Global prevalence of potentially pathogenic short-tandem repeats in an epilepsy cohort

Claudia Moreau[1], Jacques L. Michaud[2,3], Fadi F. Hamdan[2,4], Joanie Bouchard[1], Vincent Tremblay[1], Guy A. Rouleau[5,6], Berge A. Minassian[7,8], Patrick Cossette[9,10], Simon L. Girard[1]
For the Alzheimer's Disease Neuroimaging Initiative*

1. Department of Fundamental Sciences, University of Quebec in Chicoutimi, Chicoutimi, Canada
2. CHU Sainte-Justine Research Center, Montreal, Quebec, Canada
3. Department of Neurosciences and Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada
4. Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada
5. Montreal Neurological Institute, McGill University, Montreal, Canada
6. Department of Neurology and Neurosurgery, McGill University, Montreal, Canada
7. Department of Pediatrics, Hospital for Sick Children and University of Toronto, Toronto, Canada;
8. Department of Pediatrics, University of Texas Southwestern, Dallas, TX, USA
9. CHUM Research Center, Montreal, Canada
10. Department of Neurosciences, University of Montreal, Montreal, Canada

Corresponding author:
Simon Girard, PhD
Université du Québec à Chicoutimi
Département des sciences fondamentales, Office P4-2130
555, boulevard de l'Université
Chicoutimi (Québec) G7H 2B1
418 545-5011 ext 2595
1 800 463-9880 ext 2595 (toll free)
simon2_girard@uqac.ca

*Control data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

This study aims to decipher the role of short tandem repeats (STRs) of epilepsy patients. Whole genome short-read sequencing data of 752 epileptic patients and controls was used to look for known STR expansions associated to increased risk of neurodevelopmental diseases or epilepsy and to try to identify new STR expansions associated to an increased risk of developing epilepsy. Results show one hit of particular interest on ARX gene that could be causal for one EE patient. It is also shown that the risk threshold for many of the trinucleotide-repeat diseases fall in the tail of the distribution where it is not possible to distinguish between at risk and not at risk individuals. We also looked at new possible STRs associated to epilepsy in genes previously known to play a role in epilepsy and in whole genome genes, but further validation and development of new predicting tools are needed to be able to predict their association or their functional impact.

Keywords

Short tandem repeats, epilepsy, neurodevelopmental diseases

Introduction

Epilepsy affects ~3% of individuals; half of these cases start during childhood. Although monogenic forms of the disease have been reported(Beck *et al.*, 1994; Baulac *et al.*, 1999; Cossette *et al.*, 2002; Kalachikov *et al.*, 2002), they represent less than 2% of epilepsy cases. In the last decade, many groups have been working on different genetic and bio-statistic methods to better understand the complex genetic mechanisms underlying epilepsy(Koeleman, 2018). Recently, a large GWAS on epilepsy identified 16 loci associated with the disease, many of which were already known or suspected(The International League Against Epilepsy Consortium on Complex Epilepsies, 2018). Moreover, copy number variants were studied in the context of epilepsy(Olson *et al.*, 2014; Hamdan *et al.*, 2017; Monlong *et al.*, 2018). Polygenic risk scores have also been used to try to better understand the complexity of the disease(Leu *et al.*, 2019; Moreau *et al.*, 2020). Despite these efforts, there is still a substantial missing heritability component in epilepsy genetics(Thomas and Berkovic, 2014).

There is a growing evidence that short tandem repeats (STRs) affect gene expression(Fotsing *et al.*, 2019) and play a role in genetic disorders(Pellegrini *et al.*, 2012; Bolton *et al.*, 2013) notably in epilepsy(Lalioti *et al.*, 1997; Cen *et al.*, 2018; Ishiura *et al.*, 2018; Corbett *et al.*, 2019; Florian *et al.*, 2019). Because of their higher mutation rate, STRs offer a different level of resolution at which to study kinship and trait variations among individuals. However, until recently, most STR expansions were not detectable in next-generation short-read sequencing datasets because they exceeded the reads' length and therefore, were not included in genetic studies(Bahlo *et al.*, 2018). Fortunately, this has recently changed and such methods with accompanying software are now available(Dolzhenko *et al.*, 2017; Tang *et al.*, 2017; Tankard *et al.*, 2017; Dashnow *et al.*, 2018). Here we chose TREDPARSE for its high accuracy and to be able to compare our results with the 12,632 samples that were tested for STR expansion diseases in the original study.

In this study we will use TREDPARSE on whole genome short-read sequencing data of epileptic patients and controls to look for known STR expansions associated to increased risk of neurodevelopmental diseases or epilepsy and to try to identify new STR expansions associated to an increased risk of developing epilepsy.

