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], Berge A. Minassian[5,6], Patrick Cossette[7,8], Simon L. Girard[1]

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. Department of Pediatrics, Hospital for Sick Children and University of Toronto, Toronto, Canada;
6. Department of Pediatrics, University of Texas Southwestern, Dallas, TX, USA
7. CHUM Research Center, Montreal, Canada
8. Department of Neurosciences, University of Montreal, Montreal, Canada

Corresponding author:
Simon Girard, PhD
Department of Fundamental Sciences
University of Quebec in Chicoutimi, Office P4-2130
555, boulevard de l'Universite
Chicoutimi (Quebec) G7H 2B1
418 545-5011 ext 2595
1 800 463-9880 ext 2595 (toll free)
simon2_girard@uqac.ca

Abstract

This study aims to decipher the role of short tandem repeats (STRs) in epilepsy patients. Whole genome short-read sequencing data of 738 epileptic patients was used to look for known STR expansions associated with increased risk of neurodevelopmental diseases or epilepsy using three different software. Results show one hit of particular interest on *ARX* gene associated with Early Infantile Encephalopathic Epilepsy that could be causal for one patient with developmental and epileptic encephalopathy. However, we show that the different software do not agree on most of the calls above the threshold and that experimental validation is still needed for diagnostic, although these algorithms could prove useful for pre-selection of samples to be validated.

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(1–4), they represent less than 2% of epilepsy cases. In the last decade, many studies have been conducted on different genetic and bio-statistic methods to better understand the complex genetic mechanisms underlying epilepsy(5). Recently, a large GWAS on epilepsy identified 16 loci associated with the disease, many of which were already known or suspected(6). Moreover, copy number variants were studied in the context of epilepsy(7–9). Polygenic risk scores have also been used to try to better understand the complexity of the disease(10,11). Despite these efforts, there is still a substantial missing heritability component in epilepsy genetics(12).

There is growing evidence that short tandem repeats (STRs) affect gene expression(13) and play a role in genetic disorders(14,15) notably in epilepsy(16–20). 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(21). Fortunately, this has recently changed and such methods with accompanying software are now available(22–26).

In this study we used TREDPARSE(25), Expansion Hunter(22) and GangSTR(26) on whole genome short-read sequencing data of epileptic patients to look for known STR expansions associated with increased risk of neurodevelopmental diseases or 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.

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 Developmental Epileptic Encephalopathy (DEE) 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) (27).

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.

Patients with clinical and EEG characteristics meeting the 1989 ILAE syndrome definitions for IGE were included. 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 DEE 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 DEE 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.

In this study we used WGS of 218 IGE from 186 families (between one and three individuals per extended family), 222 NAFE from 198 families (ranging from one to four members) and 298 DEE from 297 families.

Libraries preparation and whole-genome sequencing

Samples were sequenced at 30X coverage 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). The 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 run 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 a 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 using BWA-MEM version 0.7.10 to create a Binary Alignment Map file (.bam).

Analyses

We looked at the number of repeats in patients at 29 trinucleotide-repeat diseases (summary in (21,25)) and six STR expansions previously shown to be associated with an increased risk of epilepsy (table S1) using TREDPARSE(25), Expansion Hunter(22) and GangSTR(26) with default parameters.

We considered hits above the threshold if two out three software called an expansion(28). When this was possible, we genotyped the parents of the patient to validate the call. We additionally simulated reads of known repeats number around the expansion threshold, aligned them using BWA-MEM as described in(26) and applied the algorithms on the resulting bams.

Results

The distribution of the number of repeats in patients for 29 trinucleotide-repeat diseases associated with neurological conditions is presented in figure 1 and for six STR expansions associated with an increased risk of developing epilepsy is shown in figure 2 for the three software. The filled dots denote alleles above the reported risk threshold. Table 1 shows the number of patients above the risk threshold for the three software. We can clearly see a difference in the number of expansions called by the three algorithms, and that Expansion Hunter called much more expansions than the other two software.

