BRAIN
ORIGINAL ARTICLE

Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia

Zhongbo Chen,[1,2,3] Arianna Tucci,[4,†] Valentina Cipriani,[4,†] Emil K. Gustavsson,[1,2,3] Kristina Ibañez,[4,5] Regina H. Reynolds,[2,3] David Zhang,[2,3] Letizia Vestito,[3,4,6] Alejandro Cisterna García,[7] Siddharth Sethi,[1,8] Jonathan W. Brenton,[2,3] Sonia García-Ruiz,[2,3] Aine Fairbrother-Browne,[2,3,9] Ana-Luisa Gil-Martinez,[1,2,3] Genomics England Research Consortium, Nick Wood,[10] John A. Hardy,[1,11,12,13,14] Damian Smedley,[4] Henry Houlden,[15] Juan Botía[1,7] and Mina Ryten[2,3]

[†]These authors contributed equally to this work.

Improvements in functional genomic annotation have led to a critical mass of neurogenetic discoveries. This is exemplified in hereditary ataxia, a heterogeneous group of disorders characterised by incoordination from cerebellar dysfunction. Associated pathogenic variants in more than 300 genes have been described, leading to a detailed genetic classification partitioned by age-of-onset. Despite these advances, up to 75% of patients with ataxia remain molecularly undiagnosed even following whole genome sequencing, as exemplified in the 100 000 Genomes Project.

This study aimed to understand whether we can improve our knowledge of the genetic architecture of hereditary ataxia by leveraging functional genomic annotations, and as a result, generate insights and strategies that raise the diagnostic yield.

To achieve these aims, we used publicly-available multi-omics data to generate 294 genic features, capturing information relating to a gene's structure, genetic variation, tissue-specific, cell-type-specific and temporal expression, as well as protein products of a gene. We studied these features across genes typically causing childhood-onset, adult-onset or both types of disease first individually, then collectively. This led to the generation of testable hypotheses which we investigated using whole genome sequencing data from up to 2182 individuals presenting with ataxia and 6658 non-neurological probands recruited in the 100 000 Genomes Project.

Using this approach, we demonstrated a high short tandem repeat (STR) density within childhood-onset genes suggesting that we may be missing pathogenic repeat expansions within this cohort. This was verified in both childhood- and adult-onset ataxia patients from the 100 000 Genomes Project who were unexpectedly found to have a trend for higher repeat sizes even at naturally-occurring STRs within known ataxia genes, implying a role for STRs in pathogenesis. Using unsupervised analysis, we found significant similarities in genomic annotation across the gene panels, which suggested adult- and childhood-onset patients should be screened using a common diagnostic gene set. We tested this within the 100 000 Genomes Project by assessing the burden of pathogenic variants among childhood-onset genes in adult-onset patients and vice versa. This demonstrated a significantly higher burden of rare, potentially pathogenic variants in conventional childhood-onset genes among individuals with adult-onset ataxia.

Our analysis has implications for the current clinical practice in genetic testing for hereditary ataxia. We suggest that the diagnostic rate for hereditary ataxia could be increased by removing the age-of-onset partition, and through a modified screening for repeat expansions in naturally-occurring STRs within known ataxia-associated genes, in effect treating these regions as candidate pathogenic loci.

Received June 27, 2022. Revised October 24, 2022. Accepted December 22, 2022. Advance access publication January 10, 2023

1 Department of Neurodegenerative Disease, Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK
2 Department of Genetics and Genomic Medicine, Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, UK
3 NIHR Great Ormond Street Hospital Biomedical Research Centre, University College London, London WC1N 1EH, UK
4 William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK
5 Genomics England, Queen Mary University of London, London EC1M 6BQ, UK
6 Ear Institute, University College London, London WC1X 8EE, UK
7 Department of Information and Communications Engineering, University of Murcia, 30003 Murcia, Spain
8 Astex Pharmaceuticals, Cambridge Science Park, Cambridge, UK
9 Department of Medical and Molecular Genetics, School of Basic and Medical Biosciences, King's College London, London WC2R 2LS, UK
10 Department of Clinical and Movement Neurosciences, Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK
11 Reta Lila Weston Institute, Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK
12 UK Dementia Research Institute, University College London, London WC1E 6BT, UK
13 NIHR University College London Hospitals Biomedical Research Centre, London W1T 7DN, UK
14 Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong SAR, China
15 Department of Neuromuscular Disease, Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK

Correspondence to: Professor Mina Ryten
Department of Genetic and Genomic Medicine
Great Ormond Street Institute of Child Health
University College London
30 Guilford Street London WC1N 1EH, UK
E-mail: mina.ryten@ucl.ac.uk

Keywords: hereditary ataxia; functional genomics; transcriptomics; repeat expansion disorders; rare disease

Introduction

Over the last two decades, there has been significant progress in the diagnosis of neurogenetic diseases.[1–3] Despite this, approximately half of patients presenting with a probable genetic cause for a neurological disorder remain undiagnosed,[1,4] with the most clinically and genetically heterogeneous disorders presenting the greatest challenge.[5,6] One such archetypal heterogeneous neurogenetic condition is hereditary ataxia. These are a group of neurodegenerative disorders characterized by the clinical syndrome of progressive incoordination due to cerebellar dysfunction[7,8] with a prevalence of approximately 1.5 to 4.9 per 100 000 persons.[9] To date, variants in more than 300 genes have been discovered to be associated with ataxia taking us away from Greenfield's patho-anatomical and Harding's clinico-genetic classifications.[10] Despite this shift towards a detailed molecular classification, diagnostic rates remain relatively low.[8,11,12] Even when using whole genome sequencing (WGS) in a highly-selective cohort in the Genomics England 100 000 Genomes Project, the diagnostic yield for hereditary ataxia was only 21% among singletons and 32% in family trios.[13] This could be explained both by the existence of as yet undiscovered causative genes and the incomplete screening of known genes in the appropriate patients.

The current genetic evaluation strategy in clinical practice for hereditary ataxia involves partitioning patients by age-of-onset; in the UK, this is employed by the 100 000 Genomes Project[14] and NHS England National Genomic Test Directory (https://www.england.nhs.uk/publication/national-genomic-test-directories/). Diagnostic-grade panels of genes for ataxia are divided into childhood-onset, adult-onset and cerebellar hypoplasia categories.[14] This is also reflected in the Childhood Ataxia and Cerebellar Group of the European Paediatric Neurology Society guidelines, which suggests a specific evaluation algorithm (including genetic testing) for early-onset ataxia.[15] In practice, the existence of these separate panels means that patients with adult-onset ataxia are seldom screened for genes typically associated with childhood-onset disorders such as Joubert syndrome and childhood-onset patients are rarely screened for adult-onset variants such as pathogenic repeat expansions typically associated with some late-onset spinocerebellar ataxia (SCA).[16] It is difficult to assess with confidence whether this age-based classification is justified. Most genetic variants associated with ataxia have only recently been recognized and consequently reported in just a handful of cases.[6] This makes it challenging to determine whether this age-of-onset division reflects the true biology of disease or arises from presentation bias.

With increasing quantities of publicly-available functional multi-omic annotations that assign biological meaning to genomic regions, this has become a tractable question.[17] Through the application of multi-omics technologies and computational tools, it is possible to produce increasingly precise and granular annotations that operate at both a tissue- and cell-type-specific level, providing insights into the shared biology of disease-associated genes.[18,19] This has been key for neurological diseases and importantly, advances in this field have already improved our interpretation of gene-phenotype associations and driven genetic discovery.[18,20] Therefore, we questioned whether the improvements in the breadth and depth of functional

multi-omic annotations would enable us to identify commonalities and hidden biological relationships among genes causing ataxia. Furthermore, we investigated whether this information could be used to improve our understanding of the underlying genetic architecture of these diseases and potentially inform novel testing strategies that could increase diagnostic yield.

In this study, we leveraged the critical mass of genes discovered in ataxia together with a curated set of approximately 300 multi-omic genic features to determine whether ataxia genes can be characterized on a genomic level to identify hidden patterns that explain their biology. We also used this analysis to determine whether the clinical division of ataxia by age-of-onset is reflected in genomic annotations with implications for the current diagnostic strategy. Finally, we assessed the potential of alternative testing strategies using WGS of 2182 probands presenting with ataxia recruited in the 100 000 Genomes Project. Using this approach, we gained further insight into the pathogenic mechanisms underlying hereditary ataxia and highlighted potential bottlenecks to diagnosis.

Materials and methods

Defining a list of genes associated with hereditary ataxia

To identify genes known to be associated with hereditary ataxia across all ages, we used three resources: (i) Genomics England PanelApp, a publicly-available crowdsourced tool to standardise gene panels[14]; (ii) Hereditary Ataxia GeneReviews, a regularly updated international point-of-care resource on inherited medical conditions[7]; and (iii) the Online Mendelian Inheritance in Man (OMIM) database.[21] We extracted '*green*', diagnostic-grade genes from PanelApp considering the four gene panels of cerebellar hypoplasia (v.1.41), adult-onset ataxia (v.2.8), childhood-onset ataxia (v.6.22) and hereditary ataxia (v.1.2.05). By combining PanelApp, GeneReviews and OMIM resources, we identified a total of 318 unique genes (workflow shown in Fig. 1). We noted that there were discrepancies between the resources in the expected age-of-onset. For example, *SCYL1* was classified in PanelApp as adult-onset but has been described as a childhood-onset disorder in OMIM (spinocerebellar ataxia autosomal recessive 21. MIM:616719).[21] Thus, to standardize information for genes of interest, we extracted typical age-of-onset data in an automated manner from OMIM 'Text' section (http://api.omim.org),[21] and also manually curated information from OMIM 'Clinical Synopsis' that reviews reported cases. Thus, the 318 genes identified were classified into: (i) adult-onset ($n = 23$, typical age-of-onset ≥18 years); (ii) childhood-onset ($n = 213$, typical age-of-onset <18 years); and (iii) overlap-onset genes ($n = 82$, causing both childhood- and adult-onset disease) (Supplementary Table 1). Given that we sought to differentiate between genes associated with ataxia and genes not known to cause ataxia, we defined the set of remaining protein-coding genes as controls ($n = 17{,}323$, Ensembl v.72).[22]

Extracting clinical phenotype information

To capture genic information about the clinically-heterogeneous phenotypes associated with hereditary ataxia genes, we used data provided within the OMIM catalogue (http://api.omim.org)[21] and the Human Phenotype Ontology (HPO) database (https://hpo.jax.org/app/Build 1271)[23] as the latter also incorporated additional information from Orphanet.[24] We extracted HPO and OMIM terms associated with each gene.

Genic feature generation

We curated a total of 294 genic features leveraging publicly-available multi-omics datasets capable of providing genome-wide information. The genic features captured information in four main categories relating to: (i) gene structure and complexity; (ii) genetic variation including evolutionary features; (iii) gene expression and co-expression; and (iv) protein product of a gene. A full list of genic features with their corresponding source is provided in Supplementary Table 2.