Data and methods

This study was approved by the CHUM research Center (CRCHUM) ethics committee and written informed consent was obtained for all patients and controls.

Phenotyping of patients

The epilepsy cohort was composed of families with at least three affected individuals with Idiopathic Generalized Epilepsy (IGE), Non-Acquired Focal Epilepsy (NAFE) or epileptic encephalopathy previously collected in CHUM Research Center in Montreal and in the Hospital for Sick Children in Toronto as part of the Canadian Epilepsy Network (CENet) and diagnosed by neurologists. The clinical epilepsy phenotype is defined based on the Classification of the Epilepsy Syndromes established by the International League against Epilepsy (ILAE) (Berg *et al.*, 2010).

More specifically for NAFE, patients were at least five years of age and have experienced at least two unprovoked seizures in the six months prior to starting treatment, an MRI scan of the brain that did not demonstrate any potentially epileptogenic lesions, other than mesial temporal sclerosis.

For IGE, patients with clinical and EEG characteristics meeting the 1989 ILAE syndrome definitions for childhood absence epilepsy, juvenile absence epilepsy, juvenile myoclonic epilepsy, or IGE not otherwise specified. An MRI of the brain was not required for participation. All patients were at least four years of age at the time of diagnosis. In IGE, we also included patients with Jeavons syndrome, which is an idiopathic generalized form of reflex epilepsy characterized by childhood onset, unique seizure manifestations, striking light sensitivity and possible occurrence of generalized tonic-clonic seizures.

The Epileptic Encephalopathy (EE) patients were also recruited by the CENet group at three centers: the Sainte-Justine University Hospital Center in Montreal, the Toronto Western Hospital, and the Hospital for Sick Children in Toronto. They included subjects with diverse EE phenotypes. The criteria used for the selection of these individuals were as follows: (1) intractable epilepsy defined as an absence of response to two appropriate and well-tolerated anti-epileptic therapies (AEDs) over a 6-month period and an average of at least one focal, generalized tonic-clonic, myoclonic, tonic, atonic, or absence seizure or epileptic spasm per month during the period of poor control; (2) ID or global developmental delay; (3) absence of malformations or focal and multifocal structural abnormalities on brain MRI; and (4) absence of parental consanguinity and family history of epilepsy, ID, or autism in first-degree relatives. Most of the EE patients were sequenced in trios and EE parents are used as controls.

Libraries preparation and whole-genome sequencing

Samples were sequenced at Genome Quebec Innovation Center in Montreal. gDNA was cleaned using ZR-96 DNA Clean & ConcentratorTM-5 Kit (Zymo) prior to being quantified using the Quant-iTTM PicoGreen dsDNA Assay Kit (Life Technologies) and its integrity assessed on agarose gels. Libraries were generated using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's recommendations. Libraries were quantified using the Quant-iTTM PicoGreen dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument. The libraries were denatured in 0.05N NaOH and diluted to 8pM using HT1 buffer. The clustering was done on an Illumina cBot and the flowcell was ran on a HiSeq 2 500 for 2x125 cycles (paired-end mode) using v4 chemistry and following the manufacturer's instructions. A phiX library was used as control and mixed with libraries at 0.01 level. The Illumina control software used was HCS 2.2.58 and the real-time analysis program used was RTA v. 1.18.64. bcl2fastq v1.8.4 was used to demultiplex samples and generate fastq reads. The filtered reads were aligned to reference Homo_sapiens assembly b37. Each readset was aligned to creates a Binary Alignment Map file (.bam).

Control data

ADNI data used as controls in the present analysis was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Analyses

TREDPARSE software(Tang *et al.*, 2017) was used to identify STRs from .bam files of 752 patients (40 unclassified epilepsy, 187 IGE, 210 NAFE, 315 EE), 396 EE trio parents and 204 ADNI controls. We looked at the number of repeats in patients and controls at 29 trinucleotide-repeat diseases (table S1), seven STR expansions previously shown to be associated to an increased risk of epilepsy (table S2) and 8,732 human STRs(Willems *et al.*, 2017) in 365 genes previously associated to epilepsy (gene list in supporting text 1) as well as 131,330 STRs in 16,495 genome wide genes. Only STRs that were outside centromeres, telomeres and low mappability regions and STRs with at least 98% genotyped samples were kept.

