Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences

Terence Gall-Duncan,1,2,3 Nozomu Sato,1,3 Ryan K.C. Yuen,1,2 and Christopher E. Pearson1,2

1 Program of Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario MSG 1L7, Canada; 2 Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

Expansions of gene-specific DNA tandem repeats (TRs), first described in 1991 as a disease-causing mutation in humans, are now known to cause >60 phenotypes, not just disease, and not only in humans. TRs are a common form of genetic variation with biological consequences, observed, so far, in humans, dogs, plants, oysters, and yeast. Repeat diseases show atypical clinical features, genetic anticipation, and multiple and partially penetrant phenotypes among family members. Discovery of disease-causing repeat expansion loci accelerated through technological advances in DNA sequencing and computational analyses. Between 2019 and 2021, 17 new disease-causing TR expansions were reported, totaling 63 TR loci (>69 diseases), with a likelihood of more discoveries, and in more organisms. Recent and historical lessons reveal that properly assessed clinical presentations, coupled with genetic and biological awareness, can guide discovery of disease-causing unstable TRs. We highlight critical but underrecognized aspects of TR mutations. Repeat motifs may not be present in current reference genomes but will be in forthcoming gapless long-read references. Repeat motif size can be a single nucleotide to kilobases/unit. At a given locus, repeat motif sequence purity can vary with consequence. Pathogenic repeats can be “insertions” within nonpathogenic TRs. Expansions, contractions, and somatic length variations of TRs can have clinical/biological consequences. TR instabilities occur in humans and other organisms. TRs can be epigenetically modified and/or chromosomal fragile sites. We discuss the expanding field of disease-associated TR instabilities, highlighting prospects, clinical and genetic clues, tools, and challenges for further discoveries of disease-causing TR instabilities and understanding their biological and pathological impacts—a vista that is about to expand.

More than 30 years ago, in 1991, expansions of DNA tandem repeats (TRs) at particular loci were first shown to cause human diseases, termed repeat expansion diseases (Kremer et al. 1991; LaSpada et al. 1991; Oberlé et al. 1991; Verkerk et al. 1991; Yu et al. 1991). After an initial period of successive identifications of similar trinucleotide repeat expansions (Pearson et al. 2005; López Castel et al. 2010), the rate of TR-associated disease discovery slowed as the limitations of technological methods reduced the ability to detect more complex pathogenic repeat expansions. However, recent technological advances in both DNA sequencing techniques and computational analysis have again increased the speed of discovery, with 17 new disease-causing and risk-associated TR expansions being published between 2019 and 2021 (Fig. 1; Table 1; Corbett et al. 2019; Cortese et al. 2019; Demaerel et al. 2019; Florian et al. 2019; Ishiura et al. 2019; LaCroix et al. 2019; Sone et al. 2019; Tian et al. 2019; van Kuilenburg et al. 2019; Yeetong et al. 2019; Katsumata et al. 2020; Ruggieri et al. 2020).

Recent repeat expansion identifications also highlight the importance of functional and clinical aspects of TR expansions in the discovery process. Expansions of TRs of the same repeat unit motifs can cause diseases with similar phenotypes independent of their genetic loci, supporting a gain-of-function pathogenesis hypothesis. For example, several SCAs caused by expansions of CAG repeat motifs present with similar motor phenotypes despite their expansions occurring within different genetic loci. More recently, careful analysis and categorization of clinical manifestations served as essential tools in recent discoveries where the same expansion mutations occurring at different genomic loci all resulted in benign adult familial myoclonic epilepsy (BAFME), also known as familial adult myoclonic epilepsy (FAME) (Ishiura et al. 2018, 2019; Corbett et al. 2019; Florian et al. 2019; Yeetong et al. 2019). Further, other new discoveries also reminded us of relatively underrecognized loss-of-function mechanisms which may precipitate pathogenesis. In the cases of Desbuquois dysplasia 2 (DBQD2; also known as Baratela-Scott syndrome) and glutaminase deficiency (GD), expansions of GC-rich TR sequences in the promoter

“These authors contributed equally to this review.
Corresponding author: christopher.pearson@sickkids.ca
Article published online before print. Article and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.299330.120.

© 2022 Gall-Duncan et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

32:1–27 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/22; www.genome.org

www.genome.org
regions cause pathogenic transcriptional suppression (LaCroix et al. 2019; van Kuilenburg et al. 2019). Further exploration of suspected loss-of-function mechanisms may therefore be beneficial in understanding how repeat mutations elicit pathogenesis.

While there have been a number of excellent reviews on diseases associated with TR expansions, they are mainly focused on disease mechanisms (Haman 2018; Rodriguez and Todd 2019). As such, considering the success of recent studies in the identification of new repeat disease motifs, the focus of this review is to highlight how DNA sequencing technologies and analytic approaches, coupled with clinical and biological assessment, facilitate repeat disease mutation discovery and our understanding of pathogenic mechanisms. We begin with a brief overview of the history of repeat disease gene discovery, with an emphasis on how these discoveries facilitated further discovery. Next, we will explore how new technologies are making more difficult

---

**Figure 1.** Overview of disease-associated repeat discovery by year, with colored inserts specifying the major technological breakthroughs that were used to make these discoveries. (AD) Alzheimer disease, (ALS/FTD) amyotrophic lateral sclerosis/frontotemporal dementia, (ASD) autism spectrum disorder, (BAFME) benign adult familial myoclonic epilepsy, (BD) bipolar disorder, (BPES) blepharophimosis, ptosis, and epicanthus inversus syndrome, (CANVAS) cerebellar ataxia, neuropathy, vestibular areflexia syndrome, (CCD) congenital central hypoventilation syndrome, (DBDQ2) Desbuquois dysplasia 2, (DM) myotonic dystrophy, (DRPLA) dentatorubropallidoluysian atrophy, (EDM1) multiple epiphyseal dysplasia, (EIEE1) epileptic encephalopathy, early infantile, 1, (EPM1) epilepsy, progressive myoclonus-1, (FEC) Fuchs endothelial corneal dystrophy-3, (FRDA) Friedreich’s ataxia, (FSHD) facioscapulohumeral muscular dystrophy, (FXTAS) fragile X axia/tremor syndrome, (GD) glutaminase deficiency, (HD) Huntington disease-like 2, (HFG) hand-foot-genital syndrome, (HPE) holoprosencephaly 5, (LOAD) late-onset Alzheimer disease, (MJD) Machado-Joseph disease, (NID) neural intraneurallineurial inclusion disease, (OPDM) oculopharyngodistal myopathy, (OPMD) oculopharyngeal muscular dystrophy, (OPML) oculopharyngeal myopathy with leukoencephalopathy, (PSACH) pseudoachondroplasia, (RCPS) Richieri-Costa-Pereira syndrome, (SBMA) spinal bulbar muscular atrophy, (SCA) spinocerebellar ataxia, (SMD) skeletal muscle disease, (SPD1) synpolydactyly-1, (S2) schizophrenia, (XPD) X-linked dystonia-parkinsonism, (22q11DS) 22q11 deletion syndrome. It has been concluded that FAME, BAFME, FEME, FCTE, and ADCME are the same clinical entity even if genetically heterogeneous—we use the acronym BAFME here as it is the most used acronym associated with the disease. The nonfolate-sensitive rare fragile sites FRA10B and FRA16B, caused by expanded AT-rich repeats, are not listed herein (see Table 1).

---

**Nomenclature**;

| Repeat motif; Tandem repeats (TRs) | Unit Size | Array size (in a normal population) |
|-----------------------------------|-----------|-----------------------------------|
| Macrosatellite                    | Several kb| Up to several hundred kb          |
| Satellite                         | 5–171 bp  | 100 kb to several Mb              |
| Minisatellite                     | 5–64 bp   | 1- to 20-kb range                 |
| Microsatellite                    | 1–4 bp    | Hundreds of bp                    |

*It is suggested to avoid the term “short tandem repeat (STR)” as the definition of “short” varies between studies. The ever-increasing number of TRs with units of almost any length will add further confusion as to definitions of “short,” “medium,” and “long.” Rather, “tandem repeat (TR), with a motif of X nucleotides” is preferred. The term “variable tandem repeat (VNTR)” also demands definition of the motif size.
## Table 1. Disease-associated repeat discovery by year

| Disease | Repeat unit | Gene, loci (fragile site) | Location of mutation | Inheritance or association | Reference | Publication date (D/M/Y) |
|---------|-------------|---------------------------|----------------------|---------------------------|-----------|------------------------|
| LGS     | CGG or CTG  | DIP2B or ATXN80S          | 5′ UTR (DIP2B)       | Association               | Qaiser et al. 2021 | 14/09/2021 |
| OPDM3   | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Yu et al. 2021    | 09/03/2021 |
| HMN     | GGC         | VWA1                      | Exon                 | AR                        | Pagnamenta et al. 2021 | 18/01/2021 |
| ALS     | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Yuan et al. 2020   | 12/2020    |
| ALS     | 69-mer TR   | WDR7                      | Intron               | Association               | Course et al. 2020 | 03/08/2020 |
| ASD     | Variable repeats of 2–20 bp | 2588 different loci | Variable             | Association               | Trost et al. 2020  | 27/07/2020 |
| SMD     | 99-mer TR   | PLN4                      | Exon                 | AD                        | Ruggieri et al. 2020 | 25/05/2020 |
| LOAD    | GGC         | GIPC1                     | 5′ UTR               | AD/AR                     | Deng et al. 2020   | 14/05/2020 |
| OPDM2   | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Yu et al. 2021    | 09/03/2021 |
| OPML    | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Tian et al. 2019; Ishiura et al. 2019; Sone et al. 2019 | 06/06/2019 |
| NIID    | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Tian et al. 2019; Ishiura et al. 2019; Sone et al. 2019 | 22/07/2019 |
| GD      | GCA         | GLS                       | 5′ UTR               | AR                        | van Kullenburg et al. 2019 | 11/04/2019 |
| CANVAS  | GGGAA       | RFC1                      | Intron               | AR                        | Cortese et al. 2019; Rafehi et al. 2019 | 24/06/1994 |
| BD and SCZ | 30-mer TR | CAGA/CACA                 | Intron               | X-linked                  | Song et al. 2018  | 09/08/2018 |
| AD      | GGC         | NOTCH2NLC                 | 5′ UTR               | AD                        | Ishiura et al. 2019 | 05/03/2018 |
| BAFME1  | TTICTA and TTTA | TNRC6A                | Intron               | AD                        | Ishiura et al. 2019 | 05/03/2018 |
| BAFME6  | TTICTA and TTTA | TNRC6A                | Intron               | AD                        | Ishiura et al. 2019 | 05/03/2018 |
| BAFME7  | TTICTA and TTTA | TAF2                | Intron               | AD                        | Ishiura et al. 2019 | 05/03/2018 |
| XDP     | CTCCTCT     | TAF1                      | Intron               | X-linked                  | Bragg et al. 2017 | 11/12/2017 |
| SCA37   | ATTC        | DAB1                      | Intron               | AD                        | Seixas et al. 2017 | 06/07/2017 |
| ID      | CGG         | ZNF713                    | 5′ UTR               | Association               | Metsu et al. 2014a | 04/09/2014 |
| ID      | CGG         | AFE3 (FRA2A)              | Promoter             | Association               | Matsu et al. 2014b | 24/04/2014 |
| RCP5    | 18- or 20-bp TR | EIF4A                 | 5′ UTR               | AR                        | Favaro et al. 2014 | 19/12/2013 |
| FECO3   | CTC         | TCF4                      | Intron               | AD                        | Wielen et al. 2012 | 21/11/2012 |
| ALS/FTD, others | GGGGCC  | C9orf72                   | Intron               | AD                        | Renton et al. 2011; Dejesus-Hernandez et al. 2011 | 21/09/2011 |
| SCA36   | GCC         | NOP56                     | Intron               | AD                        | Kobayashi et al. 2011 | 16/06/2011 |
| SCA31   | TGGAA/TTCCA | BEAN1/TK2                | Intron               | AD                        | Sato et al. 2009  | 29/10/2009 |
| ID      | CGG         | DIP2B (FRA12A)            | 5′ UTR               | AD                        | Winnepennickx et al. 2007 | 12/12/2006 |
| ID, SCZ | CGG         | C1orf80 (FRA11A)          | 5′ UTR               | Association               | Debaker et al. 2007 | 14/12/2007 |
| ID, SCZ | CGG         | FRA10A1 (FRA10A)          | 5′ UTR               | Association               | Sarafidou et al. 2004 | 23/04/2004 |
| CCHS    | Polyanaline-coding GCN | FOX2                 | Exon                 | AD                        | Amiel et al. 2003 | 17/03/2003 |
| BPES    | Polyanaline-coding GCN | FOX2                 | Exon                 | AD                        | De Baere et al. 2003 | 14/01/2003 |
| EIEE1   | Polyanaline-coding GCN | ARX                  | Exon                 | X-linked                  | Strømme et al. 2002 | 11/03/2002 |
| FXTAS   | CGG (premutation range) | FMRI (FRA1)            | 5′ UTR               | X-linked                  | Hagerman et al. 2001; Howard-Peebles 1980 | 10/07/2001 |
| HDL2    | CGC         | JPH3                      | Alt-exon             | AD                        | Holmes et al. 2001 | 05/11/2001 |
| DM2     | CCTG        | CNBP                      | Intron               | AD                        | Liquori et al. 2001 | 03/08/2001 |
| HFG     | Polyanaline-coding GCN | HOXA13                | Exon                 | AD                        | Goodman et al. 2000 | 05/06/2000 |

