Supplementary Information

S1 Table 1. All known disease-causing STR loci made of standard type of repeats.

Gene	**Reference and location**	**Motif**	**Disease**	**Inheritance**	**Normal range**	**Intermediate range**	**Pathogenic cut-off**	**References**
AFF2 (FMR2)	chrX:148500638-148500683 (5′ UTR)	CCG	FRAXE mental retardation (FRAXE)	XLR	6–25	–	≥200	[1]
AR	chrX:67545316-67545385 (Coding)	CAG	Spinal and bulbar muscular atrophy (SBMA)	XLR	17–35	–	≥40	[2–4]
ATN1	chr12:6936728-6936773 (Coding)	CAG	Dentatorubral-pallidoluysian atrophy (DRPLA)	AD	7–23	–	≥49	[5,6]
ATXN1	chr6:16327635-16327722 (Coding)	CAG	Spinocerebellar ataxia 1 (SCA1)	AD	6–32	36–38	≥39	[7,8]
ATXN2	chr12:111598950-111599019 (Coding)	CAG	Spinocerebellar ataxia 2 (SCA2)	AD	13–31	32–34	≥35	[9–11]
ATXN3	chr14:92071010-92071040 (Coding)	CAG	Spinocerebellar ataxia 3 (SCA3)	AD	12–44	45–55	≥56	[12,13]
ATXN7	chr3:63912685-63912715 (Coding)	CAG	Spinocerebellar ataxia 7 (SCA7)	AD	4–35	28–35	≥34 or ≥36	[14,15]
ATXN8OS	chr13:70139383-70139428 (3′ UTR)	CTG	Spinocerebellar ataxia 8 (SCA8)	AD	2–37	38–79$^{\$}$	≥80	[16–18]
ATXN10	chr22:45795354-45795424 (Intron)	ATTCT	Spinocerebellar ataxia 10 (SCA10)	AD	10–29	30–799$^{\$}$	≥800	[19]
C9orf72	chr9:27573528-27573546 (Intron)	GGGGCC	Amyotrophic lateral sclerosis and/or frontotemporal dementia (FTDALS1)	AD	1–23	20–30*	≥31*	[20,21]
CACNA1A	chr19:13207858-13207897 (Coding)	CAG	Spinocerebellar ataxia 6 (SCA6)	AD	4–18	19	≥21	[22–24]
CBL	chr11:119206289-119206322 (5′ UTR)	CCG	Jacobsen syndrome (JBS)	NI	8–80	85–100	≥101	[25]
CNBP (ZNF9)	chr3:129172576-129172656 (Intron)	CCTG	Myotonic dystrophy 2 (DM2)	AD	≤29	30–54	≥55	[26,27]
COMP	chr19:18786034-18786049 (Coding)	GAC	Multiple epiphyseal dysplasia (MED)	AD	5	–	6	[28]
			Pseudoachondroplasia (PSACH)	AD	5	–	4 & 7	[28,29]
DIP2B	chr12:50505003-50505024 (5′ UTR)	CGG	FRA12A mental retardation (MRFRA12A)	AD	6–23	–	≥270	[30]

DMD	chrX:31284557-31284605 (Intron)	GAA	Duchenne muscular dystrophy (DMD)	XLR	11–33	–	≥59	[31]
DMPK	chr19:45770204-45770264 (3′ UTR)	CTG	Myotonic dystrophy 1 (DM1)	AD	5–34	35–49	≥50	[32–34]
FMR1	chrX:147912050-147912110 (5′ UTR)	CGG	Fragile X Syndrome (FXS)	XLD	6–40	41–54	≥201	[35,36]
			Fragile X tremor/ataxia syndrome (FXTAS)	XLD	6–40	41–54	≥55 ≤200	[35,36]
FXN	chr9:69037286-69037304 (Intron)	GAA	Friedreich ataxia (FRDA)	AR	5–30	–	≥70	[37]
GIPC1	chr19:14496041-14496074 (5′ UTR)	GGC	Oculopharyngodistal myopathy 1 (OPDM2)	AD	12–32	–	≥73	[38]
GLS	chr2:190880872-190880920 (5′ UTR)	GCA	Glutaminase deficiency GD)	AR	5–26	–	≥680	[39]
HTT	chr4:3074876-3074933 (Coding)	CAG	Huntington's disease (HD)	AD	9–26	27–35	≥36	[40–42]
JPH3	chr16:87604287-87604329 (3′ UTR)	CTG	Huntington disease-like 2 (HDL2)	AD	6–27	–	≥41	[43]
LRP12	chr1:149390805-149390831 (5′ UTR)	CGG	Oculopharyngodistal myopathy 1 (OPDM1)	AD	13–45	–	≥93	[44]
NOP56	chr20:2652733-2652757 (Intron)	GGCCTG	Spinocerebellar ataxia 36 (SCA36)	AD	3–14	–	≥650	[45,46]
NOTCH2NLC (NBPF19)	chr1:149390802-149390829 (5′ UTR)	GGC	Neuronal intranuclear inclusion disease (NIID)	AD	7–39	–	≥90	[44]
			Hereditary essential tremor type 6 (ETM6)	AD	4–41	–	≥60	[47]
NUTM2B-AS1	chr10:79826377-79826403 (Intron)	CGG	Oculopharyngeal myopathy with leukoencephalopathy 1 (OPML1)	AD	3–16	–	≥700	[44]
PABPN1	chr14:23321472-23321490 (Coding)	GCG	Oculopharyngeal muscular dystrophy (OPMD)	AD or AR	6	–	≥7	[48]
PPP2R2B	chr5:146878728-146878758 (5′ UTR)	CAG	Spinocerebellar ataxia 12 (SCA12)	AD	7–31	–	≥51	[49,50]
PRDM12	chr9:130681606-130681639 (Coding)	CCG	Neuropathy, hereditary sensory and autonomic, type VIII (HSAN8)	AR	≤14	–	≥18	[51]
TBP	chr6:170561907-170562015 (Coding)	CAG	Spinocerebellar ataxia 17 (SCA17)	AD	25–42	–	≥43	[52,53]