To look at outlier patients' repeat expansions in epilepsy related genes and in whole genome genes, we first calculated interquartile range (IQR) in both controls' sets (EE parents and ADNI). Second, for each patient we measured the number of IQR above quantile 75 (q75). Finally, we kept only STRs for which : 1) there was at least one patient with 5, 10, 20, 30, 40 or 50 IQR above q75 and 2) specific to patients so that the ratio of the number of repeats of the highest control on the highest patient was less than 0.6. To see if the epilepsy patients have a higher burden of outlier and specific STRs in epilepsy genes than controls, we performed 1,000 permutations and performed the same analysis recording for each permutation the number of STRs that met the criteria. P-values were assessed by counting the number of permutations for which we had equal or more such STRs than in the real data.

Validation

We conducted in silico validation using WGS of 10 monozygotic twin pairs. For each trinucleotide-disease and epilepsy associated STR expansion, we looked at both alleles of twins from the same family to see if they had the same number of repeats or if they were at one or two repeats distance. We also compared the 20 alleles of both twins in the same family for 293,987 genome wide STRs. The twins have been sequenced on the same platform as our epileptic patients and EE controls.

Results

Using TREDPARSE we were able to identify several STRs in epilepsy patients and in both control groups. We first evaluated the distribution of the number of repeats of the EE, other epilepsy patients and controls for 29 trinucleotide-repeat diseases often associated to neurological conditions (Fig. 1). We observed for many of these diseases that the risk threshold seems to be at the tail of the distribution.

Table 1 presents the repeat diseases for which there are individuals above the risk threshold. We additionally compare in this table the prevalence or carrier frequency of our cohorts with the 12,632 individuals in the original TREDPARSE study(Tang *et al.*, 2017) and the known prevalence. The prevalence in our cohorts is almost always higher than the TREDPASRSE controls and the known prevalence. However, this might be partly due to the smaller sample size imprecision.

We then tested the distribution of the number of repeats of our patients and controls for seven STR expansions that have been shown to play a role in epilepsy (Fig. 2). While most of these expansions are large and exceeds the TREDPARSE detection threshold, we can see for ARX gene (causing EIEE1) one homozygote boy with his heterozygote mother being above the risk threshold. For the TNRC6A gene, we could only look at the first part of the huge and complex expansion described in (Ishiura *et al.*, 2018). Again, we can see that the risk threshold for the first part of the STR expansion is at the tail of the distribution.

To assess for the accuracy of the method given that most of our patients above the risk threshold fall in the tail of the distribution, we ran TREDPARSE on 10 monozygotic twins' pairs for 293,987 genome wide STRs. We can see in Fig. 3 that most STRs below 50 bp length are concordant for both twins of the 10 families. After 75 bp length, we observe a clear decrease in the number of allele matches.

The next step was to look for new STR aberrant expansions that could be associated with the epilepsy phenotype. First, we wanted to see if the patients had a higher burden of STR expansions in 365 genes previously associated to epilepsy. To do so, we measured IQR of controls and looked at the number of IQR above q75 for each patient in these genes. We restricted the results to STRs for which the highest control was much lower than the highest patient (ratio < 0.6) also to account for the uncertainty of the method especially in STRs longer than 125 bp (Fig. 3). Table 2 presents the STRs that meet these criteria for different IQR thresholds. To see if the

STR expansions burden in genes associated to epilepsy is higher than expected by chance in our patients, we performed 1,000 permutations and recorded for each one the number of STRs matching the criteria. The p-value is the number of permutations for which there was equal or more STRs matching the criteria than the real data.

Lastly, we wanted to see if it would be possible using this method to identify completely new STRs aberrant expansions that could play a role in epilepsy in 16,495 genome wide genes. For this analysis we took an IQR above q75 threshold of 50. Resulting STR locations are shown in table 3 and the phenotypes, functions or diseases associated to these genes are presented in table S5.

Discussion

First, we tested known STR expansion diseases associated to an increased risk of developing neurodevelopmental conditions as comorbidities and shared genetic mechanisms are frequent between these disorders and epilepsies(Pal *et al.*, 2016; Takumi and Tamada, 2018). Many of our epilepsy cases, EE and even controls are over the risk threshold for some of these expansion diseases (Fig. 1). We report in more details the STR expansions for which there were individuals above the risk threshold in table 1. Note that the carrier frequency in our samples is almost always higher than what was reported for 12,632 samples in (Tang *et al.*, 2017) and the known prevalence or carrier frequency, except for FRAXE. This might be because of the much smaller sample size in the present study. This might also be due to a population specific founder effect in a well-documented founder population, the Quebec province, where most of our patients and trio parents come from(Scriver, 2001), with the exception of the ADNI control group. In fact, some of these expansion diseases have already been demonstrated to be more frequent in Quebec (or in particular Quebec regions which are known for successive regional founder events) as a result of the founder effect(s), OPMD(Brais *et al.*, 1998), DM1(Yotova *et al.*, 2005) and FRDA(Bouchard *et al.*, 1979). For these later three, we do not report a higher prevalence in our patients, what could be explained by the regional distribution.

