Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phenome wide analysis of neurological disease associated short tandem repeats in UK Biobank

Johnathan Cooper-Knock

j.cooper-knock@sheffield.ac.uk

University of Sheffield https://orcid.org/0000-0002-0873-8689

Andrew Strange

University of Sheffield https://orcid.org/0000-0002-2004-150X

Elham Alhathli

Faculty of Nursing, Taif University,

Omar Abed

University of Sheffield

Thomas Julian

Division of Evolution, Infection and Genomics, School of Biological Sciences, University of Manchester

Marwa Hamdan

University of Sheffield

Andrew Sharp

Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0003-1415-5129

Bharati Jadhav

Arunachalam Soma

Pamela Shaw

The University of Sheffield https://orcid.org/0000-0002-8925-2567

Article

Keywords: short tandem repeat, repeat expansion, phenome-wide association study, UK Biobank, myotonic dystrophy type 1, gynaecological cancer

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11478873/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no duality of interest

Abstract

Short tandem repeat (STR) expansions cause neurological and neuromuscular disorders, but their phenotypic spectrum in unselected carriers is incompletely characterised. We performed a phenome-wide association study of 48 STR variants at 41 neurological disease loci, genotyped with ExpansionHunter in 484,561 whole-genome sequenced UK Biobank participants. Carriers were defined by published pathogenic thresholds or, where thresholds exceeded the 150-bp read length, as the top 0.01% of repeat length. Hospital inpatient ICD-10 diagnoses were tested by Firth logistic regression adjusted for age, sex and 10 genetic principal components, requiring at least five carrier cases, with false discovery rate (FDR) correction. We identified 145 associations at FDR < 0.05 across 12 loci, 85 of them novel, most at *DMPK* (101) and *TCF4* (17). *DMPK* expansion carriers (n = 324) had increased odds of gynaecological cancer (odds ratio [OR] 3.03, 95% confidence interval 1.46 to 5.53; nine carrier cases), including ovarian (OR 4.67, 1.75 to 9.93) and endometrial cancer (OR 3.58, 1.34 to 7.63), and of benign colonic neoplasm (OR 4.20, 1.95 to 7.82). Incident gynaecological cancer occurred in 4.7% of female carriers versus 1.4% of non-carriers over a median 13.8 years (hazard ratio 3.59, 1.79 to 7.17). *TCF4* carriers (n = 38,676) had lower odds of harmful tobacco use (OR 0.90, 0.85 to 0.94), chronic obstructive pulmonary disease (OR 0.89, 0.84 to 0.94) and alcohol withdrawal (OR 0.61, 0.48 to 0.77). Population-scale STR genotyping recovers established disease spectra; the excess of gynaecological cancer in *DMPK* carriers supports evaluating surveillance in myotonic dystrophy type 1.

Introduction

Short tandem repeat (STR) expansions cause disorders including Huntington's disease (*HTT*), myotonic dystrophy types 1 (*DMPK*) and 2 (*CNBP*), spinocerebellar ataxias and Fuchs endothelial corneal dystrophy (*TCF4*). Their effects may extend beyond these canonical syndromes.

Pleiotropy is established at some loci: *ATXN2* CAG repeats increase amyotrophic lateral sclerosis risk,[1] and polyglutamine expansion disorders are associated with reduced cancer incidence.[2] UK Biobank combines population-scale sequencing with linked diagnoses, enabling systematic investigation of less frequent or incompletely penetrant associations.[3]

Our neurological-disease associated STR PheWAS recovered established disease associations and identified a number of novel associations: *DMPK*-associated neoplasia, ophthalmic associations at *CNBP* and *CSTB*, muscle pathology at *TBP*, reduced odds of respiratory-related diagnoses at *TCF4*, and repeat-length and allele-dose effects on canonical phenotypes.

Materials and methods

UK Biobank cohort

The study included 489,727 UK Biobank participants with GATK-processed whole-genome sequencing.[4] Participants were aged 40 to 69 years at recruitment.[3,5] Age, genetic sex, the first 10 genetic principal components (GPCs) and hospital inpatient ICD-10 diagnoses were obtained from the 2026 phenotype release. The primary analysis included 484,561 participants with complete GPC data (Supplementary Table S1).

ExpansionHunter genotyping

ExpansionHunter v5 estimates allele lengths from spanning reads for short repeats and from flanking and in-repeat reads for longer tracts, with lower precision.[6,7] The public catalogue[8] was restricted to 48 variants at 41 genes (Supplementary Table S2). *C9orf72* was excluded because its neurodegenerative associations have been studied separately in UK Biobank.[9]

Pathogenic thresholds

Carrier status required the longer called allele present within an individual to meet the published STRipy threshold[10] (Supplementary Table S2) except where the threshold was longer than the 150-bp sequencing read, in which case we moved to outlier-based carrier calling. Within-carrier analyses examined continuous repeat length, distinguishing alleles countable within 150-bp sequencing reads from longer, inferred alleles with lower sizing precision.