TCF4	chr18:55586155-55586227 (Intron)	CTG	Fuchs endothelial corneal dystrophy-3 (FECD3)	AD	≤31	–	≥50	[54,55]
XYLT1	chr16:17470907-17470922 (5′ UTR)	GGC	Desbuquois dysplasia-2 (DBQD2)	AR	9–20	–	≥110	[56]
ZIC3	chrX:137566826-137566856 (Coding)	GCC	X-linked VACTERL syndrome (VACTERLX)	XLR	10	–	≥12	[57]

The coordinates are in hg38 assembly and 0-based. Asterisk (*) denotes data where the range vary in literature. $Denotes a range specified by us based on the findings in literature from multiple sources. Normal, intermediate and pathogenic ranges were derived from the articles marked in the references field. AD – autosomal dominant, AR – autosomal recessive, XLD – X-linked dominant, XLR – X-linked recessive, NI – not inherited.

S1 Table 2. All known disease-causing STR loci made of imperfect GCN type of repeats.

Gene	Reference and location	Motif	Disease	Inheritance	Normal range	Intermediate range	Pathogenic cut-off	References
ARX	†chrX:25013529-25013565 ‡chrX:25013649-25013697 (Coding)	GCN	Developmental and epileptic encephalopathy-1 (DEE1)	XLR	12† 16‡	–	≥20† ≥23‡	[58]
			Partington syndrome (PRTS)	XLR	12†	–	≥20†	[58]
			X-linked mental retardation with or without seizures (MRXARX)	XLR	12† 16‡	–	≥20† ≥18‡	[59]
FOXL2	chr3:138946020-138946062 (Coding)	GCN	Blepharophimosis, ptosis, and epicanthus inversus syndrome (BPES)	AD	14	–	≥19	[60,61]
HOXA13	+chr7:27199924-27199966 †chr7:27199825-27199861 ‡chr7:27199678-27199732 (Coding)	GCN	Hand-foot-genital syndrome (HFG)	AD	14+ 12† 18‡	–	≥22+ ≥18† ≥24‡	[62–64]
HOXD13	chr2:176093058-176093103 (Coding)	GCN	Synpolydactyly (SPD)	AD	15	–	≥22	[65]
PHOX2B	chr4:41745971-41746031 (Coding)	GCN	Central hypoventilation syndrome (CCHS)	AD	20	–	≥24	[66]
RUNX2	chr6:45422750-45422801 (Coding)	GCN	Cleidocranial dysplasia (CCD)	AD	17	–	≥27	[67]
SOX3	chrX:140504316-140504361 (Coding)	GCN	X-linked mental retardation (XLMR)	XLR	15	–	≥26	[68]
			X-linked panhypopituitarism (PHPX)	XLR	15	–	≥22	[69]
TBX1	chr22:19766762-19766807 (Coding)	GCN	Tetralogy of Fallot (TOF)	AD	15	–	≥25	[70]
ZIC2	chr13:99985448-99985493 (Coding)	GCN	Holoprosencephaly 5 (HPE5)	AD	15	–	≥25	[71]

Given coordinates are in hg38 assembly and 0-based. +, † and ‡ are used to distinguish values for different tracts in a locus. Normal, intermediate and pathogenic ranges were derived from the articles marked in the references field. AD – autosomal dominant, XLR – X-linked recessive.

S1 Table 3. All known disease-causing STR loci made of nested and replaced types of repeats.

Gene	Reference and location	Motif	Disease	Inheritance	Normal range	Intermediate range	Pathogenic cut-off	References
BEAN1	chr16:66490398-66490466 (Intron)	TGGAA *(TAAAA)*	Spinocerebellar ataxia 31 (SCA31)	AD	0	–	≥500#	[72]
DAB1	chr1:57367043-57367118 (Intron)	ATTTC *(ATTTT)*	Spinocerebellar ataxia 37 (SCA37)	AD	0	–	≥31	[73,74]
RFC1	chr4:39348424-39348485 (Intron)	AAGGG *(AAAAG)*	Cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS)	AR	0§	–	≥400	[75,76]
SAMD12	chr8:118366816-118366914 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 1 (FAME1)	AD	0	–	≥100	[77–79]
STARD7	chr2:96197066-96197121 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 2 (FAME2)	AD	0	–	≥274	[80]
MARCH6	chr5:10356347-10356407 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 3 (FAME3)	AD	0	–	≥668#	[81]
YEATS2	chr3:183712187-183712222 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 4 (FAME4)	AD	0	–	≥1000#	[82]
TNRC6A	chr16:24613439-24613529 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 6 (FAME6)	AD	0	–	≥1100#	[78]
RAPGEF2	chr4:159342526-159342616 (Intron)	TTTCA *(TTTTA)*	Familial adult myoclonic epilepsy 7 (FAME7)	AD	0	–	≥60	[78,79]

