Genetics inMedicine

www.nature.com/gim

Check for updates

BRIEF COMMUNICATION

Systematic analysis of short tandem repeats in 38,095 exomes provides an additional diagnostic yield

Bart P. G. H. van der Sanden[1,2], Jordi Corominas[1], Michelle de Groot[1], Maartje Pennings[1], Rowdy P. P. Meijer[1], Nienke Verbeek[3], Bart van de Warrenburg[4], Meyke Schouten[1], Helger G. Yntema[1], Lisenka E. L. M. Vissers[1,2], Erik-Jan Kamsteeg[1,6]✉ and Christian Gilissen[1,5,6]✉

PURPOSE: Expansions of a subset of short tandem repeats (STRs) have been implicated in approximately 30 different human genetic disorders. Despite extensive application of exome sequencing (ES) in routine diagnostic genetic testing, STRs are not routinely identified from these data.

METHODS: We assessed diagnostic utility of STR analysis in exome sequencing by applying ExpansionHunter to 2,867 exomes from movement disorder patients and 35,228 other clinical exomes.

RESULTS: We identified 38 movement disorder patients with a possible aberrant STR length. Validation by polymerase chain reaction (PCR) and/or repeat-primed PCR technologies confirmed the presence of aberrant expansion alleles for 13 (34%). For seven of these patients the genotype was compatible with the phenotypic description, resulting in a molecular diagnosis. We subsequently tested the remainder of our diagnostic ES cohort, including over 30 clinically and genetically heterogeneous disorders. Optimized manual curation yielded 167 samples with a likely aberrant STR length. Validations confirmed 93/167 (56%) aberrant expansion alleles, of which 48 were in the pathogenic range and 45 in the premutation range.

CONCLUSION: Our work provides guidance for the implementation of STR analysis in clinical ES. Our results show that systematic STR evaluation may increase diagnostic ES yield by 0.2%, and recommend making STR evaluation a routine part of ES interpretation in genetic testing laboratories.

Genetics in Medicine (2021) 23:1569–1573; https://doi.org/10.1038/s41436-021-01174-1

INTRODUCTION

Short tandem repeats (STRs), also called microsatellites, constitute ~3% of the human genome and are scattered throughout.[1,2] STRs vary in size but are commonly defined as tandemly repeated nucleotide motifs of 2–12 base pairs in length.[3] Expansions of a subset of STRs have been implicated in approximately 30 different human genetic disorders. A large group of repeat expansions cause different forms of spinocerebellar ataxias, while others cause Huntington disease (OMIM 143100), fragile X syndrome (OMIM 300624), or myotonic dystrophies (e.g., OMIM 160900 and OMIM 602668).[4]

Next-generation sequencing (NGS) has proven to be of great diagnostic value in clinical practice.[5,6] Over the last few years, several different tools for genome-wide genotyping of STRs from short read sequencing data, and mainly genome sequencing data, were developed.[3,7–13] Despite the extensive application of exome sequencing (ES) in routine diagnostic genetic testing and the many STR detection studies being published, STRs were not routinely identified from these data, and large-scale assessments of the diagnostic potential from detecting STRs from ES data have not yet been performed. In this study, we assess the clinical utility of detecting STR expansions in ES data for patients with (suspected) rare genetic disorders based on a cohort of 38,095 clinical exomes. We provide guidance for the implementation of an STR detection workflow in routine diagnostic clinical ES analysis with minimal additional analytic burden, and we show that it increases the ES diagnostic yield in our cohort of 38,095 clinical exomes.

MATERIALS AND METHODS

Tool selection on validation cohort

Based on the ability to call STRs from short read NGS data, we initially considered three different STR detection tools: STRetch,[8] ExpansionHunter,[9] and GangSTR.[7] We evaluated the performance using a validation cohort of 11 patients with known pathogenic repeat expansion allele(s) in six different genes (Supplementary Table 1). These expansion alleles were previously identified using conventional genetic testing by fragment length analysis of polymerase chain reaction (PCR) or repeat-primed PCR (RP-PCR) fragments. The three tools were locally installed using their GitHub installation page and were run according to the developer's instructions.

Selection of STR sites

For this study, we selected 24 well-described disease-causing STR loci from literature (Supplementary Table 2). Subsequently, after running ExpansionHunter on the validation cohort, we removed the loci without sequencing coverage ($n = 5$), which resulted in a list of 19 STR loci of interest (Supplementary Table 2).