Gene structure and complexity

We extracted information relating to gene structure from Ensembl v.72.[25] This included: gene length; number of unique transcripts within a gene; number of exon-exon junctions and the gene's GC content among others. Specific to the pathogenic mechanisms of ataxia, we used resources generated through application of Tandem Repeats Finder[26] to the human reference genome (GRCh38)[27] in HipSTR (https://github.com/HipSTR-Tool/HipSTR-references) to create a gene-based metric of short tandem repeat (STR) density, size, number of nucleotides within each STR, and location of STRs as annotated by Ensembl v.72 across the entire genome.[22] This was complemented by extraction of information on all STRs genome-wide associated with expression of nearby genes (eSTRs).[28] We generated additional annotations to reflect the presence of other repetitive elements using the RepeatMasker (http://www.repeatmasker.org) reference panel,[29,30] including short interspersed nuclear (SINE)/Alu elements, retroposon/SVAs, and long interspersed nuclear (LINE)/L1 elements (full list in Supplementary Table 2).

Genetic variation including evolutionary information

Measures of genetic variation were collated from existing large population databases and the related resources.[31] We used LoFTool (gene intolerance score based on loss-of-function (LoF) variants),[32] EvoTol (measures a gene's intolerance to mutation using evolutionary conservation of protein sequences),[33] RVIS (intolerance scoring system of a gene's functional variation),[34] gnomadpLI (LoF score from gnomAD such that a pLI closer to 1 indicates that the gene or transcript cannot tolerate protein-truncating variation)[31] among other metrics. We also derived features to capture evolutionary information about a gene. Using the phastCons20 score, a measure of inter-species conservation between primates[35] together with context dependent tolerance score (CDTS), a measure of intra-species constraint,[36] we generated the genic density of constrained, non-conserved genomic regions (CNCRs), which represent human-lineage-specific element annotation.[37]

Gene expression and co-expression

We leveraged publicly-available transcriptomic data to capture information on tissue-specific, cell-type-specific and temporally-relevant expression. Using temporal expression data generated from RNA-sequencing of human organ development over 23 time-points from 4 weeks post-conception to the sixth decade of life, we obtained a measure of developmental gene expression (developmentally-dynamic expression pleiotropy index (0–1) with a value of 1 indicating the most dynamically-repressed expression).[38] Tissue-specific gene expression and co-expression features were extracted and downloaded from G2PML (https://github.com/juanbot/G2PML), a machine learning tool for predicting disease-associated genes based on genic features.[25] For tissue-

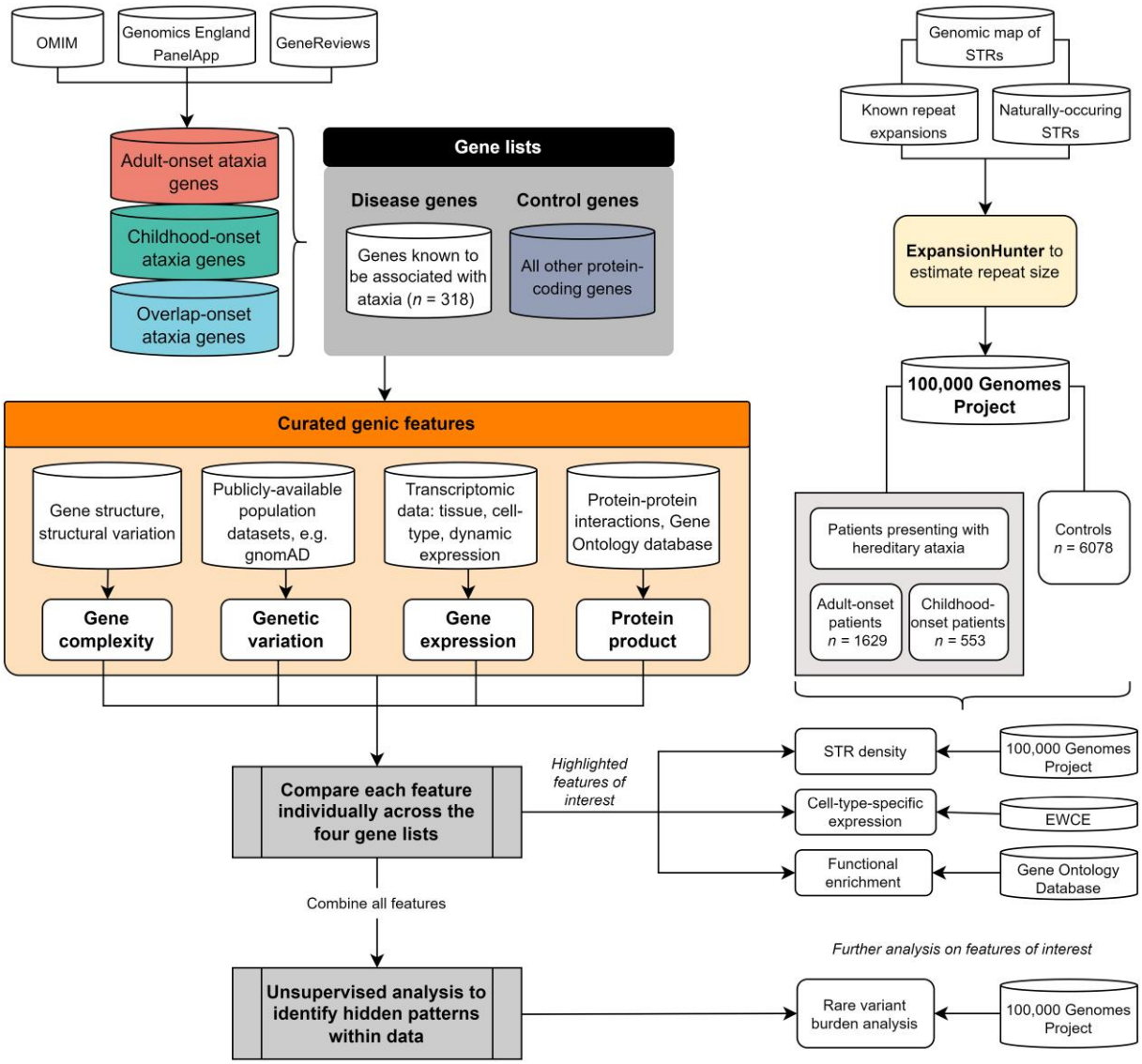

Figure 1 Overall workflow of study. Genic information is captured across categories of genetic variation, gene structure/complexity, gene expression and co-expression and protein-product of a gene, and compared across the four gene lists of: (i) adult-onset ataxia genes; (ii) childhood-onset ataxia genes; (iii) overlap-onset ataxia genes, defined as those associated with both childhood- and adult-onset, when mutated; (iv) other protein-coding genes not known to cause ataxia (control 'not ataxia' genes). The gene lists were extracted primarily from Genomics England PanelApp, but also GeneReviews and OMIM. The age-of-onset definition was derived primarily from OMIM to reduce bias. Genic features were first compared individually across the four gene lists then combined together through unsupervised clustering analysis. Individual genic features were also highlighted and put through further analyses including expression-weighted cell-type enrichment (EWCE) for cell-type-specific expression and functional gene ontology (GO) enrichment. Further verification of the results from functional genomic annotation were verified in whole genome sequencing data of patients with ataxia recruited to the 100 000 Genomes Project through rare variant burden analysis and short tandem repeat (STR) analysis.

specific expression, using already-processed Reads Per Kilobase Million (RPKM) data from 47 Genotype-Tissue Expression Project (GTEx)[39] tissues, gene expression data were filtered for genes with RPKM >0.1 and corrected for batch effects, age, sex and RNA integrity number using ComBat[40] as employed by G2PML.[25] Residuals of these linear regression models were used to calculate tissue-specific expression and to construct gene co-expression networks (WGCNA)[41] for each tissue.[25] A gene was defined as having tissue-specific expression if expression in that tissue was five-fold higher than the mean across all tissues. For each gene in a network, the Module Membership (MM) for a gene was defined as the correlation of its residual gene expression and the eigen-gene of the module to which it belonged.[25] A gene was defined as having tissue-specific MM at 3.5-fold higher than across all other GTEx tissues.[25] For cell-type-specific expression, we used single cell RNA-sequencing data from mice.[42] We generated features incorporating information from specificity matrices of expression data.[43] We used all level 2 and level 3 cell types within and outside the CNS. We then focused on cerebellar-specific expression (CNS superset level 2).[42]

Curated data on the protein product of a gene and its function

The Gene Ontology (GO)[44,45] and STRING databases were used to obtain information on the function of a gene's protein product and the

number of its protein-protein interactions.[46] We used STRING parameters with confidence of >500 and a maximum of eight interactions per gene.[47] We used g:Profiler[48] for GO enrichment analysis, with the g:SCS multiple testing correction method. In addition, we used the Full Spectrum of Intolerance to Loss-of-function (FUSIL) categorization of genes, which was based on both viability and phenotyping screens performed on knock-out mice, and essentiality screens carried out on human cell lines.[49] This provided five gene categories relating to the essentiality of their function: cellularly lethal (CL); developmentally lethal (DL); subviable (SV); viable with significant phenotype/s (VP); or viable with no significant phenotypes detected (VN).[49]

Statistical analysis

For each feature generated, we compared the feature across the four gene lists, namely adult-onset, childhood-onset, overlap-onset ataxia and control genes. For continuous variables, we used Wilcoxon rank sum test to compare the means of the metrics between two groups, taking a two-tailed P-value < 0.05 as significant. For categorical variables, we used chi-squared tests to assess statistically significant ($P < 0.05$) differences between the distributions of two groups. We used pair-wise comparison P-values due to the imbalance in the sizes of the gene lists. However, we also provide false discovery rate (FDR) corrected P-values that account for all six comparisons between the gene lists. We have outlined statistical analyses of other methods within the relevant sections. All analyses were carried out in R (v.4.0.5).