Table 2 presents the individuals for whom there are at least two software calls above the threshold for one STR. There are only four such STRs, three associated with neurodevelopmental diseases (FXTAS, SCA2, SCA8) and one associated with Early Infantile Encephalopathic Epilepsy (EIEE1, OMIM #308350). Table 2 also compares the carrier frequency of our patients with the 12,632 individuals in the original TREDPARSE study(25) and the reported prevalence when this information was available. The carrier frequency for the four expansions in our cohort is higher than the carrier frequency in the 12,632 individuals previously reported except for FXTAS where our patients' carrier frequency is lower than the reported carrier frequency, but still higher than TREDPARSE controls. Of note, we report one hemizygote (using TREDPARSE) male DEE being above the risk threshold for the *ARX* gene known for its association with EIEE1 compared to 0 carrier in the TREDPARSE study. We could validate that this STR expansion was transmitted by his (heterozygote) mother using TREDPARSE although the mother call was not validated using Expansion Hunter and the patient was called heterozygote (both alleles above threshold).

To further look at the accuracy of TREDPARSE and Expansion Hunter calls for four STRs above the threshold, we simulated reads of known repeat numbers (figure 3). The threshold is represented by the blue dashed lines. We can see that there are no false positives for SCA2 and EIEE1 even very close to the threshold.

Discussion

We tested known STR expansion diseases associated with an increased risk of developing neurodevelopmental conditions and epilepsy in 738 epilepsy patients. We found patients above the risk threshold for four out of the 29 STR expansion diseases and the six epilepsy related STR expansions tested. For these, the carrier frequency in our patients is higher than what was reported for 12,632 samples in the study(25) and also higher than the known prevalence or carrier frequency except for FXTAS (table 2). This is interesting since comorbidities and shared genetic mechanisms are frequent between neurological disorders and epilepsies(29,30). Of course, this might also be due to the small sample size in the present study.

Moreover, our patients are from the Quebec population which is a well-documented founder population (31). In fact, some of these expansion diseases have already been documented as being more frequent in particular Quebec regions as a result of the

founder effect(s). Of note, OPMD(32), DM1(33) and FRDA(34). For these three, we do not report any patient above the risk threshold with at least two software. However, Eastern Quebec regions are known for successive regional founder events that could explain frequency changes, but our patients are mainly from the Western Quebec area which is not one of those regions with a marked founder effect. Consequently, we do not think that the higher carrier frequency observed in the present study for the STR expansion diseases can be explained by the founder effect.

For the recessive *ARX* gene STR expansion causing EIEE1, we see a hemizygote DEE boy (with his heterozygote mother) who is above the risk threshold confirmed by Expansion Hunter although less convincing. The boy has been diagnosed with DEE in the neonatal period and was successfully treated with clonazepam. His DEE was accompanied by microcephaly and the 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 DEE trios sequenced for the whole genome(9). Consequently, we believe that the STR expansion in the *ARX* gene is causal.

Nevertheless, the distribution of the number of repeats for the three algorithms does not show a high correlation. There were many calls above the threshold that were not validated by another software as shown in tables 1 and 2. Validations with two out of three software, were found only between Expansion Hunter and TREDPARSE. None of the GangSTR calls was validated by another program. Moreover, the distribution of repeats using Expansion Hunter show a lot of hits above the threshold compared to both other software (figure 1). To further measure the accuracy of TREDPARSE and Expansion Hunter at the four STR sites showing an expansion, we simulated reads of known repeats' number (figure 3). Expansion Hunter and GagnSTR software have previously been reported as very accurate for huge expansions (over 1000 repeats) (Figure S2 in (26)) while TREDPARSE is less accurate above 100-200 repeats. However, the thresholds for most expansions are below 100 repeats. The goal here was to simulate around the threshold to see how the software performs into that repeats' range. The dashed blue lines represent the thresholds. For SCA2 and EIEE1, the simulations reflect the truth even at one repeat around the threshold, the software did not call false positives. On the other hand, for FXTAS and SCA8, it seems more likely that the expansion calls would in fact be false positives around the threshold.