Samples

In this study, two cohorts were analyzed, using ES data obtained as part of routine diagnostic workup of patients with rare diseases suspected to have genetic origin (https://gdnm.nl/ for an overview of all 30 clinically and genetically heterogeneous disorders included). All DNA samples were sequenced on an Illumina HiSeq 2000 instrument in combination with Agilent version 4 enrichment kit, or on Illumina HiSeq 4000 combined with an Agilent version 5 enrichment kit. More

[1]Department of Human Genetics, Radboud university medical center, Nijmegen, The Netherlands. [2]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands. [3]Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [4]Department of Neurology, Donders Institute for Brain, Cognition and Behaviour, Radboud university medical center, Nijmegen, The Netherlands. [5]Radboud Institute for Molecular Life Sciences, Nijmegen, The Netherlands. [6]These authors contributed equally: Erik-Jan Kamsteeg, Christian Gilissen. ✉email: erik-jan.kamsteeg@radboudumc.nl; christian.gilissen@radboudumc.nl

than 95% of samples had at least 75× median coverage across the enrichment kit targets. To optimize our STR detection approach, we used a de-identified ES cohort of 2,867 patients presenting with a movement disorder of suspected genetic origin, further referred to as the movement disorders cohort. For this cohort we only looked at STRs in genes associated with movement disorders (Supplementary Table 3). Subsequently, we detected STRs from our entire ES cohort of 38,095 anonymized samples, for which we used all selected STR loci of interest (Supplementary Table 2). The latter cohort also included the movement disorders cohort, but identical calls between the two cohorts were discarded to prevent double counting. In addition, this cohort contained both patients as well as unaffected individuals. We analyzed ES data as previously reported.[14]

Analysis of the cohorts

Genotypes of each expansion locus were then compared to the locus specific expansion thresholds, which were set according to existing literature (Supplementary Table 2). Genotypes were classified in three different groups: normal range, gray zone range, and pathogenic range. The gray zone is defined as either uncertain whether there is an actual expansion, whether they are pathogenic, or whether there is incomplete penetrance.

Confirmation of STRs

Alignments of the calls exceeding the gray zone threshold were manually curated for sequencing coverage, read mapping quality, and taking into account the specifics of the particular repeat structure, using the GraphAlignmentViewer.[15] All likely aberrant expansion alleles, except for expansion alleles in *TCF4* (See Supplementary Table 8), were tested using PCR and fragment length analysis using GeneScan.[16] When this only detected one normal allele size or indicated a pathogenic expansion allele, we also performed repeat-primed (RP-PCR) to confirm the GeneScan result or to make sure we did not miss a larger allele due to allelic dropout of normal PCR. RP-PCR for the *DMPK* and *CNBP* genes was performed as described.[16] For other loci, primer binding sites and test conditions are available upon request.

Genetic diagnosis of patients

For the confirmed expansion alleles in the movement disorders cohort, the short clinical descriptions of the patients were compared to the OMIM phenotype of the concomitant disorders by a trained clinical laboratory geneticist.

RESULTS

ExpansionHunter demonstrated the highest sensitivity by detecting 10 (91%) of 11 known pathogenic repeat expansion alleles compared to GangSTR (73%) and STRetch (18%) (Supplementary Figure 1; Supplementary Table 4). We decided to use ExpansionHunter in our study, based on its higher sensitivity and specificity combination for the detection of aberrant STR lengths in our validation cohort.

Systematic analysis of STRs in movement disorders cohort

To optimize our approach for detecting aberrant expansion alleles from ES data using ExpansionHunter in a diagnostic setting, we first applied the tool to the 2,867 patients in the movement disorders cohort, which yielded 91 aberrant STR alleles in 88 (3.0% of 2,867) different patients (Fig. 1a, c; Supplementary Table 5 and 6). Manual curation allowed us to discard 53 expansion alleles, leaving 38 alleles with a likely aberrant STR length (Fig. 1c, Supplementary Figure 3). From the 38 samples with a likely repeat expansion, validation by PCR and/or RP-PCR technologies confirmed the presence of aberrant expansion allele(s) in 13 (34% validation rate) (Fig. 1c, d; Supplementary Table 5 and 6). For the 53 alleles that were discarded after manual inspection, we also performed validations to exclude false negative calls, but all of these samples showed repeat sizes within the normal range based on PCR and fragment length analysis.