The higher prevalence in our samples might also be due to the risk thresholds that have not been tested in all populations and sequencing platforms and to the uncertainty of the method. There are three categories in the STR expansions where we see one or more individual(s) above the threshold (table 1): 1) The expansions for which our samples and the TREDPARSE controls are strikingly higher than the known prevalence (CCD, SCA1 and SCA17) 2) The ones where we and TREDPARSE controls have more or less the same prevalence (DM1, OPMD, SCA2 and SCA8) and 3) The expansions where we seem to have more samples over the risk threshold than the TREDPARSE controls: DM2 for which the prevalence is not known and FXTAS for which we have the same carrier frequency than the literature.

Now if we look at the distribution of the repeats for all of these in Fig. 1, we can see that for DM1, OPMD, FXTAS, SCA1, SCA17 and SCA2, the risk threshold is in the upper part of the distribution, not strikingly dissociated from the other samples who are also in the upper tail of the distribution, but right under the risk threshold. Having a closer look at the reproducibility of the TREDPARSE calls using 10 twin pairs (table S3), we see a clear tendency of the software accuracy to decrease while the length of the STR increases. With this in mind, we think that the most interesting hits in our patients are CCD for one IGE patient and DM2 for one EE patient.

STR expansions have also been described as being associated to an increased risk for different epilepsy types(Lalioti *et al.*, 1997; Cen *et al.*, 2018; Ishiura *et al.*, 2018; Corbett *et al.*, 2019; Florian *et al.*, 2019). Many of these STR expansions are huge and complex compared to the hg19 reference STR (STARD7, MARCH6, SAMD12), sometimes covering hundreds of extra repeats, which makes it very hard to study using TREDPARSE. Whenever possible (TNRC6A, CSTB, ARX, RAPGEF2), we looked at the hg19 reference STR to assess for rare outlier STR expansion in our patients (Fig. 2).

For the recessive ARX STR expansion, which is associated to epileptic encephalopathy, we see a homozygote EE boy (with his heterozygote mother) who is above the risk threshold. The recessive expansion is located on the X chromosome, which would explain why the heterozygote mother is not affected. Her son has been diagnosed with EE in neonatal period and was successfully treated with clonazepam. The EE was combined to microcephaly and degenerative process of white matter. The patient was negatively tested for the STXBP1 gene in single gene testing and we did not find any de novo variant for this patient in a study including 200 EE trios sequenced for whole genome(Hamdan *et al.*, 2017). Consequently, we believe that the present STR variant in the ARX gene is causal.

As for the TNRC6A gene, we could only look at the first part of the disease expansion(Ishiura *et al.*, 2018). In consequence, this portion alone of the complex STR expansion is probably not enough to explain the disease. Moreover, it falls again in the TREDPARSE detection limit as we can see in table S4 by the discordance of many twin pairs.

Given that for many of the repeat expansions described in Fig. 1 and 2, a change of one or two repeats can make a difference for categorizing an individual as at risk or not, we wanted to further test for the accuracy of the method to call exactly the number of repeats. We ran TREDPARSE on 10 monozygotic twins' pairs for 293,987 genome wide STRs and looked at the concordance between them. The line representing the proportion of STRs matching perfectly for the 20 alleles in each bin length drops rapidly if we match only the exact repeats' number (Fig. 3C), but if we look at the match ± two repeats (Fig. 3A), it starts to drop around 50 bp, and at 100 bp length, only half of the twins' alleles are matching at ± two repeats. This tells us that TREPARSE is accurate ± few repeats so that we can't say for sure if an individual is at risk based on one or two repeats difference, but it becomes possible if we have a more drastic change in the STR length between the at risk and not at risk individuals.

The next analysis is based on this assumption. We looked at STRs where we could see at least one outlier patient and for which the highest control was much lower than the highest patient (see methods). We expected to see a higher burden of such STRs in genes previously related to epilepsy. It turns out that we do not. For all IQR above q75 thresholds, the permutations p-values tell that the number of STRs observed in the real data is not higher than expected by chance (table 2). This might be due to the poor accuracy of the method when we get to very long STRs. This might also be due to another limitation of the STR detection method that sticks to known STRs present in the catalog since we may miss some STRs that are not yet described. However, it may also be because our patients have causal variants outside of the genes that have been previously associated to epilepsy.