(continued)
Table 1. Continued

| Disease                | Repeat unit | Gene, loci (fragile site) | Location of mutation | Inheritance or association | Reference                                                                 | Publication date (D/M/Y) |
|------------------------|-------------|---------------------------|----------------------|----------------------------|---------------------------------------------------------------------------|--------------------------|
| SCA10                  | ATTCT       | ATXN10                    | Intron               | AD                         | Matsuura et al. 2000                                                      | –/10/2000                |
| PSACH/EDM1             | GAC         | COMP                      | Exon                 | AD                         | Deere et al. 1999                                                        | 14/07/1999               |
| SCA17                  | CAG         | TBP                       | Exon                 | AD                         | Koide et al. 1999; Imbert et al. 1994                                    | –/06/1994                |
| SCA12                  | CAG         | PEP2R28                   | 5′ UTR               | AD                         | Holmes et al. 1999                                                      | –/12/1999                |
| SCA8                   | CTG/CAG     | ATXNB05/ATXN8             | 3′ UTR/Exon          | AD                         | Koob et al. 1999                                                      | 01/04/1999               |
| HPES                   | Polyalanine-coding GCN | ZIC2              | Exon                 | AD                         | Brown et al. 1998                                                      | 01/10/1998               |
| Cancer, ND             | CAG/CAA     | AIB1/SRC-3/RAC3           | Exon                 | Somatic; Amplified gene (1–23 copies) | Shirazi et al. 1998; Anzick et al. 1997                                 | –/07/2008; 15/08/1997    |
| Nondisease-associated  | 16- to 52-bp TR AT-rich motif | FRA108         | Chr 10q25             | n/a                        | Hewett et al. 1998                                                      | 01/05/1998               |
| OPMD                   | GCG         | PABPN1                    | Exon                 | AD                         | Brais et al. 1998                                                      | 01/02/1998               |
| SCZ                    | CAG         | KCN3                      | Exon                 | AD                         | Chandry et al. 1998                                                    | –/01/1998                |
| CCD                    | Polyalanine-coding GCN | RUNX2              | Exon                 | AD                         | Mundlos et al. 1997                                                    | 30/05/1997               |
| EPM1                   | CCCCCGCCCCC GCG | C3TB              | Promoter             | AR                         | Lalioti et al. 1997; Virtaneva et al. 1997; Laforenire et al. 1997 | 24/04/1997; 01/04/1997; 01/03/1997 |
| Nondisease-associated  | 33-bp TR AT-rich motif | RNA922 (FRA16B)         | Intron               | n/a                        | Yu et al. 1997                                                        | 07/02/1997               |
| SCA7                   | CAG         | ATXN7                     | Exon                 | AD                         | Lindblad et al. 1996                                                    | 01/10/1996               |
| SPD1                   | Polyalanine-coding GCN | HOXD13              | Exon                 | AD                         | Akarsu 1996                                                            | 01/07/1996               |
| FA                     | GAA         | FXN                       | Intron               | AR                         | Campuzano et al. 1996                                                    | 08/03/1996               |
| SCA2                   | CAG         | ATXN2                     | Exon                 | AD                         | Imbert et al. 1996; Sanpei et al. 1996; Pulst et al. 1996; Trotter et al. 1995; Pulst et al. 1993 | 01/11/1996; 01/11/1996; 01/11/1996; 23/11/1995; 01/09/1993 |
| FXPO1                  | CGG         | FMR1 (FRAXA)              | 5′ UTR               | X-linked                   | Conway et al. 1995                                                      | 29/07/1995               |
| Jacobsen syndrome      | CGG         | CBL2 (FRA11B)             | 5′ UTR               | Isolated cases             | Jones et al. 1994                                                      | 01/12/1994               |
| Fragile X syndrome F   | CGG         | TMEM185A (FRAXF)          | 5′ UTR               | Isolated cases             | Ritchie et al. 1994; Parrish et al. 1994; Koide et al. 1994; Orr et al. 1993 | 01/12/1994; 01/11/1994; 01/01/1994; 01/07/1993 |
| DRPLA/HRS              | CAG         | ATN1                      | Exon                 | AD                         | Nagafuchi et al. 1994; Koide et al. 1994; Orr et al. 1993               | 01/01/1994; 01/01/1994; 01/07/1993 |
| SCA1                   | CAG         | ATXN1                    | Exon                 | AD                         | –                                                                        | 1993–1998                |
| “no disease”           | CAG/CTG     | CTG18.1/ERDA1 (see FECD3/TCF4) | Intron               | AD                         | –                                                                        | 1993–1998                |
| Fragile X syndrome E   | CCG         | AF2 (FRAXE) & FMR2        | Promoter and 5′ UTR  | X-linked                   | Gu et al. 1996; Knight et al. 1993; The Huntington’s Disease Collaborative Research Group | 01/05/1996; 01/07/1993; 26/03/1993 |
| HD                     | CAG         | Htt                       | Exon                 | AD                         | Bell 1941; van Deutekom et al. 1993; Wijmenga et al. 1992                | –/01/1941; 01/12/1993; 01/09/1992 |
| FSHD                   | 3.4-kb D4Z4 macrosatellite | Chr 4q35 | Noncoding             | AD                         | van Deutekom et al. 1993; Wijmenga et al. 1992                          | 01/12/1993               |
| DM1                    | CTG         | DMPK                      | 3′ UTR               | AD                         | Fu et al. 1992; Mahadevan et al. 1992; Brook et al. 1992; Aslanidis et al. 1992; Buxton et al. 1992; Harley et al. 1992; Höweler et al. 1989; Bell 1941 | 06/03/1992; 06/03/1992; 21/02/1992; 06/02/1992; 06/02/1992; 01/06/1989; –/01/1941 |
| SBMA                   | CAG         | AR                        | Exon                 | AD                         | La Spada et al. 1991                                                    | 04/07/1991               |

(continued)
sequences in the genome accessible and discuss the need for further development of analytical tools. Lastly, we will highlight how some of the recent findings identified relatively underrecognized clinical and mechanistic features of TR-expansion-related disorders, which should not be overlooked as future research aims to improve our understanding of repeat diseases and their underlying mechanisms. By covering these topics, we attempt to provide guidance for future investigations into TRs and their roles in physiological and disease processes through the integration of technology and biological understanding.

Part I: Technological advances and repeat disease mutation discovery

Historical overview of disease-associated repeat expansion discovery

The initial discoveries in the early 1990s were trinucleotide repeats, namely a CGG repeat in the 5’ UTR of FMR1 (Kremer et al. 1991; Oberlé et al. 1991; Pieretti et al. 1991; Verkerk et al. 1991; Yu et al. 1991), a polyglutamine-coding CAG repeat in the AR gene (La Spada et al. 1991), and a CTT repeat in the 3’ UTR of DMPK (Aslanidis et al. 1992; Brook et al. 1992; Buxton et al. 1992; Fu et al. 1992; Harley et al. 1992; Mahadevan et al. 1992). Expansions of these repeats caused fragile X syndrome (FXS), spinal and bulbar muscular atrophy (SBMA), and myotonic dystrophy type 1 (DM1), respectively. It was later found that tetra- (Liquori et al. 2001), penta- (Matsuura et al. 2000; Sato et al. 2009), hexa- (DeJesus-Hernandez et al. 2001), and hepta- (Kawaguchi et al. 1994; Koide et al. 1994; Nagafuchi et al. 1994) nucleotide repeats caused myotonic dystrophy type 2 (DM2), Friedrich ataxia (FRAXA), and fragile site-associated ovarian insufficiency (FXPOI), respectively. These repeat expansions have been associated with a wide spectrum of neurological, neuromuscular, and ophthalmological disorders, highlighting the importance of repeat expansion diseases in genetic counseling and clinical practice.

Table 1. Continued

| Disease                        | Repeat unit | Gene, loci (fragile site) | Location of mutation | Inheritance or association | Reference | Publication date (D/M/Y) |
|-------------------------------|-------------|---------------------------|----------------------|----------------------------|-----------|-------------------------|
| Fragile X syndrome A          | CGG         | FMR1 (FRAXA)              | 5’ UTR               | X-linked                   | Fu et al. 1991; Pieretti et al. 1991; Kremer et al. 1991; Verkerk et al. 1991; Yu et al. 1991; Oberlé et al. 1991; Heitz et al. 1991; Bell et al. 1991; Vincent et al. 1991; Warren et al. 1987; Nussbaum et al. 1986; Sutherland et al. 1985; Pembrey et al. 1985; Sherman et al. 1985; Sherman et al. 1984; Lubs 1969; Martin and Bell 1943 | 20/12/1991; 23/08/1991; 21/06/1991; 31/05/1991; 24/05/1991; 24/05/1991; 08/03/1991; 22/02/1991; 14/02/1991; 24/07/1987; 31/05/1991/01/01/1986; –/10/1985; –/08/1985; 01/04/1985; 01/01/1984; –/05/1969; 07/10/1943 |

Abbreviations: AD, Alzheimer disease; ALS/FTD, amyotrophic lateral sclerosis/frontotemporal dementia; ASD, autism spectrum disorder; BAFME, benign adult familial myoclonic epilepsy; BD, bipolar disorder; BPE5, blepharophimosis, ptosis, and epicanthus inversus syndrome; CANSAS, cerebellar ataxia, neuropathy, vestibular areflexia syndrome; CCD, cleidocranial dysplasia; CCHS, congenital central hypoventilation syndrome; DBDQ2, Desbuquois dysplasia 2; DM, myotonic dystrophy; DRPLA, dentatorubropallidolysian atrophy; EDMD1, multiple epiphyseal dysplasia; EEE1, epileptic encephalopathy, early infantile, 1; EPM, epilepsy, progressive myoclonus; FECD3, Fuchs endothelial corneal dystrophy-3; FA, Friedreich ataxia; FSHD, facioscapulohumeral muscular dystrophy; FXPOI, Fragile X-associated primary ovarian insufficiency; FXTAS, fragile X ataxia/tremor syndrome; GD, glutaminase deficiency; HRS, Haw River syndrome; HD, Huntington disease; HDL2, Huntington disease-like 2; HFG, hand-foot-genital syndrome; HMN, hereditary motor neuropathy; HPS5, holoproencephaly 5; ID, intellectual disability; LOAD, late-onset Alzheimer disease; MJD, Machado-Joseph disease (aka SCAD); NIID, neuronal intranuclear inclusion disease; OPMD, oculopharyngodistal myopathy; OPLM, oculopharyngeal muscular dystrophy; OPML, oculopharyngeal myopathy with leukoencephalopathy; PSACH, pseudoachondroplasia; RCP5, Richieri-Costa-Pereira syndrome; SMBA, spinal bulbar muscular atrophy (aka Kennedy’s disease); SCA, spinocerebellar ataxia; SMD, skeletal muscle disease; SPDP1, synpolydactyly-1; SCZ, schizophrenia; XPD, X-linked dystonia-parkinsonism; 22q11DS, 22q11 deletion syndrome.

Abbreviations: AD, Alzheimer disease; ALS/FTD, amyotrophic lateral sclerosis/frontotemporal dementia; ASD, autism spectrum disorder; BAFME, benign adult familial myoclonic epilepsy; BD, bipolar disorder; BPE5, blepharophimosis, ptosis, and epicanthus inversus syndrome; CANSAS, cerebellar ataxia, neuropathy, vestibular areflexia syndrome; CCD, cleidocranial dysplasia; CCHS, congenital central hypoventilation syndrome; DBDQ2, Desbuquois dysplasia 2; DM, myotonic dystrophy; DRPLA, dentatorubropallidolysian atrophy; EDMD1, multiple epiphyseal dysplasia; EEE1, epileptic encephalopathy, early infantile, 1; EPM, epilepsy, progressive myoclonus; FECD3, Fuchs endothelial corneal dystrophy-3; FA, Friedreich ataxia; FSHD, facioscapulohumeral muscular dystrophy; FXPOI, Fragile X-associated primary ovarian insufficiency; FXTAS, fragile X ataxia/tremor syndrome; GD, glutaminase deficiency; HRS, Haw River syndrome; HD, Huntington disease; HDL2, Huntington disease-like 2; HFG, hand-foot-genital syndrome; HMN, hereditary motor neuropathy; HPS5, holoproencephaly 5; ID, intellectual disability; LOAD, late-onset Alzheimer disease; MJD, Machado-Joseph disease (aka SCAD); NIID, neuronal intranuclear inclusion disease; OPMD, oculopharyngodistal myopathy; OPLM, oculopharyngeal muscular dystrophy; OPML, oculopharyngeal myopathy with leukoencephalopathy; PSACH, pseudoachondroplasia; RCP5, Richieri-Costa-Pereira syndrome; SMBA, spinal bulbar muscular atrophy (aka Kennedy’s disease); SCA, spinocerebellar ataxia; SMD, skeletal muscle disease; SPDP1, synpolydactyly-1; SCZ, schizophrenia; XPD, X-linked dystonia-parkinsonism; 22q11DS, 22q11 deletion syndrome.

*It has been concluded that FAME, BAFME, FEME, FCTE, and ADCME are the same clinical entity even if genetically heterogeneous—we use the acronym BAFME here as it is the most used acronym associated with the disease.*
the identification of the CGG expansion mutation responsible for the cytogenetic fragile site, FRAXA, which until then had been the main diagnostic marker of FXS (Lubs 1969). While detailed coverage is beyond the scope of this review, an appreciation of the approaches used is relevant. Among a variety of competing potential theories (not all covered here), one was the hypothesis that an unstable repeat sequence would be the cause of the fragile site FRAXA and disease FXS. Those initial suspicions, hypothesizing the involvement of an unstable amplified repeat tract, were based upon the biology of chromosomal fragile site induction and the puzzling genetics of the disease (Sutherland et al. 1985; Nussbaum et al. 1986; Horii et al. 1988). Among the first experimental evidence supporting the involvement of an unstable DNA sequence were cytogenetic observations of chromosomal instability at the fragile site in rodent-human somatic cell hybrids; a reagent subsequently cloned, sequenced, and localized cytogenetically (FISH) the causative CGG expansion mutation (Warren et al. 1985). Warren and colleagues, citing the repeat-hypothesis proposed in 1985 (Sutherland et al. 1985; Ledbetter et al. 1986; Nussbaum et al. 1986; Horii et al. 1988), concluded “… that the fragile X site is a reiterated DNA sequence of variable length, the longest length being found in fully penetrant males and the shortest in phenotypically normal individuals…” Frailty in this region of the X has been shown to support this model in that normal, transmitting, and affected male X chromosomes (in somatic cell hybrids) show increasing frequencies of fragility… [T]he observation of reduced chromosome fragility at the translocation junctions lends support for the model of the fragile X site as a reiterated DNA sequence.” In 1991, the concept of genetic instability was further supported by variably slow-migrating DNAs on Southern blots—suspected as amplified repeats (Oberlé et al. 1991; Yu et al. 1991) and soon after revealed as a CGG expansion (Fu et al. 1991; Kremer et al. 1991; Pieretti et al. 1991; Verkerk et al. 1991). This was possible through using cytogenetics/FISH, coupled with somatic cell hybrids for FRAXA breakpoint mapping, AluPCR, and positional cloning, which together permitted identification of the CGG expansion in FMR1 (Warren et al. 1987; Bell et al. 1991; Heitz et al. 1991; Kremer et al. 1991; Vincent et al. 1991). Cytogenetics/FISH and molecular genetics are still required to validate the molecular mapping of fragile sites (Warren et al. 1987; Bell et al. 1991; Heitz et al. 1991; Kremer et al. 1991; Vincent et al. 1991). For specific details of FRAXA/CGG/FMR1 discoveries, we refer readers to a focused review, published during that early time (Oostra and Verkerk 1992). Indeed, it seems that the advances of the fragile X research, discovering a repeat expansion as the genetic cause for a disease with unusual inheritance patterns (Sherman paradox), incomplete penetrance, and strong parent-of-origin effects, paved the way for repeat-centered efforts for many of the other diseases. The localization of the mutant regions of other repeat diseases involved the use of many mapping techniques, including radiation-reduced hybrids, flow-sorted chromosome libraries, CpG island screens, exon trapping, exon amplification, and use of cosmid/yeast artificial chromosome libraries (La Spada et al. 1991; Aasland et al. 1992; Buxton et al. 1992; The Huntington Disease Collaborative Research Group 1993). The probing of positionally mapped disease regions for suspected repeat tract length variations led to the discovery of many of the other diseases that similarly showed unusual inheritance patterns and parent-of-origin effects. It was predicted that the mutation causing DM1, which showed strong genetic anticipation, similar to the Sherman paradox of FXS, could be caused by an unstable repeat (Höweler et al. 1989; Sutherland et al. 1991). Following the discovery of the DM1 mutation as an expanded CTG repeat based upon its tight association with genetic anticipation, it was predicted that HD and SCAs (known then as olivopontocerebellar ataxias) would be caused by gene-specific repeat expansions (Caskey et al. 1992; Harper et al. 1992). Following the explanation of HDs genetic anticipation by a CAG expansion (Snell et al. 1993; Trotter et al. 1994), the connection was solidified, and it was predicted that SCAs, bipolar disorder/schizophrenia, and non-FXS linked autism could also be caused by repeat expansions (Pulst et al. 1993; Ross et al. 1993). Each of these predictions, to some degree, turned out to be true for numerous SCAs, associatively for at least one form of bipolar disorder/schizophrenia (ACNA1C) (Song et al. 2018), and most recently, autism spectrum disorder (ASD) (Trost et al. 2020). It is notable that SBMA does not show obvious genetic anticipation nor high levels of repeat instability, and the discovery of the CAG expansion in the androgen receptor in affected families was a further extension of the already known polymorphism of the repeat in the unaffected population (Lubahn et al. 1988; Tilley et al. 1989; Edwards et al. 1991, 1992; La Spada et al. 1991).