Conclusion

In this study, we performed a survey of STRs in epilepsy patients. We found four expansions on STRs previously associated with neurodevelopmental disease or epilepsy. Interestingly, we found one hit on the *ARX* gene that could be causal for one DEE patient. Nevertheless, we note a lack of consensus among the different software used and an uncertainty around the risk threshold. These software could not be used as diagnostic tools, but for some STRs such as SCA2 and EIEE1, they could prove useful as pre-diagnostic tools to select cases for further experimental validation.

Data availability statement

Raw whole genome sequences of a subset of the epilepsy patients for which we have appropriate consent have been deposited in the European Genome-phenome Archive, under the accession code EGAS00001002825. The STR data calculated using the three software will be available upon request.

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.

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's resources allocation.

Competing interests

We declare no competing interests

Supplemental Material

Table S1: Description of STRs previously identified in epilepsy disorders.

References

1. 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(2):184–9.
2. 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(4):1078–85.
3. 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(1–2):95–9.
4. 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(3):335–41.
5. Koeleman BPC. What do genetic studies tell us about the heritable basis of common epilepsy? Polygenic or complex epilepsy? Neuroscience Letters. 2018. p. 10–6.
6. 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(1):5269.
7. 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(6):943–58.
8. 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;
9. 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(5):664–85.
10. Leu C, Stevelink R, Smith AW, Goleva SB, Kanai M, Ferguson L, et al. Polygenic burden in focal and generalized epilepsies. Brain. 2019;
11. 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;
12. Thomas RH, Berkovic SF. The hidden genetics of epilepsy-a clinically important new paradigm. Nat Rev Neurol. 2014 May;10(5):283–92.
13. 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;
14. 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;
15. 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;
16. 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;

17. 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;
18. 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;
19. 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. [Internet]. Vol. 386, Nature. 1997. p. 847–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9126745
20. 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;
21. 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;
22. 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;
23. 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;
24. 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;
25. 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;
26. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;
27. 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(4):676–85.
28. Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am J Hum Genet. 2018;
29. Pal DK, Ferrie C, Addis L, Akiyama T, Capovilla G, Caraballo R, et al. Idiopathic focal epilepsies: The lost tribe. Epileptic Disorders. 2016.
30. Takumi T, Tamada K. CNV biology in neurodevelopmental disorders. Current Opinion in Neurobiology. 2018.
31. Scriver CR. Human Genetics : Lessons from Quebec Populations. Annu Rev Genomics Hum Genet. 2001;2:69–101.
32. 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;
33. 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;

34. 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;

Table 1 : Number of individuals above the threshold for 15 repeat expansion diseases and 2 STRs associated with epilepsy

STR	Expansion Hunter	TREDPARSE	GangSTR
ALS	0	0	4
BPES	24	0	0
CCD	80	1	0
CCHS	24	0	1
DM2	32	1	0
FXTAS	10	3	0
HFG	181	0	0
HPE5	95	0	1
OPMD	0	0	1
SCA1	0	10	9
SCA17	2	10	5
SCA2	1	1	0
SCA3	0	0	2
SCA8	3	1	0
SD5	50	0	0
EIEE1	4	1	0
ULD	0	0	1

Table 2 : STRs above threshold for at least two software

STR	Phenotype	Sex	inheritance	Risk Threshold	Hg19 repeats	Patient TRED	Patient Exp Hunter	Father TRED	Mother TRED	Father Exp Hunter	Mother Exp Hunter	Present study count (n=738)	TRED controls count[a] (n=12,632)	Reported prevalence[b] (/100,000)
FXTAS	GGE	M	XLD	55	20	44-58	55-55					1	2	*Carrier freq : 300-500*
SCA2	NAFE	F	AD	33	23	22-34	22-34					1	4	1,5
SCA8	DEE	M	AD	80	15	12-80	12-92					1	3	0,5
EIEE1	DEE	M	XLR	20	10	20-20	20-54	10-10	20-10	10-1	10-1	1	0	NA

[a] The TREDPARSE numbers of at risk individuals were taken from table 1 (25).
[b] Prevalence (or carrier frequency for FXTAS) was reported along with references in supplementary table S1 (25).

