A GCC repeat expansion in *AFF3* is a significant cause of intellectual disability

Bharati Jadhav,[1#] Paras Garg,[1#] Joke J. F. A. van Vugt,[2] Kristina Ibanez Garikano[3,4], Delia Gagliardi[3,4], William Lee,[1] Alejandro Martin-Trujillo,[1], Scott L. Gies,[1] Mafalda Barbosa[1], Miten Jain[5,+], Henry Houlden[4], Benedict Paten[5], Genomics England Research Consortium, Project MinE ALS Sequencing Consortium, Jan Veldink,[2] Arianna Tucci[3], Andrew J. Sharp[1,*]

[1] *Department of Genetics and Genomic Sciences and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount, New York, NY 10029*

[2] *Department of Neurology, UMC Utrecht Brain Center, Utrecht University, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands*

[3] *William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK,*

[4] *Department of Neuromuscular Diseases, Institute of Neurology, University College London, London, UK,*

[5] *UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, 95064*

[#] *These authors contributed equally to this work.*

[+] *Present address: Northeastern University, 360 Huntington Ave, Boston, MA 02115*

** Address for correspondence: Andrew J. Sharp, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount, Hess Center for Science and Medicine, 1470 Madison Avenue, Room 8-116, Box 1498, New York, NY 10029 USA. Telephone: +1-212-824-8942, Fax: +1-646-537-8527, Email: andrew.sharp@mssm.edu*

Abstract

GC-rich tandem repeat expansions (TREs) are often associated with DNA methylation, gene silencing and folate-sensitive fragile sites, and underlie several congenital and late-onset disorders. Through a combination of DNA methylation profiling and tandem repeat genotyping, we identified 24 methylated TREs and investigated their effects on human traits using PheWAS in 168,641 individuals from the UK Biobank, identifying 156 significant TRE:trait associations involving 17 different TREs. Of these, a GCC expansion in the promoter of *AFF3* was linked with a 2.4-fold reduced probability of completing secondary education, an effect size comparable to several recurrent pathogenic microdeletions. In a cohort of 6,371 probands with neurodevelopmental problems of suspected genetic etiology, we observed a significant enrichment of *AFF3* expansions compared to controls. With a population prevalence that is at least 5-fold higher than TREs causing fragile X syndrome, *AFF3* expansions represent a significant cause of neurodevelopmental delay in humans.

Keywords: Tandem repeat, PheWAS, Fragile site, Repeat expansion, Epimutation

INTRODUCTION

Tandem repeat expansions (TREs) are a class of recurrent mutation in which a tandemly repeated DNA motif, e.g., poly(CGG), undergoes expansion of repeat number to unusually large size. Expansions of tandem repeats (TRs) were first recognized as a mechanism of genetic disease in the early 1990s[1–3], and since this initial discovery, more than 50 inherited disorders, almost all of which are dominant in nature, are now known to be caused by TREs[4]. One class of TRE that has been linked with a variety of both congenital and late-onset disorders are composed of GC-rich TR motifs typically located within gene promoters or 5'UTRs[5]. The best exemplar of these is the CGG TRE within the 5'UTR of *FMR1* that causes fragile X syndrome, one of the most common genetic causes of intellectual disability and autism. Although this repeat is highly polymorphic in the normal population, in individuals with fragile X syndrome this TR expands to >200 copies, becoming highly methylated, accompanied by epigenetic silencing of *FMR1* and the presence of a cytogenetically-visible chromosomal fragile site at Xq28 after which the syndrome is named[6]. In addition, intermediate between common polymorphic length variation of the TR and the much rarer, methylated "full mutation" expansions that cause fragile X syndrome, unmethylated moderate-sized expansions (55-200 copies) of the *FMR1* CGG TR are frequently identified in relatives of probands with fragile X syndrome. Termed "premutations" due to their propensity to further expand upon meiotic transmission, these shorter expanded alleles drive increased expression of *FMR1* and are now recognized as a cause of two other disorders, namely fragile X premature ovarian insufficiency (FXPOI) and fragile X tremor ataxia syndrome (FXTAS) that occur in ~5-30% of carriers[7]. Since the discovery of the molecular basis of *FMR1*-related disorders, other similar promoter-associated GC-rich TREs have been implicated as causal variants underlying a variety of conditions[8–17], raising the possibility that the human genome may harbor additional similar TREs that contribute to a range of pathologies.

Previously, we showed that the use of DNA methylation profiling represents a powerful method to identify rare methylated TREs[18]. Here, we applied this approach to analyze DNA methylation profiles from 5,750 individuals for whom PCR-free Illumina genome sequencing (GS) data were available, identifying 24 GC-rich promoter-associated TRs that become methylated when expanded. Given that expansions of similar GC-rich repeats have previously been implicated in multiple human disorders, we hypothesized that these

TREs represent strong candidates for modifiers of human disease risk. To investigate this, we used GS data to directly genotype these loci and performed a phenome-wide association study for TREs in 168,641 individuals from the UK Biobank, identifying numerous associations with human traits. Based on our findings, we further investigated the role of TREs within *AFF3* as a contributor to neurodevelopmental disorders in the UK 100,000 Genomes cohort, confirming this TRE as a clinically significant cause of intellectual disability.

RESULTS

Identification of repeat expansions associated with local hypermethylation

Profiling the methylomes of 5,750 samples, we identified a total of 2,124 differentially methylated regions (DMRs) (Supplementary Tables 1 and 2). Using ExpansionHunter Denovo to analyze available GS data, we performed genome-wide screening for TREs in these individuals, identifying 34 candidate TREs that overlapped hypermethylated DMRs. We then utilized ExpansionHunter to genotype each candidate TR that overlapped a hypermethylated DMR. Where expanded repeats exceed the sequencing read length, ExpansionHunter generates estimated allele sizes based on indirect metrics, such as the number of in-repeat reads. For GC-rich TREs that often form strong secondary structures, this can lead to uncertainty in the estimated genotypes and difficulty in distinguishing between unmethylated premutation alleles and methylated full mutations[19,20]. Despite this limitation, we observed many TR loci where individuals with local hypermethylation tended to have the longest TR alleles at that locus in the entire cohort. Using permutation analysis, 24 of the 34 loci tested showed a significant co-occurrence of hypermethylation and unusually long TR alleles in the same individuals (Figure 1), thus providing strong statistical evidence that the observed hypermethylation at each of these loci resulted from expansion of the underlying TR (Supplementary Table 3). At some of these 24 loci, a few individuals with hypermethylation were genotyped by ExpansionHunter as having two TR alleles within the normal range (Figure 1). We postulate that these may represent either a failure of ExpansionHunter to identify TREs (false negatives), or alternatively hypermethylation of the locus for reasons unrelated to the TR. In support of the former hypothesis, examination of read alignments[21] in these

individuals showed a small number of reads composed of long TR tracts, providing suggestive evidence of TREs that were missed by ExpansionHunter (Supplementary Figure 1). At the remaining 10 of 34 loci tested, permutation testing showed no association between the presence of hypermethylation and long TR alleles, and these were excluded from further analysis.

As expected, the 24 loci identified included many TRs that have previously been reported or suspected to undergo hypermethylation when expanded[12,13,15,18,22–28], including several that are associated with human disease. However, we also identified several methylated TREs that have not been previously described (Supplementary Table 3). In addition to these, we also observed methylated GGC expansions within the 5'UTR of *GIPC1*, a locus where unmethylated and, thus presumably shorter TREs have recently been reported to cause oculopharyngodistal myopathy 2 (OPDM2)[14], and methylated expansions of a GGCCTG repeat intronic within *NOP56* that cause SCA36[29], thus implicating epigenetic effects of this TRE. While the majority of these methylated TREs were triplet repeats (e.g., CGG or GGC motifs), a few were composed of longer TR motifs, including two 16mers.

Confirmation of repeat expansions by long-read sequencing

We generated long-read GS to confirm the presence of three previously unreported TREs in samples that showed long TR alleles as genotyped by ExpansionHunter. PacBio GS confirmed expansions of a 16mer located in the 5' UTR of *DPYSL4* (chr10:132,186,843-132,186,915), which consisted of 51-55 copies of the consensus motif CCGGGGGCGGGGCCTG, and an expansion located within the 5' UTR of *FAM193B* (chr5:177,554,490-177,554,531) consisting of 1,066-1,436 copies of a CCG repeat.

Using nanopore long-read GS, a single spanning read confirmed a TRE consisting of 473 copies of a GCC repeat located in the bidirectional promoter/5' UTR region of *WBP4* and *ELF1*. Multiple other reads from the expanded allele produced very low quality sequence upon traversing the TR, which appeared to occur in a unidirectional fashion, a phenomenon we also observed with PacBio GS at *DPYSL4* and previously at other methylated GC-rich TREs that presumably have strong secondary structures that inhibit sequencing[18]. Despite this, we performed phasing of nanopore reads based on presence/absence of the TRE-containing haplotype

and then analyzed the signal profiles, demonstrating that the expanded TR allele was highly methylated, while the normal TR allele was largely unmethylated (Figure 2).

Notably, although ExpansionHunter correctly identified individuals carrying unusually long TR alleles at all three of these loci, the estimated TR lengths provided by ExpansionHunter were considerably shorter than the alleles observed directly with spanning long reads.

Gene expression analysis of TREs in the GTEx cohort

To assess the functional effects of methylated TREs on gene expression, we utilized DNA methylation and RNAseq data generated by the GTEx project[30,31]. We identified 10 individuals with hypermethylation overlapping TREs at *ADARB1*, *C11orf80*, *CSNK1E*, *FRA10AC1*, *PCMTD2* and *ZNF713* (Supplementary Figure 2). Consistent with the promoter methylation associated with these TREs, these individuals showed highly biased allelic expression (Supplementary Figure 3), and, in all but one case, this was accompanied by unusually low expression of the associated gene, thus indicating that these TREs result in allelic silencing in cis. The sole exception was an individual with moderately increased methylation of *ZNF713* who exhibited unusually high expression of this gene across nearly all tissues (Supplementary Figure 4).