To have a more global picture of outlier STRs specific to our patients, we applied the IQR 50 above q75 threshold to genome wide STRs in more than 16,000 genes. 12 STRs met these criteria (table 3). A brief summary of gene function(s) and associated disease(s) are presented for these genes in table S5. On 12 STRs, we found six genes that could be related to neurodevelopmental conditions or could act on relevant mechanisms for epilepsy such as ion binding (MYT1L, PARD3B, QTRT2, ADAMTS16, CDH18, CAPN3). Of course, these STRs would need to be validated and further predictions on pathogenicity would also need to be done before to make any statement on the association of these STRs with epilepsy.

Conclusion

In this study we performed a survey of STRs in epilepsy patients. We found interesting hits on already known neurodevelopmental disease causing STRs. Especially, we found one hit on ARX gene that could be causal for one EE patient. Importantly, we have shown that the risk threshold for many of the trinucleotide-repeat diseases would need more samples and/or more accurate methods and validation to be adjusted for each population or particular cases since many of them fall in the tail of the distribution where it is not possible to distinguish between at risk and not at risk individuals. We also looked at new possible STRs associated to epilepsy in genes previously identified to play a role in epilepsy and in whole genome genes, but further validation and development of new predicting tools are needed to be able to predict their association or their functional impact on epilepsy.

Data availability statement

Raw whole genome sequences of a subset of the epilepsy patients for which we have appropriate consent has been deposited in the European Genome-phenome Archive, under the accession code EGAS00001002825. The STR data calculated using TREDPARSE will be available upon request. The R scripts to generate figures and tables are available at https://bitbucket.org/claudia_moreau/str_epilepsy/src/master/.

Acknowledgments

We are thankful to Compute Canada/Calcul Québec for the access to storage and computing resources. We are extremely grateful to all patients and their families for participating in this research.

Data collection and sharing for ADNI control samples was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding information

This work was supported by funding from Genome Quebec / Genome Canada as well as from the CIHR (#420021). It was also made possible by Compute Canada resources allocation.

Competing interests

We declare no competing interests

Supplemental Material

Table S1: Description of STRs in trinucleotide-repeat diseases.

Table S2: Description of STRs that have been identified in epilepsy disorders.

Table S3: Number of twin pairs matching for each STR expansion disease.

Table S4: Number of twin pairs matching for each STR expansion related to an increased risk of epilepsy

Table S5: Information on genes identified in table 3

Supporting Text 1: 365 epilepsy related gene list

References

Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Research 2018

Baulac S, Gourfinkel-An I, Picard F, Rosenberg-Bourgin M, Prud'homme J-F, Baulac M, et al. A Second Locus for Familial Generalized Epilepsy with Febrile Seizures Plus Maps to Chromosome 2q21-q33. Am J Hum Genet 1999; 65: 1078–85.

Beck C, Moulard B, Steinlein O, Guipponi M, Vallee L, Montpied P, et al. A nonsense mutation in the α4 subunit of the nicotinic acetylcholine receptor (CHRNA4) cosegregates with 20q-linked benign neonatal familial convulsions (EBNI). Neurobiol Dis 1994; 1: 95–9.

Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, Van Emde Boas W, et al. Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009. Epilepsia 2010; 51: 676–85.

Bolton KA, Ross JP, Grice DM, Bowden NA, Holliday EG, Avery-Kiejda KA, et al. STaRRRT: A table of short tandem repeats in regulatory regions of the human genome. BMC Genomics 2013

Bouchard JP, Barbeau A, Bouchard R, Paquet M, Bouchard RW. A Cluster of Friedreich's Ataxia in Rimouski, Québec. Can J Neurol Sci / J Can des Sci Neurol 1979

Brais B, Bouchard JP, Xie YG, Rochefort DL, Chrétien N, Tomé FMS, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet 1998

Cen Z, Jiang Z, Chen Y, Zheng X, Xie F, Yang X, et al. Intronic pentanucleotide TTTCA repeat insertion in the SAMD12 gene causes familial cortical myoclonic tremor with epilepsy type 1. Brain 2018

Corbett MA, Kroes T, Veneziano L, Bennett MF, Florian R, Schneider AL, et al. Intronic ATTTC repeat expansions in STARD7 in familial adult myoclonic epilepsy linked to chromosome 2. Nat Commun 2019

Cossette P, Liu L, Brisebois K, Dong H, Lortie A, Vanasse M, et al. Mutation of GABRA1 in an autosomal dominant form of juvenile myoclonic epilepsy. Nat Genet 2002; 31: 184–9.