Because of the 25 false positives and 68% specificity of the manual curation of the alignment graphs, we reanalyzed the graphs of the false positive calls and found that these graphs showed lower quality sequence flanking the STRs, more misaligned bases, and lower read depth (Supplementary Figure 2 and 3). Therefore, to optimize our workflow, we decided to improve the manual curation process of the STRs called by ExpansionHunter in the remainder of the full cohort.

Aberrant STR calls in remainder of full cohort

Using the optimized approach for aberrant STR length detection, we systematically analyzed our full diagnostic ES cohort (Fig. 1b). Among the 38,095 ES samples, we identified 1,130 aberrant expansion alleles in 1,117 different samples (2.9% of 38,095). Applying our stricter manual curation yielded 167 likely aberrant expansion alleles in 167 different samples (0.4% of 38,095) (Supplementary Table 7 and 8). Validation by PCR and/or RP-PCR confirmed the presence of 93 aberrant expansion alleles (56%) (Fig. 1e, Supplementary Table 8). For the 93 confirmed expansion alleles, 48 were above the pathogenic repeat threshold, while 45 were in the gray/premutation zone. Notably, the application of our improved manual curation significantly increased the validation rate for the likely aberrant expansion alleles compared to those achieved for the movement disorders cohort ($P = 0.009$, two-sided Fisher's exact test) (Supplementary Table 6 and 8).

Diagnostic implications of confirmed STR events

For 13 of the 2,867 samples from the movement disorders cohort, an aberrant expansion allele was detected by ExpansionHunter and confirmed by PCR and RP-PCR (Fig. 1c). Of these, one expansion allele was in *ATXN1*, four in *ATXN3*, three in *ATXN7*, two in *HTT*, two in *NOP56* and one in *PPP2R2B* (Fig. 1d), causing spinocerebellar ataxia 1, 3, 7, 36, and 12 and Huntington disease respectively (OMIM 164400, OMIM 109150, OMIM 164500, OMIM 614153, OMIM 604326, OMIM 143100). Given the diverse genetic and clinical composition of our cohort, we evaluated whether the genetic findings correspond with the phenotype of the patient. For 7 of the 13 samples the genotype was compatible with the phenotypic description and led to a genetic diagnosis for these patients (Figs. 1c, 2; Supplementary Table 9). For the other six confirmed expansion alleles, the genotype was not compatible with the clinical phenotype description. Based on this we estimate the additional diagnostic yield in this cohort of prescreened movement disorder patients to be 0.2% (7/2,867) (95% CI [0.093–0.45%], Bayesian binomial test).

For the 93 confirmed expansions in the full cohort, the genotype could not be compared with the phenotypic description, since the samples in this cohort were anonymized to prevent the detection of incidental findings.

DISCUSSION

In this study, we report a systematic analysis of STRs in a large clinical ES data set.

Notwithstanding the good diagnostic results we obtained with ExpansionHunter, we also show that adding a manual curation step to the workflow can be of great importance for decreasing the false discovery rate (Supplementary Figure 4). In the future, the implementation of a robust quality control score for STRs may make this curation obsolete, which would lower the risk of interpersonal interpretation and increase the diagnostic robustness. However, the number of ExpansionHunter calls that were manually curated was only 91 for the movement disorders cohort (1 in 33 samples) and 1,130 for the full cohort (1 in 33 samples) and this means that the burden of additional analysis in routine diagnostics can be minimized, with only ~3% of the samples requiring dedicated follow-up. The additional computational