The coordinates are in hg38 assembly and 0-based. Coordinates for the reference region denotes the STR location of endogenous repeats. The motif field indicates the pathogenic motif with the non-pathogenic reference motif in parentheses. Normal, intermediate and pathogenic ranges were derived from the articles marked in the references field. §The normal range is set to 0 because the pathogenic motif has not been found in the healthy population, except in one alleles of the CANVAS patients. #Size of the whole allele (including non-pathogenic repeats). Number of repeats in the pathogenic cut-off column for many diseases has been calculated from the reported allele size assuming they are composed solely of pentanucleotide repeats. AD – autosomal dominant, AR – autosomal recessive.

Text 1. Methods used for validation of STRipy and population-wide genotyping. This section describes methods used for validating STRipy and creating the STRs database.

Method of simulating samples for validation.

Creation of samples with artificial length of repeats was done in two main steps. Firstly, a template sequence was created based on the reference genome, where the known STR sequence was swapped with a newly created one. The start and end positions of the STR sequence were derived from the literature, assessed manually and in some cases the end position was adjusted in a way that it ends before at least two consecutive repeats of the non-pathogenic motif was present (coordinates shown in S1 Table 1–3). Secondly, this template was used to simulate reads that were then aligned to the reference genome.

We wrote a custom-made program that takes the genomic coordinates of a known pathogenic STR locus and extracts out 2 kb flanking sequence before and after the repeated sequence. Then, a new stretch of repeats was created with specific length, which was then merged with the flanking sequences to form over a 4 kb sequence and saved into a new FASTA file. For all the disorders that are caused by imperfect GCN types of repeats (stretches of the same amino acid but with an inconsistent repeat sequence coding alanine), we generated stretches of motifs that had one base randomly chosen. For example, to simulate polyalanine stretches we used GCN as the motif and replaced "N" randomly with either "A", "T", "G" or "C" in every repeat present in the sequence.

To determine genotyping accuracy across increasing length of repeats, we first simulated homozygous alleles that included the pathogenic motif from 60 bp until the 2100 bp limit was reached where each simulation increased the length of the allele by one repeat unit (in the range of 20–700 trinucleotide-, 15–525 tetranucleotide-, 12–420 pentanucleotide and 10–350 hexanucleotide repeats). In the case of the nested type of repeats, we created short alleles without the pathogenic motif. Secondly, for heterozygous alleles we designed the shorter allele to be fixed length (60 bp) and increased the length of the second allele from 60 bp until 2100 bp, similarly to the process of simulating homozygous repeats. For alleles of the nested type of repeats, we simulated alleles based on the formulas in S1 Table 4, where *exp* for the non-pathogenic repeat unit was a random number between 2 and 10 and for pathogenic motif the repeat was in increasing length. For example, to simulate FAME4 disorder, a normal AAAAT repeats was created which length was randomly chosen to be between 2 and 10 repeats, followed by 12 to 240 repeats of pathogenic AAATG motif to create a 60 to 2100 bp pathogenic locus.

S1 Table 4. Formulas used to simulate alleles of replaced and nested type repeats.

Gene	Disease	Repeat type	Formula for allele simulation
RFC1	CANVAS	Replaced	$(\mathrm{AAAAG})_{11}[\mathrm{AAGGG}]_{\mathrm{exp}}$
BEAN1	SCA31	Nested	$(\mathrm{TCAC})_{1}[\mathrm{TGGAA}]_{\mathrm{exp}}(\mathrm{TAGAA})_{\mathrm{exp}}(\mathrm{TAAAATAGAA})_{\mathrm{exp}}$
DAB1	SCA37	Nested	$(\mathrm{AAAAT})_{60\text{-}79}[\mathrm{GAAAT}]_{31\text{-}75}(\mathrm{AAAAT})_{58\text{-}90}$
MARCH6	FAME1	Nested	$(\mathrm{TTTTA})_{12}[\mathrm{TTTCA}]_{\mathrm{exp}}$
RAPGEF2	FAME2	Nested	$(\mathrm{TTTTA})_{\mathrm{exp}}[\mathrm{TTTCA}]_{\mathrm{exp}}(\mathrm{TTTTA})_{\mathrm{exp}}$
SAMD12	FAME3	Nested	$(\mathrm{AAAAT})_{\mathrm{exp}}[\mathrm{TGAAA}]_{\mathrm{exp}}$
STARD7	FAME4	Nested	$(\mathrm{AAAAT})_{\mathrm{exp}}[\mathrm{AAATG}]_{\mathrm{exp}}$
TNRC6A	FAME6	Nested	$(\mathrm{TTTTA})_{22}[\mathrm{TTTCA}]_{\mathrm{exp}}(\mathrm{TTTTA})_{\mathrm{exp}}$
YEATS2	FAME7	Nested	$(\mathrm{TTTTA})_{\mathrm{exp}}[\mathrm{TTTCA}]_{\mathrm{exp}}$

Motif in parentheses is the non-disease causing and motif in square brackets is the pathogenic one. The number after parentheses or square brackets shows the number of repeats of that motif, exp denotes an expansion which length is not specified and will be randomly assigned.