Outlier-based carrier calling

ExpansionHunter genotyping of STR length is relatively inaccurate beyond the 150-bp sequencing read length. To identify associations when the published pathogenic threshold cannot be accurately inferred because it is > 150-bp, we defined carrier status as participants in the top 0.01% of longer-allele repeat counts across the cohort.[11] At *DMPK*, whose threshold of 50 CTG repeats equals the read length, the top 0.1% was also examined.

Genotype quality control

Calls were assessed with the ExpansionHunter Classifier, a random forest model using read quality, GC content and motif purity.[11] The classifier evaluates call reliability, not pathogenicity. *TCF4* and *CBL* were absent from its catalogue. Locus-specific validation rates are reported in Supplementary Table S3.

PheWAS

For each variant-ICD-10 association pair, we tested for a significant association by Firth logistic regression (logistf[12,13]) adjusting for age, sex and GPCs 1 to 10. 5,166 participants missing both GPCs

and genetic sex were excluded, leaving 484,561.

To provide sufficient power for statistical testing, association tests were limited to least one case, one non-case and five carrier cases[11,14] leaving 19 testable variants. Benjamini-Hochberg FDR correction was applied within each variant across eligible diagnoses,[15] rather than study-wide. Related ICD-10 codes were inspected descriptively when only one code in a diagnostic block was significant.

In the assessment of whether *DMPK* carrier status was associated with any neoplasm we joined multiple ICD-10 codes into a single statistical test. To assess whether this result was a false positive resulting from the large number of ICD-10 codes, *DMPK* carrier status permuted across 484,561 participants in 500 iterations. Each permutation repeated the primary neoplasm analysis, including eligibility assessment and the five-carrier-case minimum. The empirical P value was the proportion of permutations yielding at least the observed number of nominally significant neoplasm codes.

Pooled neoplasm endpoints

We conducted a separate analysis of pooled neoplasm codes to determine whether cancer observed associations were specific to certain cancer categories. We extracted cancer-related traits in the UK Biobank, including data from the cancer registry, medical ICD codes, cause of death and survey questions, including self-reported cancer. We considered an individual as a cancer case if they had one or more positive indications from any of these data sources. We organized these diverse traits into groupings to define individual cancer types (e.g. Thyroid cancer, liver cancer, breast cancer etc.) and further clustered them into larger groupings representative of major categories of cancer (e.g. Hematologic, Gastrointestinal, etc). A complete list of cancer traits and their groupings is shown in Supplementary Table S4.

Cancer time-to-event

First inpatient cancer diagnosis dates were derived from matched ICD-10 and date fields. Five prespecified endpoints were examined: malignant neoplasms (C00 to C97 excluding C44), gynaecological cancer (C51 to C58), colorectal cancer (C18 to C21), benign colorectal neoplasms (D12) and other malignant neoplasms, with FDR correction across endpoints. Cox regression was applied to test for differences in time-to-diagnosis with exclusion of prevalent cases at UK Biobank recruitment and censoring at death or hospital follow-up end. Models adjusted for sex (except female-only analyses) and GPCs 1 to 10. The cancer analysis frame comprised 484,091 participants, including 324 *DMPK* carriers (171 female), before endpoint-specific exclusions. Fine-Gray and ever-diagnosed logistic models provided sensitivity analyses. In addition, among individuals diagnosed with cancer, age at first diagnosis was compared between carriers and non-carriers using an unadjusted Wilcoxon rank-sum test (all malignancies in both sexes; ovarian cancer and endometrial cancer in women only).

STR length comparison within carriers

At the 11 STR loci contributing an association to the PheWAS analysis, longer-allele repeat count was tested for association with diagnoses with at least five carrier cases using Firth regression adjusting for age, sex and GPCs 1 to 10 with within-locus FDR correction. Effects were reported per repeat and per ten repeats. X-linked *ARX* was excluded. This analysis was limited to STR length < 150-bp where ExpansionHunter calls are accurate.

Sensitivity analyses

Threshold sensitivity was assessed at the prespecified STRipy cut-off and one and two repeat units higher.

Results

PheWAS identifies established and candidate STR-associated phenotypes

The primary PheWAS identified 145 associations across 12 loci at FDR < 0.05 in 484,561 participants (Fig. 1; Supplementary Tables S5 and S6). Nineteen variants met the testing criteria, contributing 13,191 tests (Supplementary Table S7). Associations were not distributed evenly across tested loci. The top four by count were: *DMPK* 101 associations, *TCF4* 17, *CACNA1A* 11 and *HTT* five (Fig. 2).

