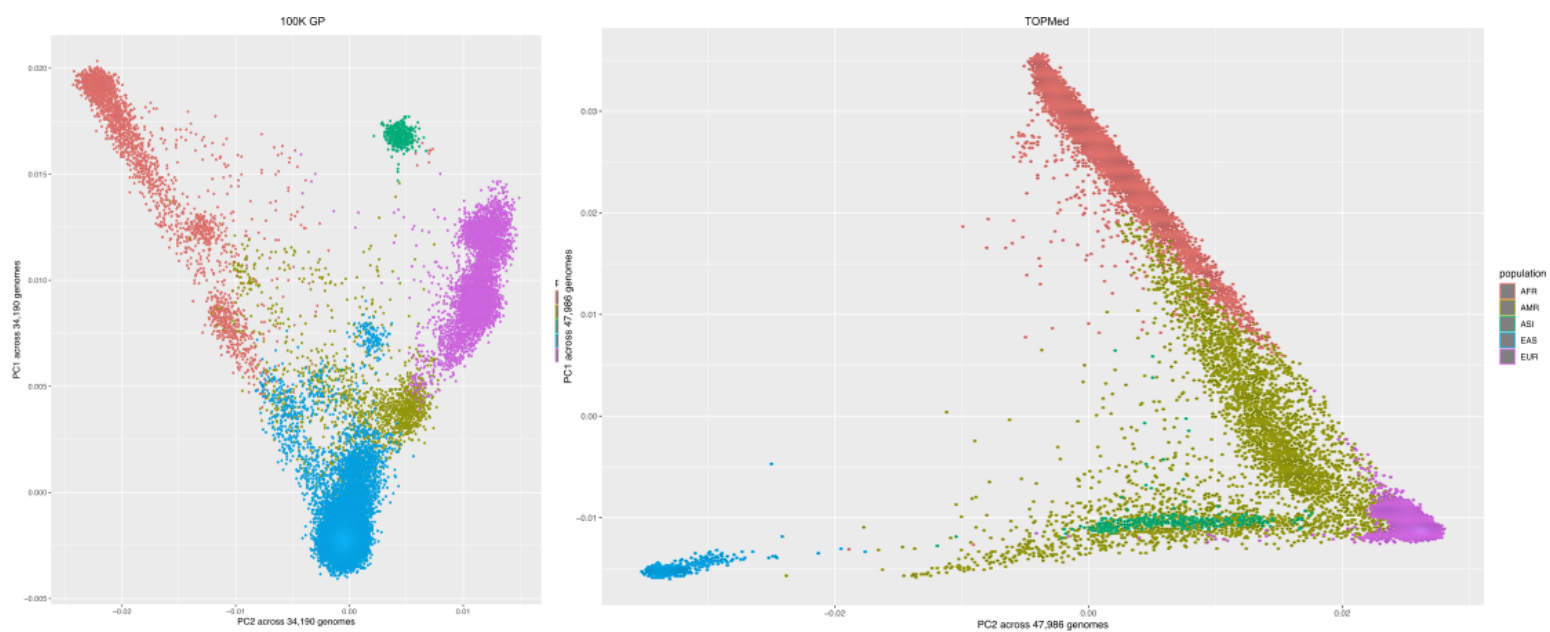


## SUPPLEMENTARY FIGURES

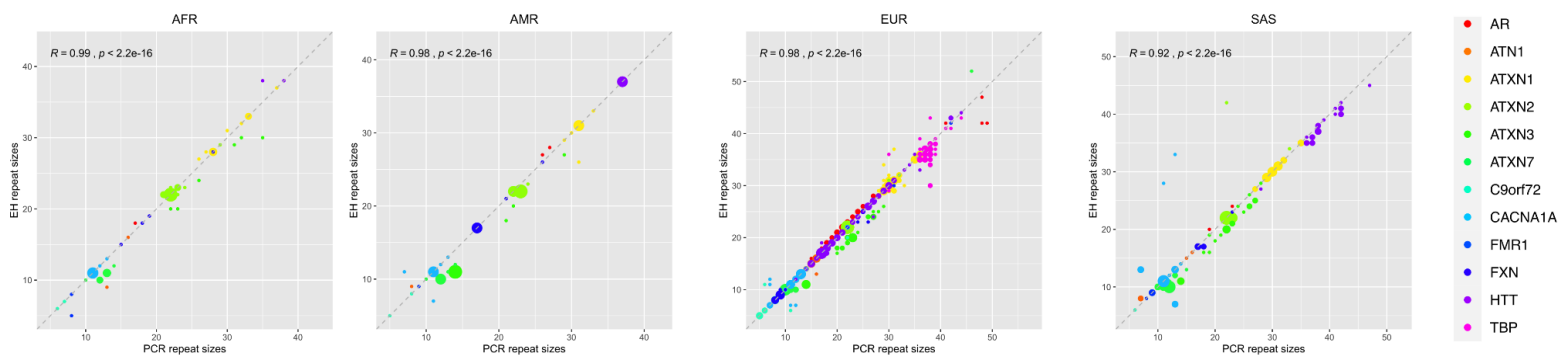
**Figure S1:** Pyramid demographics in (A) the 100K GP and (B) TOPMed cohorts.