We generated paired-end reads in the FASTQ format from the template (FASTA) files with the ART (version MountRainier 2016-06-05) next-generation sequencing read simulator tool [83]. Illumina HiSeqX PCR-free profile was used to simulate 150 bp reads with the mean fragment size of 450 bp and standard deviation 50 bp. Sequences were simulated to have total coverage of 50x, both for homozygous and heterozygous samples. All FASTQ files were aligned on the UCSC GRCh38.p12 reference genome with BWA-MEM v0.7.17 [84] and indexed with Samtools v1.10 [85].

Genotyping samples with STRipy

We used STRipy with ExpansionHunter v4.0.2 as the genotyper. STRipy was performed on each of the aligned sample (both homozygous and heterozygous) and results saved into a separate file that was analysed with R (paper_analysis.R) [86]. We ran STRipy by using the default parameters, except switching on the option to use alternative contigs.

We used root mean square error (RMSE) to determine genotyping accuracy for both, heterozygous and homozygous samples by comparing predicted and true repeat lengths as described by Mousavi et al. (2019). The $\langle x_1^i, x_2^i \rangle$ denotes a diploid genotype of sample i and is ordered by length such that $x_1^i \leq x_2^i$. To compare the actual $X = \{\langle x_1^1, x_2^1 \rangle, \langle x_1^2, x_2^2 \rangle \ldots \langle x_1^n, x_2^n \rangle\}$

and predicted $Y = \{\langle y_1^1, y_2^1\rangle, \langle y_1^2, y_2^2\rangle \ldots \langle y_1^n, y_2^n\rangle\}$ genotypes, then RMSE is defined by the following formula [87]:

$$\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(y_j^i - x_j^i)^2}{2n}}$$

Population-wide data creation

We used 2,504 samples from the DRAGEN reanalysis of the 1000 Genomes Dataset (1kGP-DRAGEN, https://registry.opendata.aws/ilmn-dragen-1kgp) to genotype all known pathogenic loci in all samples. These samples were sequenced with the Illumina NovaSeq 6000 system to produce 150 bp paired-end reads of at least 30x coverage. All BAM files in the dataset were realigned to hg38 assembly by using the Illumina DRAGEN v3.5.7b (https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html).

All samples in Amazon Web Services S3 instance: s3://1000genomes-dragen/data/dragen-3.5.7b/hg38_altaware_nohla-cnv-anchored were analysed. ExpansionHunter v4.0.2, with our complete variant catalogue that included all genomic loci, including all tracts in genes. We applied a filter to exclude results where there are less than 5 spanning reads. In case of no spanning reads, then the criterion was minimum of 50 combined flanking and in-repeat reads to increase confidence of determined genotypes. To genotype replaced and nested types of repeat loci, we created a custom-made script (genotype-nested.py) [86] which analysed spanning reads in the realigned BAM files outputted from ExpansionHunter and determined the length of pathogenic repeats in the these reads by using previously described genotyping model [88].

Next, we determined all samples where the allele was over the read length of these samples and genotyped these samples again by using a catalogue that included pre-defined off-target regions to enable genotyping long alleles. To do this, we simulated one sample for each locus whose coverage was increased to 100x and we shifted down the quality score of every read by 10 to increase sequencing error rate to 1% of the default profile. Following alignment, we determined all regions where reads aligned to and merged overlapping and nearby regions within 100 bp into one single entry and adding 50 bp on both ends. These coordinates were specified as off-target regions in the variant catalogue. Off-target regions that included or overlapped with the reference region of the disease or regions on alternative contigs were removed from the list beforehand.

S1 Table 5. Genotypes of three tracts in HOXA13 gene in the ten real biological samples estimated by ExpansionHunter and STRipy.

	ExpansionHunter			STRipy		
#	HOXA13_1	HOXA13_2	HOXA13_3	HOXA13_1	HOXA13_2	HOXA13_3
1	14/14	12/12	18/18	14/14	12/12	18/18
2	14/14	12/12	18/18	14/14	12/12	18/18
3	14/14	12/66	18/18	14/14	12/12	18/18
4	14/66	12/70	18/18	14/14	12/12	18/18
5	14/14	12/12	18/18	14/14	12/12	18/18
6	14/14	12/12	18/18	14/14	12/12	18/18
7	14/80	12/73	18/18	14/14	12/12	18/18
8	14/14	12/12	18/18	14/14	12/12	18/18
9	14/69	12/12	18/18	14/14	12/12	18/18
10	14/14	12/12	18/18	14/14	12/12	18/18

List of references

1. Knight SJL, Flannery AV, Hirst MC, Campbell L, Christodoulou Z, Phelps SR, et al. Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell. 1993 Jul;74(1):127–34. doi: 10.1016/0092-8674(93)90300-F

2. Spada AR La, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature. 1991 Jul;352(6330):77–9. doi: 10.1038/352077a0

3. Grunseich C, Kats IR, Bott LC, Rinaldi C, Kokkinis A, Fox D, et al. Early onset and novel features in a spinal and bulbar muscular atrophy patient with a 68 CAG repeat. Neuromuscul Disord. 2014;24(11):978–81. doi: 10.1016/j.nmd.2014.06.441