While these initial discoveries used technologies that were mostly not repeat-specific, following the discoveries of several CAG/CTG expansions, a series of methodological protocols was developed to detect expansions of this trinucleotide repeat without knowledge of their genomic loci: Repeat Expansion Detection (RED) (Schalling et al. 1993), Direct Identification of Repeat Expansion and Cloning Technique (DIRECT) (Sanpei et al. 1996), and Repeat Analysis, Pooled Isolation, and Detection of expanded trinucleotide repeat clones (RAPID) (Koob et al. 1998). These protocols were based on completely novel ideas in the era of positional cloning and were used to identify several new disease loci caused by unstable repeat expansions: RED brought about the discoveries of spinocerebellar ataxia type 7 (SCA7) (Lindblad et al. 1996), SCA12 (Holmes et al. 1999), Huntington disease-like 2 (HDL2) (Holmes et al. 2001; Margolis et al. 2001) and CAG/CTG18.1 (Breschel et al. 1997); DIRECT led to the identification of the ATXN2 mutation (Sanpei et al. 1996) and RAPID to the SCA8 repeat expansion (Koob et al. 1998). It should also be noted that the discoveries of the CAG mutations causative for SCA2 and SCA7 were immensely facilitated by detection of polyQ aggregates with a monoclonal antibody, which predicted expansions in SCA2 through detection of expansions in extracts of SCA2 patient cells (Trotter et al. 1995).

The continuous discovery of new TR expansion mutations in the 1990s fully leveraged the power of the Human Genome Project, as the huge numbers of sequence-tagged site (STS) markers that became available enabled fine mapping of the disease loci. Today, locus mapping can be done with high-density SNP typing using microarrays (Gentelen and Chee 1999), which facilitates the completion of linkage analysis more rapidly than ever before. The reference sequence of the human genome and the variation database made “resequencing” approaches possible. Following this, the first decade of the 21st century witnessed rapid development of new DNA sequencing technologies, now called second-generation sequencing (or next-generation sequencing: NGS). Three of the first to be widely used were Illumina’s GA/HiSeq Systems, 454 Life Sciences’ 454 System and Applied Biosystems’ Sequencing by Oligo Ligation Detection (SOLID) (van Dijk et al. 2014). Together with the completion of draft human genome sequence, these high-throughput systems contributed to the high number of genomic variations discovered to result in various human phenotypes. The new sequencing technologies led to an
increasing number of gene discoveries for Mendelian conditions from 2010 onward (Bamshad et al. 2019). However, due to limitations in analytical tools available to handle repeat sequences and the technical weaknesses associated with fidelity and processivity of DNA polymerases, it took another decade for new sequencing technologies to begin to enhance identifying disease-causing TR expansions (Fig. 1).

**Bioinformatic algorithms are unleashing the potential of NGS for repeat disease discovery**

Application of NGS approaches in this field was first published in 2011 in one of two papers that reported the discovery of the GGGGCCC repeat expansion in the C9orf72 gene associated with familial amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD) (Renton et al. 2011). Massive parallel paired-end sequencing by HiSeq 2000 permitted rapid data collection, but the expanded repeat was identified through manual inspection and realignment of the sequence data in the candidate region, which was only possible because the linkage block had been narrowed down to a region of 232 kb. Validation of the C9orf72 repeat expansion in 2011 required the use of Southern blotting (DeJesus-Hernandez et al. 2011), a method still required in 2020 for validation of repeat expansions (Trost et al. 2020). Even more recently, in a 2019 study, biallelic expansions of an (AAGGG)N repeat in the intron of RFC1 were identified as responsible for cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS) by a similar method—visual analysis of assembled short-read sequencing (SRS) data generated by HiSeq 4000 system within a 1.7-Mb candidate region (Cortese et al. 2019). Despite these success stories in the use of NGS, over this same time period, several repeat disease mutation discoveries still depended on conventional methods despite the technological advances. For example, the 2009 identification of the mutation that causes SCA31 was achieved solely by "traditional" methods: bacterial artificial chromosome (BAC)-based cloning, Sanger sequencing, and Southern blot, with targeted shotgun resequencing (Sato et al. 2009).

The techniques employed by the Human Genome Project notoriously struggled with TR stretches and, in fact, there are still regions with long TRs yet to be correctly assembled—for example, the classical satellite repeats I-IV (Miga 2015). Shorter simple tandem repeats (STR), with 1- to 6-bp motif units, were also underrecognized, with recent bioinformatic assessment showing that these STRs comprise 6.77% of the human genome—more than twice what was initially predicted (Shortt et al. 2020). To this day, repeats pose significant hurdles for even NGS, ranging from technical hurdles including difficulties of bacterial cloning, PCR amplification, and sequence read size limits, to computational hurdles including misalignment and omission of the repeat or flanks from the reference genome (LaCroix et al. 2019). The high GC-content of TRs, and of the deletion-prone regions in which they are often embedded, can account for hindering the identification of a disease-causing mutation, as highlighted by the recent discovery of a GC-rich 10-mer repeat with compound heterozygous deletions (Pagnamenta et al. 2021) and of biallelic deletions in the VWA1 gene (Deschauer et al. 2021). VWA1 had been identified as the disease-causing gene of neuromyopathy and loss-of-function supported by animal models, but the mutation remained veiled by technical hurdles. Similarly, the identification of repeat expansions in the CSTR gene of either a 15-mer, an 18-mer (Virtaneva et al. 1997), or a dodecamer (12-mer) minisatellite repeat expansion (LaFrenière et al. 1997; LaIoti et al. 1997). Further analysis confirmed a dodecamer composed exclusively of G and C residues (Lalioti et al. 1997). From this, it is clear that both first generation and NGS and analysis of NGS data struggle to detect TRs, with recent advances demonstrating that much more sophisticated bioinformatic tools are necessary for their detection (van der Sanden et al. 2021).

One major technical issue is that disease-associated expanded TR tracts (whose tracts are, for the most part, much shorter than those in satellite DNA) have been tough obstacles for Illumina's widely used HiSeq systems that depend on assembly of short-read sequences (typically paired 150-base reads). This is because reads filled entirely or partially with expanded simple repeat sequences cannot be assembled accurately and because sequences derived from TRs are associated with a higher sequencing error rate due to their low complexity and/or due to their high GC-content, thereby further promoting misassembly (Benjamini and Speed 2012).

One way to overcome this is to increase the sequencing read length so that flanking sequences are encompassed within the read to allow for simpler alignment (Ummat and Bashir 2014). This is especially needed after the 454 System—which had relatively long read lengths among second generation sequencers, reaching up to 1 kb—was discontinued. The recent discovery of CGG repeat expansions in neuronal intranuclear inclusion disease (NIID) was achieved with long-read whole-genome sequencing (WGS), using either Pacific Biosciences' PacBio RS II (average read length exceeding 10 kb; polymerase-based) or Oxford Nanopore Technologies' PromethION (potential read length of more than 2 Mbp; synthesis-free) (Sone et al. 2019). Long-read sequencing (LRS) in particular is a burgeoning opportunity for repeat disease gene discovery and will be covered in more detail in the following section.

Another critical technological advancement that is facilitating greater ability to study repeat sequences is the production of high-throughput sequencing data without PCR amplification (Scior et al. 2011; Hommelsheim et al. 2015). PCR amplification of repeat units is especially difficult because polymers may not proceed completely through the tract, thus generating fragmented reads of pure repeat sequences which cannot be assembled accurately with high confidence or are inappropriately assembled (e.g., as artificial inversions). Further, shorter repeats within the genome will also show PCR bias as they are more easily polymerized through than longer repeats, thereby biasing the coverage within the genome. On a per-sequence basis, artifacts and bias are also prevalent due to the high error rate of polymerases within repeat sequences, the truncated products acting as preferentially polymerized templates during subsequent rounds of polymerization, and the misalignment of primers within repetitive templates (Scior et al. 2011; Benjamini and Speed 2012; Hommelsheim et al. 2015). The development of Illumina's amplification-free sequencing technology (Kozarewa et al. 2009; Kozarewa and Turner 2011) made it possible to obtain massive SRS data with unbiased coverage of the genome. Using this technology, one of the most challenging sequences for PCR, a CGG repeat expansion, was discovered to cause NIID and related diseases (Ishiura et al. 2019), independently of the LRS-based study (Sone et al. 2019).

Perhaps the largest leap in advancement and a necessary component of identifying expanded repeats within sequencing data is the development of data analysis algorithms to correctly detect expanded TRs in a locus-specific manner from SRS data. A series of TR-expansion gene discoveries (Ishiura et al. 2018, 2019) was made with the help of TRHist (Doi et al. 2014), an algorithm specifically developed to detect and correctly annotate TR expansions.
Their identification of expanded TTCCA repeats within long TTTTA repeats causing BAFME1 was also facilitated by long-read sequencing technologies. The recent discovery of GCA repeat expansions in GD (van Kuilenburg et al. 2019) was done with the help of ExpansionHunter (EH) (Dolzhenko et al. 2017), another example of software for detection of TR expansions from SRS WGS data.

In addition to TRhist and EH, there are several other algorithms for detecting TR expansions from short-read WGS data, including STRetch (Dashnow et al. 2018), GangSTR (Mousavi et al. 2019), exSTra (Tankard et al. 2018), and TREDPARSE (Tang et al. 2017). Most of these algorithms require catalogs of all previously found repeat motifs within the reference genome. They leverage such information to detect expansions of these motifs at specified locations in the sequencing reads aligned to the reference. Initially, these algorithms were not well suited for accurate detection of TR expansions in regions with complex repeat configurations due to alignment complications and improper detection of individual repeat motifs within complex repeat tracts. Examples include: the expanded CCTG repeat causative for DM2, which is located immediately adjacent to CA and CAGA repeat tracts (Liquori et al. 2001), the expanded CAG repeat associated with Huntington disease (HD), which is followed by a CCG repeat tract (The Huntington Disease Collaborative Research Group 1993), and the complex polyalanine-coding GCN repeat which is expanded in congenital central hypoventilation syndrome (CCHS) (Amiel et al. 2003). In the face of these obstacles, a couple of improvements were made to EH. The new version (EH ver. 3.0.0) can now handle complex TR expansions with the help of catalog data from complex TR loci (Dolzhenko et al. 2019) and has been shown to be able to accurately genotype the polyalanine repeat in the CCHS-causing PHOX2B gene. EH and other catalog-based algorithms are updated to add more complex repeats to their catalogs as they are discovered; however, this requires the identification of a repeat prior to its inclusion in the catalog. As a consequence of this, these catalog-based algorithms can effectively be used to find repeats that fall within known motifs but are blind to motifs that have yet to be identified.

In the same vein, another major challenge is that some TR expansion disorders arise by “insertion” of new expanded repeat units into another repeat, such as the TGGAA repeat that appears at the TAAAA repeat locus in SCA31 (Sato et al. 2009) and has been shown to be able to accurately genotype the polyalanine repeat in the CCHS-causing PHOX2B gene. EH and other catalog-based algorithms are updated to add more complex repeats to their catalogs as they are discovered; however, this requires the identification of a repeat prior to its inclusion in the catalog. As a consequence of this, these catalog-based algorithms can effectively be used to find repeats that fall within known motifs but are blind to motifs that have yet to be identified.

Another example of “insertion” of new pathogenic repeats into already existing repeats of distinct sequences is the RFC1 repeats. The pathogenic repeat motif [(AAAGG)400–2000 or (ACAGG)exp] must be present homozgyously to cause CANVAS, but when present heterozygously, a nondisease state arises. In contrast, the nonpathogenic motifs, even expanded are [(AAAAAG)11 or (AAAAAG)exp] and (AAAGG)exp]. The recessive aspect of this mutation, and the change of the repeat motif at the same locus relative to the nonaffected population, suggests that this is a highly polymorphic repeat. That the disease-causing motifs include a seemingly a limited subset of sequences suggests that this repeat sequence is at the core of CANVAS disease (Akçimen et al. 2019), as with SCA31 (Sato et al. 2009).

The reference genome is missing either the repeat and/or some of its flanking sequences for numerous repeat-expandable genes, including CANVAS, SCA31, SCA37, BAFME1, 2, 3, 4, 6, 7, and DBQD. This is likely due to the inability of the methods used to handle the repeat. For example, the XYLT1 CCG repeat and its flanking sequences could not be easily obtained by PCR amplification of the G-rich promoter from a healthy individual (devoid of the CCG expansion) without highly specialized conditions, and the authors suspected G-quadruplex structures as the problem (Faust et al. 2014), also a likely source for its absence in the reference genome. These sequences are unstable in bacterial vectors used for the initial sequencing of the reference genome. Retrospectively, it is understandable that the reference genome was missing the repeat and its flanking sequences. In fact, this had previously been observed for FMR1, where two reports found different sequences flanking the repeat, derived from a clone from a normal X Chromosome; it was concluded that “…the sequences missing in the Kremer report (Kremer et al. 1991) are likely an artifact of the numerous cloning steps involved in preparation of the template and further underscore the instability of the region in heterologous hosts” (Fu et al. 1991).