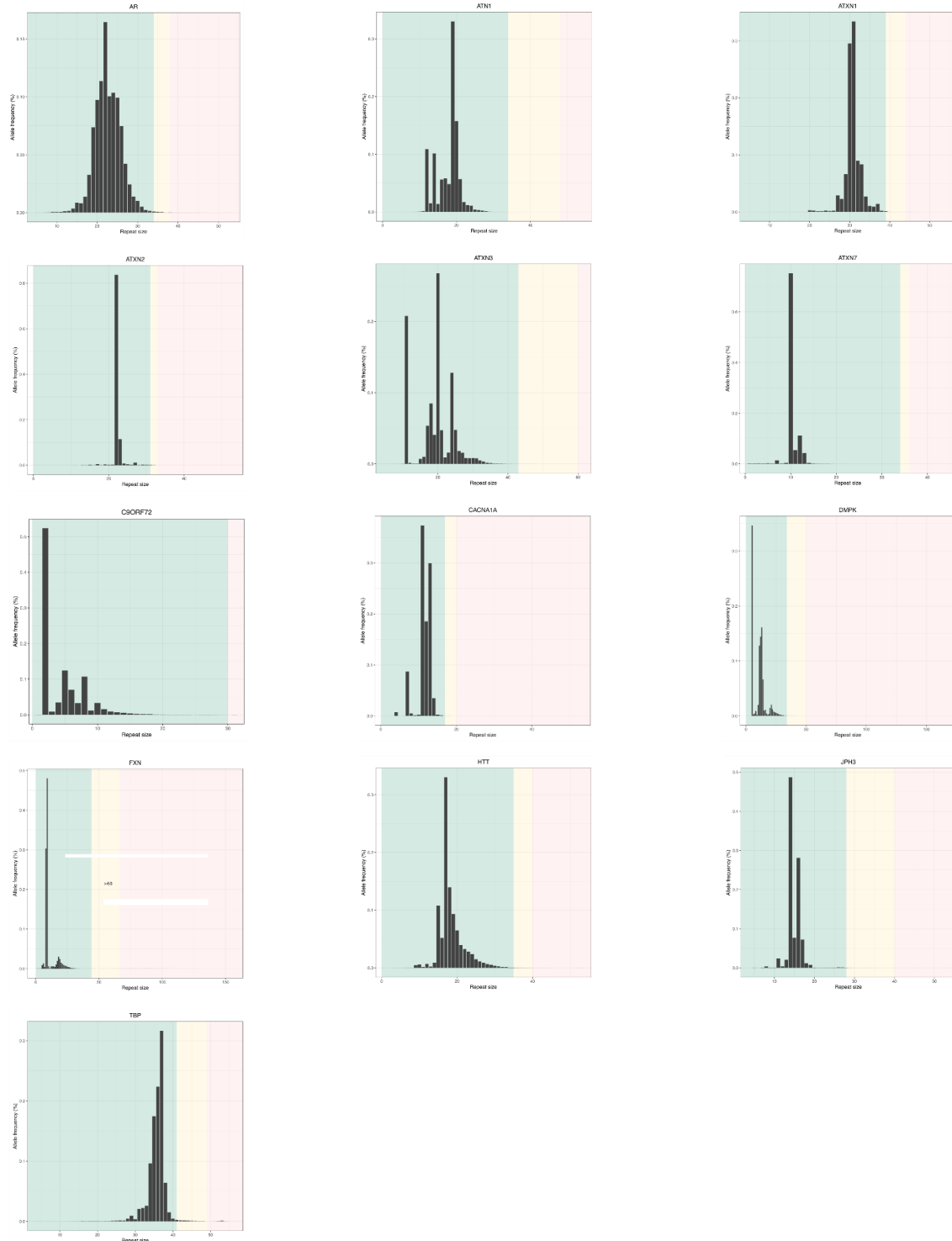
**Figure S2.** First 2 principal components (PCs) derived from PCA on A) the 100K GP and B) TOPMed samples (see **online materials** for more information).