4. Udd B, Juvonen V, Hakamies L, Nieminen A, Wallgren-Pettersson C, Cederquist K, et al. High prevalence of Kennedy's disease Western Finland - Is the syndrome underdiagnosed? Acta Neurol Scand. 1998;98(2):128–33. doi: 10.1111/j.1600-0404.1998.tb01732.x

5. Nagafuchi S, Yanagisawa H, Sato K, Shirayama T, Ohsaki E, Bundo M, et al. Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nat Genet. 1994 Jan;6(1):14–8. doi: 10.1038/ng0194-14

6. Komure O, Sano A, Nishino N, Yamauchi N, Ueno S, Kondoh K, et al. Dna analysis in hereditary dentatorubral-pallidoluysian atrophy: Correlation between cag repeat length and phenotypic variation and the molecular basis of anticipation. Neurology. 1995;45(1):143–9. doi: 10.1212/WNL.45.1.143

7. Orr HT, Chung M, Banfi S, Kwiatkowski TJ, Servadio A, Beaudet AL, et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993 Jul;4(3):221–6. doi: 10.1038/ng0793-221

8. Zühlke C, Dalski A, Hellenbroich Y, Bubel S, Schwinger E, Bürk K. Spinocerebellar ataxia type 1 (SCA1): Phenotype-genotype correlation studies in intermediate alleles. Eur J Hum Genet. 2002;10(3):204–9. doi: 10.1038/sj.ejhg.5200788

9. Riess O, Laccone FA, Gispert S, Schöls L, Zühlke C, Menezes Vieira-Saecker AM, et al. SCA2 trinucleotide expansion in German SCA patients. Neurogenetics. 1997;1(1):59–64. doi: 10.1007/s100480050009

10. Velázquez Pérez L, Cruz GS, Santos Falcón N, Enrique Almaguer Mederos L, Escalona Batallan K, Rodríguez Labrada R, et al. Molecular epidemiology of spinocerebellar ataxias in Cuba: Insights into SCA2 founder effect in Holguin. Neurosci Lett. 2009;454(2):157–60. doi: 10.1016/j.neulet.2009.03.015

11. Sanpei K, Takano H, Igarashi S, Sato T, Oyake M, Sasaki H, et al. Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. Nat Genet. 1996 Nov;14(3):277–84. doi: 10.1038/ng1196-277

12. McLoughlin HS, Moore LR, Paulson HL. Pathogenesis of SCA3 and implications for other polyglutamine diseases. Neurobiol Dis. 2020;134(July 2019):104635. doi:

10.1016/j.nbd.2019.104635

13. Ashizawa T, Öz G, Paulson HL. Spinocerebellar ataxias: prospects and challenges for therapy development. Nat Rev Neurol. 2018;14(10):590–605. doi: 10.1038/s41582-018-0051-6

14. Stevanin G, Giunti P, David G, Belal S, Dürr A, Ruberg M, et al. De novo expansion of intermediate alleles in spinocerebellar ataxia 7. Hum Mol Genet. 1998;7(11):1809–13. doi: 10.1093/hmg/7.11.1809

15. Magaña JJ, Tapia-Guerrero YS, Velázquez-Pérez L, Cerecedo-Zapata CM, Maldonado-Rodríguez M, Jano-Ito JS, et al. Analysis of CAG repeats in five SCA loci in Mexican population: Epidemiological evidence of a SCA7 founder effect. Clin Genet. 2014;85(2):159–65. doi: 10.1111/cge.12114

16. Gupta A, Jankovic J. Spinocerebellar ataxia 8: Variable phenotype and unique pathogenesis. Park Relat Disord. 2009;15(9):621–6. doi: 10.1016/j.parkreldis.2009.06.001

17. Schöls L, Bauer I, Zühlke C, Schulte T, Kölmel C, Bürk K, et al. Do CTG expansions at the SCA8 locus cause ataxia? Ann Neurol. 2003;54(1):110–5. doi: 10.1002/ana.10608

18. Brusco A, Cagnoli C, Franco A, Dragone E, Nardacchione A, Grosso E, et al. Analysis of SCA8 and SCA12 loci in 134 Italian ataxic patients negative for SCA1-3, 6 and 7 CAG expansions. J Neurol. 2002 Jul 1;249(7):923–9. doi: 10.1007/s00415-002-0760-y

19. Teive HAG, Munhoz RP, Arruda WO, Raskin S, Werneck LC, Ashizawa T. Spinocerebellar ataxia type 10 - A review. Park Relat Disord. 2011;17(9):655–61. doi: 10.1016/j.parkreldis.2011.04.001

20. Kaivola K, Kiviharju A, Jansson L, Rantalainen V, Eriksson JG, Strandberg TE, et al. C9orf72 hexanucleotide repeat length in older population: normal variation and effects on cognition. Neurobiol Aging. 2019;84:242.e7-242.e12. doi: 10.1016/j.neurobiolaging.2019.02.026

21. Rutherford NJ, Heckman MG, DeJesus-Hernandez M, Baker MC, Soto-Ortolaza AI, Rayaprolu S, et al. Length of normal alleles of C9ORF72 GGGGCC repeat do not influence disease phenotype. Neurobiol Aging. 2012;33(12):2950.e5-2950.e7. doi: 10.1016/j.neurobiolaging.2012.07.005

22. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, et al. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the α1A-voltage-dependent calcium channel. Nat Genet. 1997 Jan;15(1):62–9. doi: 10.1038/ng0197-62

23. Riess O. SCA6 is caused by moderate CAG expansion in the alpha1A-voltage-dependent calcium channel gene. Hum Mol Genet. 1997 Aug 1;6(8):1289–93. doi: 10.1093/hmg/6.8.1289

24. Mariotti C, Gellera C, Grisoli M, Mineri R, Castucci A, Di Donato S. Pathogenic effect of an intermediate-size SCA-6 allele (CAG)19 in a homozygous patient. Neurology. 2001 Oct 23;57(8):1502–4. doi: 10.1212/WNL.57.8.1502

25. Jones C, Penny L, Mattina T, Yu S, Baker E, Voullaire L, et al. Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene CBL2. Nature. 1995;376(6536):145–9. doi: 10.1038/376145a0

26. Radvanszky J, Surovy M, Polak E, Kadasi L. Uninterrupted CCTG tracts in the myotonic dystrophy type 2 associated locus. Neuromuscul Disord. 2013;23(7):591–8. doi: 10.1016/j.nmd.2013.02.013

27. Bachinski LL, Czernuszewicz T, Ramagli LS, Suominen T, Shriver MD, Udd B, et al. Premutation allele pool in myotonic dystrophy type 2. Neurology. 2009;72(6):490–7. doi: 10.1212/01.wnl.0000333665.01888.33

28. Délot E, King LM, Briggs MD, Wilcox WR, Cohn DH. Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. Hum Mol Genet. 1999;8(1):123–8. doi: 10.1093/hmg/8.1.123

29. Hecht JT, Nelson LD, Crowder E, Wang Y, Elder FFB, Harrison WR, et al. Mutations in exon 17B of cartilage oligomeric matrix protein (COMP) cause pseudoachondroplasia. Nat Genet. 1995 Jul;10(3):325–9. doi: 10.1038/ng0795-325

30. Winnepenninckx B, Debacker K, Ramsay J, Smeets D, Smits A, FitzPatrick DR, et al. CGG-repeat expansion in the DIP2B gene is associated with the fragile site FRA12A on chromosome 12q13.1. Am J Hum Genet. 2007;80(2):221–31. doi: 10.1086/510800

31. Kekou K, Sofocleous C, Papadimas G, Petichakis D, Svingou M, Pons RM, et al. A dynamic trinucleotide repeat (TNR) expansion in the DMD gene. Mol Cell Probes. 2016;30(4):254–60. doi: 10.1016/j.mcp.2016.07.001

32. Jaspert A, Fahsold R, Grehl H, Claus D. Myotonic dystrophy: Correlation of clinical symptoms with the size of the CTG trinucleotide repeat. J Neurol. 1995;242(2):99–104. doi: 10.1007/BF00887824

33. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell. 1992;68(4):799–808. doi: 10.1016/0092-8674(92)90154-5

34. Gutiérrez Gutiérrez G, Díaz-Manera J, Almendrote M, Azriel S, Eulalio Bárcena J, Cabezudo García P, et al. Clinical guide for the diagnosis and follow-up of myotonic dystrophy type 1, MD1 or Steinert's disease. Neurol (English Ed. 2020;35(3):185–206. doi: 10.1016/j.nrleng.2019.01.008

35. Mishra SK, Khosa S, Trikamji B, Khanli HM, Scheuner MT, Jamal NI, et al. A case of Fragile X-associated tremor/ataxia syndrome (FXTAS). J Clin Neurosci. 2019;66:269–70. doi: 10.1016/j.jocn.2019.05.044

36. Aziz M, Stathopulu E, Callias M, Taylor C, Turk J, Oostra B, et al. Clinical features of boys with fragile X premutations and intermediate alleles. Am J Med Genet - Neuropsychiatr Genet. 2003;121 B(1):119–27. doi: 10.1002/ajmg.b.20030

37. Pandolfo M. Molecular basis of Friedreich ataxia. Mov Disord. 2001;16(5):815–21. doi: 10.1002/mds.1162

38. Deng J, Yu J, Li P, Luan X, Cao L, Zhao J, et al. Expansion of GGC Repeat in GIPC1 Is Associated with Oculopharyngodistal Myopathy. Am J Hum Genet. 2020;106(6):793–804. doi: 10.1016/j.ajhg.2020.04.011

39. van Kuilenburg ABP, Tarailo-Graovac M, Richmond PA, Drögemöller BI, Pouladi MA, Leen R, et al. Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS. N Engl J Med. 2019 Apr 11;380(15):1433–41. doi: 10.1056/NEJMoa1806627

40. Testa CM, Jankovic J. Huntington disease: A quarter century of progress since the gene discovery. J Neurol Sci. 2019;396(April 2018):52–68. doi: 10.1016/j.jns.2018.09.022

41. Read AP. Huntington's disease: testing the test. Nat Genet. 1993 Aug;4(4):329–30. doi: 10.1038/ng0893-329

42. Semaka A, Collins JA, Hayden MR. Unstable familial transmissions of Huntington disease alleles with 27-35 CAG repeats (intermediate alleles). Am J Med Genet B Neuropsychiatr Genet. 2010 Jan 5;153B(1):314–20. doi: 10.1002/ajmg.b.30970