Known associations act as a positive control

Canonical associations were recovered (**Supplementary Figure S1**; Supplementary Table S6). *DMPK* showed the broad myotonic dystrophy type 1 spectrum,[16] including muscular, ocular, cardiac, respiratory and gastrointestinal manifestations, weight loss, falls and ventilator dependence. *CNBP* was associated with myotonic disorders, consistent with myotonic dystrophy type 2.[17]

HTT (1,052 carriers) was associated with Huntington's disease and dementia;[18] *TCF4* (38,676 carriers) with corneal dystrophy, oedema and transplantation;[19] and *CACNA1A* (77 carriers) with ataxia, gait abnormalities and falls.[20] Canonical diagnoses at *ATXN2* (43 carriers), *FMR1* (1,631) and *CSTB* (241) had fewer than five carrier cases. *ATXN3* had only three pathogenic-range carriers and was not testable.

Novel associations beyond established phenotypes

Figure 3A shows selected associations beyond canonical phenotypes; Supplementary Table S5 provides classifications and statistics.

Neoplastic associations

All FDR-significant neoplastic associations occurred at *DMPK* (324 carriers, 171 female; all participants, adjusted for sex): ovarian cancer (C56; OR 4.67, 95% CI 1.75 to 9.93; FDR = 0.019; five carrier cases, all female), endometrial cancer (C541; OR 3.58, 1.34 to 7.63; FDR = 0.047; five cases, all female), benign colonic neoplasm (D126; OR 4.20, 1.95 to 7.82; FDR = 0.0053; eight cases), and secondary retroperitoneal or peritoneal malignancy (C786; OR 3.08, 1.35 to 5.92; FDR = 0.037; seven cases; Fig. 3B; **Supplementary Figure S2**). Three further codes were nominally significant: C772 (OR 2.75, P = 0.029; six cases), benign rectal neoplasm (D128; OR 2.13, P = 0.031; ten cases) and uterine leiomyoma (D259; OR 1.96, P = 0.025; 14 cases, all female). Another 32 nominal neoplasm associations had fewer than five carrier cases. Across 500 permutations, the number of nominally significant *DMPK*-neoplasm ICD-10 associations was significantly less than observed (mean = 1.57 versus observed = 7, empirical P = 0.008). Pooled, all malignant neoplasms were not associated with *DMPK* carrier status (OR 1.04, 95% CI 0.76 to 1.40; P = 0.79), whereas gynaecological cancer was (C51 to C58; OR 3.03, 1.46 to 5.53; FDR = 0.027; nine carrier cases; Supplementary Table S4).

In a female-only analysis, *DMPK* carriers showed increased hazard of incident gynaecological cancer (HR 3.59, 95% CI 1.79 to 7.17; P < 0.001; FDR-adjusted P = 0.0015): eight events among 170 female carriers (4.7%) versus 1.4% of 260,954 female non-carriers over median follow-up of 13.8 years. Estimates were similar with competing mortality (sub distribution HR 3.47, 1.73 to 6.98) and ever-diagnosed status (OR 2.89, 1.47 to 5.66). Incident ovarian cancer was increased (four events, below the five-case minimum; HR 4.60, 1.72 to 12.26; P = 0.0023). Age at first diagnosis was similar for all malignancies (both sexes; 65.5 versus 65.6 years; P = 0.76) and ovarian cancer (female only; 65.5 versus 64.4; P = 0.64).

Ophthalmic associations

Macular degeneration was associated with *CNBP* STR carrier status (H353; OR 2.8, FDR = 0.028; 17 cases among 290 carriers), and visual field defects with *CSTB* (H534; OR 8.8, FDR = 0.034; five cases among 241 carriers, all with one expanded allele).

Injury, musculoskeletal and other systemic associations

CACNA1A associations included gait and mobility abnormalities (R268; OR 5.5, FDR = 0.001; nine cases), repeated falls (R296; OR 4.5, FDR = 0.004; nine), unspecified falls (W199; OR 6.1, FDR = 0.004; six) and falls after bumping against an object (W180; OR 15.8, FDR < 0.001; seven), consistent with downstream effects of ataxia. *TBP* was associated with myalgia (M7910; OR 14.2, FDR = 0.006; five cases among 618 carriers). *ARX* was associated with inflamed lower-extremity varicose veins (I831; OR 7.8, 95% CI 2.9 to 16.6; FDR = 0.049; five cases among 216 carriers).

Tobacco-related, respiratory and alcohol-related associations at TCF4

TCF4 carriers had lower odds of alcohol withdrawal (F103; OR 0.61, 95% CI 0.48 to 0.77; FDR = 0.0043; 74 cases), chronic obstructive pulmonary disease (J449; OR 0.89, 0.84 to 0.94; FDR = 0.0074; 1,478), harmful tobacco use (F171; OR 0.90, 0.85 to 0.94; FDR = 0.0062; 1,800) and asthma (J459; OR 0.93, 0.90 to 0.96; FDR = 0.018; 3,551). Alcoholic cirrhosis (K703; OR 0.60, P = 0.00023, FDR = 0.052; 47 cases) and

alcohol dependence (F102; OR 0.80, P = 0.0011, FDR = 0.14; 229) were nominally significant (Supplementary Table S7).