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: Detecting and discovering pathogenic short tandem repeat expansions. Genome Biol 2018

Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, Van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res 2017

Florian RT, Kraft F, Leitão E, Kaya S, Klebe S, Magnin E, et al. Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. Nat Commun 2019

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, et al. The impact of short tandem repeat variation on gene expression. Nat Genet 2019

Hamdan FF, Myers CT, Cossette P, Lemay P, Spiegelman D, Laporte AD, et al. High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. Am J Hum Genet 2017; 101: 664–85.

Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat Genet 2018

Kalachikov S, Evgrafov O, Ross B, Winawer M, Barker-Cummings C, Martinelli Boneschi F, et al. Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features. Nat Genet 2002; 30: 335–41.
Koeleman BPC. What do genetic studies tell us about the heritable basis of common epilepsy? Polygenic or complex epilepsy? Neurosci Lett 2018: 10–6.
Lalioti MD, Scott HS, Buresi C, Rossier C, Bottani a, Morris M a, et al. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature 1997; 386: 847–51.
Leu C, Stevelink R, Smith AW, Goleva SB, Kanai M, Ferguson L, et al. Polygenic burden in focal and generalized epilepsies. Brain 2019
Monlong J, Girard SL, Meloche C, Cadieux-Dion M, Andrade DM, Lafreniere RG, et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genet 2018
Moreau C, Rébillard R-M, Wolking S, Michaud J, Tremblay F, Girard A, et al. Polygenic risk scores of several subtypes of epilepsies in a founder population. Neurol Genet 2020
Olson H, Shen Y, Avallone J, Sheidley BR, Pinsky R, Bergin AM, et al. Copy number variation plays an important role in clinical epilepsy. Ann Neurol 2014; 75: 943–58.
Pal DK, Ferrie C, Addis L, Akiyama T, Capovilla G, Caraballo R, et al. Idiopathic focal epilepsies: The lost tribe. Epileptic Disord 2016
Pellegrini M, Renda ME, Vecchio A. Tandem repeats discovery service (TReaDS) applied to finding novel cis-acting factors in repeat expansion diseases. BMC Bioinformatics 2012
Scriver CR. Human Genetics : Lessons from Quebec Populations. Annu Rev Genomics Hum Genet 2001; 2: 69–101.
Takumi T, Tamada K. CNV biology in neurodevelopmental disorders. Curr Opin Neurobiol 2018
Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, et al. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet 2017
Tankard R, Bennett M, Degorski P, Delatycki M, Lockhart P, Bahlo M. Detecting tandem repeat expansions in cohorts sequenced with short-read sequencing data. Detect known repeat Expans with Stand Protoc next Gener Seq Towar Dev a single Screen test Neurol repeat Expans Disord 2017
The International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun 2018; 9: 5269.
Thomas RH, Berkovic SF. The hidden genetics of epilepsy-a clinically important new paradigm. Nat Rev Neurol 2014; 10: 283–92.
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
Yotova V, Labuda D, Zietkiewicz E, Gehl D, Lovell A, Lefebvre JF, et al. Anatomy of a founder effect: Myotonic dystrophy in Northeastern Quebec. Hum Genet 2005

Table 1: Number of STR expansions over the risk threshold by category of samples.

STR expansion	Controls (/100 000)	EE (/100 000)	Epilepsies (/100 000)	TREDPARSE controls(Tang *et al.*, 2017) (/100 000)	Known prevalence[a] (/100 000)
CCD	1 (167)	0	1 (229)	5 (40)	0.1
DM1	1 (167)	0	0	15 (119)	0.5-18.1
DM2	0	1 (317)	0	0	
FXTAS	6 (1,000)	1 (317)	2 (458)	2 (16)	(carrier freq) 300-500
OPMD	1 (167)	0	0	8 (63)	1
SCA1	13 (2,167)[b]	5 (1,587)[b]	5 (1,144)	26 (206)	1
SCA17	3 (500)	5 (1,587)	5 (1,144)	52 (412)	0.2
SCA2	0	0	1 (229)	4 (32)	1.5
SCA8	0	1 (317)	0	3 (24)	0.5

The TREDPARSE family numbers were taken from table 1(Tang *et al.*, 2017) on 12,632 samples. XLMR and SBMA are not reported in this table because the samples are heterozygote and these two expansion diseases are of recessive inheritance.

[a] Prevalence was taken in supplementary table S1 of (Tang *et al.*, 2017)

[b] Including two father-child pairs

Table 2: Number of STRs in epilepsy genes

IQR above q75	STRs in epi. genes	Permutations p-value
5	57	0.339
10	37	0.253
20	8	0.909
30	4	0.556
40	3	0.660
50	1	0.608

Number of STRs in 365 epilepsy genes (on a total of 8,732 STRs) that meets criteria for different IQR above q75 thresholds and ratio max control / max patient < 0.6. P-values are shown for 1,000 permutations.