43. Holmes SE, O'hearn E, Rosenblatt A, Callahan C, Hwang HS, Ingersoll-Ashworth RG, et al. A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. Nat Genet. 2001;29(4):377–8. doi: 10.1038/ng760

44. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. Nat Genet. 2019;51(8):1222–32. doi: 10.1038/s41588-019-0458-z

45. Kobayashi H, Abe K, Matsuura T, Ikeda Y, Hitomi T, Akechi Y, et al. Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet. 2011;89(1):121–30. doi: 10.1016/j.ajhg.2011.05.015

46. García-Murias M, Quintáns B, Arias M, Seixas AI, Cacheiro P, Tarrío R, et al. "Costa da Morte" ataxia is spinocerebellar ataxia 36: Clinical and genetic characterization. Brain. 2012;135(5):1423–35. doi: 10.1093/brain/aws069

47. Sun QY, Xu Q, Tian Y, Hu ZM, Qin LX, Yang JX, et al. Expansion of GGC repeat in the human-specific NOTCH2NLC gene is associated with essential tremor. Brain. 2020;143(1):222–33. doi: 10.1093/brain/awz372

48. Brais B, Bouchard J-P, Xie Y-G, Rochefort DL, Chrétien N, Tomé FMS, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet. 1998 Feb 1;18(2):164–7. doi: 10.1038/ng0298-164

49. Bahl S, Virdi K, Mittal U, Sachdeva MP, Kalla AK, Holmes SE, et al. Evidence of a Common Founder for SCA12 in the Indian Population. Ann Hum Genet. 2005 Sep;69(5):528–34. doi: 10.1046/j.1529-8817.2005.00173.x

50. Srivastava AK, Choudhry S, Gopinath MS, Roy S, Tripathi M, Brahmachari SK, et al. Molecular and clinical correlation in five Indian families with spinocerebellar ataxia 12. Ann Neurol. 2001;50(6):796–800. doi: 10.1002/ana.10048

51. Chen Y-C, Auer-Grumbach M, Matsukawa S, Zitzelsberger M, Themistocleous AC, Strom TM, et al. Transcriptional regulator PRDM12 is essential for human pain perception. Nat Genet. 2015 Jul 25;47(7):803–8. doi: 10.1038/ng.3308

52. Silveira I, Miranda C, Guimarães L, Moreira M-C, Alonso I, Mendonça P, et al. Trinucleotide Repeats in 202 Families With Ataxia. Arch Neurol. 2002 Apr 1;59(4):623. doi: 10.1001/archneur.59.4.623

53. Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito M, et al. A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: A new polyglutamine disease? Hum Mol Genet. 1999;8(11):2047–53. doi: 10.1093/hmg/8.11.2047

54. Zarouchlioti C, Sanchez-Pintado B, Hafford Tear NJ, Klein P, Liskova P, Dulla K, et al. Antisense Therapy for a Common Corneal Dystrophy Ameliorates TCF4 Repeat Expansion-Mediated Toxicity. Am J Hum Genet. 2018 Apr;102(4):528–39. doi: 10.1016/j.ajhg.2018.02.010

55. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, et al. A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. PLoS One. 2012;7(11):5–12. doi: 10.1371/journal.pone.0049083

56. LaCroix AJ, Stabley D, Sahraoui R, Adam MP, Mehaffey M, Kernan K, et al. GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. Am J Hum Genet. 2019;104(1):35–44. doi: 10.1016/j.ajhg.2018.11.005

57. Wessels MW, Kuchinka B, Heydanus R, Smit BJ, Dooijes D, De Krijger RR, et al. Polyalanine expansion in the ZIC3 gene leading to X-linked heterotaxy with VACTERL association: A new polyalanine disorder? J Med Genet. 2010;47(5):351–5. doi: 10.1136/jmg.2008.060913

58. Strømme P, Mangelsdorf ME, Shaw MA, Lower KM, Lewis SME, Bruyere H, et al. Mutations in the human ortholog of Aristaless cause X-linked mental retardation and epilepsy. Nat Genet. 2002;30(4):441–5. doi: 10.1038/ng862

59. Bienvenu T, Poirier K, Friocourt G, Bahi N, Beaumont D, Fauchereau F, et al. ARX, a novel Prd-class-homeobox gene highly expressed in the telencephalon, is mutated in X-linked mental retardation. Hum Mol Genet. 2002;11(8):981–91. doi: 10.1093/hmg/11.8.981

60. Nallathambi J, Moumné L, Baere E, Beysen D, Usha K, Sundaresan P, et al. A novel polyalanine expansion in FOXL2: The first evidence for a recessive form of the blepharophimosis syndrome (BPES) associated with ovarian dysfunction. Hum Genet. 2007;121(1):107–12. doi: 10.1007/s00439-006-0276-0

61. De Baere E, Beysen D, Oley C, Lorenz B, Cocquet J, De Sutter P, et al. FOXL2 and BPES: Mutational hotspots, phenotypic variability, and revision of the genotype-phenotype correlation. Am J Hum Genet. 2003;72(2):478–87. doi: 10.1086/346118

62. Innis JW, Mortlock D, Chen Z, Ludwig M, Williams ME, Williams TM, et al.

Polyalanine expansion in HOXA13: Three new affected families and the molecular consequences in a mouse model. Hum Mol Genet. 2004;13(22):2841–51. doi: 10.1093/hmg/ddh306