Threshold sensitivity

Raising thresholds by one and two repeats retained 137 and 124 of the 145 primary associations, respectively. At one repeat higher, eight associations were lost: personal history of other diseases (Z864) at *CBL*, visual field defects (H534) at *CSTB*, ventricular tachycardia (I472), hypertrophy of uterus (N852) and pain in throat and chest (R073) at *DMPK*, unspecified injury of abdomen, lower back and pelvis (S399) and degenerative disease of the nervous system (G319) at *HTT*, and myalgia (M7910) at *TBP*. All except R073 and G319 fell below five carrier cases rather than failing FDR correction. At two repeats higher, 13 further associations were lost: all 11 at *CACNA1A*, where only three carriers remained, pulmonary embolism with acute cor pulmonale (I260) at *HTT*, which fell below five carrier cases, and calculus of gallbladder with other cholecystitis (K801) at *DMPK* (FDR = 0.050).

Outlier-based carrier calling

At loci where the canonical disease length was longer than 150bp (see Methods), we performed outlier analysis. At the top 0.01% cut-off, rank-based calling yielded 189 associations. These included myotonic disorders at *DMPK* (G711; OR 28,704, 95% CI 14,292 to 61,995; FDR < 0.001; 38 cases), Huntington's disease at *HTT* (G10; OR 18,425, 95% CI 9,241 to 38,469; FDR < 0.001; 29 cases) and hereditary ataxia at *CACNA1A* (G119; OR 449, 95% CI 235 to 803; FDR < 0.001; 13 cases).

Allele dosage

Six autosomal loci had at least five individuals with bi-allelic expansions in the pathogenic range: *TCF4*, *FXN*, *ATXN8OS*, *ATXN1*, *HTT* and *TBP*. Excluding double-expanded carriers retained 143 of 144 primary associations. *TCF4* corneal scars and opacities (H183) was the single loss (pooled FDR < 0.001; single-allele FDR = 0.070; Supplementary Table S8).

At *TCF4*, two versus one expanded allele (809 versus 37,867 carriers) was associated with corneal dystrophy (H185; OR 4.23, 95% CI 2.96 to 5.86; FDR < 0.001; 38 double-expanded cases) and transplantation (Z947; OR 5.22, 2.67 to 9.24; FDR = 0.036; 11 cases). Other corneal changes (H182; OR 5.86, 2.13 to 13.11; five cases) and opacity (H184; OR 1.63, 0.87 to 2.76; 12 cases) did not meet the FDR threshold. *TCF4* was the only locus with a fitted within-carrier dosage contrast; no diagnosis at *ATXN8OS*, *HTT* or *TBP* had five double-expanded cases.

Repeat length within carriers

Within-carrier analyses comprised 7,599 tests at 12 loci (Supplementary Table S9). Length effects occurred at *DMPK*, *HTT* and *TCF4*. At *DMPK*, 47 of 333 associated phenotypes met FDR < 0.05, spanning myotonic, cardiac, respiratory and ocular disease, dysphagia and aspiration. At *HTT*, effects per CAG repeat were OR 3.09 for Huntington’s disease (95% CI 2.47 to 4.01; 50 cases), 2.51 for Huntington’s dementia (1.82 to 3.77; ten) and 1.78 for aspiration pneumonitis (1.37 to 2.38; 13). At *TCF4*, the OR for

corneal dystrophy was 1.10 per ten CTG repeats (1.06 to 1.15; 480 cases). No other locus showed a significant length effect.

Carrier alleles at *CACNA1A*, *ATXN2*, *HTT* and almost all *TBP* carriers were countable within 150-bp reads. Supplementary Table S9 flags extrapolation beyond observed ranges and sparse analyses.

Discussion

This STR PheWAS identified 145 associations at 12 loci and recovered canonical disease spectra, particularly at *DMPK*, *TCF4*, *CACNA1A* and *HTT*. Non-detection at other loci may reflect short-read limitations or rarity: *RFC1* pathogenic expansions exceed the length of our 150-bp sequencing reads leading to inaccurate calling via ExpansionHunter, while *DIP2B* and *JPH3* pathogenic alleles are exceptionally uncommon.[14]

We observed *DMPK* neoplastic associations which extend observations from clinical myotonic dystrophy cohorts, including ovarian and endometrial cancer[21,22] and colorectal polyps,[23] to genetically identified carriers. Gynaecological cancer was observed in ~ 5% of genetic carriers which could justify a screening program. *DMPK* STR expansion was not associated with an all-malignancy endpoint but we note a large number of nominally significant cancer related associations which were directionally consistent (**Supplementary Figure S2**), which could suggest that our all-malignancy test was under powered in this cohort.