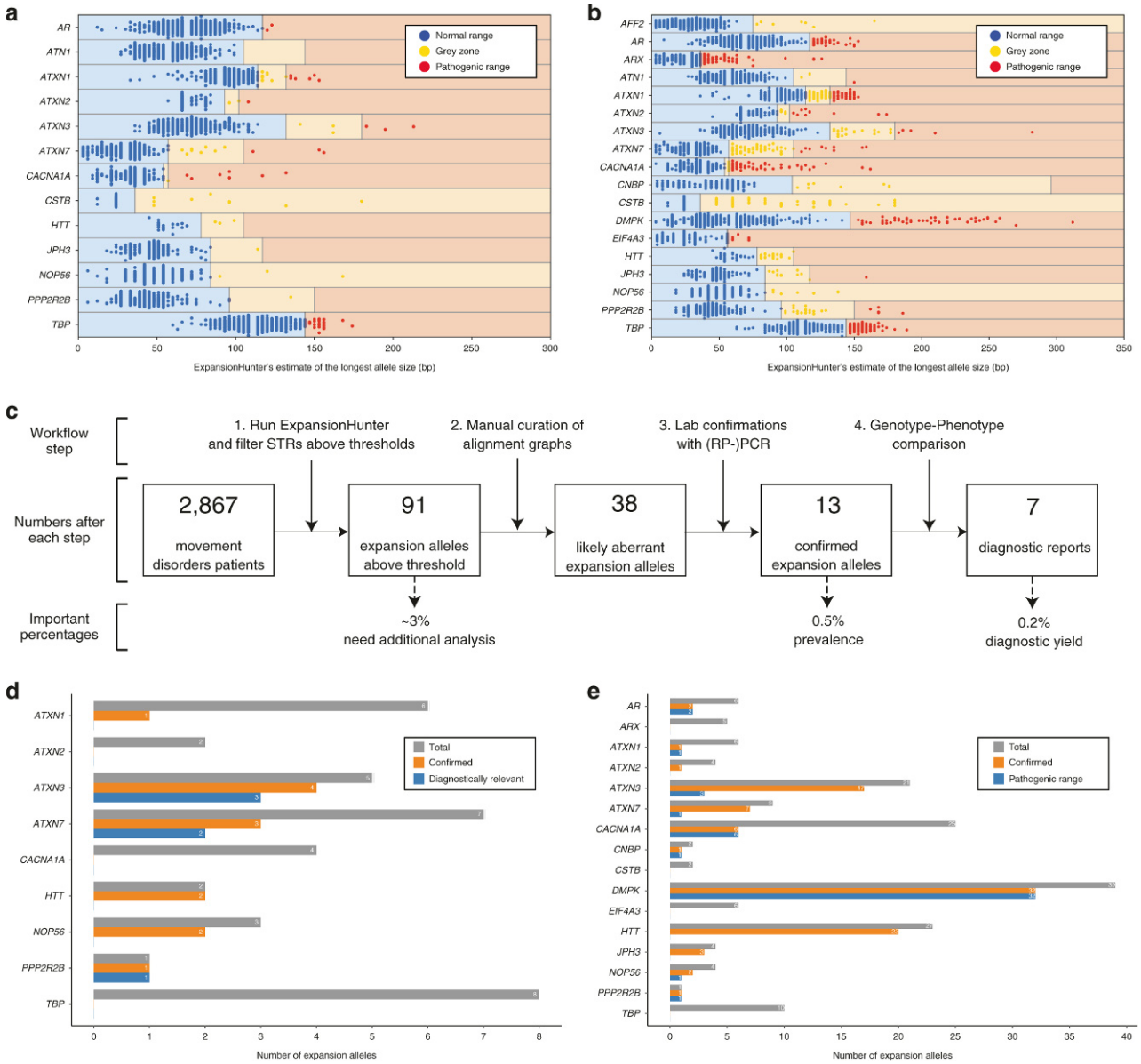

Fig. 1 Results of short tandem repeat (STR) size detection in two cohorts of exome sequencing (ES) samples and confirmations of alleles that exceeded the thresholds. a Allele sizes for 2,867 ES samples of patients with movement disorders. Only movement disorder associated genes were analyzed. All samples in the movement disorders cohort with an STR length above the thresholds are listed in Supplementary Table 5. **b** Allele sizes for all loci of interest for the full cohort of 38,095 ES samples. All samples in the full cohort with an STR length above the thresholds are listed in Supplementary Table 7. **c** The workflow that we used to analyze the ES data of the 2,867 patients presenting with a movement disorder. The different steps are projected in the top part. The numbers of STR expansion alleles that were left after each step are presented in the middle and in the bottom part three concluding percentages are displayed. **d** Validation rate per gene for the movement disorders cohort. **e** Validation rate per gene for all 38,095 clinical ES samples. RP-PCR repeat-primed polymerase chain reaction.

burden to run ExpansionHunter is only 10–15 minutes per exome on a single compute core. Therefore, this should not hamper diagnostic implementation. Complementary experimental validations will however remain necessary to confirm detected expansion alleles.

The prevalence of aberrant expansion alleles was higher in our movement disorders cohort compared to our diagnostic ES cohort (0.5% vs. 0.2% respectively), likely due to the fact that the ES cohort also contained samples of unaffected individuals (e.g., parents). In addition, it is known that STRs play a major role in the disease etiologies of movement disorders (mainly spinocerebellar ataxias), and therefore this cohort may be enriched for STR expansions.[3] Still, the diagnostic yield for the movement disorders cohort was only 0.2% (7/2,867), which is likely due to the clinical prescreening of this group of patients for repeat expansions.