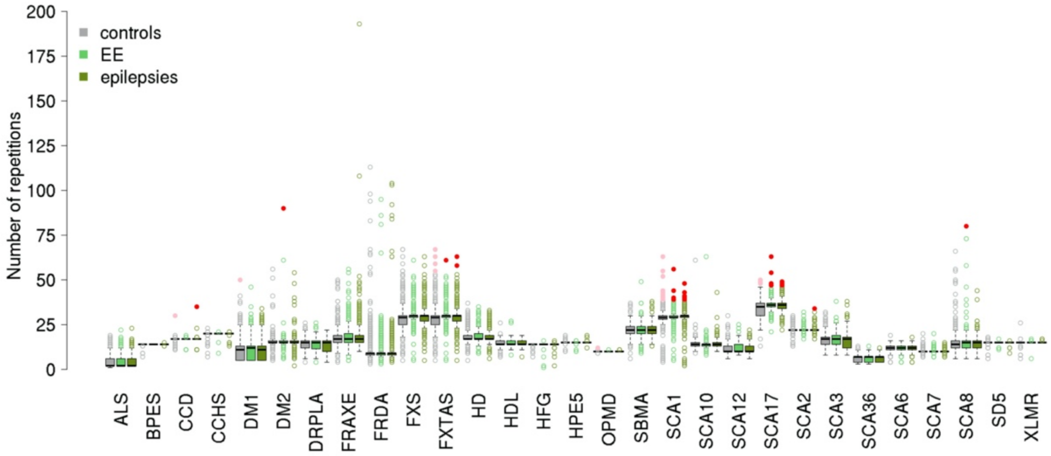
Table 3: STR expansions with 50 IQR above q75 in whole genome genes

Chr	Start	Max length (bp)	Hg19 ref length (bp)	Gene	Epilepsy phenotype
2	2299626	158	22	MYT1L	4 EE, 4 NAFE
2	205615313	704	64	PARD3B	1 IGE
3	99574943	875	40	CMSS1, FILIP1L	1 IGE
3	113802116	285	81	QTRT2	2 IGE
5	5151710	132	26	ADAMTS16	1 Jeavons
5	20574062	160	12	CDH18	1 NAFE

5	77575610	130	16	AP3B1	1 EE
7	105368301	132	14	ATXN7L1	1 Jeavons
7	137631454	158	22	CREB3L2	4 NAFE
15	42690855	132	24	CAPN3	1 NAFE
16	75015460	130	16	WDR59	2 IGE
19	7119637	132	14	INSR	1 EE

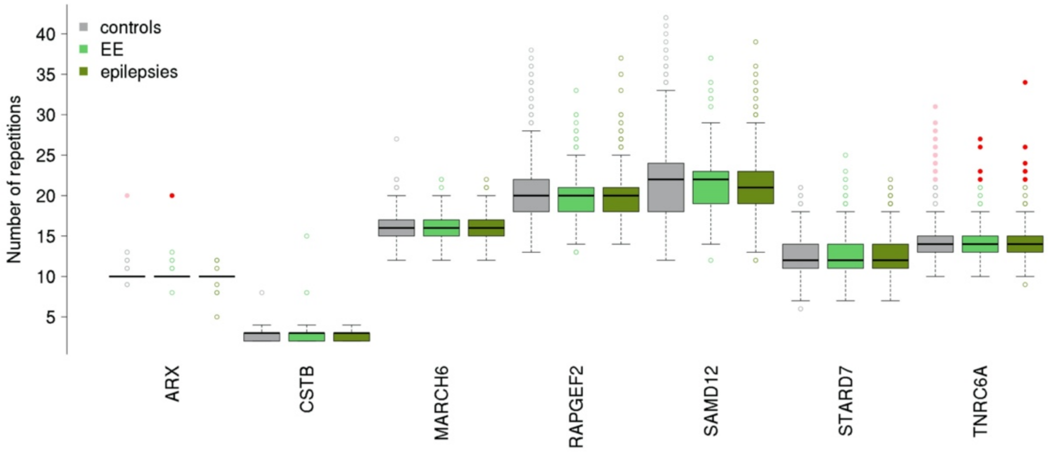
STR expansions with at least one patient with 50 IQR above q75 and the highest control / the highest patient ratio less than 0.6 in 16,495 genome wide genes. Note that all STRs are located in introns.