63. Utsch B, Becker K, Brock D, Lentze MJ, Bidlingmaier F, Ludwig M. A novel stable polyalanine [poly(A)] expansion in the HOXA13 gene associated with hand-foot-genital syndrome: proper function of poly(A)-harbouring transcription factors depends on a critical repeat length? Hum Genet. 2002 May 4;110(5):488–94. doi: 10.1007/s00439-002-0712-8

64. Goodman FR, Bacchelli C, Brady AF, Brueton LA, Fryns JP, Mortlock DP, et al. Novel HOXA13 mutation and the phenotypic spectrum of hand-foot-genital syndrome. Am J Hum Genet. 2000;67(1):197–202. doi: 10.1086/302961

65. Albrecht AN, Kornak U, Böddrich A, Süring K, Robinson PN, Stiege AC, et al. A molecular pathogenesis for transcription factor associated poly-alanine tract expansions. Hum Mol Genet. 2004;13(20):2351–9. doi: 10.1093/hmg/ddh277

66. Sivan Y, Zhou A, Jennings LJ, Berry-Kravis EM, Yu M, Zhou L, et al. Congenital central hypoventilation syndrome: Severe disease caused by co-occurrence of two PHOX2B variants inherited separately from asymptomatic family members. Am J Med Genet Part A. 2019;179(3):503–6. doi: 10.1002/ajmg.a.61047

67. Mundlos S, Otto F, Mundlos C, Mulliken JB, Aylsworth AS, Albright S, et al. Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell. 1997;89(5):773–9. doi: 10.1016/S0092-8674(00)80260-3

68. Laumonnier F, Ronce N, Hamel BCJ, Thomas P, Lespinasse J, Raynaud M, et al. Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. Am J Hum Genet. 2002;71(6):1450–5. doi: 10.1086/344661

69. Woods KS, Cundall M, Turton J, Rizotti K, Mehta A, Palmer R, et al. Over- and underdosage of SOX3 is associated with infundibular hypoplasia and hypopituitarism. Am J Hum Genet. 2005;76(5):833–49. doi: 10.1086/430134

70. Rauch R, Hofbeck M, Zweier C, Koch A, Zink S, Trautmann U, et al. Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. J Med Genet. 2010;47(5):321–31. doi: 10.1136/jmg.2009.070391

71. Brown L, Paraso M, Arkell R, Brown S. In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: Alanine tract expansion modulates DNA binding and transactivation. Hum Mol Genet. 2005;14(3):411–20. doi: 10.1093/hmg/ddi037

72. Sato N, Amino T, Kobayashi K, Asakawa S, Ishiguro T, Tsunemi T, et al. Spinocerebellar Ataxia Type 31 Is Associated with “Inserted” Penta-Nucleotide Repeats Containing (TGGAA)n. Am J Hum Genet. 2009;85(5):544–57. doi: 10.1016/j.ajhg.2009.09.019

73. Corral-Juan M, Serrano-Munuera C, Rábano A, Cota-González D, Segarra-Roca A, Ispierto L, et al. Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. Brain. 2018;141(7):1981–97. doi: 10.1093/brain/awy137

74. Seixas AI, Loureiro JR, Costa C, Ordóñez-Ugalde A, Marcelino H, Oliveira CL, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. Am J Hum Genet. 2017;101(1):87–103. doi: 10.1016/j.ajhg.2017.06.007

75. Cortese A, Tozza S, Yau WY, Rossi S, Beecroft SJ, Jaunmuktane Z, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. Brain. 2020;143(2):489–90. doi: 10.1093/brain/awz418

76. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yau WY, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019 Apr 29;51(4):649–58. doi: 10.1038/s41588-019-0372-4

77. 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;141(8):2280–8. doi: 10.1093/brain/awy160

78. 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;50(4):581–90. doi: 10.1038/s41588-018-0067-2

79. Lei XX, Liu Q, Lu Q, Huang Y, Zhou XQ, Sun HY, et al. TTTCA repeat expansion causes familial cortical myoclonic tremor with epilepsy. Eur J Neurol. 2019;26(3):513–8. doi: 10.1111/ene.13848

80. 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;10(1):1–10. doi: 10.1038/s41467-019-12671-y

81. 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;10(1):1–14. doi: 10.1038/s41467-019-12763-9

82. Yeetong P, Pongpanich M, Srichomthong C, Assawapitaksakul A, Shotelersuk V, Tantirukdham N, et al. TTTCA repeat insertions in an intron of YEATS2 in benign adult familial myoclonic epilepsy type 4. Brain. 2019;142(11):3360–6. doi: 10.1093/brain/awz267

83. Huang W, Li L, Myers JR, Marth GT. ART: A next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4. doi: 10.1093/bioinformatics/btr708

84. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Preprint]. 2013 [cited 2021 May 1]. Available from: https://arxiv.org/abs/1303.3997v2

85. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. doi: 10.1093/bioinformatics/btp352

86. Halman A. STRipy’s supplementary code and results [dataset]. Zenodo; 2021. doi: 10.5281/zenodo.4939980

87. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019 Sep 5;47(15):e90–e90. doi: 10.1093/nar/gkz501

88. 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 Nov;27(11):1895–903. doi: 10.1101/gr.225672.117