We observed an association between *CNBP* STR expansion and macular degeneration which extends beyond its recognised syndrome of myotonic dystrophy type 2, although retinal abnormalities have been described in myotonic dystrophy type 1.[24]

Our results suggest that *TCF4* STR expansion is associated with reduced tobacco and alcohol consumption, and correspondingly with lower risk of exposure-related diseases such as asthma and liver cirrhosis. This pattern is consistent with behavioural modification than with direct molecular protection, but worth further exploration in future works.

Limitations

Our study has several limitations. First, ExpansionHunter sizing from 150-bp short reads becomes less precise for longer alleles; as a result, loci whose pathogenic alleles exceed the read length were studies using outlier-based calling. Second, phenotypes were drawn from hospital inpatient ICD-10 codes. These are incomplete, and they may be subject to ascertainment bias: carriers with a clinical diagnosis, particularly of myotonic dystrophy, are likely to have more hospital contact and investigation. Third, UK Biobank participants are predominantly of European ancestry, aged 40 to 69 at recruitment and healthier than the general population. This may lead to underestimation of severe or early-onset phenotypes.

Conclusions

Population-scale STR genotyping in UK Biobank recovered the established disease spectra of neurological repeat expansions and identified 85 novel associations. *DMPK* expansion carriers had a threefold excess of gynaecological cancer, which supports evaluation of gynaecological cancer surveillance in myotonic dystrophy type 1. Replication in independent sequenced cohorts with linked primary care and cancer registry data, and long-read sizing at loci whose pathogenic alleles exceed the short-read length, should follow.

Declarations

Funding

This work was supported by NIHR (NF-SI-0617-10077 to P.J.S). J.C.K, P.J.S, and A.Str are supported by the NIHR Sheffield Biomedical Research Centre (JS-BRC-1215-20017). J.C.K is supported by the Lister Institute, the MND Association (894 – 791, 899 – 792 and 913 – 792), the Robert Packard Center for ALS Research (2006781541), TargetALS and The Medical Research Council (MRC) Clinical Research Training Fellowship (MR/Z504105/1, to T.H.J.)

Ethical approval

UK Biobank has ethical approval from the North West Multi-centre Research Ethics Committee as a Research Tissue Bank (21/NW/0157). All participants provided written informed consent at recruitment. This research was conducted using UK Biobank under application number 68812.

Competing interests

The authors declare no competing interests.

Funding:

National Institute for Health and Care Research (NIHR); Lister Institute; Motor Neurone Disease Association; Robert Packard Center for ALS Research; TargetALS; Medical Research Council (MRC). Grant numbers are given in the Funding statement.

Author contributions

J.C.K. and A.Str. conceptualised the study. A.Str. performed the analyses with assistance from E.A., T.J., A.Sha., B.J., M.H. and A.Som. O.A. reviewed established STR associations. A.Str. drafted the manuscript with input from all authors; J.C.K. and P.J.S. supervised the work and revised the manuscript.

Acknowledgements

This research was conducted using the UK Biobank Resource under application number 68812. We thank the UK Biobank participants.

Data availability

UK Biobank data are available to approved researchers on application to UK Biobank (https://www.ukbiobank.ac.uk). This study used application 68812. Results of all 13,191 association tests are provided in Supplementary Table S7.

References

1. Elden AC, Kim HJ, Hart MP, Chen-Plotkin AS, Johnson BS, Fang X, et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature. 2010;466(7310):1069–1075.
2. Ji J, Sundquist K, Sundquist J. Cancer incidence in patients with polyglutamine diseases: a population-based study in Sweden. Lancet Oncol. 2012;13(6):642–648.
3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
4. UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. Nature. 2025;645(8081):692–701.
5. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209.
6. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27(11):1895–1903.
7. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35(22):4754–4756.
8. Illumina. ExpansionHunter variant catalogues (repeat-expansion specification and coordinates). https://github.com/Illumina/ExpansionHunter (accessed 30 June 2026).
9. Gao J, Douglas AGL, Chalitsios CV, Scaber J, Talbot K, Turner MR, et al. Neurodegenerative disease in C9orf72 repeat expansion carriers: population risk and effect of UNC13A. Brain. 2025;148(11):3865–3871.
10. Halman A, Dolzhenko E, Oshlack A. STRipy: a graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. Hum Mutat. 2022;43(7):859–868.