Expression-weighted cell-type enrichment

Expression-weighted cell-type enrichment (EWCE) was used to determine whether ataxia genes have higher expression within particular cell types than would be expected by chance (https://github.com/NathanSkene/EWCE).[43] We used the adult-, childhood- and overlap-onset ataxia gene lists as input with specificity matrices calculated for level 2/3 cell types,[42] as well as level 2 cell types from the superset containing cerebellar-specific cell types. We controlled for transcript length and GC-content in the bootstrap lists where EWCE was run with 10 000 bootstrap replicates. Genes without a 1:1 mouse:human ortholog were excluded. P-values for multiple testing were corrected using the Benjamini–Hochberg method over all cell types and the three gene sets. At the cerebellar-specific cell type level, P-values were only corrected for the number of gene sets given the granularity of cell-types at this level. To assess the contribution of specific genes to cell-type-specific expression, we obtained the mean expression for the gene of interest within the cell type divided by its expression across all cell types.[43]

Unsupervised analysis of all features

In order to compare the utility of all features in classifying ataxia genes, first, we used recursive feature elimination (caret R package[50]) to remove redundant features, defined as those with Pearson's correlation >0.9 between two features (visualized using corrplot package[51]). This approach also helped to account for the imbalance in gene list sizes by taking a random sample (50%) of disease genes and an equal number of bootstrap-selected control genes from the larger control set. The recursive feature elimination then took a random subset of input features based on feature importance and assessed the number of features with the highest accuracy as defined by Cohen's κ statistic. It then fitted a generalized linear model for the best set of features. The feature elimination operates in a k-fold cross-validation manner to improve the accuracy of the fit of the model for unseen samples. We extracted the minimum proportion of times that a feature was selected as most relevant out of the total number of iterations in the repeat variable (parameter r). In this case, we used features that had an $r \geq 0.8$, indicating that they appeared more than 40 times out of the 50 repetitions performed. Taking these selected features, we used uniform manifold approximation and projection (UMAP) to investigate any hidden patterns within the data defined using these salient features.[52]

Validation within 100 000 Genomes Project participants

The 100 000 Genomes Project is a UK programme to assess the value of WGS in patients with rare diseases.[13,16,53] Participants were identified by healthcare professionals and researchers from 13 genomic medicine centres in England as having an undiagnosed rare disease and recruited with consent after approval from the national research ethics committee.[13,16,53] WGS was carried out as per previously-described methods, with variant and sample quality control as outlined.[13,16] The rare variant burden testing pipeline was also developed through the 100 000 Genomes Project diagnostics pipeline as previously described.[13]

Rare genetic variant burden testing

To test whether adult- and childhood-onset hereditary ataxia patients should be screened for mutations using a common gene set, a case-control gene burden analysis was adapted from an analytical framework used within the rare disease component of the 100 000 Genomes Project.[13,16,53] Cases were defined as all probands recruited under the clinical indications: 'hereditary ataxia', 'cerebellar hypoplasia' or 'pontine tegmental cap dysplasia'. For the corresponding control group, we used all other probands aged ≥40 years at recruitment excluding individuals recruited under relevant neurological or related disease categories ($n = 6658$) (exclusion criteria for controls in Supplementary Table 3). Cases were defined as childhood-onset (<18 years) ($n = 306$) and adult-onset (≥18 years) ($n = 816$). We used the recorded age-of-onset, where available, or age at recruitment as proxy. Using these definitions, four different combinations of rare genetic variant burden analyses were performed as summarized in Supplementary Table 4: (i) typically childhood-onset genes in adult-onset cases versus in controls; (ii) adult-onset genes in childhood-onset cases versus in controls; (iii) overlap-onset genes in adult-onset cases versus in controls; and (iv) overlap-onset genes in childhood-onset cases versus in controls, with the latter two analyses used for controlled comparison as we would expect the find overlap-onset variants in both childhood- and adult-onset cases. We defined controls as those over 40 years of age to ensure that any large repeat expansions in the control group are less likely to be pathogenic.

The sets of rare variants for the gene burden analyses were obtained running Exomiser[54] on all probands' WGS data to filter coding variants that are rare (minor allele frequency < 0.1%, for dominant, and <1%, for recessive variants in gnomAD[31] v2.1.1 and v3.1.1 as well as within the local 100 000 Genomes Project cohort and segregated with disease status (where family information was available). Gene-based enrichment of rare variants in cases was assessed using one-sided Fisher's exact test under four scenarios: (i) enrichment of rare, predicted LoF variants; (ii) enrichment of rare, predicted pathogenic variants (Exomiser variant score >0.8); (iii) enrichment of rare, predicted pathogenic variants in a

constrained coding region; and (iv) enrichment of rare, *de novo* variants. For the latter, only trios or larger families where *de novo* calling was possible were considered. To maintain statistical validity, the analysis was limited to those disease-gene associations where relevant variants were seen in at least four probands over the entire cohort of cases and controls. Benjamini–Hochberg method was used to correct for multiple testing; an overall FDR *P*-threshold of 0.05 was used for claiming significant gene-disease associations accounting for the total number of case-control gene burden tests under all four scenarios, i.e. 360.

To support this, we then reviewed these patients with ataxia for apparent incongruities between the age of disease onset and typical age-of-onset category associated with the diagnostic pathogenic variant on the final formal genetic report issued.

Short tandem repeat sizes in participants with ataxia

We investigated whether naturally-occurring STRs within the 318 known ataxia genes harboured differences in repeat size distributions between individuals presenting with ataxia and controls recruited in the 100 000 Genomes Project. Ataxia cases were any probands presenting with ataxia, cerebellar hypoplasia or pontine tegmental cap dysplasia either within the enrolled disease group or as an HPO term (i.e. ataxia as a primary symptom or as part of a more complex syndrome) (adult-onset $n = 1629$, childhood-onset $n = 553$) and non-neurological controls aged ≥40 years ($n = 6078$) from all recruited probands defined as per Supplementary Table 4. There were 580 fewer control samples used for this analysis compared to the burden analysis due to exclusion of the participants recruited into the pilot phase of the 100 000 Genomes Project. STR genotyping was performed using ExpansionHunter v.3.1.2[55] using methods as previously described[16] at HipSTR reference loci for naturally-occurring trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides located in the exons or 5′UTRs of the 318 genes across all cases and controls.[27] Using this definition, we studied a total of 197 STRs across 107 unique genes with seven being known pathogenic repeats. We then ranked the maximum repeat size for each STR and partitioned the repeat sizes into bins for controls and for cases. Taking the top 1% repeat sizes in controls and the top 1% in ataxia cases, partitioned by age-of-onset, we compared the mean differences in the repeat sizes. We applied the same testing strategy to STRs known to cause repeat expansion disorders, and which would be expected to have a higher repeat size in ataxia cases compared to controls.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its Supplementary material. Sources for publicly-available data used for generating the genic features are shown in Supplementary Table 2. Code is available through https://github.com/ZhongboUCL/hereditary-ataxia-functional-genomics.

Results

Childhood- and adult-onset ataxia differ in pathogenic variant type and phenotypic presentations

We analysed 318 genes known to cause hereditary ataxia when mutated, classified as: (i) adult-onset ($n = 23$); (ii) childhood-onset ($n = 213$); and (iii) overlap-onset genes (associated with either childhood- or adult-onset disease, $n = 82$). We confirmed that childhood-onset genes showed a higher proportion of biallelic autosomal inheritance compared with adult-onset (chi-squared $P = 2.329 \times 10^{-7}$) and overlap-onset genes ($P = 6.094 \times 10^{-7}$). In contrast, a significantly higher proportion of pathogenic repeat expansions cause adult-onset (34.8%) compared with childhood-onset disease (0.5%) (chi-squared $P = 3.205 \times 10^{-14}$) (Supplementary Fig. 1). Using known phenotypic associations captured within OMIM and HPO[23] showed that childhood-onset ataxia genes had the highest mean number of associated HPO terms per gene (Fig. 2A) and affected a significantly larger number of distinct body systems compared with adult-onset (Wilcoxon rank sum $P = 5.400 \times 10^{-4}$) and overlap-onset genes ($P = 1.800 \times 10^{-4}$) (Fig. 2B and Supplementary Fig. 1C). In summary, we confirmed that childhood-onset ataxia genes were more likely to be autosomal recessive, less likely to be associated with repeat expansion disorders and tended to manifest in multiple systems compared with adult-onset genes, in line with current clinical perception.

Ataxia genes contain an increased density of repetitive elements

We leveraged the increasing availability of functional genomic annotation to expand our understanding of genes associated with hereditary ataxia. This analysis was performed by comparing each individual genic feature from our collation of ~300 functional multi-omic annotations across the four gene lists (results of all comparisons in Supplementary Table 5). We began by focusing on measures of gene complexity and structure. Using this approach, we saw increased overall complexity amongst genes associated with ataxia compared to the control set. Ataxia genes harbour more transcripts and junctions per gene, suggesting that splicing variants could contribute to pathogenesis (Fig. 2C and D). However, we noted that gene complexity was a feature of all disease gene sets and did not distinguish between childhood-, adult- and overlap-onset ataxia genes (Supplementary Fig. 2).

Next, we expanded our analysis to consider the impact of STRs on the complexity of ataxia gene structure. Although pathogenic STR expansions are already known to be an important disease-causing mechanism for adult-onset hereditary ataxia[56] (Supplementary Fig. 1 and Fig. 3A), we questioned whether a higher genic density of non-pathogenic STRs could also be a distinguishing property of ataxia genes. Generating a genomic map of STR elements,[27] we found that the majority of intragenic STRs and eSTRs reside in the intron (95.13% and 94.43% respectively) (Fig. 3A). Furthermore, we noted the existence of 1143 naturally-occurring intragenic CAG repeats which could be considered candidate loci for further interrogation in an unsolved cohort (Fig. 3A). Focusing on ataxia genes, we found that this gene set harboured a higher number of STRs per gene (median 34.5 and 27 STRs per adult- and childhood-onset gene respectively) than control genes (median 16 STRs per gene) (Fig. 3B). Surprisingly, this was evident when comparing childhood-onset genes to non-ataxia genes ($P = 0.008$), although only adult-onset genes had a significantly higher trinucleotide repeat density than controls (Fig. 3C). This trend extended to eSTRs, defined as STRs associated with variable expression of nearby genes proportional to their repeat length.[28] We found that childhood-onset genes had a higher number of associated eSTRs compared to control genes ($P =$