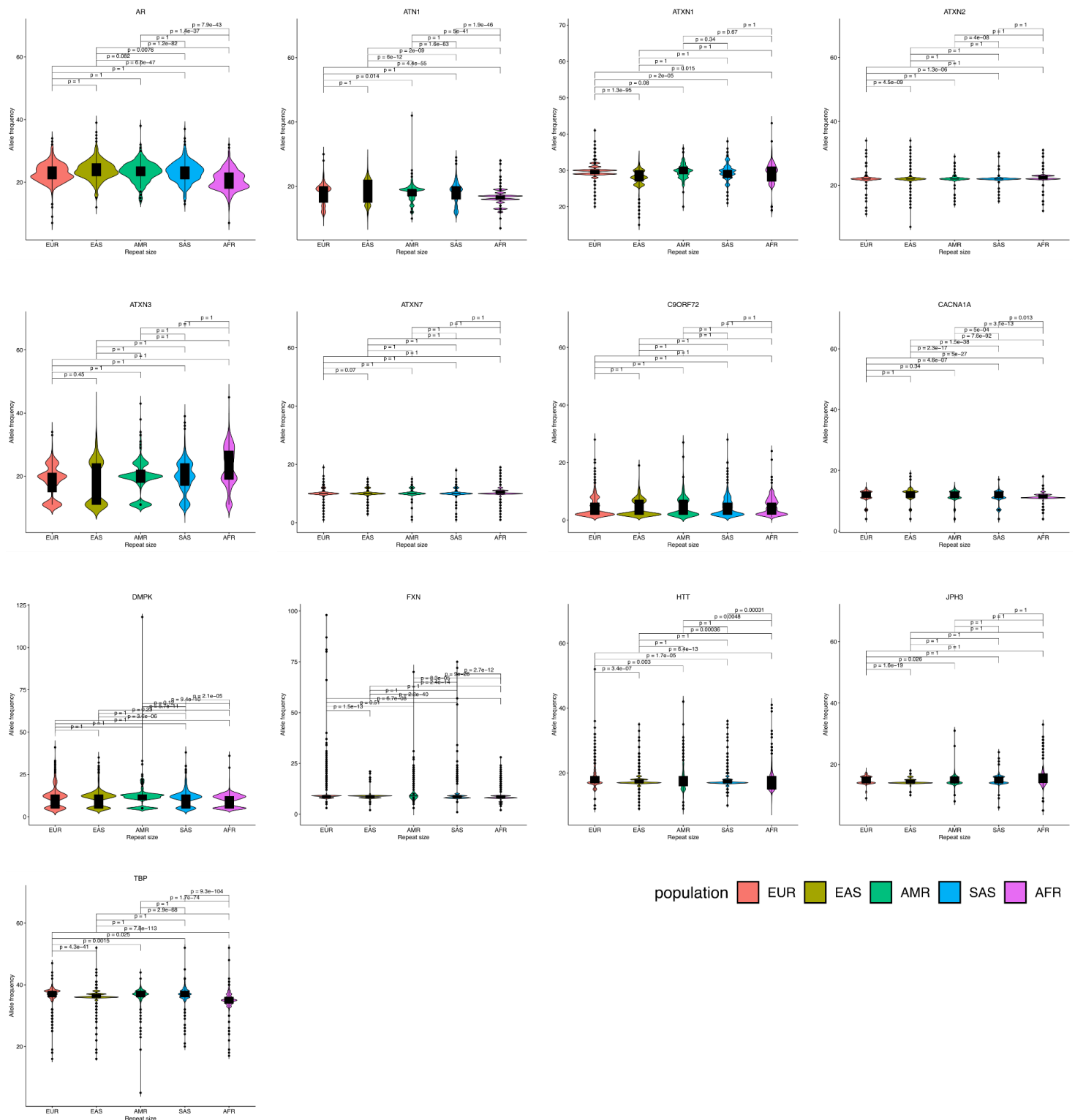
**Figure S3.** Experimental estimations of RE sizes using PCR vs genotypes generated by EH v3.2.2, split by super-population (samples of EAS ancestry were not tested). Points indicate the RE size estimated by both PCR and EH v3.2.2. We show the R correlation coefficient calculated using Pearson's equation.