Phenome-wide association study of repeat expansions in the UK Biobank

As multiple GC-rich TREs are known to underlie a variety of human disorders[1,8–17], we next sought to identify traits associated with the presence of GC-rich TREs through phenome-wide association studies (PheWAS). We used ExpansionHunter to analyze Illumina GS data from 168,641 unrelated individuals of European ancestry from the UK Biobank, generating genotypes for 28 TRs, representing the 24 TREs associated with hypermethylation identified here in addition to four other GC-rich TRs from the literature where rare methylated TREs have been reported (Supplementary Table 3). In order to optimize our PheWAS pipeline to detect associations with expanded repeat alleles where the pathogenic threshold might vary widely among different loci, we performed outlier analysis using six different stringencies at each TR locus (Supplementary Table 4, Supplementary Figure 5) to identify sets of unusually long TR alleles representing putative TREs (Supplementary Table 5) and performed association analysis of expanded repeat alleles with 10,615 traits.

We identified 156 significant unique TRE:trait pairwise associations involving 17 different TREs at a threshold of 10% FDR (Supplementary Tables 6 and 7), indicating diverse negative effects of these TREs on human health. The most significant and numerous of these associations (n=93) occurred with expansions of the GGGGCC repeat within *C9orf72* that is known as a major cause of amyotrophic lateral sclerosis (ALS)[12,13]. We identified 350 individuals with a *C9orf72* TR allele >30 copies, indicating a prevalence of potentially pathogenic *C9orf72* TREs of 1 per 482 individuals (0.21%), in agreement with two prior studies that reported frequencies of 0.15%[32] and 0.19%[20]. Based on the frequency of this mutation in ALS[33], this carrier frequency is significantly above expectation and suggests pleiotropy and/or reduced penetrance, similar to a recent finding in Kennedy's disease[34]. Although UK Biobank data does not include a specific phecode for ALS, all of the observed traits associated with *C9orf72* expansions were consistent with the known symptoms, effects and treatments of ALS, including multiple measures of reduced brain volume, increased frequencies of depression, dysphagia, urinary tract infections, imaging studies of the central nervous system, and higher death rates attributed to motor neuron disease and respiratory failure[35] (Figure 3, Supplementary Table 6).

In addition to these known pathologies, we also identified many other significant associations with TREs. For example, a GCC expansion in the promoter of *AFF3*, which was previously suggested as potentially linked with neurodevelopmental problems[23], showed a strong association with reduced educational attainment, with a 2.1-fold lower probability of obtaining O-levels (exams normally taken at age 16) ($p=9.4x10^{-11}$) and a 2.4-fold lower probability of obtaining a university degree ($p=9.6x10^{-11}$). Individuals with this *AFF3* TRE also showed significantly reduced height, lung function and physical activity, showed altered alcohol use as adults, had lower fluid intelligence and word interpolation abilities, and reported worse overall health compared to controls, indicating negative effects of this expansion on multiple physical and mental traits.

Similarly, a GGC TRE in the 5'UTR of *DIP2B*, which is known to cause silencing of *DIP2B*, the FRA12A folate-sensitive fragile site and was previously identified in several individuals with neurocognitive problems[24], was associated with significantly reduced height, lung capacity, vitamin E levels, altered alcohol use and changes in the microstructure of brain white matter from diffusion MRI, accompanied by a nominal association (p=0.0002) for lower age of completing full time education.

Other notable signals we identified (Figure 3, Supplementary Figure 6) include:

(i) Expansions of a CGG repeat in the 5'UTR of *CBL* associated with a significant increase in blood platelet levels ($p=1.1x10^{-14}$). Consistent with the known link between coding variants in *CBL* and juvenile myelomonocytic leukemia[36], we also observed that carriers of this TRE showed a nominal association with chemotherapy treatment for neoplasms (p=0.0005).

(ii) Expansions of a GCC repeat in the 5'UTR of *LINGO3* associated with reduced volume of gray matter in multiple regions of the cerebellum (peak $p=8.8x10^{-9}$), suggesting potential links of this TRE with neurodegeneration.

(iii) Expansions of a GCC repeat in the 5'UTR of *ADARB1* associated with recent changes in speed/amount of moving or speaking ($p=3.3x10^{-7}$), suggesting potential links with depression or late-onset neurological disease.

(iv) Expansions of a GCY repeat in the 5'UTR of *TBC1D7* associated with significantly reduced performance on the numeric addition test ($p=9.2x10^{-6}$) and altered microstructure of brain white matter from diffusion MRI ($p=2.1x10^{-5}$).

(v) Expansions of a CGC repeat in the 5'UTR of *CSNK1E* associated with reduced alcohol intake, increased body fat, and increased incidence of diseases of the circulatory system.

(vi) Expansions of a CGC repeat in the 5'UTR of *BCL2L11* associated with blood eosinophil and neutrophil levels.

Underlying data distributions for significant associations are shown in Supplementary Figures 7-15.

AFF3 expansions are enriched in unsolved probands with intellectual disability in The 100,000 Genomes Project

The 100,000 Genomes Project (100kGP)[37] recruited 90,190 individuals with a range of rare diseases of unknown etiology or cancer, including 7,871 probands with a variety of neurodevelopmental disorders (https://re-docs.genomicsengland.co.uk/current_release/). Illumina GS was performed on each individual and, as of January 31st 2023, pathogenic or likely pathogenic mutations had been identified in 19% of those recruited. Given the observed association in UKB between a TRE in *AFF3* and educational attainment, we hypothesized that TREs of *AFF3* might represent the pathogenic mutation in some patients with neurodevelopmental disorders. We genotyped the GCC TR in *AFF3* in this cohort and compared the frequency

of long alleles in the 6,371 unsolved cases with a primary referral reason of intellectual disability to 8,794 controls, approximately matched for ancestry. While the distributions of *AFF3* repeat alleles in the common polymorphic range (99% of control alleles were ≤38 copies) were indistinguishable between cases and controls, we observed a significant excess of long *AFF3* alleles in probands in whom no other pathogenic mutation had been identified from GS (3.9-fold enrichment for alleles ≥61 copies, p=0.002, Two-tailed Fisher's exact test) (Figure 4), indicating this TRE as the likely pathogenic mutation in these individuals (Supplementary Figure 16, Supplementary Table 8). As a further indicator of causality, the enrichment of *AFF3* expansions in idiopathic neurodevelopmental disorders occurred exclusively in probands where no other presumptively pathogenic mutation had been identified. Considering expanded *AFF3* alleles as those genotyped by ExpansionHunter as ≥60 copies of GGC (Figure 1), 17 of 6,371 unsolved probands had *AFF3* expansions, compared to none of the 1,500 probands in whom other pathogenic mutations had been identified by GS (p=0.056, Two-tailed Fisher's exact test). Thus, our observations indicate that GCC expansions of *AFF3* represent the causal genetic lesion in ~0.3% of patients with neurodevelopmental disorders, although this is potentially an underestimate due to the limitations of accurately sizing long GC-rich TR alleles from Illumina GS data[19,20].

Based on data provided by the referring clinical centers, in addition to the primary referral reason of intellectual disability/global developmental delay, probands with *AFF3* expansions presented with a variety of other recurrent phenotypes. Of these, speech and motor delays were observed in 59%, seizures in 41%, behavioral disturbances in 35%, generalized hypotonia in 24%, and a variety of dysmorphic features and congenital anomalies at lower frequencies (Supplementary Table 9).

Relative impact and modifiers of AFF3 expansions on educational attainment

To assess the relative effect size of the *AFF3* TRE on educational attainment in comparison to other mutations that impact neurodevelopment, we used read depth from GS to identify several recurrent microdeletion/duplications in the set of 168,641 UK Biobank individuals used for PheWAS (Supplementary Figure 17). Among these, expansions of *AFF3* exert a negative effect on educational attainment that has an effect size comparable to that seen in carriers of del(22q11.21), del(1q21.1) and del(15q13.3) (Figure 5).

Given the apparent phenotypic heterogeneity among individuals with *AFF3* expansions, with some obtaining university degrees, we investigated possible modifiers of penetrance. Using a large set of SNVs identified from GWAS[38], we calculated a PRS for educational attainment in UK Biobank participants. Among individuals with *AFF3* expansions, we observed significantly higher PRS in those who obtained A-levels (exams normally taken at age 18) compared to those that did not (p=0.02, Two-tailed unpaired t-test) (Supplementary Figure 18). Partitioning individuals with *AFF3* expansions based on quintiles of PRS showed a clear influence of genetic background on educational attainment, with individuals in the upper quintile of PRS showing an ~2-fold higher likelihood of obtaining A-levels compared to those in the bottom quintile (Supplementary Figure 19). Similar results were also observed for several recurrent microdeletions/duplications. These observations strongly suggest that the negative effects of *AFF3* expansions and other recurrent CNVs on educational attainment are modified by polygenic background.

Population prevalence of methylated AFF3 expansions

As it can be difficult to accurately estimate the true size of expanded GC-rich TRs from Illumina GS, in order to provide a more accurate estimate of the frequency of methylated *AFF3* expansions that cause the FRA2A fragile site, we utilized DNA methylation data from 32,776 individuals profiled using the Illumina 450k and 850k arrays[18,39], identifying 38 individuals with hypermethylation of the *AFF3* locus (Supplementary Figure 20). Thus, presuming these hypermethylation events correspond to the presence of an underlying TRE, we estimate the population prevalence of methylated *AFF3* expansions as ~1 per 862 individuals (95% confidence interval, 1 per 654 to 1,266).

SNVs that associate with presence of the AFF3 repeat expansion match those found by GWAS

Multiple prior GWAS have reported SNVs at the *AFF3* locus that show significant associations with intelligence, cognitive ability and educational attainment[40]. Based on the observation that, at some TRE loci, local SNVs show significant associations with the presence of a TRE[41–47], we hypothesized that the GCC TRE in *AFF3* might be the causal variant that underlies these GWAS signals. Consistent with this, we identified hundreds of SNVs located within the *AFF3* locus that show strong association with presence of an expanded

GCC TR allele and observed that many of these are the same SNVs that have been previously identified by GWAS for intelligence, cognitive ability and educational attainment (Figure 6). These observations provide further evidence for the role of the *AFF3* TRE in neurodevelopment and indicate that a subset of GWAS signals are likely driven by underlying TR variants that preferentially occur on specific founder haplotypes.