Figure legends

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

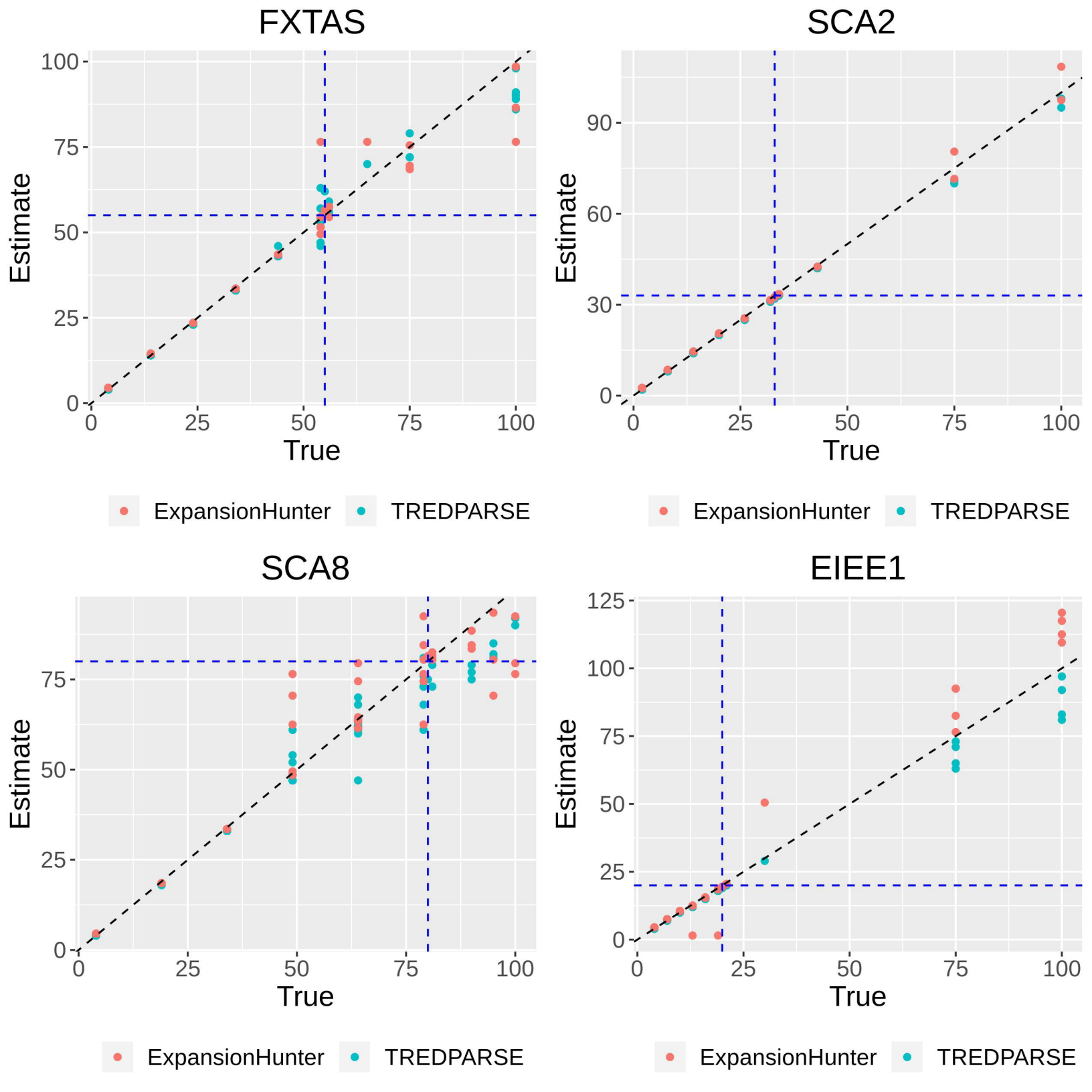
Boxplots of the number of repeats for 29 trinucleotide-repeat diseases for DEE (grey), IGE (blue) and NAFE (green). Filled dots are above the risk threshold. Graphs were cut at 100 repeats for better visibility, note that very few individuals have expansions higher than 100 repeats.