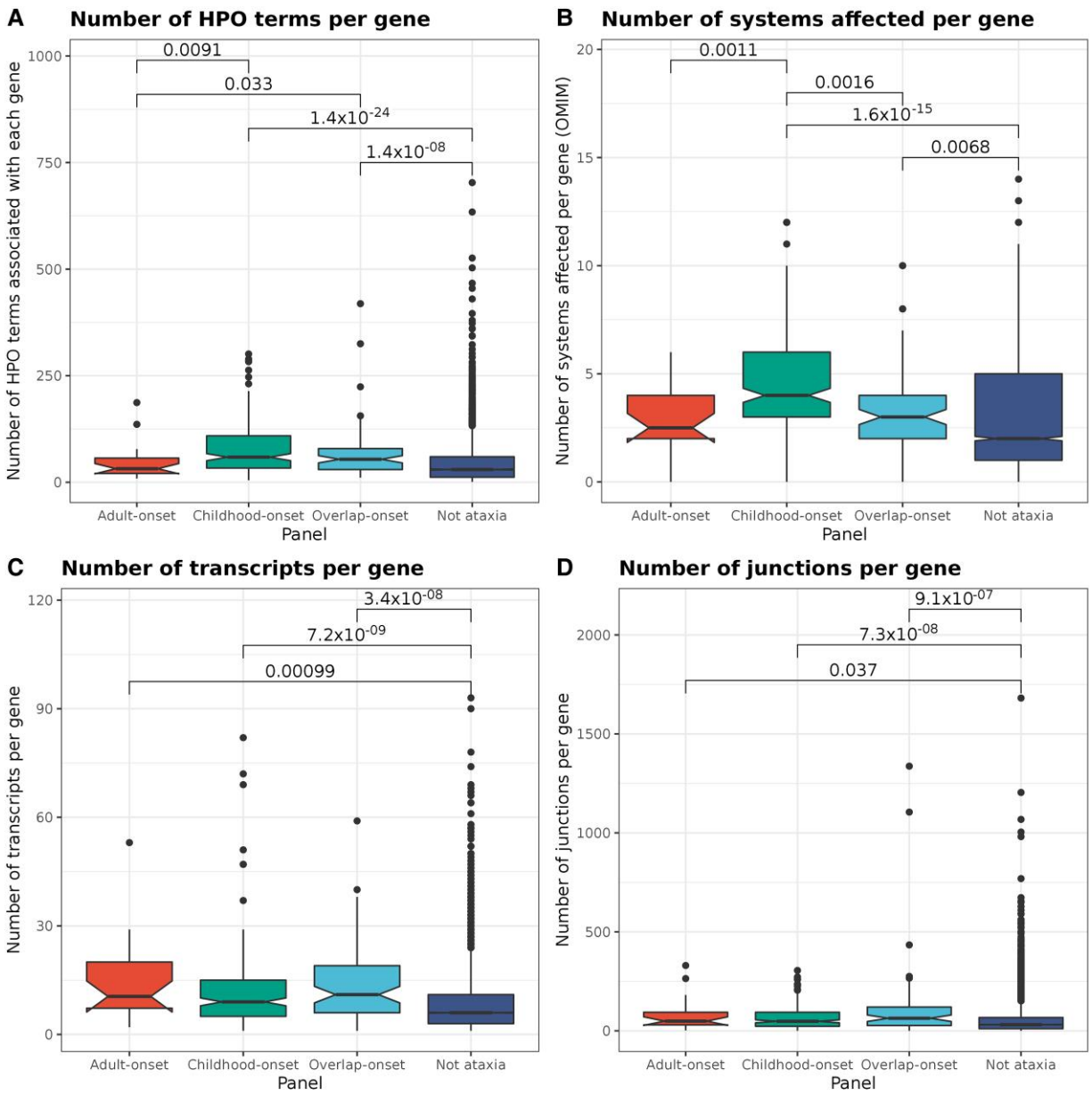

Figure 2 Comparison of phenotypes associated with genes as annotated by HPO and OMIM and gene complexity features between different gene panels. The number of known HPO terms associated with each gene is shown in **A**. The number of body systems affected associated with each gene as annotated by OMIM is shown in **B**. The number of transcripts of each known gene as annotated by Ensembl v.72 is shown in **C**. The number of annotated junctions within each gene as annotated by Ensembl v.72 is shown in **D**. Only significant Wilcoxon rank sum *P*-values (<0.05) are given for pairwise comparisons above the square brackets. The corresponding horizontal lines on the notched boxplots represents the lowest quartile, median, and upper quartile of the data. Further results are presented in Supplementary Table 5.

7.300×10^{-4}) (Fig. 3D), and that these eSTRs were detected in a higher number of tissues compared with control genes ($P = 4.900 \times 10^{-5}$) (Fig. 3E).

Using RepeatMasker's library of repetitive elements,[29,30] we identified differences in the density of other interspersed repetitive elements between the gene lists. Interestingly, both adult- and childhood-onset genes had a higher number of LINE/L1 elements per gene than controls genes (Fig. 3F), as well as SINE/Alu elements among others (Supplementary Fig. 2). Thus, our findings demonstrated that ataxia genes are structurally more complex than other genes and this extended to STRs and other repetitive elements. Unexpectedly, these findings were relevant to genes causing childhood- as well as adult-onset ataxia.

Population-based genetic variation differentiates ataxia genes

We used population-based measures of genetic variation to analyse ataxia genes. As would be expected, the findings largely reflected known differences in inheritance patterns (Supplementary Fig. 3). For example, the probability of a gene being intolerant of homozygous and missense variants (gnomadpMiss), but not heterozygous LoF variants from gnomAD data (gnomadpRec),[31] was significantly higher for childhood-onset genes compared to control genes (Supplementary Fig. 3). We also found that CNCR density, a measure of the proportion of human-lineage-specific elements within a gene,[37] was higher for both childhood- and overlap-onset

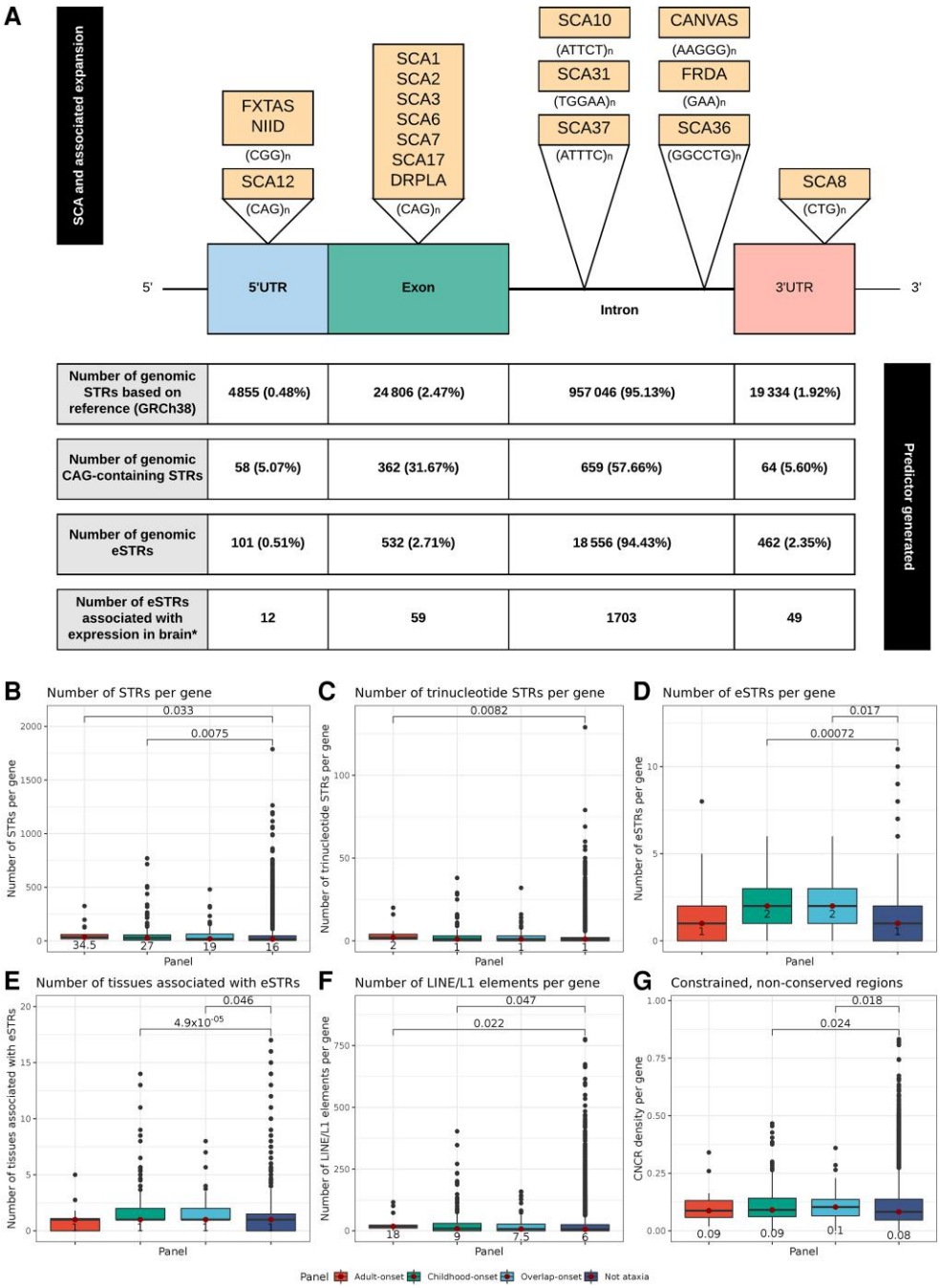

Figure 3 Summary of gene features generated by leveraging information on genomic map of known STRs and eSTRs with examples of hereditary ataxia associated with pathogenic STRs. (**A**) *Top* illustrates the location of repeat expansions within SCA and other ataxias: Fragile X-associated tremor-ataxia syndrome (FXTAS); neuronal intranuclear inclusion disease (NIID); dentatorubral-pallidoluysian atrophy (DRPLA); cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS) and Friedreich's ataxia (FRDA). *Bottom* shows locations of genomic STRs, number of genomic CAG-containing STRs are taken the HipSTR package.[27] Number of genomic eSTRs are based on analyses from Fotsing *et al.*[28] These eSTRs are taken from those within the top 28 375 eSTRs associated with a high CAVIAR (CAusal Variants Identification in Associated Regions) score for posterior probability of causality when fine-mapped against top 100 nearby SNPs. *Number of eSTRs associated with expression in brain is also derived from this work.[28] The total number of STRs/eSTRs is presented with the percentage of overall intragenic location for each STR in parentheses. (**B**) Comparison of the number of STRs within each gene (as defined within the HipSTR package) across the four gene lists. (**C**) Comparison of the number of trinucleotide STRs for each gene across the gene lists. (**D**) Comparison of the number of eSTRs per gene across the gene lists. (**E**) Comparison of the number of tissues in which eSTRs affect gene expression across the gene lists. (**F**) Comparison of the number of LINE/L1 elements per gene (as defined by RepeatMasker) across the gene lists. (**G**) Comparison of the density of CNCRs per gene across the gene lists. The CNCR density of a gene reflects the proportion of gene length that is covered by regions fulfilling criteria for constrained but not conserved sequences, such that a density of 1 signifies that the entire gene fulfils criteria for CNCRs. CNCRs are taken from Chen et *al.*[37] and reflect the regions of the genome likely to be more human-lineage-specific. Only significant Wilcoxon rank sum *P*-values (<0.05) are given for pairwise comparisons above the square brackets. The numbers below or within the boxes of the box and whisker plots represent the median values for that genic feature. The corresponding horizontal lines on the notched boxplots represents the lowest quartile, median and upper quartile of the data. Further results are presented in Supplementary Table 5.

compared with non-ataxia genes ($P = 0.024$ and $P = 0.018$, respectively) suggesting a higher density of sequences important in human-specific evolution associated with known ataxia genes (Fig. 3G).