DISCUSSION

In 1991, the recognition of a methylated repeat expansion in *FMR1* as the causal mutation underlying fragile X syndrome represented a new paradigm for the genetic basis of human disease[2,48]. However, the inherent technical difficulties of genotyping long TR alleles has meant that progress in elucidating the wider role of TREs in human traits over the past three decades has been relatively modest[4]. Here, we performed a three-part study in which we first used a combination of epigenome profiling and TR genotyping to identify 24 methylated TREs in the human genome which resemble the TRE that causes fragile X syndrome. We then performed a population-scale PheWAS to identify the effects of these TREs on diverse human traits, and finally tested their contribution to rare disease in a cohort of patients who remained refractory to genetic diagnosis despite undergoing GS. In addition to several that were previously described, we identified multiple methylated TREs that had not been previously recognized and found that many of these exert diverse effects on human traits.

Of particular note, analysis in the UKB showed that expansions of a GCC motif within *AFF3* have a strong negative effect on educational attainment, cognitive function, and multiple measures of general health. Comparison with other recurrent CNVs showed that the effect of these *AFF3* expansions on educational attainment are similar in magnitude to those caused by several clinically significant microdeletions that are a frequent cause of neurodevelopmental anomalies, such as del(22q11.2)[49], del(1q21.1)[50,51] and del(15q13.3)[52,53]. Furthermore, we showed that the phenotypic impact in individuals with the *AFF3* TRE and other recurrent CNVs is modified by polygenic background, thus giving insight into the variable penetrance associated with these events. It should be noted that the UKB recruited individuals from an apparently healthy population, creating an ascertainment bias that will lead to underestimates in the prevalence of more severe

neurocognitive deficits and morbidities that may be associated with a mutation. As such, it is likely that our results from the UKB underestimate the true impact of this TRE on cognitive function. In contrast, patients recruited by the 100kGP represent individuals that had been referred to clinical genetics centers but in whom prior genetic testing had failed to identify any causal mutations and thus represent a sample of individuals that tend to exhibit more severe developmental anomalies. Despite this, results from these two populations converged, showing that these *AFF3* TREs represent a significant cause of developmental delay. We estimate the prevalence of methylated *AFF3* expansions as ~1 per 850 individuals in the general population, a frequency that is 5- to 10-fold higher than the estimated prevalence of CGG expansions within *FMR1* that cause fragile X syndrome[1,54], and 3- to 8-fold higher than recurrent microdeletions/duplications of 1q21.1, 15q13.3 and 16p11.2[55]. We, therefore, propose that the *AFF3* TRE likely represents a relatively common cause of developmental anomalies and recommend that testing for this TRE should be considered in individuals with unexplained developmental delay.

Genotypes of GC-rich TREs exceeding the read length of Illumina GS data provided by ExpansionHunter are estimates that often show considerable inaccuracy[19,20], which creates several limitations: (i) Using genotypes from Illumina GS alone, it is difficult to distinguish large expansions from shorter premutation alleles, or to accurately estimate the prevalence of the TREs we studied. (ii) As such, our analysis potentially underestimates the true prevalence of *AFF3* expansions as a cause of intellectual disability. (iii) Where an expanded *AFF3* allele was also detected in one parent of a proband, we were unable to reliably distinguish whether these were premutation alleles that underwent further expansion into the pathogenic range upon transmission, or discern whether there was any parental bias in transmission to affected offspring[56]. Studies of larger numbers of pedigrees using more accurate genotyping methods will be needed to examine these issues.

The original report that identified a recurrent TRE at *AFF3* underlying the FRA2A fragile site showed that these expansions also caused methylation and allelic silencing of *AFF3*[23]. Our data support the notion that methylated and transcriptionally silent full mutations of *AFF3*, rather than shorter unmethylated premutation alleles, are causal for intellectual disability. However, while this potentially implicates haploinsufficiency of *AFF3* as the underlying pathogenic mechanism, current evidence is unclear. Missense mutations of *AFF3* have

been identified in KINSSHIP syndrome[57], which includes a spectrum of anomalies not observed in carriers of the *AFF3* GCC expansion. However, variants causing KINSSHIP syndrome all occur within a single protein domain thought to be important for the interaction of AFF3 with its binding partners, suggesting the syndrome results from a specific dysregulation of the AFF3 pathway. As the GCC repeat expansion in *AFF3* is intronic, an alternative possibility is that expanded alleles may exert pathology through mechanisms such as RAN-translation[58,59]. Further work will be needed to better elucidate the underlying pathological mechanisms of this TRE.

Although we utilized data from >168,000 deeply phenotyped individuals in our PheWAS, statistical power to detect associations was relatively limited for many of the loci and phenotypes we tested as most TREs and many of the traits we studied are rare. Despite this, it is notable that a high proportion of the TREs tested here (17 of 28) exhibit associations with human traits. In an era where SNV-based GWAS are one of the most common methods utilized for investigating the genetic basis of human disease, our study highlights the role of more complex recurrent variants that are poorly assayed by GWAS as potential contributors to the "missing heritability" of the human genome[60,61]. Emphasizing this, by directly studying TRs, we were able to identify the TRE at *AFF3* as showing a much stronger effect on educational attainment than nearby SNVs reported in prior GWAS. We anticipate that the increasing availability of large Biobanks with both GS and extensive phenotype data will enable many novel insights into the role of TREs in human disease and variation.

ONLINE METHODS

Quality control and data processing of methylation data

For the identification of epivariations, we utilized methylation data from 8,371 individuals from the Project MinE cohort[39]. Informed consent for genetic research was obtained from all participants, approved by the Trent Research Ethics Committee 08/H0405/60. DNA methylation profiles were generated using either the Illumina HumanMethylation450 BeadChip (henceforth called “450k array”) (n=4,474) or the Illumina Infinium MethylationEPIC BeadChip (henceforth called “850k array”) (n=3,897), using DNA extracted from peripheral whole blood. We removed any sample with >1% of autosomal probes with detection p-value >0.01 and performed Principal Component Analysis (PCA) based on β-values of all probes located on chr1, removing outlier samples. Data were then normalized, as described previously[62,63]. Briefly, raw signal intensities were subjected to color correction, background correction and quantile normalization using the Lumi package in R[64], and the normalized intensities converted into β-values, which range between 0-1, representing the methylation ratio at each measured CpG. In order to correct for inherent differences in the distribution of β-values reported by Infinium I and Infinium II probes, we applied BMIQ[65]. Data from 450k and 850k arrays were normalized independently. After normalization, we estimated the major cellular fractions comprising each blood sample directly from β-values using the method described by Houseman[66] and removed outlier samples, defined as those that showed cellular fractions either ≥99th percentile +2%, or ≤1st percentile -2% of any cell type. After all quality control and filtering steps, 7,447 samples were processed to identify epivariations, comprising 5,403 individuals with a diagnosis of amyotrophic lateral sclerosis (ALS) and 2,044 controls. Of this total, 5,750 also had available GS data.

Identification of rare epigenetic variants

In order to identify rare epigenetic variants, also termed differentially methylated regions (DMRs), we utilized a sliding window approach to compare individual methylation profiles of a single sample against all other samples. This process was performed separately using data from the 450k and 850k array platforms. We defined DMRs as regions of outlier methylation represented by multiple independent probes using the following parameters:

Hypermethylated DMRs: Any 1 kb region with at least three or more probes with β-values ≥99.5th percentile plus 0.15, and contains at least three consecutive probes with β-values ≥99.5th percentile. In addition, we required that the minimum distance spanned by probes that were ≥99.5th percentile was ≥100 bp.

Hypomethylated DMRs: Any 1 kb region with at least three or more probes with β-values ≤0.5th percentile minus 0.15, and contains at least three consecutive probes with β-values ≤0.5th percentile. In addition, we required that the minimum distance spanned by probes that were ≤0.5th percentile was ≥100 bp.

As the presence of an underlying homozygous deletion at a probe binding site can result in spurious β-values[62], we removed any DMR call in which the carrier individual reported one or more probes within the DMR with failed detection p-value (p>0.01). Finally, we removed samples with an unusually high number (n>20) of autosomal DMRs.

For the GCC repeat within *AFF3*, as only two probes on the 450k and 850k arrays show a clear increase of methylation levels in individuals with an expanded repeat allele, this locus did not meet our stringent genome-wide discovery criteria for identifying DMRs. Therefore, for this locus, we considered hypermethylation events as those where the normalized β values of probes cg14267725 and cg23118464 were both >99.8th percentile of their respective population distributions.

For the GGGGCC repeat within *C9orf72*, as the cohort comprised a large fraction of samples with ALS and, thus, is highly enriched for carriers of expanded *C9orf72* TR alleles, we considered hypermethylation events at this locus as those where the normalized β values of at least two probes within 1kb of the TR were each >99th percentile of their respective population distributions.

Identification of tandem repeat expansions in the Project MinE cohort

In order to identify putative TREs potentially linked with rare hypermethylation events, we performed two analyses. First, we utilized ExpansionHunter DeNovo[67] to identify genomic loci containing TR alleles ≥150 bp in length. The regions output from ExpansionHunter DeNovo were overlapped with DMRs identified by methylation profiling (extended by ±500 bp), and where these regions intersected within the same sample, they were taken forward for targeted genotyping using ExpansionHunter (v3.2).

In addition, we utilized a set of highly polymorphic TRs that showed ≥10 different alleles in the population based on profiling with HipSTR[68]. Based on the observation that a high rate of polymorphism in the population is a very strong predictor of unstable TRs that also undergo occasional extreme expansions[18,69], we reasoned these would be strongly enriched for possible TREs. These putatively unstable TRs were intersected with hypermethylated DMRs identified by methylation profiling (extended by ±500bp), and any overlapping TRs were then genotyped using ExpansionHunter. To identify TREs that associate with local DNA methylation, at each locus we performed 100 million permutations to assess the probability of observed distribution of methylated TR alleles. In each permutation, we randomly assigned hypermethylated samples and compared the mean TR length of individuals with hypermethylation with the observed mean, generating an estimated p-value based on the fraction of simulations where the permuted mean exceeded the observed value.