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

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<sup>1</sup>.

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`)<sup>2</sup>, 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/](http://www.cog-genomics.org/plink/2.0/))<sup>3</sup>. For relatedness, PLINK2 `--king-cutoff` ([www.cog-genomics.org/plink/2.0/](http://www.cog-genomics.org/plink/2.0/)) relationship-pruning algorithm<sup>3</sup> was used with a threshold of 0.044. These were then partitioned into 'related' (up to, and including 3<sup>rd</sup> 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 `Ntrees` 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<sup>4,5</sup>. 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<sup>6</sup>. **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)

- $\text{freq\_carrier} = \text{round}(\text{total\_unrel} / \text{total\_exp\_after\_VI\_locus}, \text{digits} = 2)$
- $\text{ci\_max} = \text{round}(\text{total\_unrel} / (\text{total\_unrel} * ((\text{total\_exp\_after\_VI\_locus} / \text{total\_unrel}) - 1.96 * \sqrt{((\text{total\_exp\_after\_VI\_locus} / \text{total\_unrel}) * (1 - \text{total\_exp\_after\_VI\_locus} / \text{total\_unrel})) / \text{total\_unrel}})), \text{digits} = 2)$
- $\text{ci\_min} = \text{round}(\text{total\_unrel} / (\text{total\_unrel} * ((\text{total\_exp\_after\_VI\_locus} / \text{total\_unrel}) + 1.96 * \sqrt{((\text{total\_exp\_after\_VI\_locus} / \text{total\_unrel}) * (1 - \text{total\_exp\_after\_VI\_locus} / \text{total\_unrel})) / \text{total\_unrel}})), \text{digits} = 2)$

## PREVALENCE ESTIMATE (x in 100,000)

$$x = 100,000 / \text{freq\_carrier}$$

$$\text{new\_freq} = 100000 * (1 / \text{frequency\_cohort2\_df\$carrier\_freq})$$

$$\text{new\_low\_ci} = 100000 * \text{low\_ci}$$

$$\text{new\_high\_ci} = 100000 * \text{high\_ci}$$

## 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<sup>7</sup>, 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<sup>8</sup>. HD onset was modelled on 246 patients from the UK's General Practice Research Database<sup>9</sup> and DM1 was modelled on a cohort of 395 patients<sup>10</sup>. 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<sup>11</sup>.

- (iii) Mortality from disease. Median survival length is approximately three years for C9orf72-ALS and ten years for C9orf72-FTD<sup>12</sup>. HD and SCA2 have a median survival of fifteen years<sup>13,14</sup>. Given that approximately 22% of patients with DM1 die over a period of 11 years, we estimated a survival of 80% after 10 years<sup>15</sup>.
- (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<sup>16</sup>. 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<sup>17</sup> for a 40-CAG repeat carrier.
- (v) Reduced penetrance, e.g., C9orf72-carriers may not develop symptoms even after 90 years of age<sup>8</sup>. Thus, age-related penetrance of C9orf72-ALS/FTD was derived from the red curve in

**Fig. 2B** reported by Murphy et al<sup>8</sup>, 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<sup>18</sup> (83.5 in 100,000). Since 4-29% of FTD patients carry a *C9orf72* repeat expansion<sup>19</sup>, 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<sup>20</sup> and *C9orf72* repeat expansion is found in 30%-50% of individuals with familial forms and in 4%-10% of people with sporadic disease<sup>21</sup>. Given that ALS is familial in 10% of cases and sporadic in 90%, we estimated the prevalence of *C9orf72*-ALS by

calculating the  $[(0.4 \text{ of } 0.1) + (0.07 \text{ of } 0.9)]$  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<sup>22</sup> to 10 in 100,000 in Europeans<sup>23</sup>, 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<sup>24</sup> 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<sup>25</sup>.

## 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  $\text{maf} \geq 0.01$  that were within  $\pm 5$  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 Tandem 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<sup>26</sup> 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<sup>27</sup>.

```
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.euroasca.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, *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<sup>28</sup>. 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<sup>28</sup>. 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](http://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).<sup>28</sup>