Transcriptomic signatures of ataxia genes

Given the phenotypic variability of hereditary ataxia, we explored the tissue- and cell type-specificity of disease genes, as well as their expression over time. The latter was performed using publicly-available temporal expression data generated by RNA-sequencing of human organ development.[38] We found that there was a significantly higher developmentally-dynamic pleiotropy index representing genes with more repressed temporal expression only within the cerebellum within childhood-onset ataxia genes compared to control genes ($P = 0.042$) suggesting timing in expression in earlier development is important (Fig. 4A).

Next, we used measures of tissue-specific expression and co-expression derived from bulk RNA-sequencing data from 47 human tissues included in GTEx[25,39] and found significant differences in tissue-specific expression between the three ataxia gene lists (Supplementary Fig. 4). Cerebellar-specific expression appeared to be most associated with overlap-onset genes but was also an important feature of childhood-onset ataxia, with no statistically significant difference in the proportion of genes with cerebellar tissue-specific expression between the two groups. Similarly, measures of gene co-expression (module membership) highlighted brain-specific tissue co-expression but did not demonstrate significant differences between disease gene sets (Supplementary Table 5).

Using EWCE,[43] together with data on single-cell gene expression profiling of both mouse CNS and non-CNS tissue,[42] we studied the cell-type-specific expression of ataxia genes. Using all cell types, we demonstrated significant enrichment of CNS glia-specific expression in childhood-onset genes (FDR $P < 1 \times 10^{-7}$) and CNS neuron-specific expression within overlap-onset genes (FDR $P = 0.018$) (Fig. 4B). Focusing specifically on cerebellar cell types, we found that childhood-onset genes exhibited cell-type-specific expression within molecular layer interneurons (FDR $P = 0.036$) and overlap-onset ataxia genes showed cell-type-specific expression within cerebellar Purkinje cells (FDR $P = 0.018$) (Fig. 4C). Of interest, childhood-onset genes with the highest cerebellar molecular layer interneuron cell-type expression were associated with ataxia syndromes manifesting partially or fully with seizures (e.g. *KCNA1* and *RORA*). In contrast, overlap-onset genes driving expression in Purkinje cells were associated with ataxia-predominant syndromes including *ITPR1* and *CACNA1G* (Supplementary Fig. 5).

Childhood-onset ataxia genes generate protein products required for viability

Since gene expression levels are often poorly correlated with protein abundance and need not reflect protein-protein interactions,[57] we assessed all disease genes using the STRING and GO databases.[45,46] Consistent with the association of childhood-onset genes with multi-system disorders, we saw a statistically significant enrichment of GO terms associated with glycosylation pathways (e.g. GO:0009101 glycoprotein biosynthetic process) and cilia (GO:0005929 cilium) (Fig. 5A). In contrast, overlap-onset genes were enriched for nervous system-associated terms (e.g. GO:0043005 neuron projection) and ion channel biological processes (GO:0098662 inorganic cation transmembrane transport) (Fig. 5B). Adult-onset ataxia genes revealed no enriched GO terms, likely reflecting the small size of this gene list. Given that not all molecular and biological processes are captured by GO, we also analysed the number of protein-protein interactions per gene using data provided by the STRING database. This demonstrated a higher number of protein-protein interactions amongst genes in overlap-onset ataxia than those causing childhood-onset ataxia ($P = 0.018$).

We extended our analysis to consider functional assessments of gene-protein products by using FUSIL categorisation.[49] FUSIL classifies genes based on cross-species integrated measures of essentiality.[49] When comparing childhood- with overlap-onset genes, we saw a significantly higher proportion of genes within the VP category (denoting genes where LoF mutations are viable, but with an abnormal phenotype in mice) in the overlap-onset group ($P = 0.024$) (Fig. 5C). However, we noted a significantly higher proportion of CL genes, where LoF mutations cause cellular lethality; $P = 2.467 \times 10^{-4}$, and DL genes, where LoF mutations cause developmental lethality; $P = 1.896 \times 10^{-17}$ amongst childhood-onset genes compared with non-ataxia genes. This reflects the developmental importance of childhood-onset genes.

A modified testing strategy for variants and STRs in known ataxia genes could improve diagnostic yield

Although we found significant differences in some genic properties across the ataxia gene sets, there were also surprising commonalities, as exemplified by the high density of non-pathogenic STRs in both childhood- and adult-onset genes. To summaries data across all 294 genic features analysed and account for redundancies in correlation between features (Supplementary Fig. 7), we used recursive feature elimination to identify the 84 most relevant annotations (Supplementary Table 6). Using these salient features, we found that while there were small clusters of childhood- and overlap-onset genes, most ataxia genes did not cluster and adult-onset genes showed no distinct classification on UMAP (Fig. 6A and Supplementary Fig. 6). This suggested that adult-onset genes are not easily distinguishable from childhood-onset genes using the genic features assessed, and that the age-of-onset division does not appear to be reflected in functional genomic annotation. This finding has important implications for genetic testing strategies. More specifically, it would suggest that genes currently considered to cause childhood-onset ataxia could be expected to also cause adult-onset disease, and STR expansion may be a more common pathogenic mechanism than expected, potentially operating in combination with known pathogenic variants.

We explicitly tested the first of these two hypotheses by measuring the burden of potentially pathogenic variants among childhood-onset genes in adult-onset ataxia patients recruited to the 100 000 Genomes Project and vice versa. The demographics of each of the groups assessed (childhood- and adult-onset ataxia cases and controls) are listed in Supplementary Table 7. Interestingly, this demonstrated a significantly higher burden of rare potentially pathogenic LoF variants in the typically childhood-onset ataxia genes *SACS* and *POLR3A* [both odds ratio (OR) 7.038, FDR $P = 0.032$, 95% CI:1.948–24.512] and rare, predicted potentially pathogenic variants (Exomiser score >0.8) in *POLR3A* (OR 3.461, FDR $P = 0.018$, 95% CI: 1.655–6.840) amongst individuals with adult-onset ataxia (Fig. 6B) (Supplementary Table 8). This was reflected on interrogation of the clinical records and the final diagnostic-grade genetic reports of ataxia patients within the 100 000 Genomes Project. We identified eight cases where

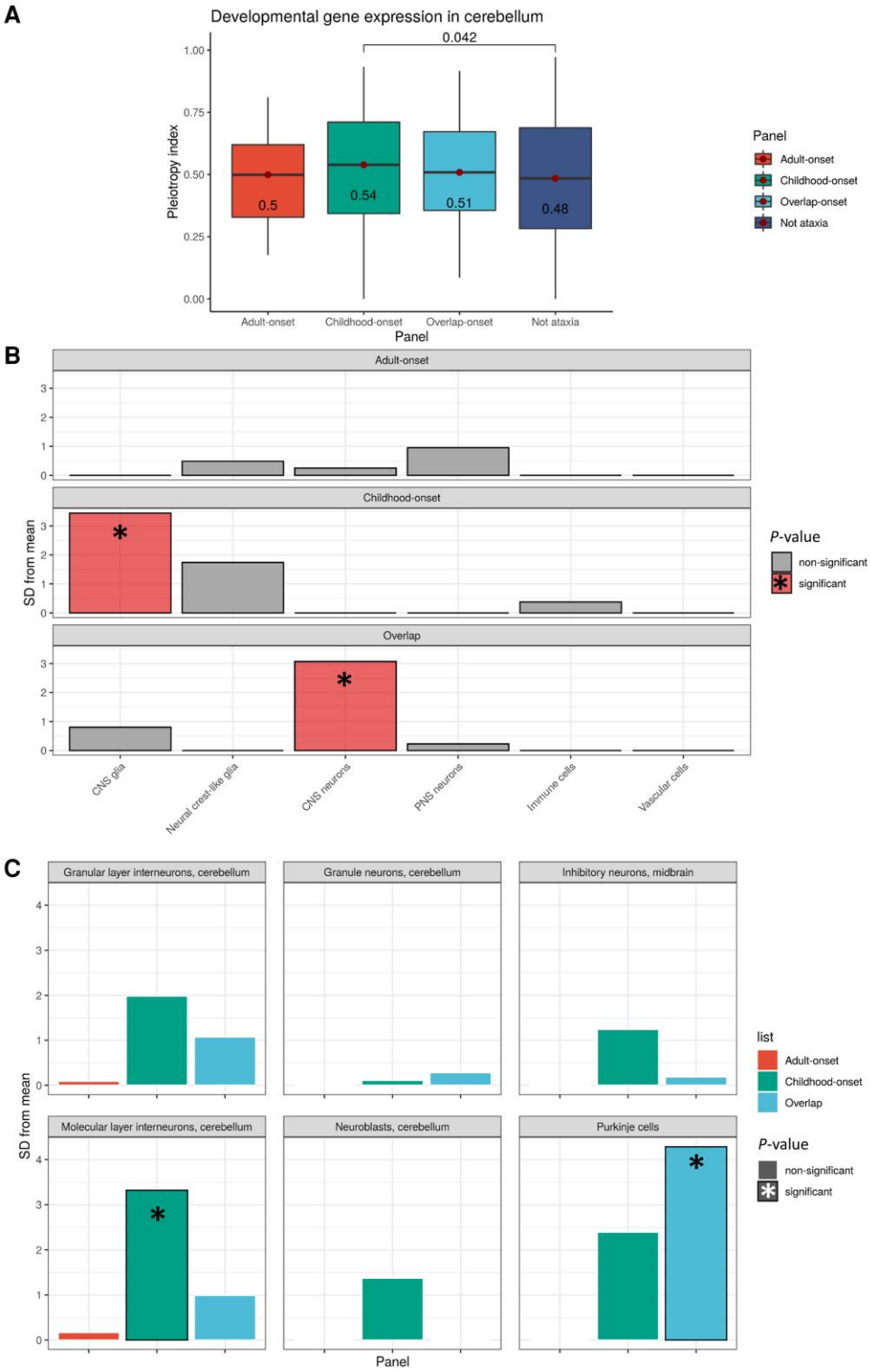

Figure 4 Comparison of markers of dynamic gene expression. (**A**) Comparison of dynamic specificity indices (where 1 represents repressed temporal expression) in the cerebellum across different gene sets. Only significant Wilcoxon rank sum *P*-values (<0.05) are given for pairwise comparisons above the square brackets. The numbers within the box of the box and whisker plots represent the median values for that genic feature. The corresponding horizontal lines on the boxplots represents the lowest quartile, median, and upper quartile of the data. Further results are presented in Supplementary Table 5. Expression-weighted cell-type enrichment results showing significantly-enriched cell-type-specific expression across two levels of cell information. (**B**) Enrichment of ataxia-associated genes (three sets of different ages of onset) in cell types from mouse single-cell RNA-sequencing data was determined using EWCE. Standard deviations (SD) from the mean indicate the distance of the mean expression of the target list from the mean expression of the bootstrap replicates. Significance at $P < 0.05$ after correction for multiple testing with the Benjamini–Hochberg method over all cell types and the three gene panels was used. CNS refers to central nervous system and PNS refers to the peripheral nervous system. (**C**) Enrichment of ataxia-associated genes within cerebellar-specific cell types of the Karolinska dataset are shown with significant *P*-values noted by an asterisk and column outline.