To some degree, the hurdles noted above may be overcome through the production of “gapless” reference genomes via long-read sequencing. These efforts are being spearheaded by the Telomere-to-Telomere (T2T) Consortium (https://sites.google.com/ucsc.edu/t2tworkinggroup), which aims to fill in the numerous gaps within the reference genome by conducting complete long-read sequencing gapless assemblies of each individual chromosome. To date, the T2T Consortium has assembled and published complete sequences for several chromosomes and have preprints of assemblies of the whole genome (Jain et al. 2018a; Miga et al. 2020; Hoyt et al. 2021; Logsdon et al. 2021; Nurk et al. 2021). The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X. As expected, many of the gaps were occupied by repeat-rich sequences such as pericentromeric regions, ribosomal DNA arrays, and large segmental duplications with high sequence similarity between duplications (Bork and Copley 2001; Eichler 2001). Such efforts reveal a massive amount of genetic information that has been impenetrably cloaked by previous sequencing efforts and hence unable to be included in
many biological assessments. Specifically, they now enable the unveiling of the roles that these highly polymorphic sequences might play in biology, evolution, natural variation, and disease. For example, the heterochromatic regions of Chromosomes 1, 9, and 16 have long been known to be composed of classical satellites (Gosden et al. 1975), and these were shown to be polymorphic in length by cytogenetics (Craig-Holmes and Shaw 1971). In individuals with the rare disorder, immunodeficiency, centromeric instability, and facial (ICF) syndrome, in addition to numerous clinical presentations, their chromosomes form complex multiralid associations at the classical satellites 2 and 3 at juxtacentromeric regions of Chromosomes 1, 9, and 16 (Xu et al. 1999). The satellite repeats at the heterochromatic region of Chr 9 are involved in pericentricimeric inversions of Chr 9 (Gosden et al. 1981) and are thought to be linked to a variety of diseases (Mohsen-Pour et al. 2021). Length variations of satellite tracts on Chromosomes 1, 9, and 16 were thought to be associated with both the multiralids in ICF and pericentricimeric inversions of Chr 9 (Gosden et al. 1981; Luciani et al. 2005). A deeper appreciation of satellite repeat tract length variations, and possibly sequence purity, gained by long-read sequencing could reveal associations of disease variation for these and other repeat-rich regions. Another huge advance is the discovery of the huge numbers of previously uncataloged repeats, definitively revealing that the repetitive content in the human genome is 53.9% in CHM13 (Hoyt et al. 2021).

To specifically address the current issue of the absence of a repeat motif in the reference genome, ExpansionHunter Denovo (EHdn) (Rafehi et al. 2019; Dolzhenko et al. 2020) was developed to roughly infer the genomic location (within ~1 kb) and repeat size of “de novo” TR unit expansions (sequence motifs not present in the reference genome) within de novo assembles of SRS data in a catalog-free manner. Independent of the work by Cortese et al., Rafehi et al. have used EHdn to identify expanded TRs in the WGS data of CANVAS-affected individuals (Cortese et al. 2019; Rafehi et al. 2019). They successfully found expanded AAGGG repeats in both alleles of an intron of the gene RFC1, where the reference sequence harbors (AAAAG)11. The first discovery of the RFC1 repeat expansion required considerable efforts that were time-consuming, as evidenced by Cortese et al. (2019), where the rapid independent discovery clearly demonstrates the strength and usefulness of EHdn. The group also showed that no other catalog-based algorithm was able to identify the complex repeat motif as it was not found within the reference genome.

More recently, EHdn, coupled with a novel outlier detection approach, led to the discovery of 2588 loci with TR expansions associated with ASD (Trost et al. 2020). This is the first time a heterogeneous complex disorder was linked to a variety of TR expansions. The reported loci are located in genes that were previously linked to ASD (such as FMR1), and many other genes that are responsible for nervous system development—a novel functional pathway for ASD that would otherwise have not been recognized by using other approaches. As much as 42.3% of the identified TRs in this study have not been previously reported. Even for the ones that were previously reported, 6% of them had at least one repeat sequence that was not present in the reference genome. These findings were bolstered by another recent publication which also identified TRs associated within a separate cohort of ASD individuals, using a novel bioinformatic tool called MontSTR (Mitra et al. 2021).

Given the substantial genetic overlap between neurodevelopmental disorders such as ASD and schizophrenia (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013; Grove et al. 2019), it is likely that TR expansions may also be involved in other related disorders. Indeed, Mitra et al. (2021) also identified TRs which clustered near GWAS signals for schizophrenia and educational attainment within their ASD cohort, and a recent study also identified repeat expansions known to be associated with monogenic neurological diseases within a separate cohort of schizophrenia patients (Mojadid et al. 2021a). These recent studies highlight the necessity of developing tools which enable reference-free assembly and interrogation of the genome. Moreover, they highlight the need to be aware of the degree of genetic variation possible and hence be broad-minded in future developments.

However, while these algorithms are necessary for the accurate detection of expanded repeats within SRS data, the threshold length for many of the disease-causing TRs is close to or beyond the typical short sequence read length of 100–150 bases. Algorithms like EH can infer repeat lengths from SRS data, but their accuracy is still not sufficient to make reliable diagnoses (Bahlo et al. 2018). As such, this key limitation arising from the short read length need to be complemented by LRS technologies. The advantages and disadvantages of different sequencing technologies are summarized in Table 2.

**Long-read sequencing is expected to unveil longer disease-associated TRs**

The development of third generation sequencing technologies, namely Pacific Biosciences’ single molecule real-time (SMRT) sequencing (Rhoads and Au 2015) and Oxford Nanopore Technologies’ nanopore sequencing (Jain et al. 2018b), made it possible to obtain long-read data. Besides their advantages in de novo assembly, structural variant analysis, and haplotype phasing, their abilities to analyze single molecules without GC bias allowed the discoveries of disease-causing stretches of GC-rich repetitive sequences as described above (Ishiiura et al. 2019). While their current cost, data generation speed, read depth, and base-calling accuracy are inferior to SRS (Midha et al. 2019), methods and analytical tools for improvements have been under development. One example is consensus circular sequencing (CSS) applied to PacBio’s SMRT sequencing (Li et al. 2014). This method obtains the consensus sequence from multiple passes of a circular single molecule made by ligating both ends of the same double-strand DNA to form the circular template. While each pass produces an error-prone read, the accuracy of consensus sequences obtained by CSS has been shown to be comparable to SRS parallel sequencing in the setting of WGS (Wenger et al. 2019). Another method that is being explored to increase coverage and depth at regions of interest via Nanopore PromethION sequencing is through selective enrichment of the region—especially useful for analysis of expanded repeats which might present with repeat length mosaicism within patients. For example, Giesselmann et al. (2019) used a CRISPR-Cas12a/Cas9 approach to enrich for C9orf72-associated GGGGCC repeats and FMR1-associated CGG repeats, increasing coverage from 10 reads to nearly 100 (Cas12a) and 1000 (Cas9) reads specifically at the repeat region. The same group also combined this approach with a novel algorithm, STRiQue, to determine repeat length and methylation status of these repeats.

Harnessing these LRS technologies with new algorithms is rapidly becoming a more common strategy, and the data generated in this way provides a powerful tool for discovery of longer and more complex repeats. The discovery of NIID by Sone et al. (2019) was facilitated by the development of tandem-genotypes (Mitsuhashi et al. 2019), an algorithm to detect expanded TRs in long-read WGS data. Prior to this, an algorithm called
Specifications for each device were obtained from manufacturers. Only main sequencing technologies currently available for detection of expanded TRs are listed, with their advantages and disadvantages.

Review paper by Midha et al. (2019).

10 Genome Research

Table 2. Sequencing technologies to detect TR expansions

| Sequencing method /sequencing system (manufacturer) | Read length | Data yield | Necessity of preamplification of templates | Accuracy | Cost | TR expansion finding algorithms |
|----------------------------------------------------|-------------|------------|------------------------------------------|----------|------|----------------------------------|
| **Sanger sequencing**                              |             |            |                                          |          |      |                                  |
| 3730xl (Thermo Fisher Scientific)                  | Up to 900 bases | ~80 kb/run | Yes                                      | High     | High |                                  |
| **Second generation sequencing**                   |             |            |                                          |          |      |                                  |
| NovaSeq 6000 (Illumina)                            | 2 × 150 bases | 300 Gb/run | Yes/No                                   | High     | Low  | TRhist                           |
| MiSeq (Illumina)                                   | 2 × 300 bases | 15 Gb/run  |                                          |          |      | STretch, GangSTR, exSTRa, TREDPARSE ExpansionHunter ExpansionHunterDenovo |
| **Third generation sequencing:** also allows analysis of epigenetic status and phasing |             |            |                                          |          |      | RepeatHMM tandem-genotypes NanoSatellite STRique RepeatHMM tandem-genotypes |
| Nanopore sequencing PromethION 48 (ONT)            | ~30 kb      | Up to 220 Gb/flow cell | No | Low | Medium | tandem-genotypes NanoSatellite STRique RepeatHMM tandem-genotypes |
| SMRT sequencing Sequel II system (PacBio)          | ~30 kb      | Up to 160 Gb/flow cell | No | Low (can be improved with HiFi reads) | Medium | tandem-genotypes |

Only main sequencing technologies currently available for detection of expanded TRs are listed, with their advantages and disadvantages. Specifications for each device were obtained from manufacturers’ web sites (Thermo Fisher Scientific; https://www.thermofisher.com/us/en/home/lifescience/sequencing/sanger-sequencing.html; Illumina: https://www.illumina.com/systems/sequencing-platforms.html; Oxford Nanopore Technologies [ONT]: https://nanoporetech.com/products/comparison; Pacific Biosciences [PacBio]: https://www.pacb.com/products-and-services/), as well as a review paper by Midha et al. (2019).

RepeatHMM was published and shown to be able to accurately measure pathogenic CAG expansions in the ATXN3 gene causing Machado-Joseph disease/SCA3 and long expansions of ATTCT repeats resulting in SCA10 (Liu et al. 2017).

LRS also enables analyses of TRs with longer repeat units that are known to be associated with various complex disorders, such as Variable Number of TRs (VNTRs). VNTRs is a broad ill-defined category of TRs ranging from 6 bp to 10 kb, such as the 99-mer repeat expansion recently discovered to cause skeletal muscle disease (Ruggieri et al. 2020), or the MUC6 VNTR which has a repeat unit size of ~507 bp and is suggested to be associated with an increased risk of Alzheimer disease (AD) (Katsumata et al. 2020; Nelson et al. 2020). Several tools have been developed to detect or genotype VNTR with short reads (Bakhitiari et al. 2018; Lu et al. 2021). However, short-read sequencing of these regions is typically difficult, often resulting in low mapping quality scores and a number of calls that fail to pass quality control filters, resulting in them being “dark and camouflaged” regions of the genome that were largely excluded from prior analysis (Nelson et al. 2020). Recent LRS studies have been instrumental in shining a spotlight on these regions. For example, a 300- to 10,000-bp VNTR in the ABCA7 gene, whose expansions are associated with increased AD risk, was recently identified from LRS data obtained by nanopore sequencing (De Roeck et al. 2019). Another example is a study made with nanopore sequencing to assess the D4Z4 macrosatellite repeat (3.3 kb/unit) number at the facioscapulohumeral muscular dystrophy 1 (FSHD1) locus (Mitsuhashi et al. 2017). Nonaffected D4Z4 alleles are polymorphic with 11–100 repeat units, whereas FSHD1-affected individuals have 10 or fewer units (Mostacciuolo et al. 2009; Pearson 2010). Mitsuhashi et al. applied LAST, a computational analysis tool developed to detect segmental duplications (Kielbasa et al. 2011) for this purpose. Although this method still needs validation, it clearly illustrated the advantages of LRS in the analyses of longer TRs. To date, little is known about diseases caused by long TRs, and LRS is expected to open the door to this field. Moving forward, the development of LRS will likely benefit from leveraging pre-existing short-read sequencing data sets and/or integration of synthetic long reads (such as single-molecule optical maps), such that three- or two-way hybrid assembly between the different assemblies can be used to assess assembly conflicts/errs, highlight misassemblies, maximize base calling accuracy, and limit false positives (Amarasinghe et al. 2020). Even in nonrepetitive sequences, this has also shown to be a powerful approach—for example, optical maps were previously shown to greatly facilitate the resolution of three erroneously linked chromosome-scale contigs derived from SMRT-based LRS in relatives of Arabidopsis thaliana (Jiao et al. 2017). However, the strength of this approach is especially potent for repeat sequences—for example, the use of three-way integration of SRS, LRS, and optical mapping data facilitated the characterization of 36 previously unidentified large repetitive regions in the Eurasian crow, most of which were complex arrays of 14-kb satellite repeats (Weissensteiner et al. 2017). These studies clearly show the potential for LRS both as an independent approach and in tandem with pre-existing approaches.
TRs elicit toxicity (Fig. 2): mechanisms of expanded repeat toxicity. To date, at least 11 different therapeutic options. As such, here we will briefly review the understanding, coupled with technological advances, greatly facilitates various mechanisms of expanded repeat toxicity. To date, at least 11 different pathogenic mechanisms have been proposed on how pathogenic TRs elicit toxicity (Fig. 2):

1. Loss-of-function (LOF) due to transcriptional silencing: for example, CGG expansions within the 5′ UTR of FMR1 in FXS (Pieretti et al. 1991; Devys et al. 1992; Knight et al. 1993) or the gene promoter in progressive myoclonus epilepsy type 1 (Lafrenière et al. 1997; Virtaneva et al. 1997) causes hypermethylation of the repeat and adjacent CpG islands and pathogenically silences gene transcription. While this mechanism has been known for some time, it has been understudied as a potential cause of other diseases where general loss-of-function is suspected. However, two of the latest repeat disease discoveries were shown to elicit pathogenicity through this mechanism: in Desbuquois dysplasia 2 (DBQD2) (LaCroix et al. 2019), the GGC repeat expansion within exon-1 is hypermethylated and this causes suppression of XYLTI transcription, and in glutaminase deficiency (van Kuilenburg et al. 2019), the GCA repeat expansion results in insufficient glutaminase (GLS) mRNA transcription. These two findings remind us of the relevance of this LOF mechanism and suggest that similar findings may follow in the future.

2. LOF due to expansions within protein-coding genes: in CCHS, expansions of polyalanine-coding GCN repeats result in impairment of the paired function of the protein (Amiel et al. 2003).

3. LOF due to expansions within introns: in Friedreich ataxia (Al-Mahdawi et al. 2008), this causes pathogenic suppression of transcription through a variety of debated mechanisms which may overlap with epigenetic changes such as hypermethylation.

4. Gain-of-function (GOF) due to toxic RNA production which impinges on normal cellular function: in DM1, DM2, and fragile X-associated tremor/ataxia syndrome (FXTAS), repeat expansions in the 3′ UTR, 5′ UTR, and introns aberrantly generate expanded repeat RNA products, which are bound by various RNA-binding proteins, often causing toxic RNA foci. The RNA-binding proteins can also be sequestered away from their proper functions, leading to their LOF (Ranum and Cooper 2006).