Seven cases in the movement disorders cohort with a confirmed STR expansion received a genetic diagnostic report, which will help them provide insight into the prognosis of the disorder and help patients and their relatives understand the cause of the disorder that is segregating in their family. For the other six confirmed expansion alleles, the genotype was not compatible with the clinical phenotype description. This is partly due to the thresholds we used for filtering the ExpansionHunter calls. Patients with a small or premutation-sized repeat expansion allele may not present with characteristic clinical features of the specific disorder yet, such as for example repeat expansions in the

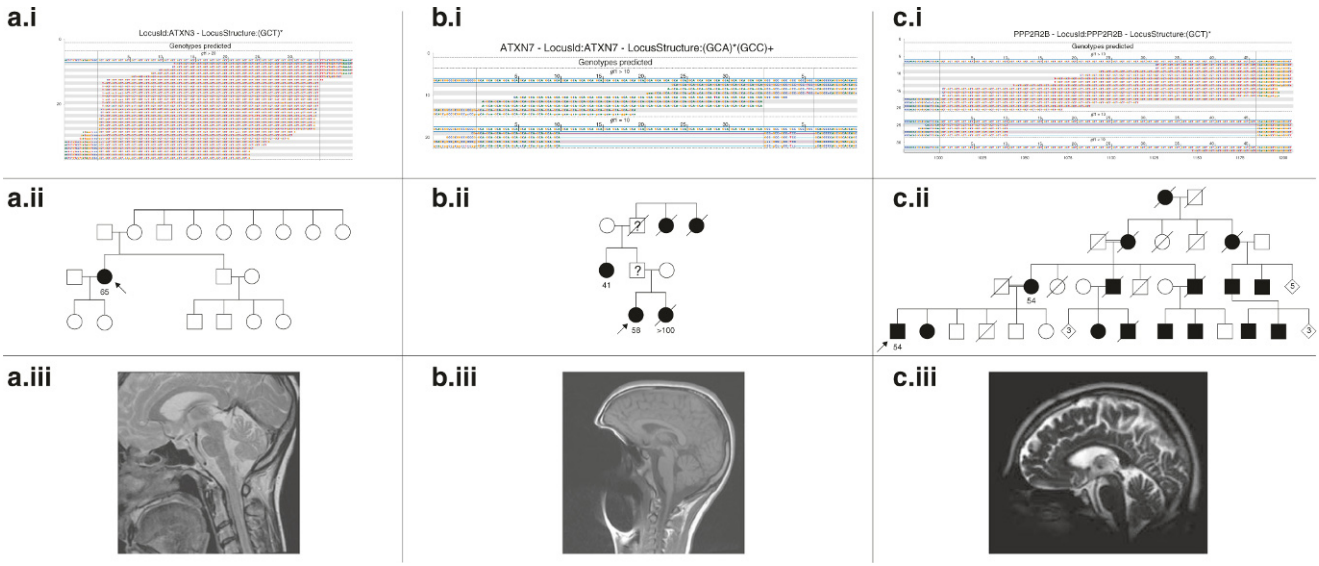

Fig. 2 Three examples of clinical compatibility between genotype and the phenotypic description of two patients from the movement disorders cohort that received a genetic diagnosis based on the short tandem repeat (STR) detection workflow, and one validation sample to compare to. For these three patients, the ExpansionHunter findings resulted in a genetic diagnosis or confirmed the previously detected expansion allele, showing the importance of routine diagnostic STR detection from ES data. The three patients are indicated with an arrow in their respective family tree. For the samples with (previously) known genotypes at the specific STR locus, the longest STR allele size in repeat units, as determined by (RP-)PCR, is depicted underneath the corresponding family tree symbol. **a.i** ExpansionHunter alignment graph depicting the detected STR expansion in *ATXN3* in this patient. **a.ii** The family tree of the family showing that the index is the only family member with an established spinocerebellar ataxia phenotype. **a.iii** The sagittal T2 MRI brain scan showing atrophy of the cerebellar vermis. **b.i** ExpansionHunter alignment graph for the second patient showing the repeat expansion in *ATXN7* for one of the validation samples. Running ExpansionHunter on this sample confirmed the presence of the previously identified expansion allele in *ATXN7* in this sample. **b.ii** The family tree of this family visualizing the individuals affected by a SCA7 phenotype. The index in this family has ataxia and cone dystrophy. In addition, her sister died at 14 months of age due to the most severe end of the disease spectrum characterized by intellectual disability, hypotonia, nephrosis, and retinitis pigmentosa. **b.iii** The sagittal T1 MRI brain scan showing absence of clear cerebellar vermian atrophy at the age of onset of the disorder for the index, despite the compatibility between the genotype and phenotypic description. **c.i** ExpansionHunter alignment showing the repeat expansion in *PPP2R2B*. **c.ii** The family tree depicting a likely dominantly inherited ataxia/tremor syndrome (in the individuals shown as solid symbols), despite the two consanguineous relations (double lines) that may have been misleading in the initial analyses. **c.iii** The sagittal T2 MRI brain scan shows mild cerebellar but also cerebral cortical atrophy.