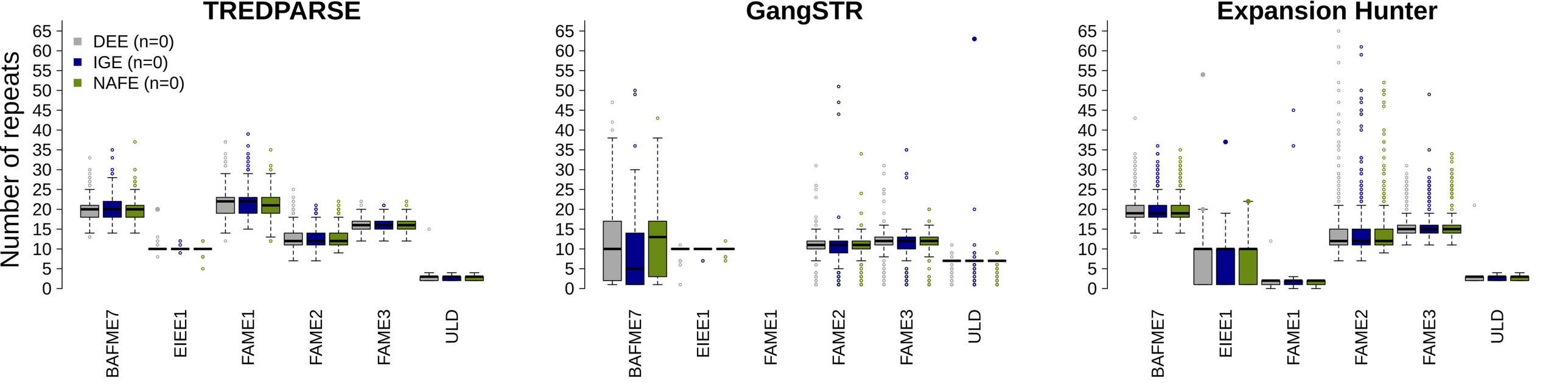
Figure 2: Number of repeats for six epilepsy related STRs for three software