Long read sequencing of putative tandem repeat expansions

Pacific Biosciences long insert libraries with the addition of barcodes were prepared for samples with epivariations at *DPYSL4* and *FAM193B*, the two samples mixed at equimolar amounts, and sequenced on a single 8M SMRT cell with the Pacific Biosciences Sequel II system in continuous long read (CLR) mode. Subreads were aligned to the hg38 human reference genome using pbmm2 v1.0.0[33] with default parameters.

We sequenced a sample with hypermethylation of the *ELF1* promoter using Oxford Nanopore Technology (ONT) to a mean depth of 12x coverage. Reads were mapped to the hg38 human reference genome using minimap2 (v2.7)[70], and the bam file sorted and indexed using samtools (v1.7)[71]. To estimate methylation levels on normal and expanded CGG repeat alleles separately, we first phased the reads using SNVs within the reads. Using nanopolish (v0.10.2)[72], we created index files to link reads with their signal level data in FAST5 files, followed by estimation of DNA methylation status at each CpG located within 2 kb of CGG TRs, requiring a minimum log likelihood ratio ≥2.5 at each site.

Expression studies of genes with methylated repeat expansions in the GTEx cohort

We utilized published DNA methylation data generated using the Illumina 850k array from 9 tissues and 424 individuals in the The Genotype-Tissue Expression (GTEx) cohort[31] to first identify outlier hypermethylation

events at the 24 methylated TREs that were identified in Project MinE samples. We then utilized corresponding RNAseq data from 948 individuals published on the GTEX portal (Version 9, https://www.gtexportal.org/) to compare normalized expression levels (TPM) and WASP-corrected allele-specific expression (ASE) data of the relevant gene in carriers of each TRE to other tissue-matched samples in the GTEx cohort.

Identification of tandem repeat expansions in the UK Biobank (UKB)

Collection of the UKB data was approved by the Research Ethics Committee of the UKB obtained under application 32568. All study participants provided informed consent, and the protocols for UKB are overseen by The UK Biobank Ethics Advisory Committee, see https://www.ukbiobank.ac.uk/ethics/. From the set of 200,000 individuals with Illumina 150 bp paired-end whole GS data, we defined 188,915 individuals of European ancestry using an approach based on principal component analysis of a set of ~64,000 high quality linkage disequilibrium (LD) pruned SNVs with minor allele frequency (MAF) >5%, as described previously[73]. In brief, GCTA (v1.93) was used to calculate the first 40 Principal Components (PCs), which were used to predict the ancestry of each individual. Since only a small minority of individuals recruited to the UK biobank cohort are of non-European ancestry and are, therefore, highly under-powered in PheWAS, we focused our analysis on those individuals of European origin. We first selected all samples predicted as European ancestry and recalculated PCs without use of a reference population using one third of individuals, before projecting the remaining two thirds of individuals onto this. We removed outlier samples using within group PCA. In addition, we kept only those samples with self-reported ancestry as "White", "British","Irish","Not listed in table", "Any other white background", "Other ethnic group", or "Prefer not to answer". Further sample-level filters were applied as follows:

1. Using pairwise kinship coefficients provided by the UK Biobank, for samples with 2nd degree relationships or higher (kinship coefficient >0.0883) we retained only a single unrelated individual.
2. We removed 199 individuals with predicted sex chromosome aneuploidy.
3. We removed 241 individuals who we were notified had withdrawn their consent for inclusion in the UK Biobank study.

After these quality control (QC) steps, we retained data for 168,641 unrelated individuals of European ancestry. Using the UK Biobank DNAnexus Research Analysis Platform, we performed genotyping of 28 GC-rich TRs from the GS data with ExpansionHunter (v5)[74] (Supplementary Table 10). From the resulting diploid genotypes output by ExpansionHunter, we utilized data only for the longer of the two alleles per sample at each locus, discarding data from the shorter allele. To avoid artifacts due to possible technical effects on genotypes, samples were then separated by sequencing center (deCODE and Sanger Center), and, for loci on the X chromosome, samples were further separated by sex. Based on the set of alleles observed at each TR locus per sample group, we calculated population statistics and used these for the determination of putative expanded alleles using six levels of stringency (named Very low, Low, Medium, High, Very High and Extremely High) each of which utilized the absolute and percentile allele size, difference from the population median, and outlier probability based on the z-score (Supplementary Table 4). An example of the output of these different stringencies of outlier calling are shown in Supplementary Figure 5.

To reduce false positive calls for TREs that often result from low read depth or poor-quality alignments of Illumina reads over TR regions[75], we developed a Random Forest classifier to recognize low-confidence TRE calls based on characteristics of the TR locus and the output metrics supporting each genotype generated by ExpansionHunter. We first trained this tool on a set of ~3,500 putative TREs from ExpansionHunter for which we performed manual curation of read pileup plots to determine true and false positives using an 80:20 split and 10-fold cross validation. Testing of the trained model indicated 93% precision and a 91% true positive rate for TREs. For each putative TRE, we generated a visualization of reads overlapping the TR locus using GraphAlignmentViewer (https://github.com/Illumina/GraphAlignmentViewer) and by manual inspection classified these as either true or false based on the number and quality of the read alignments supporting the putative TRE. In general, TREs considered as true were supported by multiple reads with few mismatches, while those considered as false were often supported by one or very few reads with multiple mismatches in the alignment (Supplementary Figure 21). We subsequently applied the trained classifier to screen all putative TREs called using each of the six outlier criteria, and any genotypes that were scored as false positive TREs were removed from further analysis. Following this automated quality filtering step, we defined each individual with a TRE that passed QC as “expanded”, and defined all other individuals as “not expanded”.

PheWAS analysis in the UK Biobank

In order to identify traits associated with the presence of TREs, we input the set of filtered binary genotype calls from the six stringencies used for identifying TREs ("expanded" versus "not expanded") into association analysis. We utilized phenotype data for individuals in the UK Biobank derived from a total of 10,615 ICD10 codes and quantitative, categorical and binary traits (Supplementary Table 11), accessed through UK Biobank application number 82094. Before performing association testing, phenotype data were processed as follows:

1. For phenotypes with multiple categories, e.g. ICD10 codes, we considered each category as a separate binary trait. If a subject had multiple entries for the same category, they were assigned as "1" (i.e. positive for that trait) if any of the instances were positive. We removed from analysis any categories labeled as "Prefer not to answer", "Unsure" or "Do not know".
2. ICD9 and ICD10 codes were merged into one using General Equivalence Mapping (GEM) files downloaded from https://www.cms.gov/Medicare/Coding/ICD10. We restricted controls to subjects who did not have a positive diagnosis for that phenotype.
3. In addition to using each individual ICD code as a separate trait, ICD codes were also grouped into higher level codes to yield additional summary traits with increased sample size. For example, the ICD10 code for Schizophrenia is F20, which is composed of 10 different sub-codes (e.g. F20.0 "Paranoid schizophrenia:, F20.2 "Catatonic schizophrenia", F20.9 "Schizophrenia, unspecified", etc), with each subcategory containing between 4 and 894 individuals. Here, we created a new summary code for "Schizophrenia_group" comprising all sub-categories under F20 which included 1,485 individuals who were positive for any subtype of schizophrenia. Similar groupings were performed for any higher level ICD code that comprised multiple subtypes.
4. For phenotypes with >2 categorical outcomes, each outcome was assigned an integer value and these were analyzed as quantitative traits. e.g. the phenotype "Alcohol intake frequency" included six categories: "Daily or almost daily", "Three or four times a week", "Once or twice a week", "One to three times a month", "Special occasions only" and "Never". Here, each sample was assigned an integer

value ranging from 1 to 6, with each value corresponding to the six ordered frequencies for this phenotype. Categorical traits with just two separate categories were considered as binary traits.

5. For quantitative traits with integer values, where a subject had multiple instances, we utilized the mean of all instances rounded to nearest integer
6. For quantitative traits with continuous values, where a subject had multiple instances, we utilized the mean of all instances.

For the 1,316 quantitative traits analyzed, we applied a rank based inverse normal transformation and required a minimum sample size of 50 phenotyped individuals and standard deviation ≥0.001 to be included in association analysis. For the 9,299 binary traits analyzed, we required a minimum sample size of at least 25 individuals with the trait and a minimum of at least five individuals that carried the TRE to be included in the analysis.

Association analysis was performed using *REGENIE*[76], incorporating covariates of sex, age, GS insert size and the top five principle components derived from analysis of SNVs to account for ancestry. For binary traits, we utilized the Saddle Point Approximation function to reduce the type I error rate. To avoid possible technical effects on TR genotypes, associations were performed separately in samples sequenced by deCODE and Sanger Center, before combining results using z-score based meta-analyses in *METAL*[77]. We required that associations showed the same direction of effect in both sub-cohorts. For loci on the X chromosome, samples were separated by both sequencing center and sex and associations performed separately in the four resulting sub-groups (sequencing center + sex) before combining results using *METAL*, requiring that at least three of the four sub-groups met the minimum sample size requirements as stated above and all showed the same direction of effect.

We applied multiple testing corrections using both a false discovery rate (FDR) and Bonferroni approach based on the total number of TRs and independent traits analyzed. However, given that many phenotypes in the UK Biobank are highly correlated (e.g. neutrophil count and neutrophil percentage), it should be noted that these multiple testing corrections tend to be overly stringent. We considered associations significant at 10% FDR (FDR q<0.1).

For creating plots of grouped PheWAS results, we assigned each trait reported by the UKB into 22 different categories based on shared physiological systems, tests or treatments. To create these groupings, phenotype codes were mapped to ICD10 chapters as listed in Supplementary Table 1 of Wang et al.[78]. Phenotypes that did not have a category assigned with these annotations were further annotated using categories provided within the UKB Data Showcase. Finally, we applied manual curation, removing categories that were composed of a very small number of traits and making reassignments to improve the consistency of groupings.