Figure 1: Number of repeats for 29 trinucleotide-repeat diseases



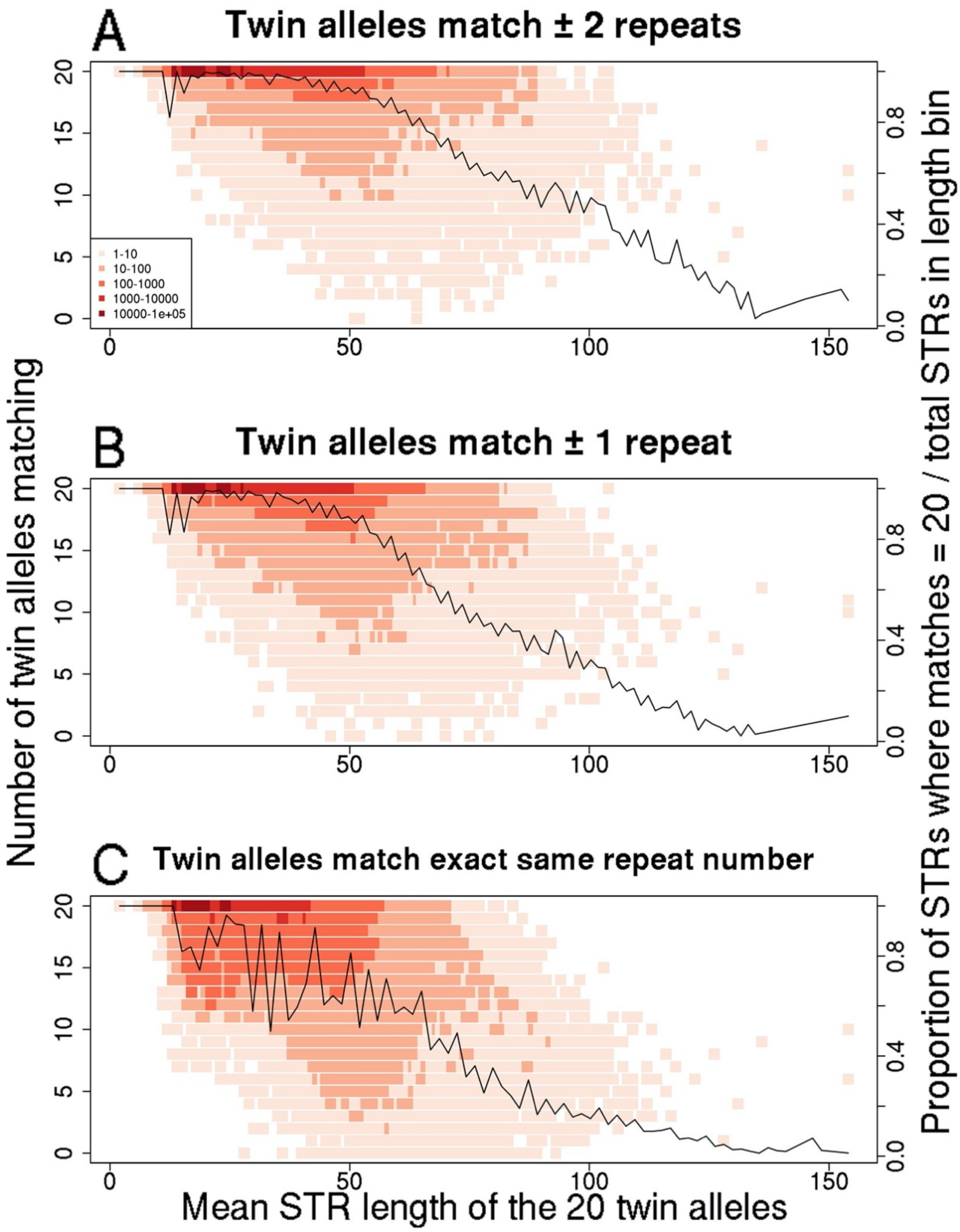
Boxplots of the number of repeats for 29 trinucleotide-repeat diseases for epilepsy cases (green), EE (light green) and controls (grey). Red and pink filled dots are above the risk threshold for cases and controls respectively.

Figure 2: Number of repeats for seven epilepsy related STRs



Boxplots of the number of repeats for seven epilepsy related STRs for epilepsy cases (green), EE (light green) and controls (grey). Red and pink filled dots are the individuals above the risk threshold for cases and controls respectively.

Figure 3: Distribution of the number of matches in monozygotic twin alleles



Distribution of the number of twin alleles matching (on a total of 20 matches) as a function of the STR mean length (bins=1). Quares represent the number of twin alleles matching (left axis) for each length (x axis) and are colored according to the density (log scale) of the STRs over all length bins. The black line (right axis) represents the proportion of STRs matching on a total of 20 alleles for each length bin (x axis).