5. Gain-of-function due to toxic proteins generated by coding expanded repeats: in the cases of polyglutamine (polyQ)-coding CAG repeat expansions such as HD and SCA3, the polyQ stretch results in protein misfolding and aggregation (Sisodia 1998). Understanding of this GOF toxic mechanism in particular has facilitated the development of therapeutics. Recently, gene silencing methods with antisense oligonucleotides (ASOs) and small interfering RNAs (siRNAs) have made considerable progress toward clinical application. One such ASO in advances stages of development is RG6042 (formally called IONIS-HTTRx; an ASO against the HTT gene developed by Ionis) which has been shown to lower mutant HTT protein levels in the cerebrospinal fluid of affected individuals and has satisfactory short-term safety profiles (Tabrizi et al. 2019). It is now entering a Phase 3 clinical trial. While gain-of-function aspects of the diseases may make approaches by ASOs/siRNAs appear attractive, caution should be paid to the contribution of other pathogenic mechanisms in the same individuals.

6. GOF due to aberrant splicing of coding repeats: in HD the expansion within the coding exon-1 disrupts correct splicing of intron-1, leading to its retention. This leads to the production of a toxic fragmented RNA product that contains the expanded repeat, which is then translated to a toxic truncated protein fragment (Sathasivam et al. 2013; Neudeker et al. 2017, 2018; Franich et al. 2019).

7. GOF due to expansion promoting aberrant retention of introns near or encompassing repeat tracts: in DM2, Fuchs endothelial corneal dystrophy 3 (FECD3) (Fautsch et al. 2021), and C9orf72-associated ALS/FTD, the repeat prevents correct splicing of introns to generate a fragmented toxic RNA product. Similar to HD (mechanism 6), the fragmented RNA products may be translated into a toxic truncated protein (Sznajder et al. 2018).

8. GOF due to repeat-associated non-ATG/AUG-mediated translation (RAN-translation): in SCA31, FXTAS, C9orf72-associated ALS/FTD, and SCA8, expanded RNA aberrantly recruits translation machinery and produces toxic peptides without needing an ATG/AUG start site. The complexity of this pathogenic mechanism is exemplified by the multiple frames that are coded within the same repetitive sequence and the fact that both strands may undergo RAN translation—thereby generating a variety of different toxic peptides (Sato et al. 2009; Zu et al. 2011; Ishiguro et al. 2017; Glineberg et al. 2018; Krans et al. 2019).

9. GOF due to up-regulation of the nonmutant protein caused by a cis-mechanism of the expanded repeat: the increased expression of the PPP2R2B gene, associated with SCA12, appears to be caused by a cis-effect of the expanded CAG tract on the PPP2R2B gene products (Lin et al. 2010). This is a regulatory mechanism that likely acts on many of the repeat-containing genes in the genome.

10. GOF due to toxic proteins produced from transcription and translation across one strand of the expanded repeat and production of toxic RNA by transcription across the opposite strand: for example, in SCA8, the CAG repeat expansion within the ATXN8 is transcribed and translated to a toxic polyglutamine protein, while transcription, but not translation, of the ATXN8 opposite strand, which has the complementary CUG expanded repeat in its 3′ UTR, produces a toxic CUG-RNA (Moseley et al. 2006).

11. GOF due to inappropriate expression of a gene encoded by the unstable repeat unit: for example, pathogenicity in FSHD1 results from the improper developmental expression of the double homeobox 4 (DUX4) gene, encompassed in the contracted 4q35 array of D4Z4 repeats (3.3 kb/unit). Each unit contains a DUX4 gene that is epigenetically activated upon contraction of the repeat array, with pathogenic levels of contraction resulting in FSHD1. This mechanism is likely to be more appreciated with the discovery of more clinical presentations associated with the variation of unstable arrays of large genesized repeat motifs. This highlights the importance of an awareness that very large repeat motifs can be unstable in a disease-relevant manner. Moreover, repeat contractions, and not just expansions, can be relevant.
From this list of potential mechanisms of pathogenicity, it is clear that repeat disease pathobiology is highly complex, with shared or similar repeat sequences and often overlapping clinical presentations. It is also noteworthy that a single TR-associated disease likely has contributions from multiple pathogenic processes. For example, loss-of-function paths can exacerbate gain-of-function paths in some diseases (Schneider et al. 2020; Pal et al. 2021). Also, the surprising findings of RAN-translation and intron-retention indicates that we need to have an open mind to the diverse ways in which these mutations can express disease. Clearly, understanding the crossplay between different pathogenic repeat sequences and their functional pathogenic outcomes bolsters our ability to determine if pathologies of unknown cause result from similar repeat sequences.

Repeat expansions and fragile sites

One noteworthy association with LOF due to transcriptional silencing (pathogenic mechanism 1) is that the epigenetic changes associated with transcriptional silencing at expanded repeats often coincide with mapped fragile sites. All molecularly mapped folate-sensitive fragile sites are caused by expanded CGG repeats and associated with aberrant CpG methylation and silencing of the associated gene (FRAXA, FRAXE, FRAXF, FRA2A, FRAXA, FRA10A, FRA11A, FRA11B, FRA12A, and FRA16A). “Rare” fragile sites (~40) are present in ≤5% of the population (with FRA16B being the most frequent rare fragile site [Felbor et al. 2003]) but can present in as few as a single individual. To this degree, recent advances toward identifying additional folate-sensitive fragile sites have included isolating epigenetic modifications of repeats (Garg et al. 2020), identification of all expandable CGG-tracts (Annear et al. 2021), and recent colocalization of these fragile sites to GC-rich repeat expansions have been found in ASD (Trost et al. 2020). However, such advances can only suggest the possible source of a fragile site; that a particular repeat is the cause of fragility must be molecularly mapped cytogenetically by FISH and genetics (Saveleva and Brueckner 2014). Although the pathogenic link between the CGG expansions in each of these 10 folate-sensitive fragile sites and their associated partially penetrant symptoms is not yet clear, what is clear is that each shows aberrant epigenetic modifications coincident with CGG expansion regardless of disease presentation. For example, like FRAXA (FMR1), CGG expansions and aberrant methylation in AFF2, ZNF713, AFF3, and DIP2B at the fragile sites FRAXE, FRA7A, FRA2A, and FRA12A, respectively, are associated with intellectual disability, albeit in a small number of families (Winneppeninckx et al. 2007; Metsu et al. 2014a,b; Correia et al. 2015). This raises the possibility that these, and other fragile sites associations with GC-rich repeat expansions, could be more definitively linked with disease. Strengthening this hypothesis is the recent finding that the human genome contains nearly 6110 CGG repeats longer than four repeat units (found on all but the Y Chromosome)—410 of them being associated with known and candidate neurodevelopment disease genes and multiple being coincident with known fragile sites (Annear et al. 2021). Of the molecularly mapped “common” fragile sites (27 of ~230), none are associated with a particular DNA sequence motif, repeat or otherwise (Irony-Tur Sinai and Karem 2019). It is likely that all folate-sensitive fragile sites, the largest group of rare fragile sites (~30/40), are due to CGG expansions (Handt et al. 2000; Felbor et al. 2003). Other rare fragile sites, like the distamycin A-inducible rare fragile site FRA16B, is caused by an expanded AT-rich 33-bp repeat motif (Yu et al. 1997), while the bromodeoxyuridine-inducible rare fragile site FRA10B is caused by expanded AT-rich 42-bp repeat motif (Hewett et al. 1998). All fragile sites are associated with chromosomal instability (deletions, rearrangements). Both FRA10B and FRA16B have been observed homozgyously in seemingly normal individuals, suggesting that at least for those individuals, at the observed ages, these expansions appear benign (Sutherland 1981; Hocking et al. 1999).

Clinical associations of specific and related repeat sequence motifs

The role of clinical geneticists cannot be underestimated. For example, as early as 1918, clinicians recognized strange transmission patterns of diseases, such as genetic anticipation in DM1 and HD.

Figure 2. Proposed mechanisms through which disease-associated repeats may exert toxicity. Multiple mechanisms may be active at a single locus. (RAN) Repeat-associated non-ATG, (UTR) untranslated region.
(Bell 1941; for review, see Höweler et al. 1989). Similar puzzling segregation was observed for FXS (Martin and Bell, 1943). Even in the face of persuasive discrediting, based upon incorrect claims of ascertainment bias (Penrose 1947; for review, see Höweler et al. 1989), the observations of the clinical geneticist persevered and were shown to be based in the genetic instability of the repeats (Fu et al. 1991; Ashizawa et al. 1992a,b; Harper et al. 1992; Snell et al. 1993; Trottier et al. 1994). Moreover, clinicians played critical roles of both characterizing, collecting patients, and participating in genetic mapping and diagnostic aspects. These contributions are necessary ingredients to identifying disease-causing mutations. Below, we highlight recent advances of the clinical aspect in discovering repeat expansion mutations.

Some pathogenic expansions cause similar disease phenotypes according to their repeat sequence motifs, independently of the functions of genes harboring the mutations (Table 1). An early example is the discovery of CGG repeat expansion in FMRI and the CCG/CGG repeat expansions at AFF2, ZNF713, AFF3, and DIP2B, which were all subsequently shown to be related to X-linked intellectual disability albeit in a small number of families (Knight et al. 1993; Winnecheninckx et al. 2007; Matsu et al. 2014a,b; Correia et al. 2015). Another example is the CAG/CTG repeat expansions causative for HD, SCA1, 2, 3, 6, 7, 8, 12, and 17 (Pearson et al. 2005), the discoveries of which were greatly assisted by the similarities in clinical presentation of particular motor phenotypes and pathologic presentation of polyQ aggregates.

A deep clinical understanding can also facilitate understanding of underlying pathogenic mechanisms and outcomes. One such example is the similarities between DM1 and DM2 that led to the identification of the DM2 mutation. The two diseases both manifest systemic symptoms, including myotonia, muscle weakness, frontal balding, cataracts, and cardiac arrhythmias (Ricker et al. 1994). Before the identification of the DM2 mutation, researchers detected nuclear RNA foci in muscle sections of DM2-affected individuals using CUG-repeat probes against DM1 RNA foci and established that the same protein (MBNL) colocalizes with these foci (Mankodi et al. 2002). These findings led to speculation that the mutation responsible for DM2 is similar to the non-coding CTG repeat expansion that causes DM1. The DM2 mutation was subsequently revealed to be a CCTG repeat expansion in an intron of the CNBP gene (Liquori et al. 2001). The shared clinical characteristics of DM1 and DM2 are considered to be caused by toxic gain-of-function of the mutations acting in trans (Ranum and Cooper 2006). Both CUG and CCUG repeat transcripts sequester their common binding proteins, such as MBNL (Mankodi et al. 2001), and as a result, RNA metabolism is disrupted in both situations. Consistent with this model, missplicing of genes such as CLC1 (Mankodi et al. 2002) and BIN1 (Fugler et al. 2011) has been observed in both diseases.

Further, an appreciation of the clinical genetics of a disease can serve as a “red-flag” for repeat expansions—for example, the explanation of the unusual inheritance patterns of FRAXA/FXS (initially known as Martin-Bell syndrome) (Martin and Bell 1943; Sherman et al. 1984, 1985), or the explanation of genetic anticipation in DM1 or HD by repeat expansions over generations. In fact, anticipation (then called anedating) had been connected with DM1 and HD families as early as 1941 by Bell, although these findings were largely unpappreciated for some time (Bell 1941). The concept was not appreciated until FXS, where repeat length directly affected the likelihood of expansion of premutations to full mutation in FXS families. This provided a mechanistic basis for the long-debated phenomenon of genetic anticipation, which in FXS is evident as incomplete penetrance accompanied by increasing likelihood of disease with subsequent generations in pedigrees (then known as the Sherman paradox). The early FRAXA studies paved the way for subsequent discoveries of repeat expansion mutations and for significant mechanistic insights.

Discoveries made in the past couple of years have strengthened the hypothesis of “repeat motif-phenotype correlation” (Ishiura et al. 2018; Ishiura and Tsuji 2020), in which toxic GOF mechanisms elicited by some expansions are the main drivers of pathogenesis, rather than altered function of the genes which contain the repeat expansion. This concept is especially useful in identifying repeat expansions which manifest similar clinical presentations. For example, in 2018, the mutation responsible for BAFME1 was identified as an insertion of an expanded TTITCA repeat into a TTITCA repeat in an intron of SAMD12 (Ishiura et al. 2018). In addition, two families manifesting indistinguishable disease phenotypes were found to have expansions of TTITCA and TTITTA repeats in introns of the TNRC6A (BAFME6) and RAPGEP2 (BAFME7) genes, respectively. As predicted by Ishiura et al. (2018), these discoveries paved the way for the identifications of other BAFME-causing repeat mutations in 2019, which have exactly the same repeat sequence in different genes: YEATS2 for BAFME4 (Ye et al. 2019), STARTD7 for BAFME 2 (Corbett et al. 2019), and MARCHF6 for BAFME3 (Florian et al. 2019). Similar to DM1 and DM2, RNA foci of UUUCA repeats were found in autopsied brain samples of BAFME1 individuals. These observations also suggest that the trans-acting RNA gain-of-function mechanism (proposed pathogenic mechanism 4) is relevant in familial adult-onset myoclonic epilepsy. The expanded repeat causing SCA37 has the same TTITCA motif within a long stretch of TTITTA repeat (reported to be “TTITCA” repeat) (Seixas et al. 2017). BAFME patients have been reported to manifest cerebellar dysfunction (Striano et al. 2009) and atrophy (Bujink et al. 2016), and in a homozygous BAFME1 patient, histopathological findings of Purkinje cell degeneration, similar to those in SCA31 (Owada et al. 2005), have been observed. These findings connect BAFMEs with SCA37, whose cardinal clinical presentation is cerebellar ataxia and atrophy (Seixas et al. 2017). The above pieces of evidence suggest that TTITCA can now be recognized as a new common motif resulting in similar disease phenotype—another addition to the ever-growing list of diseases caused by common TR motifs.

Even more recent examples were discovered in 2019, when three groups independently found that GGC repeat expansions in an intron of NOTCH2NL/CNBP/F19 genes causes NIID (Ishiura et al. 2019; Sone et al. 2019; Tian et al. 2019), a neurodegenerative disorder difficult to correctly diagnose based on clinical presentations alone (Sone et al. 2016; Okubo et al. 2019). Further reports showed that an intronic CGG repeat expansion in LOC642361/NOTCH2NL/CNBP/F19 leads to oculopharyngeal myopathy with leukoencephalopathy (OPML), and the same repeat expansion in LRP12 is responsible for oculopharyngodistal myopathy (OPDM1) (Ishiura et al. 2019). These findings were truly eye-opening for neurologists and neuropathologists, as they provided critical insight into the fundamental pathogenesis of these diseases and connected degenerative disorders of the central and peripheral nervous systems and muscles. NIID is a disease that mainly affects the brain and the peripheral and autonomic nervous systems, manifesting a variety of symptoms including dementia, tremor, cerebellar ataxia, and autonomic failure (Sone et al. 2016). On the other hand, OPDM is a type of muscular dystrophy that causes facial, bulbar, and distal weakness (Satoyoshi and Kinoshita 1977). Now, with the identification of GGC repeats as the cause of both NIID and OPDM.
diseases, and the knowledge that there are overlapping symptoms and radiological findings among them (Ishiiura et al. 2019), we have a view that there is a clinical spectrum with NiID on one end, OPDM on the other end, and OPML in the middle. There are also strong clinical similarities between NiID and FXTAS, the classic example of neurodegenerative disease caused by CGG repeat expansion (Hagerman et al. 2001). The disease spectrum suggested by these new findings is a completely new concept for physicians, pathologists, and geneticists, one that cannot have been unimaginable without the discovery of its genetic cause.