Identification of *repeat* expansions in *The 100,000 Genomes Project*

The 100,000 Genomes Project (100kGP) is a diagnostic and research study aimed at sequencing genomes from ~70,000 patients seen within the UK National Health Service with either selected rare diseases or cancers. The 100kGP was approved by the East of England Cambridge South Research Ethics Committee (reference 14/EE/1112). In the rare disease arm of the project, each proband was recruited under a specific disease, and clinical data linked to each proband was collected including the clinical signs and symptoms encoded using human phenotype ontology (HPO). From this cohort, we selected 16,665 individuals: 7,871 probands with intellectual disability (defined as individuals whose clinical data included “intellectual disability” as an HPO term), and 8,794 unrelated controls (defined as probands not recruited under any neurological disease). Ancestry assignments for each individual were performed as described above using a set of high-quality LD-pruned SNVs. We performed genotyping of the *AFF3* TR in PCR-free 150bp paired-end Illumina GS data using ExpansionHunter (v3.2)[20]. We performed manual inspection of read alignment plots for all repeat alleles genotyped as >150bp and those deemed to be false positives due to poor read support or low quality alignments over the TR were excluded. To avoid introduction of potential bias, manual classification of read alignment plots was performed by a user who was blind to the phenotype of the individual.

Calculating polygenic risk score for educational attainment

We utilized data for 3,952 SNVs identified in a prior GWAS[38] to calculate a polygenic risk score (PRS) for educational attainment for each individual in the UK Biobank. SNV genotypes were extracted from the vcf file

derived from GS data, removing any genotypes with read depth <10 or quality score <20. For each of the 3,952 SNVs for which genotypes were available, at each SNV in each individual we multiplied the SNV effect size by the number of effect alleles present, and summed these scores across all SNVs per individual to create a single numerical value per individual. The resulting PRS scores were then ranked and divided into quintiles, where quintile 1 represented individuals with the highest PRS for educational attainment.

Identification of recurrent microdeletions/microduplications

Using available GS data, we applied mosdepth to calculate the normalized read depth for each coding and untranslated region extended by ±100 bp per RefSeq gene, as described previously[73]. We selected six recurrent microdeletion/duplication regions[79] that are known to have effects on neurodevelopment[80] and occur at sufficient frequencies in the UK Biobank cohort for robust statistical analyses (1q21.1, 15q11.2, 15q13.3, 16p13.11, 16p11.2, and 22q11.2). Individuals with deletions and duplications of these six regions were identified based on being clear outliers with an estimated copy number of ~1 or ~3 through visual inspection of the median read depth of all genes located within the CNV region (Supplementary Figure 17).

Identification of SNVs that associate with the *AFF3* TRE

For the 168,641 unrelated individuals of European ancestry in the UKB utilized for PheWAS, SNV genotypes with MAF ≥0.001 located within ±500 kb of the *AFF3* TR were extracted from the vcf file derived from GS data, removing any genotypes with read depth <10 or quality score <20. We utilized REGENIE to perform association analysis between the 319 individuals we defined as carrying expansions of the GCC TR within *AFF3* at the Very Low threshold and 151,777 individuals in the bottom 90% of the population distribution by *AFF3* TR allele size, incorporating covariates of sex, age, GS insert size and the top five principle components derived from analysis of SNVs to account for ancestry. Associations were performed separately in samples sequenced by deCODE and Sanger Center, before combining results using z-score based meta-analyses in *METAL*. Results were then joined with SNVs from the GWAS Catalog with $p<5x10^{-8}$ for the following traits: Educational attainment, Educational attainment (years of education), Educational attainment (college

completion), Educational attainment (MTAG), Intelligence, Intelligence (MTAG), Cognitive ability, Cognitive ability (MTAG) and General cognitive ability.

ACKNOWLEDGMENTS

This work was supported by NIH grants AG075051, NS105781, HD103782 and NS120241 to A.J.S., NHLBI Biodata Catalyst fellowship 5120339 to A.M.T., funding from The Prinses Beatrix Spierfonds (W.OR20-08) to J.J.F.A.V., funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 772376 – EScORIAL) to J.V., 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.

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 and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

DATA AND CODE AVAILABILITY

Project MinE DNA methylation data are available from the European Genome-phenome Archive via accession EGAS00001004587

GTEx DNA methylation data, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213478

GTEx normalized expression level data, https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz

GTEx allele specific expression data, https://www.google.com/url?q=https://console.cloud.google.com/storage/browser/fc-secure-ff8156a3-ddf3-42e4-9211-0fd89da62108/GTEx_Analysis_2017-06-05_v8_ASE_WASP_counts_by_subject/&sa=D&source=docs&ust=1678389406863768&usg=AOvVaw1McwlkcmsHEds7G5NSMcf6

UKB Data Showcase, https://biobank.ctsu.ox.ac.uk/crystal/search.cgi

Information on how to access the 100kGP data by joining a Genomics England Clinical Interpretation Partnership is available online, www.genomicsengland.co.uk/join-a-gecip-domain

Code utilized for this study is available at github, as follows:

https://github.com/bharatij/ExpansionHunter_Classifier

https://github.com/bharatij/Global-Ancestry-Assignment

Author contributions

B.J., P.G., J.J.F.A.V., W.L. and K.I.G. designed bioinformatics pipelines and performed data analyses., D.G., M.J., A.M.T., S.L.G. and M.B. performed data analyses. H.H., B.P., J.V. and A.T. supervised and advised on the project. A.J.S. conceived the study, performed data analyses, supervised the project and drafted the manuscript. All authors reviewed and approved the final draft.

References

1. Kremer, E. J. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* **252**, 1711–1714 (1991).
2. Oberlé, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–1102 (1991).
3. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
4. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
5. Boivin, M. & Charlet-Berguerand, N. Trinucleotide CGG Repeat Diseases: An Expanding Field of Polyglycine Proteins? *Front. Genet.* **13**, 843014 (2022).
6. Hagerman, R. J. *et al.* Fragile X syndrome. *Nat. Rev. Dis. Primer* **3**, 17065 (2017).
7. Hagerman, P. J. & Hagerman, R. J. The fragile-X premutation: a maturing perspective. *Am. J. Hum. Genet.* **74**, 805–816 (2004).
8. Murray, A., Webb, J., Grimley, S., Conway, G. & Jacobs, P. Studies of FRAXA and FRAXE in women with premature ovarian failure. *J. Med. Genet.* **35**, 637–640 (1998).
9. Hagerman, R. J. *et al.* Intention tremor, parkinsonism, and generalized brain atrophy in male carriers of fragile X. *Neurology* **57**, 127–130 (2001).
10. Knight, S. J. *et al.* Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* **74**, 127–134 (1993).
11. Ritchie, R. J. *et al.* The cloning of FRAXF: trinucleotide repeat expansion and methylation at a third fragile site in distal Xqter. *Hum. Mol. Genet.* **3**, 2115–2121 (1994).
12. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of *C9ORF72* causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
13. Renton, A. E. *et al.* A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).

14. Deng, J. *et al.* Expansion of GGC Repeat in *GIPC1* Is Associated with Oculopharyngodistal Myopathy. *Am. J. Hum. Genet.* **106**, 793–804 (2020).

15. LaCroix, A. J. *et al.* GGC Repeat Expansion and Exon 1 Methylation of *XYLT1* Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *Am. J. Hum. Genet.* **104**, 35–44 (2019).

16. Lafrenière, R. G. *et al.* Unstable insertion in the 5' flanking region of the cystatin B gene is the most common mutation in progressive myoclonus epilepsy type 1, EPM1. *Nat. Genet.* **15**, 298–302 (1997).

17. Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in *NOTCH2NLC* associated with neuronal intranuclear inclusion disease. *Nat. Genet.* **51**, 1215–1221 (2019).

18. Garg, P. *et al.* A Survey of Rare Epigenetic Variation in 23,116 Human Genomes Identifies Disease-Relevant Epivariations and CGG Expansions. *Am. J. Hum. Genet.* **107**, 654–669 (2020).

19. Rajan-Babu, I.-S. *et al.* Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Med.* **13**, 126 (2021).

20. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).

21. Dolzhenko, E. *et al.* REViewer: haplotype-resolved visualization of read alignments in and around tandem repeats. *Genome Med.* **14**, 84 (2022).

22. Sarafidou, T. *et al.* Folate-sensitive fragile site FRA10A is due to an expansion of a CGG repeat in a novel gene, *FRA10AC1*, encoding a nuclear protein. *Genomics* **84**, 69–81 (2004).

23. Metsu, S. *et al.* FRA2A is a CGG repeat expansion associated with silencing of *AFF3*. *PLoS Genet.* **10**, e1004242 (2014).

24. Winnepenninckx, B. *et al.* CGG-repeat expansion in the *DIP2B* gene is associated with the fragile site FRA12A on chromosome 12q13.1. *Am. J. Hum. Genet.* **80**, 221–231 (2007).

25. Debacker, K. *et al.* The molecular basis of the folate-sensitive fragile site FRA11A at 11q13. *Cytogenet. Genome Res.* **119**, 9–14 (2007).

26. Jones, C. *et al.* Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene *CBL2*. *Nature* **376**, 145–149 (1995).

27. Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**,

80–86 (2020).

28. Metsu, S. *et al.* A CGG-repeat expansion mutation in *ZNF713* causes FRA7A: association with autistic spectrum disorder in two families. *Hum. Mutat.* **35**, 1295–1300 (2014).
29. Kobayashi, H. *et al.* Expansion of intronic GGCCTG hexanucleotide repeat in *NOP56* causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.* **89**, 121–130 (2011).
30. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
31. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
32. Beck, J. *et al.* Large *C9orf72* hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *Am. J. Hum. Genet.* **92**, 345–353 (2013).
33. Majounie, E. *et al.* Frequency of the *C9orf72* hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurol.* **11**, 323–330 (2012).
34. Zanovello, M. *et al.* Unexpected frequency of the pathogenic *AR* CAG repeat expansion in the general population. *Brain J. Neurol.* awad050 (2023) doi:10.1093/brain/awad050.
35. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur. J. Neurol.* **27**, 1918–1929 (2020).
36. Loh, M. L. *et al.* Mutations in *CBL* occur frequently in juvenile myelomonocytic leukemia. *Blood* **114**, 1859–1863 (2009).
37. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).
38. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54**, 437–449 (2022).
39. Hop, P. J. *et al.* Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS. *Sci. Transl. Med.* **14**, eabj0264 (2022).

40. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).

41. Crowley, J. J. *et al.* Common-variant associations with fragile X syndrome. *Mol. Psychiatry* **24**, 338–344 (2019).

42. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).