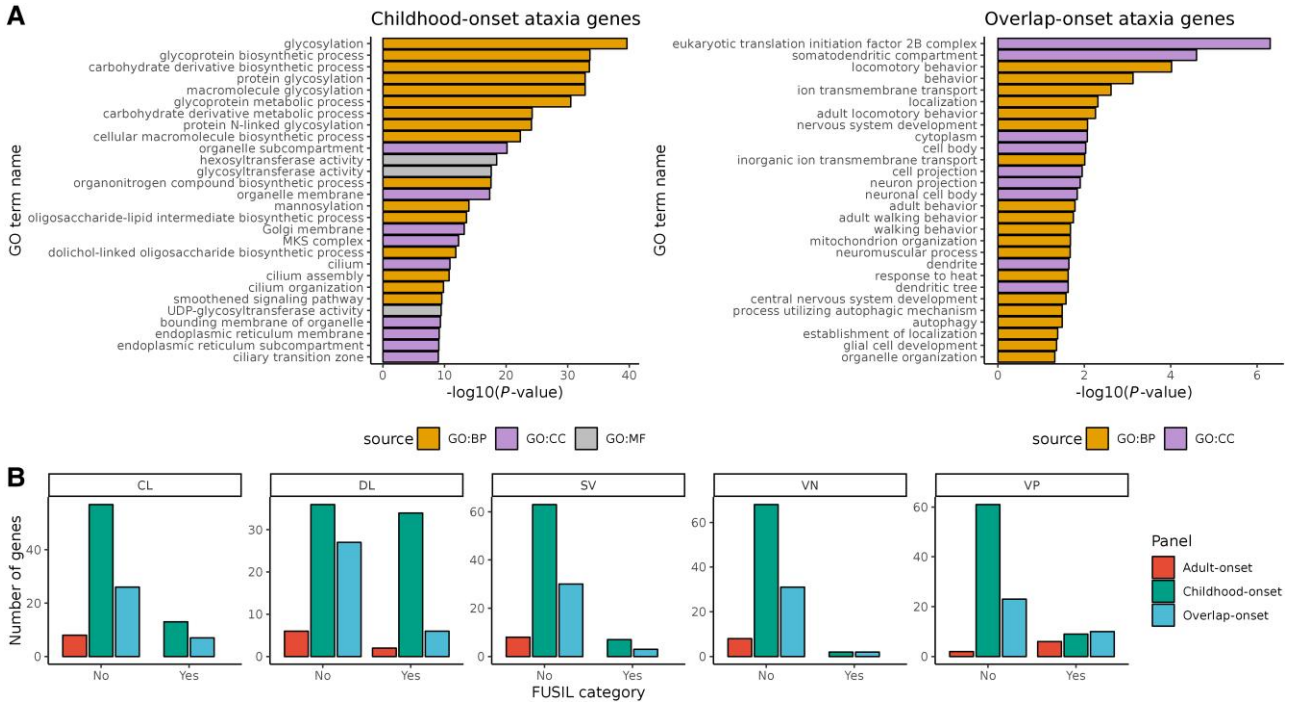

Figure 5 Enriched GO terms for childhood-onset hereditary ataxia genes (top 25 shown only) and for overlap-onset ataxia genes with associated g: SCS-corrected P-values from gene set analysis. (**A**) The source depicts the GO of the biological domain with respect to three aspects: biological process (BP), cellular component (CC) and molecular function (MF). (**B**) Bar plots of the number of genes across gene panels are shown for each FUSIL category[49]: CL, DL, SV, VP and VN. 'Yes' refers to genes that fulfil criteria for that particular FUSIL category. 'No' refers to genes that do not fulfil criteria for that particular FUSIL category.

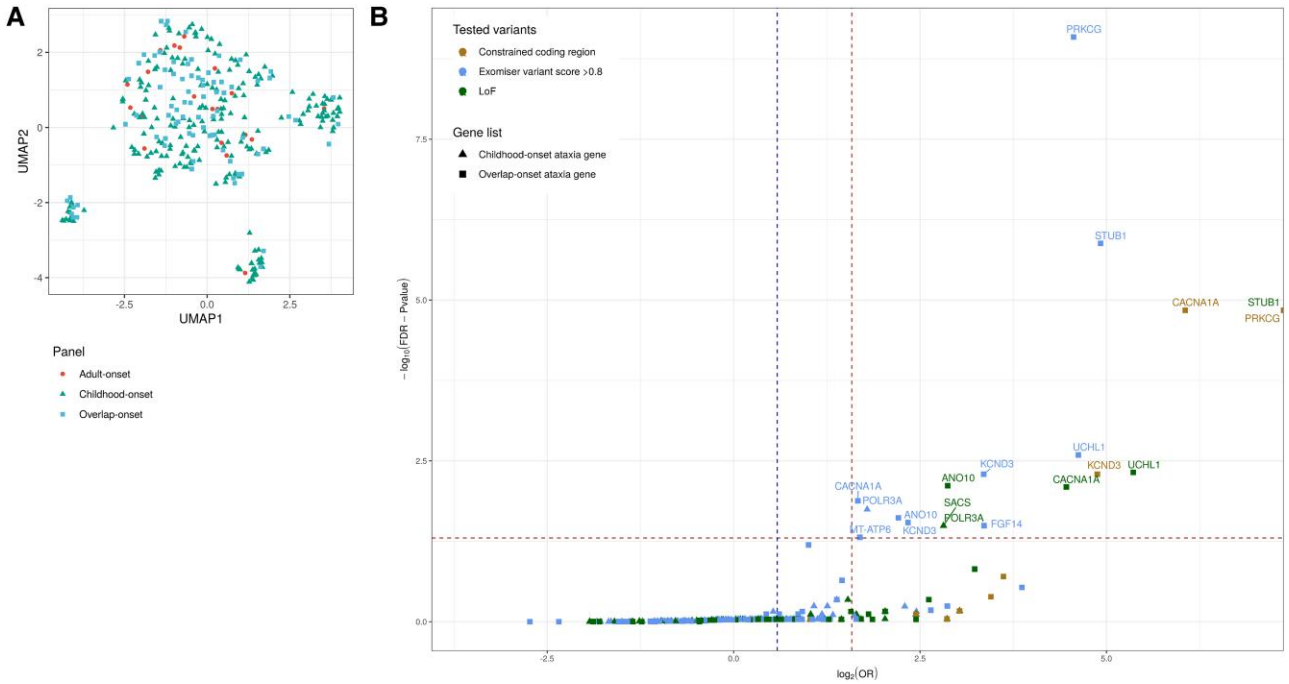

Figure 6 UMAP of all ataxia genes partitioned by age-of-onset using 84 selected genic features from recursive feature elimination. (**A**) Results are for each of the three gene panels shown in Supplementary Fig. 6. (**B**) Volcano plot depicting results from rare variant burden analysis using 100 000 Genomes Project participants. In this gene-based burden testing analysis, we assessed the number of adult-onset ataxia patients carrying variants in childhood-onset genes filtered for rare variants within constrained coding regions, or with an Exomiser score of >0.8 to indicate likely pathogenicity, or LoF variants. We also tested this burden of rare variants in overlap-onset ataxia genes which are expected to be significantly enriched within adult-onset patients. The OR is the odds of enrichment of a variant in cases over controls (defined in the 'Materials and methods' section). Benjamini–Hochberg method was used to correct for multiple testing; an overall FDR-adjusted P-value of 0.05 (horizontal dashed line) was used for claiming significant gene-disease associations taking into account the total number of case-control gene burden tests under all four scenarios analysed. The vertical dashed line on the left of the plot represents an OR of 1.5 and the other dashed vertical line represents an OR of 3.

despite the patient presenting with adult-onset disease (range 22 to 47 years of age), diagnoses were made based on a pathogenic variant typically associated with childhood-onset ataxia (in *POLR3A, SACS, PMM2, WFS1*) (Supplementary Table 9). This suggests that we may be missing diagnoses in adult patients by not screening for genes typically described to be associated with childhood-onset disorders. Furthermore, an *a priori* assumption of the age-of-onset of an associated gene may cause bias when assessing the variant pathogenicity, especially if the existing described cases are rare.

Furthermore, we assessed the potential importance of STR expansion across a wider set of ataxia genes. More specifically, given the high STR density within ataxia genes (mean of 59.5 STRs per adult-onset gene and 55.8 STRs per childhood-onset gene), we investigated whether ataxia genes not currently thought to cause disease through repeat expansions had higher repeat sizes amongst individuals with ataxia. To assess this, we studied all 190 naturally-occurring 5′UTR and exonic STRs located in 100 genes, of which 116 STRs were classified within genes of typical childhood-onset. We then compared the distribution of STR sizes in ataxia cases (either childhood- or adult-onset) and controls recruited in the 100 000 Genomes Project. We did not match the groups for sex or ethnicity (demographics reported in Supplementary Table 10) due to the large sample sizes needed to detect rare variants. Furthermore, there were no statistically significant differences between the estimated repeat sizes between female and male participants (Wilcoxon $P = 0.88$) and amongst the different ethnicities (Kruskal–Wallis $P = 0.15$). Using this approach, we found a trend for a higher maximum number of repeats in the top 1% of repeat sizes in patients presenting with both childhood- and adult-onset ataxia compared with controls (Fig. 7). This demonstrated that expansions of apparently benign STRs are associated with disease as evidenced by the trend for higher repeat size in cases over controls even at naturally-occurring STRs. While this does not definitively demonstrate the pathogenicity of specific STRs, it suggests that these STRs could be contributing to disease susceptibility potentially by operating as modifiers of disease risk. Consequently, screening for expansions at naturally-occurring STR sites in established ataxia genes not known to cause disease