11. Jadhav B, Garg P, van Vugt JJFA, Ibanez K, Gagliardi D, Lee W, et al. A phenome-wide association study of methylated GC-rich repeats identifies a GCC repeat expansion in AFF3 associated with intellectual disability. Nat Genet. 2024;56(11):2322–2332.
12. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.
13. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21(16):2409–2419.
14. Pounraja VK, Sul JH, Herman J, O'Keeffe S, Rajagopal V, Bai X, et al. Population-scale repeat expansions elucidate disease risk and brain atrophy. Nature. 2026;653(8115):796–808.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300.
16. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, 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. 1992;68(4):799–808.
17. Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science. 2001;293(5531):864–867.
18. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72(6):971–983.
19. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, et al. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy. PLoS One. 2012;7(11):e49083.
20. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, et al. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha-1A-voltage-dependent calcium channel. Nat Genet. 1997;15(1):62–69.
21. Gadalla SM, Lund M, Pfeiffer RM, Gørtz S, Mueller CM, Moxley RT 3rd, et al. Cancer risk among patients with myotonic muscular dystrophy. JAMA. 2011;306(22):2480–2486.
22. Win AK, Perattur PG, Pulido JS, Pulido CM, Lindor NM. Increased cancer risks in myotonic dystrophy. Mayo Clin Proc. 2012;87(2):130–135.
23. Alsaggaf R, St George DMM, Zhan M, Pfeiffer RM, Wang Y, Anderson LA, et al. Benign tumors in myotonic dystrophy type I target disease-related cancer sites. Ann Clin Transl Neurol. 2019;6(8):1510–1518.
24. Kimizuka Y, Kiyosawa M, Tamai M, Takase S. Retinal changes in myotonic dystrophy. Clinical and follow-up evaluation. Retina. 1993;13(2):129–135.

Figures

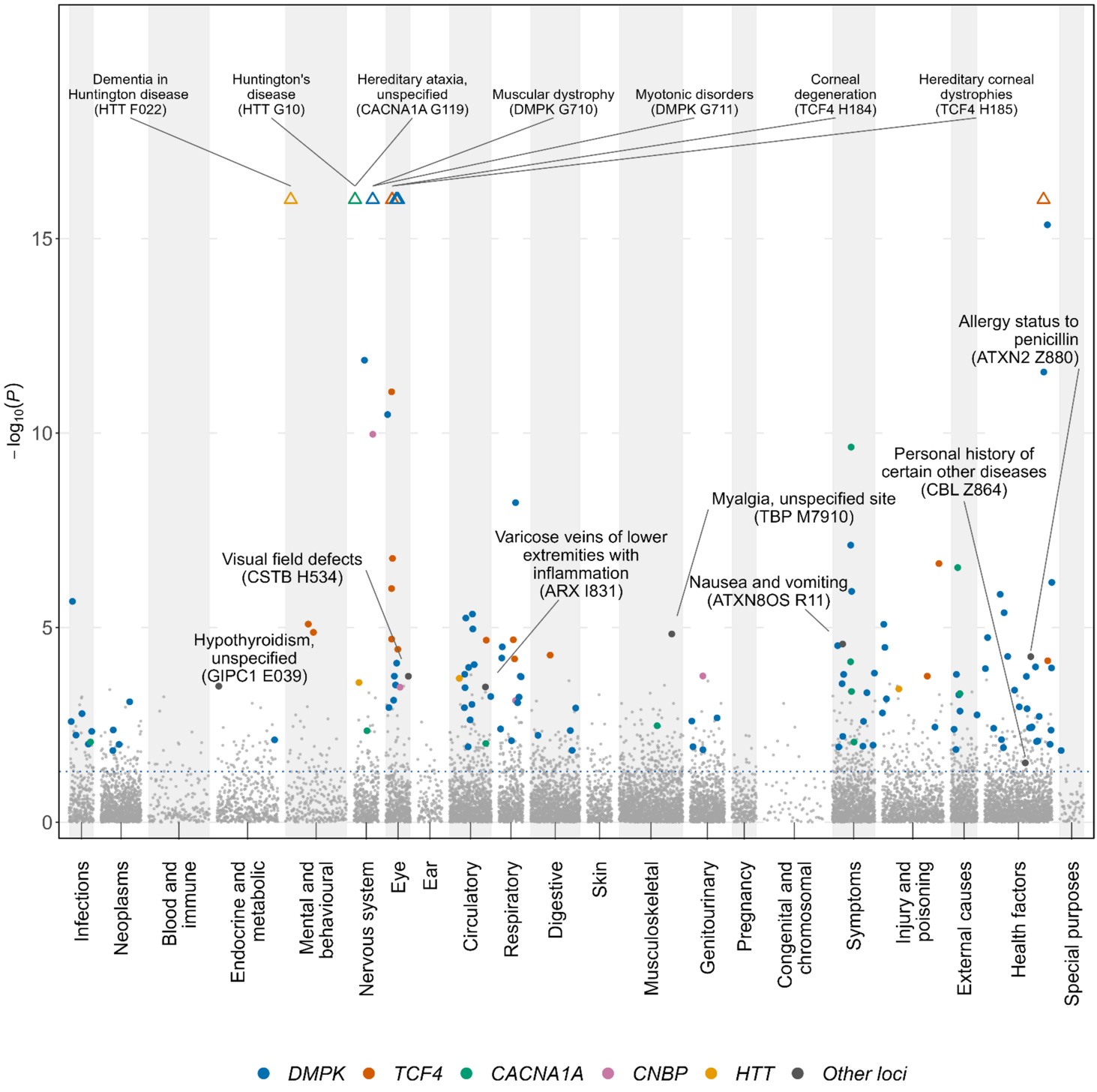


Figure 1

STR PheWAS across 13,191 eligible variant-ICD-10 tests. Points are grouped by ICD-10 chapter; associations meeting within-variant FDR <0.05 are coloured by locus. The dotted line marks nominal P = 0.05. Open triangles denote P values below the limit of double precision, plotted at the axis ceiling. Models adjust for age, sex and GPCs 1 to 10 in 484,561 participants and require five carrier cases.