43. Afshari, N. A. *et al.* Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat. Commun.* **8**, 14898 (2017).

44. Nicolas, A. *et al.* Genome-wide Analyses Identify *KIF5A* as a Novel ALS Gene. *Neuron* **97**, 1268-1283.e6 (2018).

45. Cossée, M. *et al.* Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7452–7457 (1997).

46. Squitieri, F. *et al.* DNA haplotype analysis of Huntington disease reveals clues to the origins and mechanisms of CAG expansion and reasons for geographic variations of prevalence. *Hum. Mol. Genet.* **3**, 2103–2114 (1994).

47. Bampi, G. B. *et al.* Haplotype Study in SCA10 Families Provides Further Evidence for a Common Ancestral Origin of the Mutation. *Neuromolecular Med.* **19**, 501–509 (2017).

48. Verkerk, A. J. *et al.* Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).

49. Bassett, A. S. *et al.* Clinical features of 78 adults with 22q11 Deletion Syndrome. *Am. J. Med. Genet. A.* **138**, 307–313 (2005).

50. Bernier, R. *et al.* Clinical phenotype of the recurrent 1q21.1 copy-number variant. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18**, 341–349 (2016).

51. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359**, 1685–1699 (2008).

52. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322–328 (2008).

53. van Bon, B. W. M. *et al.* Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *J. Med. Genet.* **46**, 511–523 (2009).

54. Hunter, J. *et al.* Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am. J. Med. Genet. A.* **164A**, 1648–1658 (2014).

55. Gillentine, M. A., Lupo, P. J., Stankiewicz, P. & Schaaf, C. P. An estimation of the prevalence of genomic disorders using chromosomal microarray data. *J. Hum. Genet.* **63**, 795–801 (2018).

56. Nolin, S. L. *et al.* Familial transmission of the *FMR1* CGG repeat. *Am. J. Hum. Genet.* **59**, 1252–1261 (1996).

57. Voisin, N. *et al.* Variants in the degron of *AFF3* are associated with intellectual disability, mesomelic dysplasia, horseshoe kidney, and epileptic encephalopathy. *Am. J. Hum. Genet.* **108**, 857–873 (2021).

58. Rodriguez, C. M. *et al.* A native function for RAN translation and CGG repeats in regulating fragile X protein synthesis. *Nat. Neurosci.* **23**, 386–397 (2020).

59. Green, K. M. *et al.* RAN translation at *C9orf72*-associated repeat expansions is selectively enhanced by the integrated stress response. *Nat. Commun.* **8**, 2005 (2017).

60. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).

61. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

62. Garg, P. & Sharp, A. J. Screening for rare epigenetic variations in autism and schizophrenia. *Hum. Mutat.* **40**, 952–961 (2019).

63. Barbosa, M. *et al.* Identification of rare de novo epigenetic variations in congenital disorders. *Nat. Commun.* **9**, 2064 (2018).

64. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinforma. Oxf. Engl.* **24**, 1547–1548 (2008).

65. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma. Oxf. Engl.* **29**, 189–196 (2013).

66. Houseman, E. A., Kelsey, K. T., Wiencke, J. K. & Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics* **16**, 95 (2015).

67. Dolzhenko, E. *et al.* ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).

68. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).

69. Mitsuhashi, S., Frith, M. C. & Matsumoto, N. Genome-wide survey of tandem repeats by nanopore sequencing shows that disease-associated repeats are more polymorphic in the general population. *BMC Med. Genomics* **14**, 17 (2021).

70. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

72. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

73. Garg, P. *et al.* A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. *Am. J. Hum. Genet.* **109**, 1065–1076 (2022).

74. Dolzhenko, E. *et al.* ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

75. 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).

76. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).

77. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

78. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).

79. Wetzel, A. S. & Darbro, B. W. A comprehensive list of human microdeletion and microduplication syndromes. *BMC Genomic Data* **23**, 82 (2022).

80. Kendall, K. M. *et al.* Cognitive Performance Among Carriers of Pathogenic Copy Number Variants:

Analysis of 152,000 UK Biobank Subjects. *Biol. Psychiatry* **82**, 103–110 (2017).

Figures

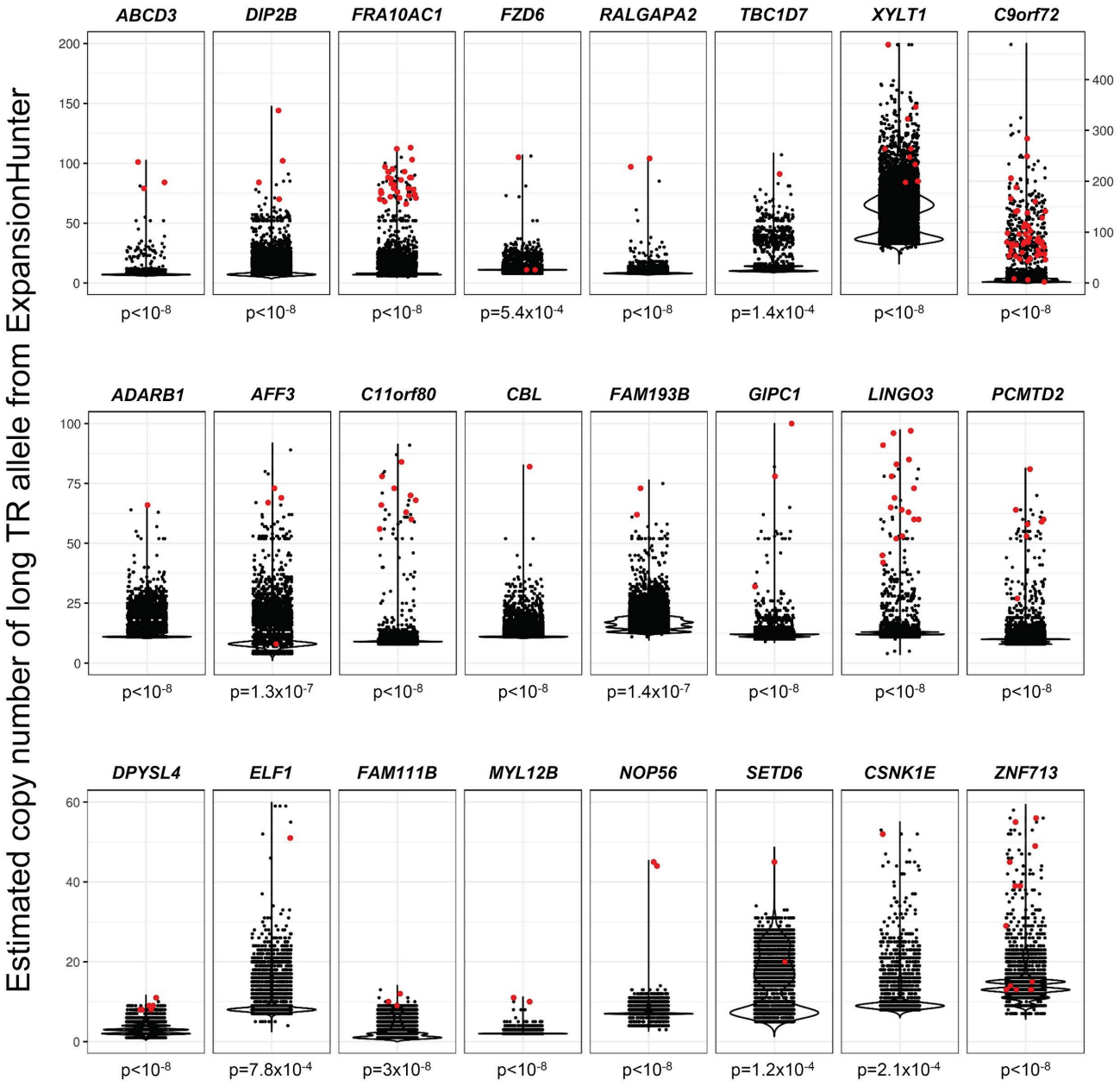

Figure 1. Confirmation of expanded GC-rich tandem repeats as the underlying cause of rare gains of methylation at 24 loci in the human genome. Each plot shows the copy number of the long TR allele in each sample derived from ExpansionHunter analysis of Illumina GS data. Individuals with hypermethylation of the locus are shown in red, individuals without hypermethylation are shown in black. P-values below each plot

show the probability that the observed distribution of individuals with hypermethylation would occur by chance, derived from 100 million permutations. Full details of each TR locus are shown in Supplementary Table 3.