If we are to accept the hypothesis that particular repeat unit sequence expansions are pathogenic and cause diseases independently of their genomic locations, it is essential to establish the possibility that these de novo TR unit expansions may also be associated with similar diseases that have yet to be identified. In the past, prototypical approaches were taken to search for trinucleotide repeat expansions in a locus-independent manner simply by looking for expansions of known disease-causing motifs (RED assay and its variations and catalog-based algorithms more recently, as described above). Now that we are equipped with high-throughput LRS technologies and computational analysis tools, we can search for TR expansions from NGS data. New discoveries are expected to follow, which will further broaden our understanding of the biological roles of TRs and their contribution to disease.

Assessing the association of repeat expansion diseases must be conducted without bias or assumptions

Several recently reported repeat disease associations occurred in unstable repeats previously thought to be clinically unimportant. For example, one of the most prevalent (affecting ∼4% of people aged 40 or over in the United States) TR expansion diseases, FECD3, was initially missed because of a clinical bias toward the expectation that the disease would present with neurodegeneration. Due to the absence of neurodegenerative phenotypes, the initial discoveries of the CTG18.1 repeat expansion (the genetic cause of FECD3 in 1993 (Schalling et al. 1993) and 1997 (Bresche1 et al. 1997) led to an incorrect assumption that the expansion was benign. It was not until 20 years later that the CTG repeat expansions within the TCF4 gene (CTG18.1) were linked with FECD3 (Wieben et al. 2012; Fautsch et al. 2021). In another case, an expanded CCG repeat reported at FRA16A, which in the heterozygous state has no obvious clinical impact (Nancarrow et al. 1994), was subsequently found to result in autosomal recessive skeletal disorder DBQS2 when CCG expansions were present in a homozygous state (LaCroix et al. 2019). Because the initial identifications of pathogenic TR expansions were mostly of dominant or X-linked diseases (with the exception of Friedreich ataxia), the possibility of recessive diseases tended to be overlooked, but the recent discoveries demonstrate that this is a groundless assumption (Cortese et al. 2019; LaCroix et al. 2019; van Kuilenburg et al. 2019; Pagnamenta et al. 2021).

Similarly, effects of TR length on varied clinical presentations must also be considered. One example is the human androgen receptor (AR), where the function of the gene product is well characterized, unlike the genes for most TR expansions. The AR gene contains an exonic polymorphic CAG repeat in which 95% of individuals inherit between 16 to 29 CAG repeats, encoding a polyglutamine tract. Nonrepeat LOF mutations of AR, and hence an absence of androgen receptor activity, lead to masculine feminization (androgen insensitivity syndrome; AIS), a non-neurological presentation. In contrast, expansions of the CAG tract (average size of ∼47 units in patients) lead to spinal bulbar muscular atrophy (SBMA). While not evident in 1991, the contrasting phenotypes of the LOF mutations with the CAG expansions provided support for the pursuit of a GOF toxicity path. Albeit that in most neurologically affected SBMA individuals, the AR is still functional, they do show signs of AIS (La Spada et al. 1991). Biochemically, the length of this encoded polyglutamine tract inversely affects AR transcriptional activity (Mhatre et al. 1993; Chamberlain et al. 1994; Kazemi-Esfarjani et al. 1995). Very large expansions (68 or 72 repeats) lead to severe clinical presentations of both SBMA and AIS consistent with reduced transactivation activity with long CAG tracts (Mhatre et al. 1993; Kazemi-Esfarjani et al. 1995; Grunseich et al. 2014; Madeira et al. 2018). Within the normal range, longer AR CAG tracts (>∼21 repeats) have been associated with male infertility, breast cancer, osteoporosis, and male-to-female transsexualism (Summers and Crespi 2008; Hare et al. 2009), while shorter tracts in the normal range have been associated with prostate cancer, head and neck cancer, colorectal cancer, cardiac disease, and cognition and behavior disorders (Summers and Crespi 2008). In cancers, somatic variation of the CAG repeat was biased toward contracted repeats (Ferro et al. 2002; Di Fabio et al. 2009), presumably due to the increased androgen sensitivity of the AR protein, with shorter tracts giving those cells a growth advantage (Mhatre et al. 1993; Kazemi-Esfarjani et al. 1995). Other studies suggest that gender incongruence/dysphoria in the transgender woman (male-to-female) population have significantly longer polymorphic CAG repeat sequences in the AR gene, which may affect antenatal androgen activity and possibly contribute to gender incongruence (D’Andrea et al. 2020). Thus, one gene, depending upon mutation type and/or TR tract length can display extremely variable clinical manifestations and demonstrates how understanding of the natural function of the gene can serve as a guide to elucidating its mechanism of disease action. For FMR1, nonrepeat-related null LOF mutations were identified following the discovery of the fragile X-associated expansion. In the case of FMR1, intragenic deletions, nonsense changes (Wöhle et al. 1992; Hirst et al. 1995; Lugtenbeel et al. 1995), and a missense mutation, 1304N (De Boulle et al. 1993) led to severe fragile X syndrome, confirming the LOF associated with the CGG expansion. Similarly, increases in FMR1 copy numbers can also lead to similar phenotypes (Rio et al. 2010). New sequencing and informatic analyses should facilitate such future pathogenic connections.

Another crucial point that must be highlighted is the delayed recognition of distinct clinical presentations and diseases associated with different expansion lengths within the FMR1 gene. Full expansion of the CGG repeat (>200 CGG) in FMR1, coupled with aberrant DNA CpG methylation, is widely known to cause the intellectual disability syndrome FXS, a phenotype recognized since 1943 (Martin and Bell 1943). It wasn’t until 58 years after the description of the first FXS pedigree that individuals with unmethylated FMR1 expansions of 50–200 CGG units were distinctly characterized within a separate disease, which presented with late onset tremor and ataxia, now known as FXTAS (Hagerman et al. 2001). Prior to this, these individuals were referred to as “premutation” or “normal transmitting males/females,” a term highlighting their limited FXS phenotypes and their ability to pass on the FXS-eliciting expansion to their children, despite their distinct clinical presentations of ataxia later in life (Loesch et al. 1994). Although ataxia had previously been observed in families with Martin-Bell syndrome/FXS, prior to the discovery of the disease-causing CGG expansions (Howard-Peebles 1980), because
the ataxia and intellectual disability did not simultaneously appear in a single individual, Dr. Howard-Peebles stated “Several family members are ataxic...[which] appears to be an unusual variety of spinocerebellar atrophy...[and] There appears to be no relationship between this disorder and X-linked mental retardation with a fragile Xq”. These early observations of FXTAS were also complicated by limited family member numbers and the multiplicity of phenotypes resulting from different expansion lengths, which has only been delineated within the early 2000s (Jacquemont et al. 2004; Rodriguez-Reyenga et al. 2009). This penetrance issue also complicated early characterization of DM1 and other diseases (Echenne and Bassez 2013; De Antonio et al. 2016; Joosten et al. 2020). Despite this, it is highly laudable that Dr. Howard-Peebles noted the ataxia in the FXS family, as it facilitated the eventual correct characterization of the ataxia as a separate disorder (Howard-Peebles 1980). While it is unlikely that this is the first time such an association was clinically observed, it may be the first published description of FXTAS in FXS families. Retrospectively, it is understandable how an association between FXS and the late-onset, slowly progressive motor symptoms of FXTAS had been overlooked for such a long time, but this should serve as a teaching lesson: ascertainment bias by human involvement may lead to missed genetic attributions of varied clinical presentations (symptoms, ages at onset, etc.) to a single TR locus. For example, a confounding factor in the case of FXS is that families (boys with FXS and their mothers) were typically under the care of pediatricians, while the grandfathers at risk for FXTAS were seen by separate specialists (neurologists, geriatricians).

A similar example occurred in cases of nonneuropsychiatric clinical presentations in FXS family members, which were met with delayed recognition and acceptance by the research community for being genetically linked to CGG expansions in FMR1. In the late 1980s to early 1990s, studies of FXS families suggested that non-FXS premutation mothers were at risk of early menopause and increased rates of dizygotic twinning, which in 1995 led to testing an association of ovarian failure in these women (Conway et al. 1995; Murray et al. 1998; Sherman 2000). A link between premutation (CGG) 55–199 lengths in FMR1 and fragile-X associated primary ovarian insufficiency (FXPOI) is now accepted, and recently a study of 1668 women has refined the risk of FXPOI: specifically, females with 85–89 repeats are at the highest risk, while those with 55–65 repeats or 120–199 repeats did not have a significantly increased risk for FXPOI compared to women without any CGG expansions <45 repeats (Allen et al. 2021). The risk of early menopause was very similar (Allen et al. 2021). A link of twinning rates and FMR1 premutations remains an enigma with inconsistent claims (for and against an association), likely due to studies that do not account for repeat size and possibly timing of twinning relative to X-inactivation (Sherman 2000; Allen et al. 2007). Thus, very distinct clinical presentations can arise from repeat expansions in a given gene, and these can critically depend upon repeat expansion size.

With the clarity of hindsight, it is easy to overlook that late-onset genetic disorders, especially those that show incomplete penetrance, are challenging to study through families—especially so in the era prior to molecular biology. A thought experiment is instructive here: without the link provided by individuals ascertained through fragile X syndrome, would it have been possible to define the partially-penetrant FXTAS or FXPOI? It seems unlikely, and this lesson should be carried with researchers into the future, especially with the expansion of molecular biology research, as similar breadths of clinically diverse presentations may also be linked to expansion size at the DM1 (Trost et al. 2020), C9orf72 (Miller et al. 2016; Van Mossevelde et al. 2017; Fredi et al. 2019; Tábuas-Pereira et al. 2019), the NIID loci (Sone et al. 2016), FMR1 (Schneider et al. 2020), or any other of the known or yet undiscovered repeat diseases. To this degree, clinicians and researchers must share observations, be open-minded and embrace the likelihood that repeat diseases are not limited to only continuums of severity of one but of potentially diverse phenotypes.

The length of repeats in one gene can predispose to distinct diseases

Recent studies have started to unravel how the nondiseased size of a TR in one disease-associated gene can be linked to the susceptibility of another distinct disease. For example, the intermediate CAG repeat lengths of ATXN2 gene has been identified as a risk factor for developing ALS (Elden et al. 2010; Conforti et al. 2012), SCA3 (Tezenas du Montcel et al. 2014), FTD (Fournier et al. 2018), and AD (Rosas et al. 2020). This association was further confirmed by a series of studies, and a meta-analysis of these data showed that an intermediate CAG repeat (30–33) allele in ATXN2 is associated with increased risk of developing ALS with the odds ratio of 4.44 (Wang et al. 2014).

Other instances of this phenomena are observed in the intermediate CAG repeat lengths of: ATXN1 as a risk factor for FTD, AD, ALS, SCA3, and SCA6; ATXN3 as a risk factor for SCA6 and SCA7; ATXN7 as a risk factor for SCA2; TBP as a risk factor for SCA7; and HTT as a risk factor for FTD, AD, and SCA3 (Conforti et al. 2012; Tezenas du Montcel et al. 2014; Rosas et al. 2020). The nonpathogenic HTT CAG tract length has also recently been associated with variable changes in risk for ASD, where longer allele lengths are associated with an enhanced ASD risk (Piras et al. 2020). Another observation is that the GT repeat length in the promoter region of the HMOX1 gene modulated the risk of human immune deficiency virus (HIV)-related central nervous system inflammation, such that shorter GT repeats are related with decreased risk of HIV encephalitis (Gill et al. 2018) and HIV-associated neurocognitive impairment (Garza et al. 2020). Increasing the complexity of these associations is the fact that in certain instances longer repeat lengths could also be protective—for example, with longer CAG lengths in HTT (within the normal range) being protective for SCA3 age of onset (Tezenas du Montcel et al. 2014).

The effects of TRs resulting in variable disease susceptibility can partially be explained by their effects on gene expression. A genome-wide search utilizing RNA-seq data of lymphoblastoid cell lines and lobSTR (Gymrek et al. 2012)—another software to analyze TR length—revealed 2060 STRs in association with gene expression (the authors coined these STRs as eSTRs) (Gymrek et al. 2016). This study further identified that 12 eSTRs are significantly associated with clinical phenotypes, including Crohn’s disease, rheumatoid arthritis, and type 1 diabetes mellitus. The findings of eSTRs are supported by a study by Quilez et al., in which they genotyped 4849 promoter-associated STRs in 120 individuals and found more than 100 STRs associated with DNA methylation and neighboring gene expression (Quilez et al. 2016). These TRs were shown to have tendencies toward overlapping with transcription factor binding sites, providing an explanation for possible biological mechanisms of action. The same group used another STR analysis software, HipSTR, to link hundreds of eSTRs with complex disorders such as schizophrenia and inflammatory bowel disease, and complex traits including height and intelligence (Foting et al. 2019). A more recently developed powerful bioinformatic
tool, adVNTR-NN, used a neural network to rapidly genotype 10,264 VNTRs in 652 individuals (Bakhtiari et al. 2021). Greatly improving processing times from previous tools, adVNTR-NN can genotype a single VNTR from 55x whole-genome data in 18 sec with high accuracy. The group found 163 VNTRs associated with regulation of proximal gene expression (designated eVNTRs) in 46 different tissues—with about 50% of these having a likely causal impact on the expression of proximal genes. Within the eVNTRs, several were associated with Alzheimer disease, obesity, and familial cancers, supporting that repeat-associated expression dysregulation is likely a contributing factor to pathogenesis.

Analysis of these secondary repeat instability effects offers key insights into disease and potentially illuminates therapeutic potential. For example, a recent study demonstrated that cancer cells with microsatellite instability arising from DNA mismatch repair deficiency incur previously unknown large-scale expansions of TA repeats (Van Wietmarschen et al. 2020). TA repeats are found genome-wide and, when expanded, they stalled replication forks, activated DNA damage response kinases, and required WRN helicase for processing. In the absence of WRN, however, expanded repeats were susceptible to cleavage by the MUS81 nuclease, leading to massive chromosome shattering and synthetic lethality in cancer cells. Nearly 15% of colorectal cancers, 20%–30% of endometrial cancers, 15% of gastric cancers, and 12% of ovarian cancers are caused by deficiency in DNA mismatch repair, supporting the development of therapeutic agents that target WRN for microsatellite instability-associated cancers.

Further studies are expected to clarify the role of TRs in complex human traits and diseases, which has been proposed to explain “missing heritability” (Hannan 2010) as demonstrated in the recent study on ASD (Trost et al. 2020).