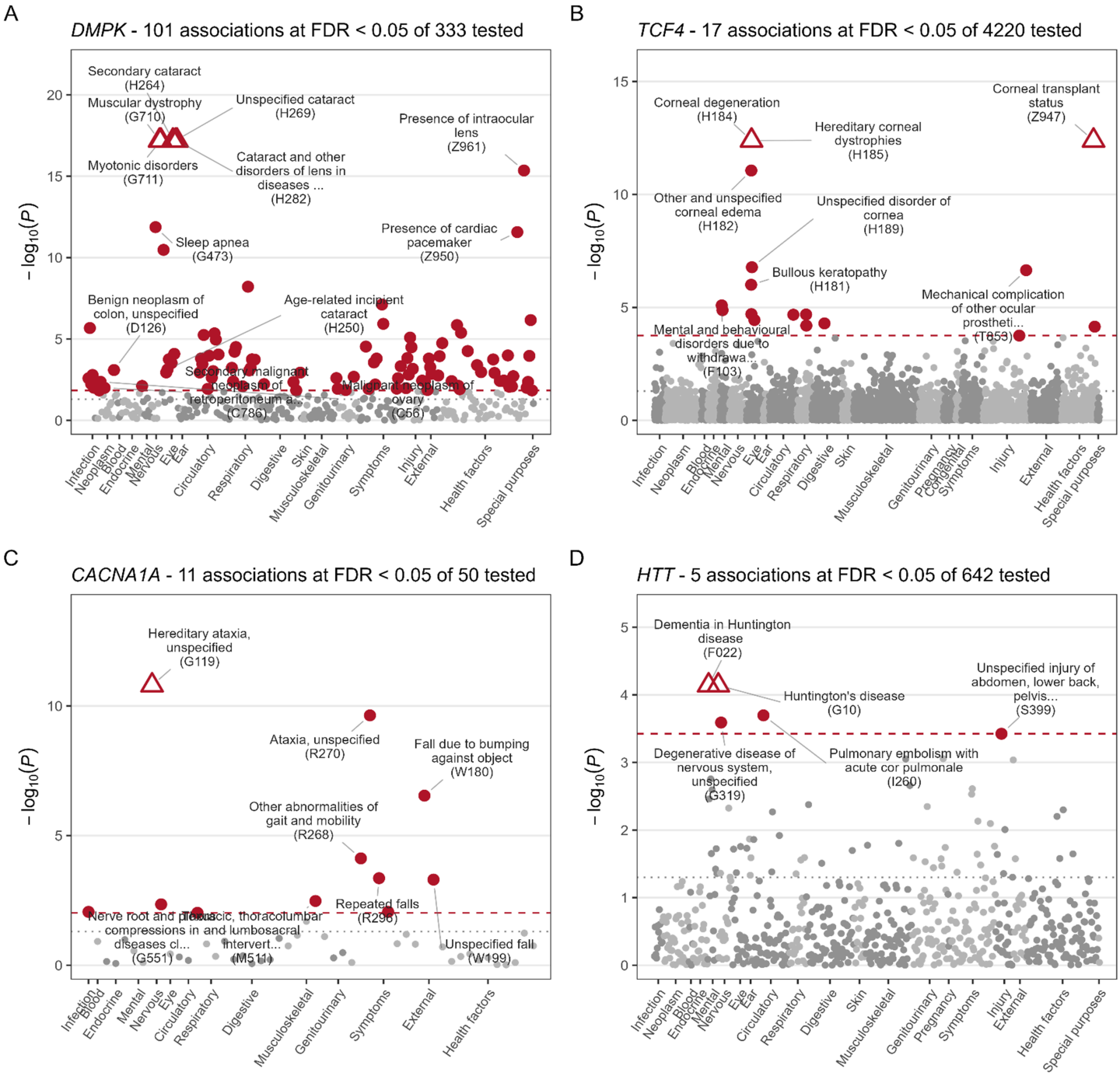


Figure 2

Association spectra at (A) *DMPK*, (B) *TCF4*, (C) *CACNA1A* and (D) *HTT*. Points represent eligible ICD-10 tests grouped by chapter. Red points meet within-variant FDR <0.05; dashed and dotted lines show the FDR and nominal P = 0.05 thresholds. Open triangles denote P values below the limit of double precision, plotted at the axis ceiling. All models adjust for age, sex and GPCs 1 to 10 in 484,561 participants.