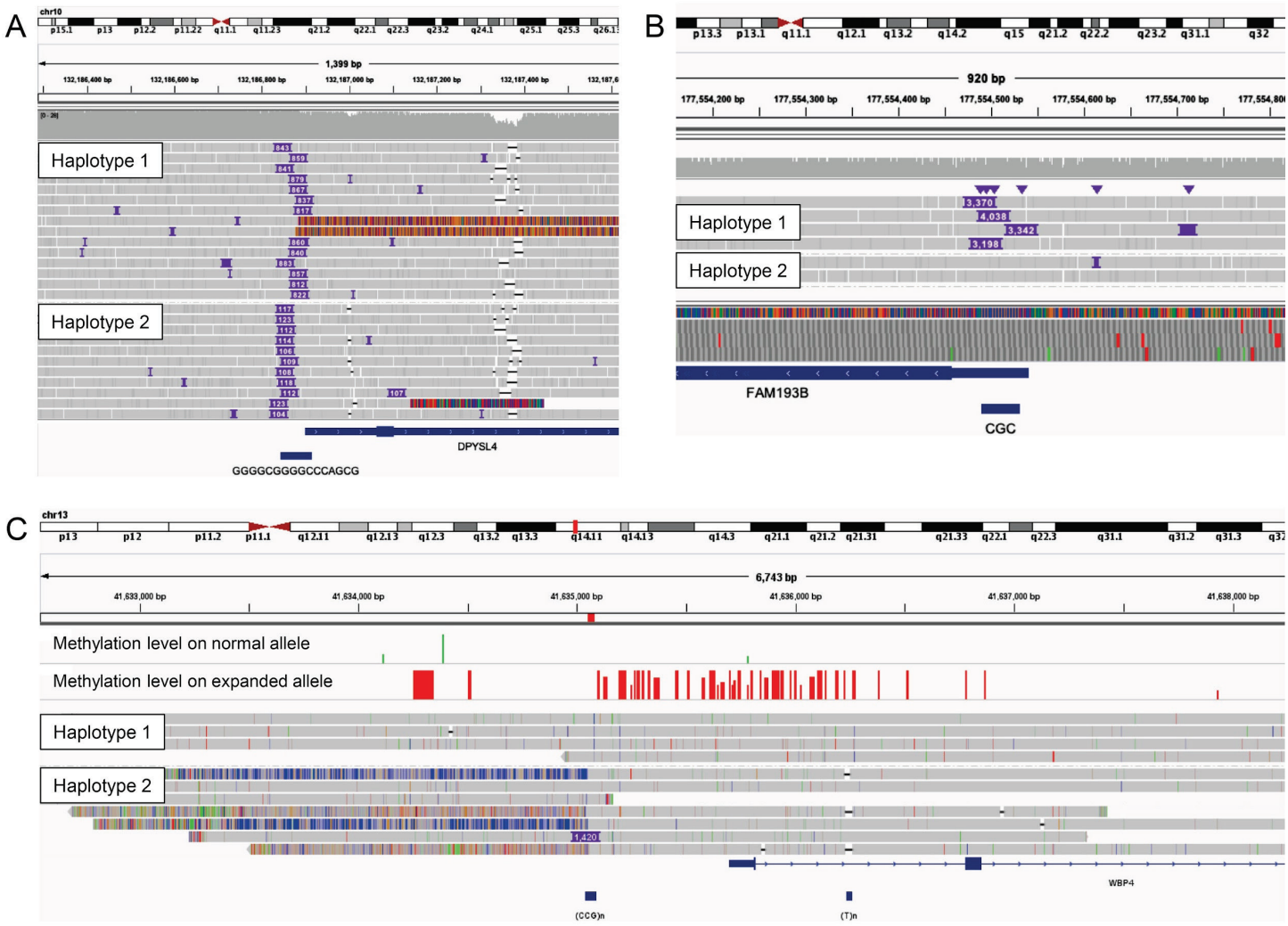

Figure 2. Validation of GC-rich repeat expansions using long-read sequencing. PacBio GS of **(A)** A 16mer TRE within the 5' UTR of *DPYSL4*, and **(B)** A CGG TRE within the 5' UTR of *FAM193B*. **(C)** Oxford Nanopore Technology GS of a CGG TRE in the bidirectional promoter/5' UTR of *WBP4* / *ELF1*. By analyzing the signal profiles of the reads after phasing, we demonstrated that the expanded allele was highly methylated, while the normal TR allele was largely unmethylated. Images show read alignments in IGV. Below each track are shown gene and simple repeat annotations.

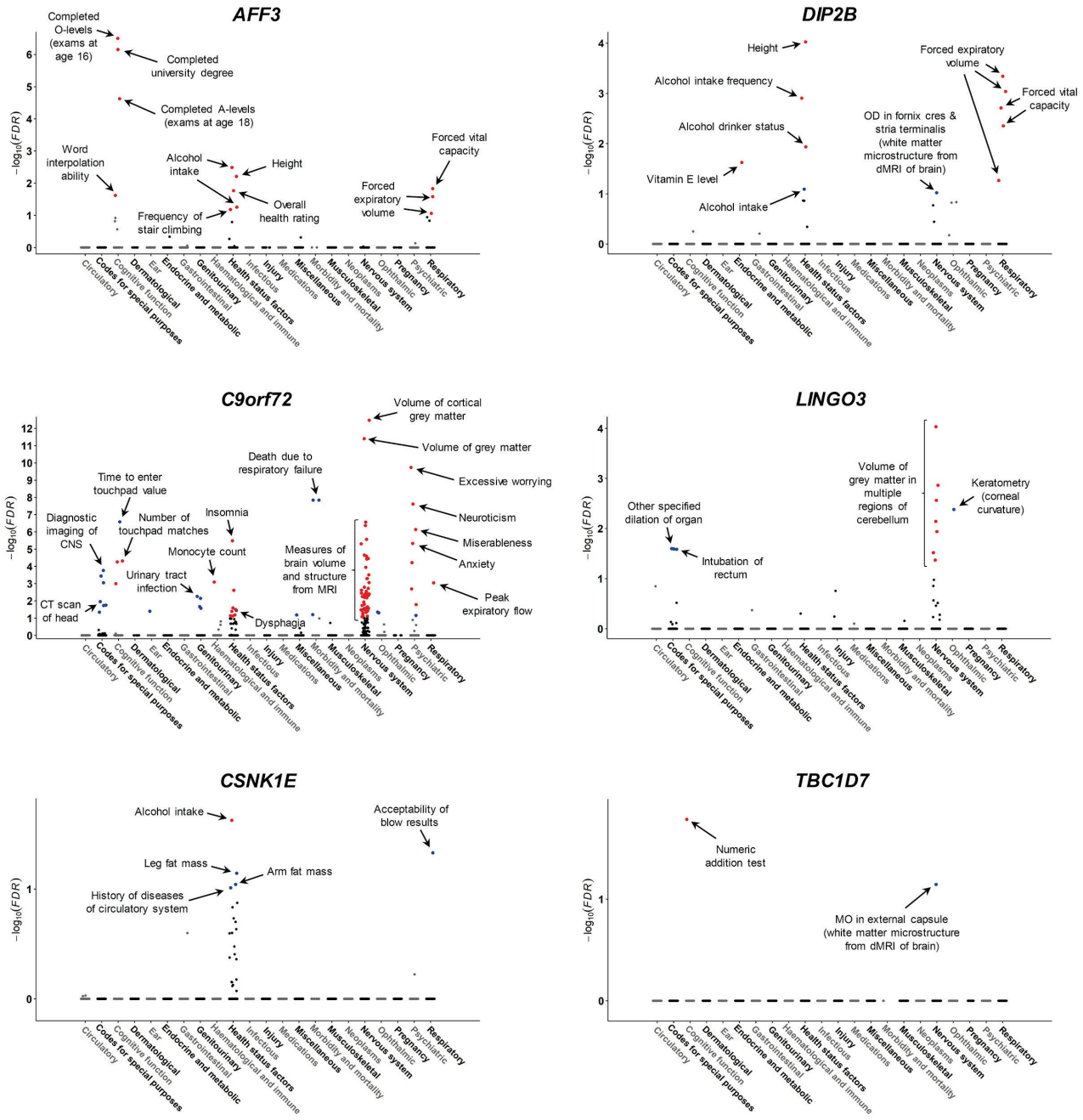

Figure 3. Results of phenome-wide association analysis for expansions of six GC-rich tandem repeats. Traits are grouped into physiological categories (x-axis), and we plot the -$\log_{10}$ FDR-corrected q-value (y-axis). Significant associations (FDR q<0.1) are individually labeled and shown in color to indicate directionality of effect (blue for positive associations, red for negative associations). Note that in some cases the trait names shown have been modified for clarity/brevity. Where a TRE:trait pair was significant under multiple definitions

for identifying TREs (see Methods), we show only the most significant result for that TRE:trait pair. Results for additional TREs are shown in Supplementary Figure 6, with complete data in Supplementary Tables 6 and 7.

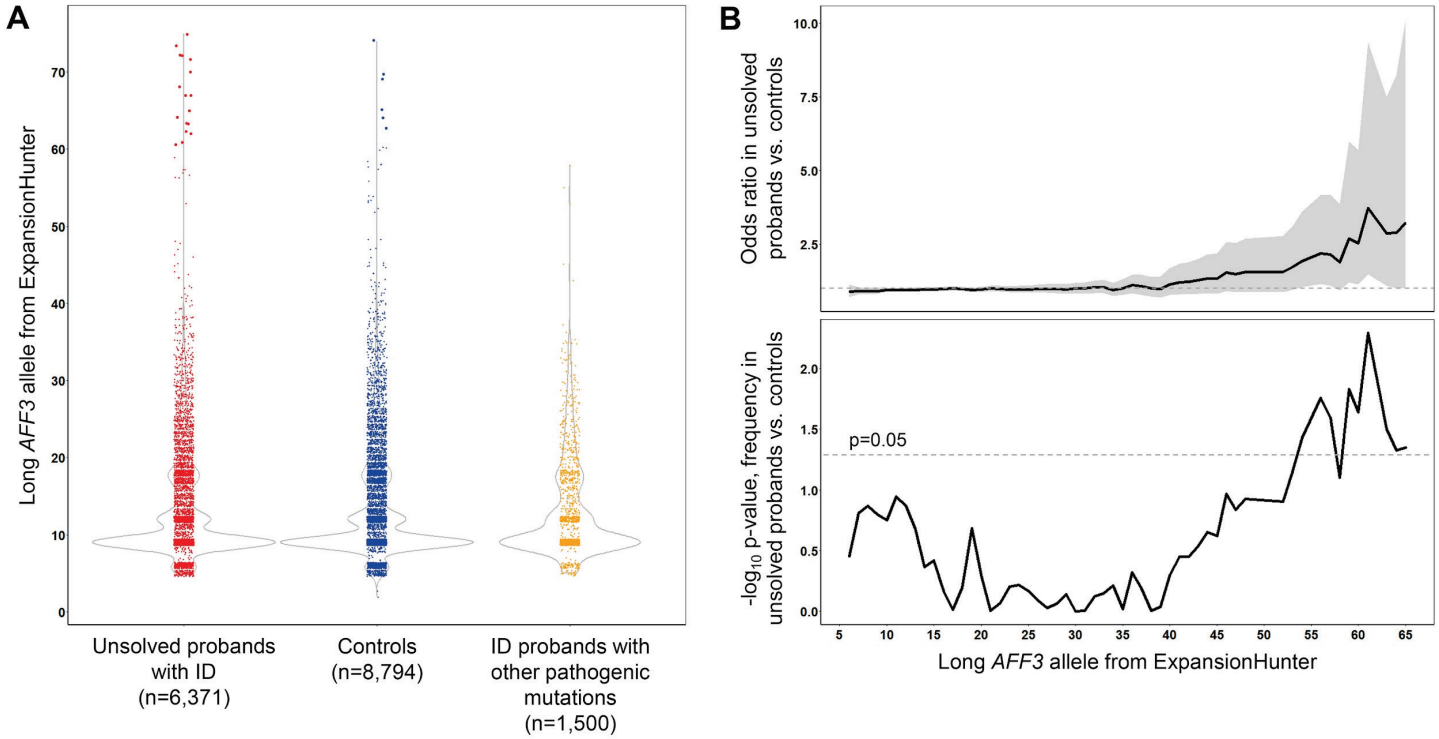

Figure 4. *AFF3* expansions are enriched specifically in unsolved probands with intellectual disability in The 100,000 Genomes Project. (A) Distributions of *AFF3* GCC TR alleles observed in previously unsolved probands with ID (red), controls (blue) and ID probands with other pathogenic or likely pathogenic mutations considered causal for their phenotype (orange) within the 100kGP cohort. TR alleles >60 copies are plotted as larger points. **(B)** Odds ratio and associated p-value for the presence of long *AFF3* TR alleles in 6,371 unsolved probands with a primary referral reason of intellectual disability compared to 8,794 controls. Peak enrichment occurs for alleles of ≥61 copies (p=0.002, Two-tailed Fisher's exact test, 3.9-fold enrichment). In the upper panel, the gray shaded region represents the 95% confidence interval of the odds ratio, while the horizontal dashed line represents no enrichment (odds ratio=1). In the lower panel, the horizontal dashed line represents the significance threshold of p=0.05. To maintain statistical robustness, for the purposes of this plot, alleles ≥65 copies were combined together as a single group.