Repeat tract purity and gene variance can be an issue for clinical awareness and research

Another major aspect of repeat disease genetics that is highly relevant to clinical awareness and understanding of pathogenesis is the concept of repeat tract purity—the presence or absence of non-repeat units within a tract of tandemly repeating motifs. Recently, it has become increasingly apparent that the presence of interruptions within expanded TR tracts affects genetic instability as well as age-of-disease-onset severity. Soon after the discovery of repeat expansions as a cause of disease, it was found that interruptions within the repeat tract stabilized these repeats against expansions, whereas loss of interruptions makes the repeat susceptible to expansion (Chung et al. 1993). Interrupted nonexpanded repeat tracts are typically associated with beneficial aspects for FXS, SCA1, SCA2, and HD (Eichler et al. 1994; Chong et al. 1995; Latham et al. 2014)—protecting against germline and somatic repeat instability and, in this manner, “protecting” against disease aspects for the gene in which the repeat resides. The purity of the FMR1 premutation CGG tracts, when diagnostically assessing the AGG interruptions in the premutation CGG expansions in FMR1 by single-molecule PacBio sequencing, allows accurate risk estimates for having a child with FXS (Ardui et al. 2018). Sequencing has many advantages over PCR-based methods and provides improved genetic counseling for women with a premutation—for example, in decisions of family planning.

Because the size of the expansion correlates with disease severity, inhibition of repeat expansions was hypothesized to drastically modulate age-of-disease-onset severity. Indeed, recent data supports this hypothesis by revealing that, in a portion of individuals with HD (Ciosi et al. 2019; Genetic Modifiers of Huntington Disease (GeM-HD) Consortium 2019; Wright et al. 2019), the absence of CAG tract purity may have strong effects on the age-of-onset, disease progression, severity, and phenotypic manifestations. For example, in HD, the polyQ-coding CAG repeat usually ends with CAACAG (which also codes for QQ) or (CAACAG)2 (which codes for QQQQ), but in some HD individuals, the interrupting CAA units are absent. The groups found that those carrying the (CAACAG)2 interruption had significantly delayed disease age-of-onset and lessened severity, and those carrying no interruption had significantly hastened age-of-onset and worsened severity, relative to those carrying a single CAACAG—despite all expressing a mutant HTT protein with the same polyQ length. These facts may suggest that pure repeat tracts are more susceptible to somatic repeat instability and thus result in earlier disease onset and more severe phenotypes. As such, correct identification of interrupted repeats within patient cohorts is essential for planning clinical trials, providing prognostic insight, and in conducting patient research.

It should also be noted that interrupted repeats can also have deleterious attributes. As mentioned in the previous section, larger nonpathogenic length repeats can affect presentation of other diseases—for example, larger ATXN1 or ATXN2 CAG tracts within the wild-type range can be associated with ALS, FTD, AD, and SCA3. These larger tracts are typically interrupted CAG tracts (Corrado et al. 2011; Yu et al. 2011; Conforti et al. 2012). While the manner by which the interruptions contribute to disease predisposition is unknown, a broader appreciation of repeat purity is clearly wanting.

While the presence of repeat tract interruptions on shorter/nonexpanded tracts has long been known (Eichler et al. 1994; Latham et al. 2014), the assessment of purity of longer/expanded tracts and complex motifs is more challenging. For example, while long-presumed to be pure, long disease-associated expansions of the myotonic dystrophy CTG tract have been shown to be interrupted with various non-CTG units, with unusual patterns. These interrupted alleles have been associated with altered predispositions to germline and somatic instabilities and may be associated with vastly altered clinical presentations (Musova et al. 2009; Braida et al. 2010; Santoro et al. 2013, 2017; Botta et al. 2017; Cumming et al. 2018; Tomé et al. 2018; Ballester-Lopez et al. 2020). While LRS is expected to reveal the inaccessible areas of long stretches of TRs, the high error rates of nanopore sequencing and SMRT sequencing technologies appear as obstacles to fine analysis and may introduce “artificial interruptions.” While CSS has been applied to some expanded TRs, such as GGGGCC repeat expansion in C9orf72 (Ebbert et al. 2018) and CGG repeat expansion in NOTCH2NL (Sone et al. 2019), its usefulness to characterize the purity of TRs still needs more validation. Current interruptions still present a major challenge for LRS that will need to be addressed as the field progresses. A full appreciation of the purity of any expanded repeat will likely lead to improvement to clinical, diagnostic, and genetic counselling. The evolution of bioinformatic tools is wanting.

In addition to repeat tract purity, another major modifier of disease is naturally occurring variants of genes which act as trans-modifiers of disease. Of note are the DNA repair gene variants known to modify disease presentation in patients, likely by modifying the level of somatic expansions at the repeat. For example, recent age-of-onset for HD GWASs have identified SNP variants in the DNA repair genes MSH3, FAN1, PMS2, LIG1, and MLH1 (Genetic Modifiers of Huntington’s Disease (GeM-HD)
Overview of nonhuman phenotype-associated repeats. Abbreviations: SCA, spinocerebellar ataxia.

| Phenotype (species)                                                                 | Repeat unit         | Gene, loci | Location of mutation | Inheritance or association | Human equivalent | Ref                          | Publication date (D/M/Y) |
|-----------------------------------------------------------------------------------|---------------------|------------|----------------------|--------------------------|------------------|------------------------------|----------------------------|
| Immunomodulation (Pinctada fucata martensi)                                        | Large tandem arrays | nAChR      | Multiple locations   | –                        | No               | Cao et al. 2021              | –/–/2021                   |
| Life span (Saccharomyces cerevisiae)                                                | Varied              | FLO11 and HPF1 | Intragenic           | –                        | No               | Barré et al. 2020            | 10/04/2020                 |
| Degreening (Malus domestica)                                                        | Serine coding repeats | ERF17 | Exon | Association | No | Han et al. 2018 | 05/02/2018 |
| Impaired growth (Arabidopsis thaliana)                                              | TTC/GAA             | IIL1       | Intron               | –                        | No               | Tabib et al. 2016, 2017; Sudeshkumar et al. 2009 | 31/08/2016 20/02/2009 |
| SCA (Canidae)                                                                      | GAA                 | ITPR1      | Intron               | AR                      | Deletions in humans cause SCA1; Das et al. 2017; Zambonin et al. 2017 | Forman et al. 2015 | 30/10/2014 |
| Epilepsy and behavioral changes (Canidae)                                           | Poly(A)             | SLC6A3     | Intron               | Association             | 40-bp VNTR variations associated with alcoholism, epilepsy, ADHD, susceptibility to Parkinson disease; Hauser et al. 2002; Ivashchenko et al. 2015; Lafuente et al. 2007 | Lit et al. 2013 | 23/12/2013 |
| Lafora disease/ EPM2 (Canidae)                                                     | GCCGCCCCCCC GC     | Epm2b      | Exon                 | AR                      | Nonrepeat LOF mutations cause Lafora disease in humans; Chan et al. 2003 | Lohi et al. 2005 | 07/01/2005 |

Overview of nonhuman phenotype-associated repeats. Abbreviations: SCA, spinocerebellar ataxia.

Consortium 2019), and a separate GWAS identified variants of MSH3 modified somatic instability and disease severity in HD and DM1 patients (Flower et al. 2019). Corroborating these findings and illuminating overlap between the different diseases, a separate GWAS identified FAN1 and PMS2 variants as significant modifiers of age-of-onset for several different CAG expansion SCAs (Bettencourt et al. 2016). These studies demonstrate the importance of these DNA repair proteins in disease pathogenesis. Furthermore, their impact is not limited to CAG/CTG disorders, as a recent GWAS in XDP (caused by a CCCTCT repeat expansion) also identified variants of MSH3 and PMS2 as significant modifiers of age-of-onset (Laabs et al. 2021). While these SNP-based approaches sifting known variants illuminate novel shared pathways that may contribute to pathogenicity, large-scale whole-genome sequencing efforts are likely to reveal novel gene variants that are significant modifiers of disease (e.g., see Deshmukh et al. 2021).

Beyond humans and beyond disease

Nonhuman organisms can display TR length variations with associated disease or biological consequences. Naturally occurring repeat length variations are implicated in disease of nonhuman organisms (summarized in Table 3) and in human nondisease phenotypes, such as height. Prominent examples include various canine diseases associated with repeat expansions, such as (1) the dodecamer GCCGCCCCCCC GC pathogenic repeat associated with a epilepsy (canine Lafora disease) in a variety of species of dogs (Lohi et al. 2005; Webb et al. 2009; Barrientos et al. 2019; Kehl et al. 2019; for review, see von Klopmann et al. 2021), (2) a 38-bp VNTR in the dopamine transporter gene, DAT/SLC6A3, associated with seizures and behavioral issues in Belgian Malinois dogs (Lit et al. 2013), and (3) the GAA repeat expansion associated with spinocerebellar ataxia in Italian Spinone dogs (Forman et al. 2015). Oddly, nonrepeat mutations in the human homologs of some of these genes, like NHLRC1 and ITPR1, cause similar disease in humans, but the human gene does not contain the unstable repeat present in the canine gene as outlined in Table 3 (Chan et al. 2003; Das et al. 2017; Zambonin et al. 2017; for review, see von Klopmann et al. 2021). While it remains puzzling that the highly unstable, pathogenic “dynamic” repeat mutations seem to be mostly confined to humans, in these particular cases, it seems that the canine disease, but not the human, is linked to repeat expansions. However, there seems to be some overlap between dogs and human, albeit controversial, with VNTR variation in DAT/SLC6A3 and behavioral presentations (Hauser et al. 2002; Lafuente et al. 2007; Ivashchenko et al. 2015). Repeat expansions, with biological consequences, have been documented in plants. For example, an expanded TTC/GAA intronic repeat within the ILL1 gene of Arabidopsis thaliana is responsible for growth defects and temperature sensitivity within a strain of the plant species (Sureshkumar et al. 2009; Tabib et al. 2016). An awareness of repeat biology in crops is only beginning—for example, different numbers of serine-encoding TCG repeats of ERF17 may regulate apple peel degreasing during ripening (Han et al. 2018). In nonvertebrates, variations in repeat length within coding sequences have been suggested as a source of speciation in honey bees (Zhao et al. 2018), as a mediator of lifespan in yeast (Barré et al. 2020), and as a regulator of immunity in pearl oysters (Cao et al. 2021). Analysis of Arabidopsis thaliana reveals a large degree of genetic variability...
variability associated with natural polymorphic variants within repeat tracts in the genome; with 95% of the 2046 STR loci tested displaying significant polymorphism (Press et al. 2018). These examples exhibit that many nonhuman disease or complex phenotypes are associated with repeat length variations and could contribute to evolutionary-selectable traits which could be beneficial or detrimental.

As our understanding of repeat sequences grows, so too will our appreciation for natural variability of complex traits in humans due to variations in repeat length. For example, a recent study analyzing the genomes of 3622 Icelanders by LRS identified a median of 22,636 structural variants per person, representing 13,353 insertions and 9474 deletions spanning a total of 10 Mb per haploid genome. While some of these variations are disease-relevant (such as the 69-mer variation within NACA, associated with atrial fibrillation), some variants were associated with nondisease complex traits, such as the 57-bp repeat within ACAN which was associated with height (Beyter et al. 2021). This is coincident with previous reports which also found associations of repeat lengths with height (Fotsing et al. 2019). Limb and skull morphological variations in dogs have already been linked with differences in repeat sizes of a variety of genes, suggesting that natural human variation could also be attributed to repeat length variations (Fondon and Garner 2004). Indeed, a recent large scale repeat length polymorphism analysis of 118 coding VNTRs in more than 400,000 individuals reveals associations of repeat lengths with nearly 800 different human trait phenotypes, including height, male pattern baldness, and hair morphology, and potentially disease-associated phenotypes such as lipoprotein concentration and kidney function (Mukamel et al. 2021). Thus, DNA repeat length variations may affect various phenotypes, not necessarily disease attributes only, thereby precipitating rapid phenotypic variations which may affect rates of natural selection. On an evolutionary scale, TR variations that may have null or deleterious effects could, with environmental change, become advantageous.

Concluding remarks—toward future discoveries

As of December 2021, there were 63 disease-associated or disease-causing unstable TR loci, at least 22 repeat motifs (not counting complex, large, and/or variable motifs), associated with >69 diseases, where some diseases are common to some of the same TR loci. Our quest for TR expansions and their association with disease is still far from complete; currently we only see the tip of the iceberg. Recent technological progress has facilitated the unveiling of TR expansions with large effect size on clinical phenotypes, but our knowledge of those with small effect size is extremely limited. The mechanism of TR expansion has been eagerly sought after, and earlier studies indicate the impact of DNA repair proteins and their naturally occurring variants (Tomé et al. 2009). This view is supported by the recent large-scale screens for disease modifiers (Moss et al. 2017; Flower et al. 2019; Genetic Modifiers of Huntington Disease (GeM-HD) Consortium 2019), which may lead to the development of disease-modifying therapies. The impact of somatic instability of TRs is now recognized for numerous neurodegenerative disorders and cancers, but noninvasive methods to evaluate its degree in various organs and tissues are still lacking. We must pause to consider how many TR expansions may yet prove to be associated with biological functions, diseases, and evolutionary change. Further identification and understanding of TRs, beyond the tip of the iceberg, will reveal a new landscape of biology and medicine.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the Canadian Institutes of Health Research (FRN175329, R.K.C.Y.; and FRN148910, C.E.P.), the Natural Sciences and Engineering Research Council (RGPIN-2016-08335, C.E.P.), the Marigold Foundation (C.E.P.), the Petroff Family Fund (C.E.P.), and Brain Canada (R.K.C.Y.; C.E.P. holds a Tier 1 Canada Research Chair in Disease-Associated Genom Instability. The authors thank the reviewers, who provided significant, fruitful, and constructive feedback for this work. Dedication: This review is dedicated to two pioneers and scientific catalysts: Professor Sir Peter S. Harper (1939–2021), an exceptional clinical geneticist, advocate, and mentor to many in the arenas of myotonic dystrophy, Huntington disease and beyond; and Professor Stephen T. Warren (1953–2021), a pioneering human molecular geneticist, ever committed to advancing the understanding of fragile X and related disorders, a leader, advocate, and mentor to many.

Note added in proof

During the proofing stage of this review, an additional preprint has revealed an association for expanded repeats and schizophrenia risk (Mojarad et al. 2021b).