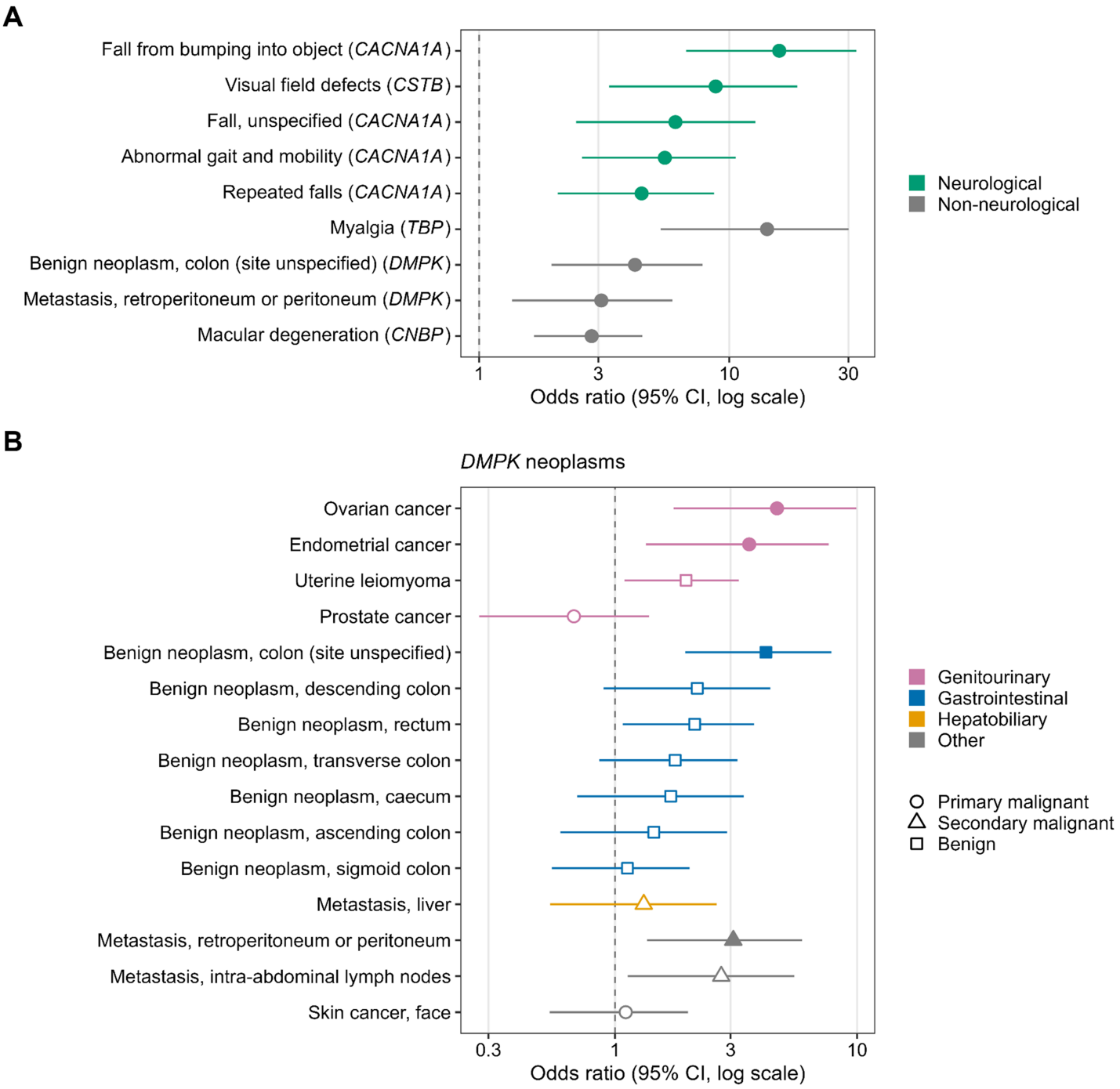


Figure 3

Novel association effect sizes. (A) Selected reported associations, coloured by neurological versus non-neurological phenotype. (B) All eligible *DMPK* neoplasm associations, coloured by organ system (genitourinary, gastrointestinal, hepatobiliary or other); shape marks primary malignant, secondary malignant or benign neoplasms. Filled points mark within-variant FDR <0.05. Panel A labels give the phenotype and locus; ICD-10 codes are in Supplementary Tables S5 and S6. Odds ratios and 95% confidence intervals are plotted on logarithmic scales; dashed lines mark OR = 1. Both panels use age-,

sex- and GPC1-10-adjusted estimates from 484,561 participants. Full results are in Supplementary Tables S5 and S6.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryTablesS1S9.xlsx