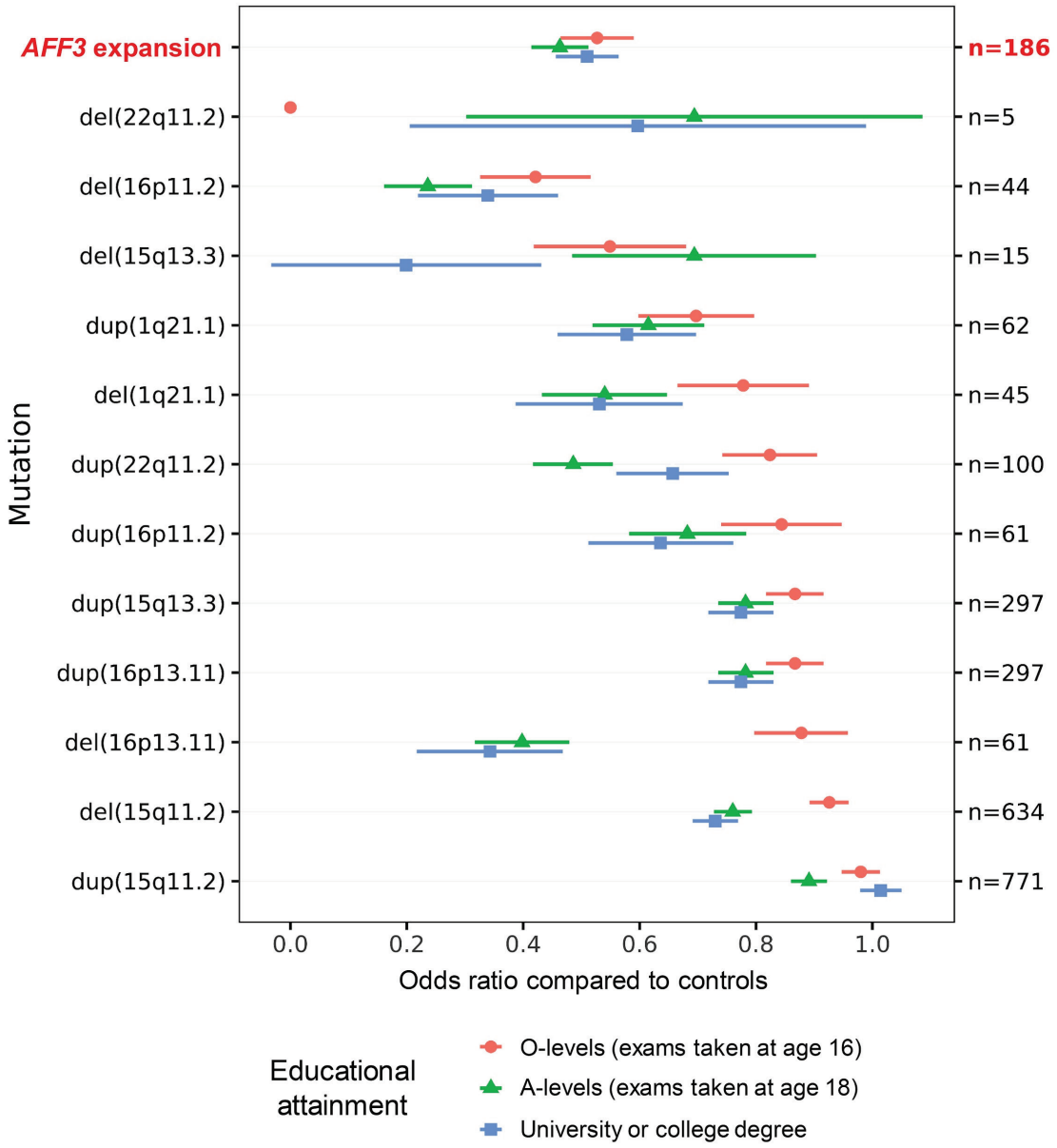

Figure 5. Relative effects of the *AFF3* expansion on educational attainment compared to recurrent microdeletions/duplications. Points show the odds ratio of each group, with horizontal lines indicating 95% confidence intervals. CNVs are ordered by odds ratio for completing O-levels (exams taken at age 16) compared to controls that do not carry any of these mutations. The number of individuals with each mutation type among the 168,641 UK Biobank samples assayed is shown on the right vertical axis. The number of methylated *AFF3* expansions is based on an estimated population prevalence of 1 per 862, with effect sizes on educational attainment based on the 186 individuals with the largest *AFF3* alleles in the UK Biobank individuals utilized for PheWAS.

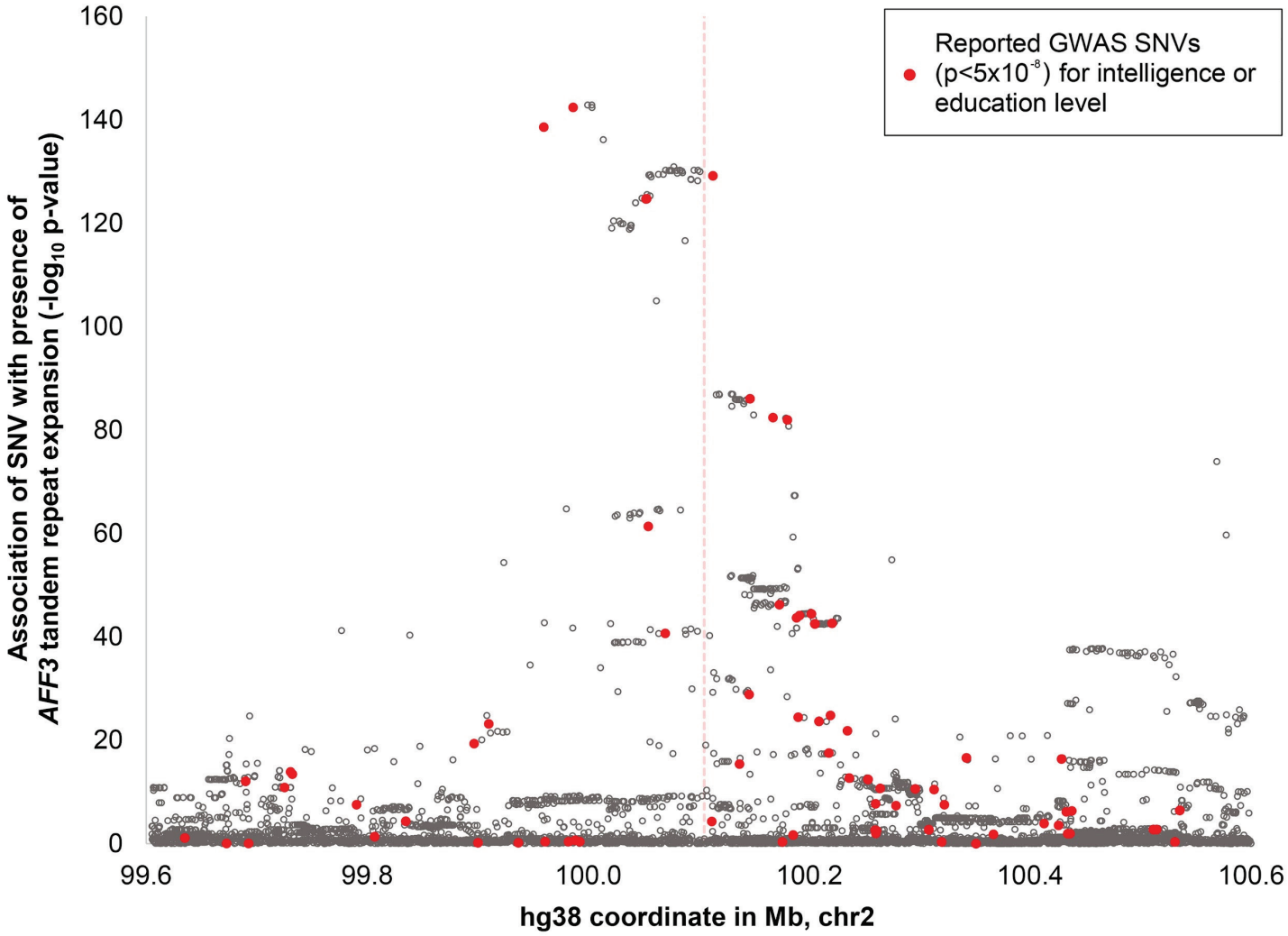

Figure 6. SNVs that associate with presence of the *AFF3* repeat expansion are the same as those identified in prior GWAS of intelligence, cognitive ability and educational attainment. We identified SNVs that associate with the *AFF3* TRE in the UKB cohort, and joined these with SNVs for intelligence, cognitive ability or education level reported in the GWAS catalog (shown as red points). The vertical dashed line in the center of the plot indicates the position of the GCC TR within *AFF3*.