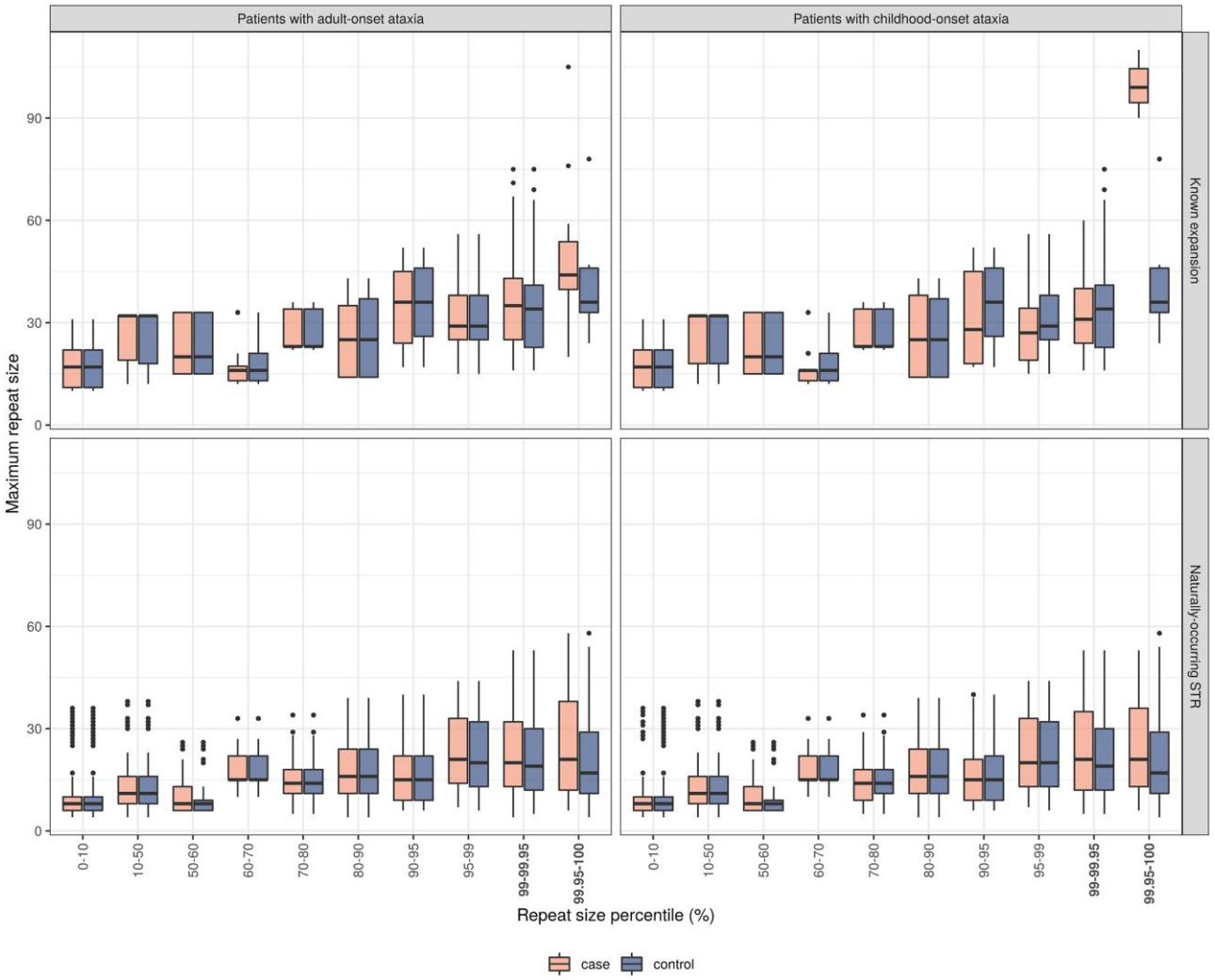

Figure 7 Maximum allelic repeat sizes estimated using ExpansionHunter at STR loci annotated by HipSTR reference database in adult patients presenting with ataxia ($n = 1629$) and patients presenting with childhood-onset ataxia ($n = 553$) compared with controls ($n = 6078$) defined as unrelated non-neurological probands recruited under the Rare Disease arm of the 100 000 Genomes Project. The repeat sizes were estimated across STRs in which repeat expansions are known to cause ataxia ('Known expansion') and across naturally-occurring STRs, not currently known to be associated with disease. The corresponding horizontal lines on the boxplots represents the lowest quartile, median and upper quartile of the data.

through repeat expansions should be prioritized in unsolved cohorts.

Discussion

In this study, we used 294 functional genomic annotations to study genes causally linked to hereditary ataxia, with the aim of identifying commonalities and differences in gene properties. We provide evidence to show that first, there is an unexpectedly high STR density within childhood-onset ataxia genes suggesting that we may be missing pathogenic repeat expansions as a disease mechanism within this cohort. Secondly, adult-onset ataxia genes cannot easily be distinguished from those causing childhood-onset disease when all genic features are considered, suggesting a common underlying biology. Thirdly and most importantly, diagnostic yield for hereditary ataxia could be improved by using a common screening gene panel and by analysing STRs in existing ataxia genes not known to harbour pathogenic expansions as demonstrated using WGS data from ~2000 individuals with ataxia in the 100 000 Genomes Project.

By applying a systems biology approach without *a priori* assumptions, we found that genes associated with ataxia have many common characteristics. Genic features such as increased transcript count did not differentiate between childhood- and adult-onset ataxia, but rather distinguished between ataxia and control genes. Unexpectedly, even STR-based measures of genic complexity were unable to distinguish between adult- and childhood-onset ataxia, as evidenced by a higher STR density in genes that currently harbour no known pathogenic repeats, including in childhood-onset ataxia genes. Furthermore, the higher eSTR density within childhood-onset ataxia genes highlighted a potential regulatory role for these elements in modulating gene expression and disease severity. The high density of intronic STRs within ataxia genes may also be a source for potentially missed pathogenic loci. Thus, our analysis provides support for the utility of screening for pathogenic repeat expansions at all ages of onset. With this in mind, we noted that five previously-undiagnosed children presenting with ataxia in the 100 000 Genomes Project were found to have repeat expansion disorders typically associated with adult-onset disease, for which they had not been initially screened.[16] Similarly, we provided data to suggest that even STRs not known to be associated with pathogenic repeat expansions tended to have higher repeat sizes in individuals with both adult- and childhood-onset ataxia as compared to controls. This highlights naturally-occurring STRs within known ataxia genes as candidates for screening in an unsolved cohort. Intriguingly, we also found a higher density of other repetitive elements such as LINE/L1 in both adult- and childhood-onset ataxia genes than controls. This finding is of interest in highlighting the potential role of LINE/L1 dysregulation in pathogenesis, in line with a recent RNA-sequencing study that demonstrated cerebellar LINE/L1 activation in driving ataxia phenotype in mouse models.[58]

Our analyses highlighted not only the potential for common pathogenic mechanisms for disease across hereditary ataxia, but also common biology. We found no clear separation of genes causing childhood- and adult-onset ataxia using UMAP to visualize genes based on a recursively-selected set of features. Most strikingly, adult-onset ataxia genes were highly scattered suggesting that there is a spectrum of disease and that genes causing childhood-onset ataxia have the potential to cause adult-onset disease. We found support for this hypothesis through the demonstration of a significantly higher burden of rare potentially pathogenic variants in the conventionally childhood-onset genes defined within OMIM:[21] *SACS* and *POLR3A* amongst individuals with adult-onset ataxia enrolled in the 100 000 Genomes Project. Furthermore, we recognize that for *SACS*, clinical reports exist of adult-onset disease presentations.[59] On review of diagnostic genetic reports, we found eight patients who presented with adult-onset ataxia being diagnosed with variants typically associated with childhood-onset disease within the 100 000 Genomes Project. However, there has not been a systematic approach to gene screening, and these findings would suggest that we could be missing diagnoses in adult-onset patients by either not screening for, or not prioritizing pathogenic variants that have been typically described in children. This may be consistent with what many clinicians suspect, highlighting limitations that could potentially arise from the use of restricted gene panels in diagnostic screening.

While our study has largely uncovered significant commonalties across childhood- and adult-onset ataxia, there were also differences. Given the cellular complexity of the cerebellum, a region that harbours more than half of all brain neurons,[60] this is unsurprising, and is already supported by evidence that different cerebellar cell types are affected in different forms of ataxia.[61] Consistent with previous gene expression analyses[62] and the cerebellum being the most frequent site of neurodegeneration on neuropathological examination of ataxia,[61] we found that ataxia genes exhibited cerebellar-specific expression and co-expression. Interestingly, we demonstrated a preferential CNS glial enrichment within childhood-onset genes compared to a CNS neuronal enrichment within overlap-onset genes. We found that CNS glial enrichment was driven by ciliopathy genes, which are known to affect early brain development, mediated in part through radial glial progenitor cells.[63] This finding was also reflected in glycosylation and cilia GO enrichment within childhood-onset genes. As would be expected, childhood-onset genes not only exhibited dynamically-repressed expression in the cerebellum when compared with control genes, but a higher proportion were also classified as cellularly or developmentally lethal. Genes causing both adult- and childhood-onset ataxia exhibited Purkinje cell-type-specific expression, driven by genes associated with 'pure' ataxia syndromes; a finding supported by previous analyses using mouse[64] and human transcriptomics.[62] The expression of childhood-onset ataxia genes was also enriched within inhibitory GABAergic molecular layer interneurons,[65] driven by genes associated with ataxia-epilepsy syndromes. This supports possible additional function for these interneurons in epileptogenesis, mirroring the role of dentate basket cells in temporal lobe epilepsy.[66]

Although this study highlighted key biological information, such as the contribution of particular cell types, and the potential importance of specific pathogenic processes such as STR expansions, our analyses were limited by the quality and availability of existing functional genomic annotation. For example, information regarding dynamic gene expression taken from human tissues[38] is limited by the resolution of bulk RNA-sequencing data as not all genes are successfully sequenced at all timepoints, thus dynamic gene expression quantification is limited. Likewise, we are also limited by the accuracy of input gene lists and age-of-onset classification which changes regularly with new cases of hereditary ataxia described. We attempted to overcome this problem by using a range of different resources across the four main gene feature categories and disease gene panels. ExpansionHunter is limited in sizing repeats much larger than the read length,[16] and thus our STR analysis may have missed large outliers of repeat sizes in the top percentiles. Despite this technical limitation, our results would

likely still stand given that the current trend for large repeat sizes may therefore be an underestimate of the true repeat size.

In summary, this study suggests that childhood- and adult-onset ataxia exist across a spectrum of disease rather than as distinct entities; a finding which would be hard to generate from clinical experience given that there are many hereditary ataxia genes each accounting for a very small number of cases. This core observation has important clinical implications for the classification of hereditary ataxia by age-of-onset. But, most importantly, it suggests that the diagnostic rate for hereditary ataxia would be expected to increase simply by removing the age-of-onset partition, and through modified screening for repeat expansions in naturally-occurring STRs within known ataxia genes, in effect treating these regions as candidate pathogenic loci.

Acknowledgements

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 other participants collected by the National Health Service as part of their care and support. We thank all participants and healthcare teams at the thirteen NHS Genomic Medicine Centres in England, where around 5000 multidisciplinary staff enrolled patients to the 100 000 Genomes Project. Patients were also enrolled to the 100 000 Genomes Project from Scotland by the Scottish Genomes Project, and across Wales and Northern Ireland.

Funding

Z.C. was supported by a clinical research fellowship from the Leonard Wolfson Foundation. A.T. is a UK Medical Research Council clinician scientist (MR/S006753/1). J.H. was supported by the UK Dementia Research Institute, which receives its funding from DRI Ltd., funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. J.H. has also been funded by the UK Medical Research Council (MR/N026004/1), Wellcome Trust (202903/Z/16/Z), Dolby Family Ventures and National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre. J.B. is supported through the Science and Technology Agency, Fundación Séneca, CARM, Spain (research project 00007/COVI/20). M.R. was supported by the UK Medical Research Council through the award of a Tenure Track Clinician Scientist Fellowship (MR/N008324/1). The views expressed are those of the authors and not necessarily those of the NHS, the funding bodies or the Department of Health and Social Care.