DMPK gene, causing myotonic dystrophy type 1. This disorder shows extreme anticipation with varying clinical features for small premutation (36–49 repeats; no disease), medium (50–150 repeats, mild and/or late onset) or large allele sizes (>150 repeats; spectrum from late onset to congenital disease).[16] In agreement, the detected pathogenic *DMPK* alleles in our anonymized cohort were all medium. Genetic diagnostic laboratories may need to consider carefully about how to manage such findings.

DATA AVAILABILITY

Data and materials are available upon request.

CODE AVAILABILITY

ExpansionHunter script is available at: https://github.com/Illumina/ExpansionHunter. GangSTR script is available at: https://github.com/gymreklab/GangSTR. STRetch script is available at: https://github.com/Oshlack/STRetch.

Received: 23 October 2020; Revised: 29 March 2021; Accepted: 30 March 2021;
Published online: 12 April 2021

REFERENCES

1. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature.* **409**, 860–921 (2001).
2. Fungtammasan, A. et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res.* **25**, 736–749 (2015).
3. Tankard, R. M. et al. Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. *Am. J. Hum. Genet.* **103**, 858–873 (2018).
4. Walker, F. O. *Huntington's disease. Lancet.* **369**, 218–228 (2007).
5. Vissers, L. et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet. Med.* **19**, 1055–1063 (2017).
6. Gilissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
7. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
8. Dashnow, H. et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).
9. Dolzhenko, E. et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
10. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
11. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods.* **14**, 590–592 (2017).
12. Tang, H. et al. Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.* **101**, 700–715 (2017).
13. Halman, A. & Oshlack, A. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. *F1000Res.* **9**, 200 (2020).
14. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
15. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics.* **35**, 4754–4756 (2019).
16. Kamsteeg, E. J. et al. Best practice guidelines and recommendations on the molecular diagnosis of myotonic dystrophy types 1 and 2. *Eur. J. Hum. Genet.* **20**, 1203–1208 (2012).

ACKNOWLEDGEMENTS

We thank Michael Eberle and Egor Dolzhenko for kindly providing the Python code for the swimlane plots. We thank Ingrid Siegelaer and Monique Gerrits for helping

with the molecular confirmations of the *HTT* allele sizes. This project was financially supported by an Aspasia grant of the Dutch Research Council (015.014.066 to L.E.L.M.V.), a VIDI grant (917-17-353 to CG) and the NWO X-omics project (184.034.019 to CG). The aims of this study contribute to the Solve-RD project (to C.G. and L.E.L.M.V.) which has received funding from the European Union's Horizon 2020 research and innovation program (number 779257).

AUTHOR CONTRIBUTIONS

B.P.G.H.v.d.S.: Methodology, Project administration, Writing—original draft, Writing—review & editing. J.C.: Methodology, Software, Formal analysis, Investigation, Data curation. M.d.G.: Methodology, Software, Formal analysis, Investigation, Data curation. M.P.: Validation, Formal analysis. R.P.P.M.: Validation, Formal analysis. N.V.: Resources, Visualization. B.v.d.W.: Resources, Visualization. M.S.: Resources, Visualization. H.G.Y.: Writing—review & editing. L.E.L.M.V.: Writing—review & editing, Funding acquisition. E.-J.K.: Conceptualization, Methodology, Project administration, Writing—original draft, Writing—review & editing, Supervision. C.G.: Conceptualization, Methodology, Project administration, Writing—original draft, Writing—review & editing, Supervision, Funding acquisition.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS DECLARATION

Patient samples, together with a basic phenotype description were anonymized. Study was approved by the institutional review board "Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen" under number 2011/188. We received and archived consent for participation/publication from every individual whose data is included in this manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41436-021-01174-1.

Correspondence and requests for materials should be addressed to E.-J.K. or C.G.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.