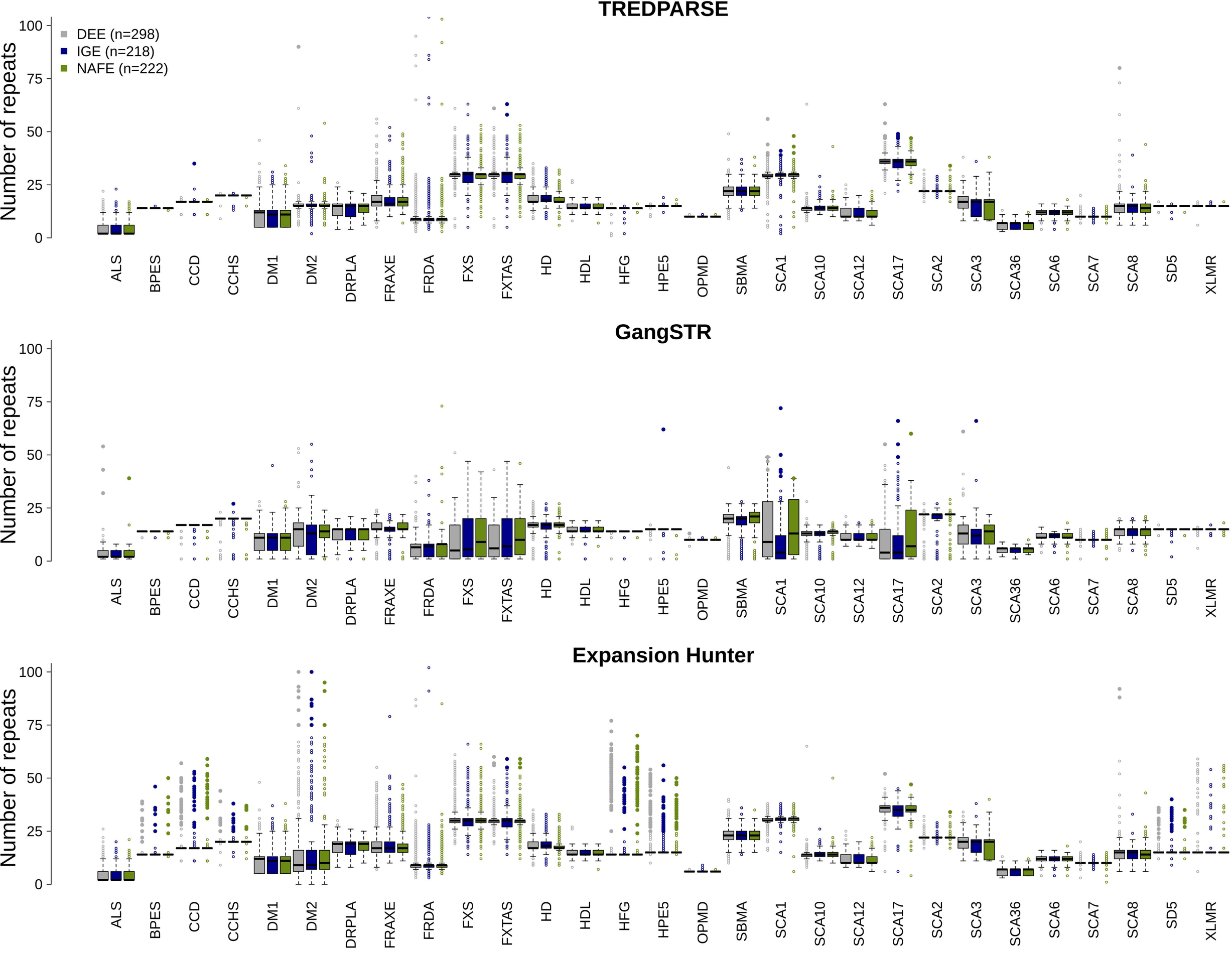
Boxplots of the number of repeats for six epilepsy related STRs for DEE (grey), IGE (blue) and NAFE (green). Filled dots are above the risk threshold. Graphs were cut at 65 repeats for better visibility, note that very few individuals have expansions higher than 65 repeats.

Figure 3: Simulations vs real data for four STR expansions for two software

Real (x axis) vs simulated (y axis) number of repeats for TREDPARSE (cyan) and Expansion Hunter (pink). The diagonal dashed line represents the truth and the blue dash lines the thresholds.


FXTAS
SCA2
SCA8
EIEE1
Estimate
True
ExpansionHunter
TREDPARSE


TREDPARSE
GangSTR
Expansion Hunter
Number of repeats
DEE (n=0)
IGE (n=0)
NAFE (n=0)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
BAFME7
EIEE1
FAME1
FAME2
FAME3
ULD


TREDPARSE
GangSTR
Expansion Hunter
DEE (n=298)
IGE (n=218)
NAFE (n=222)
Number of repeats
0
25
50
75
100
ALS
BPES
CCD
CCHS
DM1
DM2
DRPLA
FRAXE
FRDA
FXS
FXTAS
HD
HDL
HFG
HPE5
OPMD
SBMA
SCA1
SCA10
SCA12
SCA17
SCA2
SCA3
SCA36
SCA6
SCA7
SCA8
SD5
XLMR