Competing interests

The authors report no competing interests

Supplementary material

Supplementary material is available at *Brain* online.

Appendix I

Genomics England Research Consortium

Full details are provided in the Supplementary material.

Collaborators: John C. Ambrose, Prabhu Arumugam, Marta Bleda, Freya Boardman-Pretty, Jeanne M. Boissiere, Christopher R. Boustred, Clare E. H. Craig, Anna de Burca, Andrew Devereau, Greg Elgar, Rebecca E. Foulger, Pedro Furió-Tarí, Joanne Hackett, Dina Halai, Angela Hamblin, Shirley Henderson, James Holman, Tim J. P. Hubbard, Rob Jackson, Louise J. Jones, Melis Kayikci, Lea Lahnstein, Kay Lawson, Sarah E.A. Leigh, Ivonne U.S. Leong, Javier F. Lopez, Fiona Maleady-Crowe, Joanne Mason, Michael Mueller, Nirupa Murugaesu, Chris A. Odhams, Daniel Perez-Gil, Dimitris Polychronopoulos, John Pullinger, Tahrima Rahim, Pablo Riesgo-Ferreiro, Tim Rogers, Mina Ryten, Kevin Savage, Kushmita Sawant, Afshan Siddiq, Alexander Sieghart, Damian Smedley, Alona Sosinsky, William Spooner, Helen E. Stevens, Alexander Stuckey, Razvan Sultana, Simon R. Thompson, Carolyn Tregidgo, Emma Walsh, Sarah A. Watters, Matthew J. Welland, Eleanor Williams, Katarzyna Witkowska, Suzanne M. Wood, Magdalena Zarowiecki.

References

1. Boycott KM, Rath A, Chong JX, *et al*. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet*. 2017;100:695-705.
2. Németh AH, Kwasniewska AC, Lise S, *et al*. Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. *Brain*. 2013;136(Pt 10):3106-3118.
3. Rexach J, Lee H, Martinez-Agosto JA, Németh AH, Fogel BL. Clinical application of next-generation sequencing to the practice of neurology. *Lancet Neurol*. 2019;18:492-503.
4. Boycott KM, Hartley T, Biesecker LG, *et al*. A diagnosis for all rare genetic diseases: The horizon and the next frontiers. *Cell*. 2019; 177:32-37.
5. Warman Chardon J, Beaulieu C, Hartley T, Boycott KM, Dyment DA. Axons to exons: The molecular diagnosis of rare neurological diseases by next-generation sequencing. *Curr Neurol Neurosci Rep*. 2015;15:64.
6. Galatolo D, Tessa A, Filla A, Santorelli FM. Clinical application of next generation sequencing in hereditary spinocerebellar ataxia: Increasing the diagnostic yield and broadening the ataxia-spasticity spectrum. A retrospective analysis. *Neurogenetics*. 2018;19:1-8.
7. Bird TD, *et al*. Hereditary ataxia overview. In: Adam MP Ardinger HH and Pagon RA, eds. *Genereviews®*. University of Washington; 1993.
8. Jayadev S, Bird TD. Hereditary ataxias: Overview. *Genet Med*. 2013;15:673-683.
9. Ruano L, Melo C, Silva MC, Coutinho P. The global epidemiology of hereditary ataxia and spastic paraplegia: A systematic review of prevalence studies. *Neuroepidemiology*. 2014;42:174-183.
10. Harding AE. Classification of the hereditary ataxias and paraplegias. *Lancet*. 1983;321:1151-1155.
11. Klockgether T, Mariotti C, Paulsonp HL. Spinocerebellar ataxia. *Nat Rev Dis Primers*. 2019;5:24.
12. Sullivan R, Yau WY, O'Connor E, Houlden H. Spinocerebellar ataxia: An update. *J Neurol*. 2019;266:533-544.
13. Smedley D, Smith KR, Martin A, *et al*. 100,000 Genomes pilot on rare-disease diagnosis in health care — Preliminary report. *N Engl J Med*. 2021;385:1868-1880.

14. Martin AR, Williams E, Foulger RE, *et al*. Panelapp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*. 2019;51:1560-1565.
15. Brandsma R, Verschuuren-Bemelmans CC, Amrom D, *et al*. A clinical diagnostic algorithm for early onset cerebellar ataxia. *Eur J Paediatr Neurol*. 2019;23:692-706.
16. Ibañez K, Polke J, Hagelstrom RT, *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*. 2022;21:234-245.
17. Gagliano SA. It's all in the brain: a review of available functional genomic annotations. *Biol Psychiatry*. 2017;81:478-483.
18. Reynolds RH, Hardy J, Ryten M, Gagliano Taliun SA. Informing disease modelling with brain-relevant functional genomic annotations. *Brain*. 2019;142:3694-3712.
19. Li X, Li Z, Zhou H, *et al*. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52:969–9983.
20. Buniello A, MacArthur JAL, Cerezo M, *et al*. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012.
21. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian inheritance in man (OMIM). *Hum Mutat*. 2000;15:57-61.
22. Cunningham F, Allen JE, Allen J, *et al*. Ensembl 2022. *Nucleic Acids Res*. 2022;50(D1):D988-D995.
23. Köhler S, Carmody L, Vasilevsky N, *et al*. Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res*. 2019;47(D1):D1018-d1027.
24. Adler A, Kirchmeier P, Reinhard J, *et al*. Phenodis: A comprehensive database for phenotypic characterization of rare cardiac diseases. *Orphanet J Rare Dis*. 2018;13:22.
25. Botía JA, Guelfi S, Zhang D, *et al*. G2p: Using machine learning to understand and predict genes causing rare neurological disorders. *bioRxiv*. [Preprint] doi: 10.1101/288845
26. Benson G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27:573-580.
27. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14:590-592.
28. Fotsing SF, Margoliash J, Wang C, *et al*. The impact of short tandem repeat variation on gene expression. *Nat Genet*. 2019;51:1652-1659.
29. Smit AF, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0 at http://www.repeatmasker.org/.
30. Wheeler TJ, Clements J, Eddy SR, *et al*. Dfam: A database of repetitive DNA based on profile hidden markov models. *Nucleic Acids Res*. 2012;41(D1):D70-D82.
31. Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443.
32. Fadista J, Oskolkov N, Hansson O, Groop L. Loftool: A gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics*. 2016;33:471-474.
33. Rackham OJL, Shihab HA, Johnson MR, Petretto E. Evotol: A protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Res*. 2015;43:e33.
34. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;9:e1003709.
35. Siepel A, Bejerano G, Pedersen JS, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034-1050.
36. di Iulio J, Bartha I, Wong EHM, *et al*. The human noncoding genome defined by genetic diversity. *Nat Genet*. 2018;50:333-337.
37. Chen Z, Zhang D, Reynolds RH, *et al*. Human-lineage-specific genomic elements are associated with neurodegenerative disease and APOE transcript usage. *Nat Commun*. 2021;12:2076.
38. Cardoso-Moreira M, Halbert J, Valloton D, *et al*. Gene expression across mammalian organ development. *Nature*. 2019;571:505-509.
39. Consortium G, Ardlie KG, Deluca DS, *et al*. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348:648-660.
40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2006;8:118-127.
41. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC bioinformatics*. 2008;9:559.
42. Zeisel A, Hochgerner H, Lönnerberg P, *et al*. Molecular architecture of the mouse nervous system. *Cell*. 2018;174:999-1014.e22.
43. Skene NG, Grant SGN. Identification of vulnerable cell types in Major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front Neurosci*. 2016;10:16.
44. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25-29.
45. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330-D338.
46. Szklarczyk D, Gable AL, Lyon D, *et al*. STRING V11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613.
47. Quinodoz M, Royer-Bertrand B, Cisarova K, Di Gioia SA, Superti-Furga A, Rivolta C. DOMINO: Using machine learning to predict genes associated with dominant disorders. *The Am J Human Genet*. 2017;101:623-629.
48. Reimand J, Arak T, Adler P, *et al*. G:Profiler—A web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res*. 2016;44(W1):W83-W89.
49. Cacheiro P, Muñoz-Fuentes V, Murray SA, *et al*. Human and mouse essentiality screens as a resource for disease gene discovery. *Nat Commun*. 2020;11:655.
50. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28:26.
51. R package "corrplot": Visualization of a Correlation Matrix [Internet]. 2017. https://github.com/taiyun/corrplot
52. McInnes L, Healy J, Melville JJ. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. [Preprint] doi: 1802.03426
53. Caufield M, Davies J, Dennys M, *et al*. The national genomics research and healthcare knowledgebase. *figshare Dataset*. https://doi.org/10.6084/m9.figshare.4530893.v5
54. Smedley D, Jacobsen JO, Jäger M, *et al*. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc*. 2015;10:2004-2015.
55. Dolzhenko E, Deshpande V, Schlesinger F, *et al*. Expansionhunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*. 2019;35:4754-4756.
56. 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*. 2021;108:764-785.
57. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*. 2012;13:227-232.

58. Takahashi T, Stoiljkovic M, Song E, *et al*. LINE-1 activation in the cerebellum drives ataxia. *Neuron*. 2022;110:3278–3287.e8.
59. Baets J, Deconinck T, Smets K, *et al*. Mutations in SACS cause atypical and late-onset forms of ARSACS. *Neurology*. 2010;75:1181-1188.
60. Wang VY, Zoghbi HY. Genetic regulation of cerebellar development. *Nat Rev Neurosci*. 2001;2:484-491.
61. Seidel K, Siswanto S, Brunt ER, den Dunnen W, Korf HW, Rüb U. Brain pathology of spinocerebellar ataxias. *Acta Neuropathol*. 2012;124:1-21.
62. Bettencourt C, Ryten M, Forabosco P, *et al*. Insights from cerebellar transcriptomic analysis into the pathogenesis of ataxia. *JAMA Neurol*. 2014;71:831-839.
63. Guo J, Higginbotham H, Li J, *et al*. Developmental disruptions underlying brain abnormalities in ciliopathies. *Nat Commun*. 2015;6:7857.
64. Peng J, Sheng AL, Xiao Q, *et al*. Single-cell transcriptomes reveal molecular specializations of neuronal cell types in the developing cerebellum. *J Mol Cell Biol*. 2019;11:636-648.
65. Brown AM, Arancillo M, Lin T, *et al*. Molecular layer interneurons shape the spike activity of cerebellar purkinje cells. *Sci Rep*. 2019;9:1742.
66. Zhang W, Buckmaster PS. Dysfunction of the dentate basket cell circuit in a rat model of temporal lobe epilepsy. *J Neurosci*. 2009; 29:7846-7856.