References

Akarsu A. 1996. Genomic structure of HOXD13 gene: a nine polyanaline duplication causes synpolyactyly in two unrelated families. Hum Mol Genet 5: 945–952. doi:10.1093/hmg/5.7.945
Akçimen F, Ross JP, Bourassa CV, Liao C, Rochefort D, Gama MTD, Dicaire MJ, Barsottini OG, Brais B, Pedroso JL, et al. 2019. Investigation of the RCP1 repeat expansion in a Canadian and a Brazilian ataxia cohort: identification of novel conformations. Frontot 10: 1219. doi:10.3389/fgen.2019.01219
Allen EG, Sullivan AK, Marcus M, Small C, Dominguez C, Epstein MP, Charen K, He W, Taylor KC, Sherman SL. 2007. Examination of reproductive aging milestones among women who carry the FMR1 premutation. Hum Reprod 22: 2142–2152. doi:10.1093/humrep/dem148
Allen EG, Charen K, Hiss HP, Shubeck L, Amin A, He W, Nolin SL, Glicksman A, Tortora N, McKinnon B, et al. 2021. Refining the risk for fragile X-associated primary ovarian insufficiency (FXPOI) by FMR1 CGG repeat size. Genet Med 23: 1648–1655. doi:10.1038/s41436-021-01177-y
Ali-Mahdavi S, Pinto RM, Ismail O, Varshney D, Lympieri S, Sandi C, Trabzuni D, Pook M. 2008. The Friedreich ataxia GSAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. Hum Mol Genet 17: 735–746. doi:10.1093/hmg/ddm346
Amarasinghe SL, Su S, Dong X, Zappa L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21: 30. doi:10.1186/s13059-020-1935-5
Amiel J, Laudier B, Attié-Bitach T, Trang H, de Pontual L, Gener B, Trochet D, Etechevers H, Ray P, Simonneau M, et al. 2003. Polyglutamine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. Nat Genet 33: 459–461. doi:10.1093/natgen/11130
Annear DJ, Vandeweyer G, Elnick E, Sanchis-Juan A, French CE, Raymond L, Kooy RF. 2021. Abundance of polymorphic CGG repeats in the human genome suggest a broad involvement in neurological disease. Sci Rep 11: 2515. doi:10.1038/s41598-021-82050-5
Anzick SL, Kononen J, Walker RL, Azorsa DO, Tanner MM, Guan X-Y, Sauter G, Kallioniemi O-P, Trent JM, Meltzer PS. 1997. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. Science 277: 965–968. doi:10.1126/science.277.5328.965
Ardizi S, Race V, de Ravel T, Van Esch H, Devriendt K, Matthijs G, Vermeesch JR. 2018. Detecting AGG interruptions in females with a FMR1 premutation by long-read single-molecule sequencing: a 1 year clinical experience. Front Genet 9: 150. doi:10.3389/fgene.2018.00150
Ivashchenko DV, Shuvalov SA, Chuprova NA, Kibitov AO. 2015. The association of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Mol Cell 60: 285–291. doi:10.1016/j.molcel.2015.09.022

Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Mgla KH. 2018b. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol 36: 321–323. doi:10.1038/nbt.4109

Jin M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Mgla KH. 2018b. Assembly of a human centromere on the Y chromosome. Nat Biotechnol 36: 321–323. doi:10.1038/nbt.4109

Kaiser DL, Raccichini G, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid-Herrero E, et al. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res 27: 778–786. doi:10.1101/gr.213652.116

Jones C, Siljepcevic P, Marsh S, Baker E, Langdon WY, Richards RJ, Tunnacliffe A. 1994. Physical linkage of the fragile site FRA11B and a Jacobsen Syndrome chromosome deletion breakpoint in 11q12.3. 3. Mol Hum Genet 5: 2123–2131. doi:10.1038/nature00444

Joosten IBT, Hellebrekers DMEJ, de Groot BTA, Smets HJM, de Die-Smulders CEM, Faber CG, Gerrits MM. 2020. Parental repeat length instability in myotonic dystrophy type I pre- and postmutations. Eur J Hum Genet 28: 956–962. doi:10.1038/s41431-020-00601-4

Katsmata S, Fardo DW, Bachstetter AD, Artushin SC, Wang WX, Wei A, Brzezinski LJ, Nelson BG, Huang Q, Abner EL, et al. 2020. Alzheimer disease pathology-associated polymorphism in a complex variable number of tandem repeat region within the MUC6 gene, near the AP2A2 gene. J Neuropathol Exp Neurol 79: 3–21. doi:10.1093/jnen/nlz116

Kawaguchi Y, Okamoto T, Tanwaki M, Aizawa M, Inouye K, Matayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I, et al. 1994. CAG repeat in a novel gene causes fragile-X-like disease at chromosome 14q12.1. Nat Genet 8: 221–228. doi:10.1038/ng1194-221

Kazemi-Esfarjani P, Trifiro MA, Pinsky L. 1995. Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: a possible pathogenetic role for polyglutamine expansions in androgen-dependent neuro-pathies. Hum Mol Genet 4: 523–527. doi:10.1093/hmg/4.4.523

Kehl A, Cizinauskas S, Langbein-Detsch I, Mueller E. 2019. NHLRC1 dodecamer expansion in a Welsh Corgi (Pembroke) with Lafora disease. J Hum Genet 64: 673–679. doi:10.1038/s41436-019-0193-1

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res 21: 487–493. doi:10.1101/ gr.113895.110

Knight SJL, Flannery AV, Hirst MC, Campbell L, Christoukoulou Z, Pehrs SP, Pointon J, Middleton-Price HR, Barnicoat A, Pembrey ME, et al. 1993. Trinucleotide repeat amplification and hypermethylation of a Cpg island in FRAXE mental retardation. Cell 74: 127–134. doi:10.1016/0092-8674(93)90030-F

Kobayashi H, Abe K, Matsuta S, Ikeeda Y, Hitomi T, Akechi Y, Hubo T, Iwu W, Okuda H, Koizumi A. 2011. Expansion of intronic GGCCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet 89: 91–103. doi:10.1016/j.ajhg.2011.05.015

Koide R, Ikeuchi T, Onodera O, Tanaka H, Igarashi S, Endo K, Takahashi H, Kondou R, Ishikawa A, Hayashi T, et al. 1994. Unstable expansion of CAG repeat in hereditary dentato-rubral-pallidoluysian atrophy (DRPLA). J Hum Genet 39: 1–9. doi:10.1038/ng0194-9

Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito T, Yamada M, Takahashi H, Tsuchi S. 1999. A neurological disease caused by expanded CAG trinucleotide repeat in the NOP56 gene: a new polyglutamine disease? Hum Mol Genet 8: 2047–2053. doi:10.1093/hmg.8.11.2047

Koob MD, Benzow KA, Bird TD, Day JW, Moseley ML, Ranum LP. 1998. Further cloning of expanded trinucleotide repeat sequences from genomic DNA. Nat Genet 20: 72–75. doi:10.1038/ng0919-72

Koob MD, Moseley SL, Chu JT, Benzow KA, Bird TD, Day JW, Ranum LP. 1999. An untruncated CAG expansion causes a novel form of spinocerebellar ataxia (SCA6). Nat Genet 21: 379–384. doi:10.1038/77710

Kozareva I, Turner DJ. 2011. Amplification-free library preparation for paired-end illumina sequencing. Methods Mol Biol 733: 257–266. doi:10.1007/978-1-61779-089-8_18

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. Nat Methods 6: 291–293. doi:10.1038/nmeth.1313

Kranz A, Skariah G, Zhang Y, Bayly B, Todd PK. 2019. Neuropathology of RAN translation proteins in fragile-X-associated tremor/ataxia syndrome. Acta Neuropathol Commun 7: 152. doi:10.1186/s40478-019-0782-7

Kreamer DJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, Richards R. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence PCGCg. Science 252: 1711–1714. doi:10.1126/science.1675488

Laabs B-H, Klein C, Pozojevic J, Domingo A, Brügmann N, Grutz K, Rosati B, Jamora RD, Toth PK. 2021. Identifying genetic modifiers of age-associated penetrance in X-linked dystonia-parkinsonism. Nat Commun 12: 3216. doi:10.1038/s41467-021-23491-4

LaCroix AJ, Stabler D, Sahraoui R, Adam MP, Mhaffy M, Kerman M, Myers CT, Fagerstrom C, Andriotis G, Akkar VM, et al. 2019. GCC repeat

22 Genome Research www.genome.org
MECP2 gene causes a new synergistic mental retardation condition. Eur J Hum Genet 2009; 17: 1399–1404. doi:10.1038/ejhg.2009.159

Ritchie RJ, Knight SJ, Hirst MC, Grewal PK, Bobrow M, Cross GS, Davies KE. 1994. The cloning of FRAXF: trinucleotide repeat expansion and methylation at a fragile site in drosophila. Hum Mol Genet 3: 2115–2121. doi:10.1093/hmg/3.12.2115

Rodriguez CM, Todd PK. 2019. New pathologic mechanisms in nucleotide repeat expansion disorders. Neurobiol Dis 130: 104515. doi:10.1016/j.nbd.2019.10.017

Rodriguez-Rebola I, Madrigal I, Paganobarraga J, Xuncía M, Badenas C, Kulisevsky J, Gomez B, Milà M. 2009. Penetration of FMR1 premutation associated pathologies in fragile X syndrome families. Eur J Hum Genet 2009; 17: 1399–1404. doi:10.1038/ejhg.2009.51

Rosas I, Martínez-Clemente J, Claramunt J, Illán-Gala I, Dols-Cardo O, Borroni B, Almeida MR, van der Zee J, Van Broeckhoven C, et al. 2020. Role for ATXN1, ATXN2, and HTT intermediate repeats in frontotemporal dementia and Alzheimer’s disease. Neurobiol Aging 87: e139.e139-e139. doi:10.1016/j.neurobiolaging.2019.10.017

Ross CA, McNinn MG, Margolis RL, Li SH. 1993. Genes with triplet repeats: candidate mediators of neuropsychiatric disorders. Trends Neurosci 16: 254–260. doi:10.1016/0166-2236(93)90175-L

Ruggieri A, Naumenko S, Smith MA, Iannibelli E, Blasievich F, Bragato C, Gibertini S, Barton K, Vorgend M, Marcus K, et al. 2020. Multitonic elucidation of a coding 99-mer repeat-expansion skeletal muscle disease. Acta Neuropathol 140: 231–235. doi:10.1007/s00401-020-02164-4

Sanpei K, Takano H, Igarashi S, et al. 1996. Identification of the spinocerebellar ataxia type 31 is associated with (TGGAA)n. Exp Neurol 140: 344–348. doi:10.1006/exnr.2000.6410

Saterland P, Levitt BR, Landwehrmeyer GB, Wild EJ, Saft C, Barker RA, Blair L. 2005. Further segregation analysis of the fragile X syndrome with special reference to transmitting males. Hum Genet 69: 289–299. doi:10.1007/s00439017-1964-4

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaela N, Roogurala MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51: 30–35. doi:10.1038/s41588-018-0273-y

Shirazi SK, Bober MA, Coetzer AL. 1998. Polymorphic exonic CAG microsatellites in the gene amplified in breast cancer (ABI1 gene). Clin Genet 54: 102–103. doi:10.1111/j.1399-0004.1998.tb03704.x

Shortt JA, Ruggiero RP, Cox C, Wachholder AC, Pollock DD. 2020. Finding and extending ancient simple sequence repeats derived regions in the human genome. Mob DNA 11: 11. doi:10.1186/s13100-020-00206-y

Sisodia SS. 1998. Nuclear inclusions in glutamine repeat diseases: are they pernicious, coincidental, or beneficial? Cell 95: 1–4. doi:10.1016/s0092-8674(00)81743-2

Snell RG, MacMillan JC, Cheadle JP, Fenton I, Lazarou LP, Davies P, MacDonald ME, Gusella JF, Harper PS, Shaw DJ. 1993. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington’s disease. Nat Genet 4: 393–397. doi:10.1038/ng0893-393

Sone M, Kiri M, Inagaki T, Katsumara R, Takagi S, Yokoi S, Araki K, Kato T, Nakamura T, Koike H, Iwashita A, Sato T, Oyake M, Sasaki H, et al. 2016. Long-read sequencing identifies GGC repeat expansions in NOTCH2NL associated with neuronal intranuclear inclusion disease. Brain 139: 3170–3186. doi:10.1093/brain/aww249

Sone M, Miyashita S, Fujita A, Mizuguchi T, Hamaoka K, Kiri M, Koike H, Hasegawa T, Takashima H, Sugiyama H, et al. 2019. Long-read sequencing identifies GGC repeat expansions in NOTCH2NL associated with neuronal intranuclear inclusion disease. Nat Genet 51: 1215–1221. doi:10.1038/s41588-019-0459-y

Song JT, Lowe CR, Kingsley DM. 2018. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. Am J Hum Genet 103: 421–430. doi:10.1016/j.ajhg.2017.07.011

Strano P, Caramia F, Di Benedetto R, Tortora F, Zara F, Striano S. 2009. H-MLR spectroscopy indicates prominent cerebellar dysfunction in benign adult familial myoclonic epilepsy. Epilepsia 50: 1491–1497. doi:10.1111/j.1528-1167.2008.01900.x

Strome P, Mangelbosch ME, Shaw MA, Lower KM, Lewis SM, Buxhve R, Lütcherath V, Gedon AK, Wallace RH, Scheffer IE, et al. 2002. Mutations in the human ortholog of Aristaless cause X-linked mental retardation and epilepsy. Science 290: 1548–1551. doi:10.1126/science.104026-P

Suthersand MP, Krishen F, Lynch M, Pritchard M, Yu S, Richards RL, Haan EA. 1991. Hereditary unstable DNA: a new explanation for some old genetic questions? Lancet 338: 289–292. doi:10.1016/s0140-6736(91)90426-p

Szalay J, Adikari R, Gallego J, Carver P, Schmitter C, Kato T, Al-Saadi J, Hebert D, et al. 2018. Intron retention mechanisms in fragile sites. Genome Res 28: 98–108. doi:10.1101/gr.223405.117

Tabrizi SJ, Leavitt BR, Landwehrmeyer GB, Wild RJ, Cafferty SR, Blair LF, Craufurd D, Priller J, Rickards H, et al. 2019. Targeting Huntington expression in patients with Huntington’s disease. N Engl J Med 380: 227–236. doi:10.1056/NEJMoa181311

Tábuas-Pereira M, Almeida L, Almeida MR, Duarte S, Pinho J, Matos A, Negrão L, Geraldo A, Santana I. 2019. Increased risk of melanoma in CS90F12 repeat expansion carriers: a case-control study. Muscle Nerve 59: 362–365. doi:10.1002/mus.26835
Tang, H., Kinkes EF, Lippert C, Biggs WH, Fabani M, Guzmán E, Ramakrishnan S, Lavrenko V, Kakaradov B, Hou C, et al. 2017. Profiling of short- tandem-repeat disease alleles in 12,632 human whole genomes. Am J Hum Genet 101: 700–715. doi:10.1016/j.ajhg.2017.09.012

Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. 2018. Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. Am J Hum Genet 103: 858–873. doi:10.1016/j.ajhg.2018.10.015

Tezenas du Montcel S, Durr A, Figueroa KP, Ichikawa Y, Brussino A, Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. 2018. Intermediate CAG repeat expansion in the ATXN2 gene is a unique genetic risk factor for ALS—a systematic review and meta-analysis of observational studies. PLoS One 9: e0135534. doi:10.1371/journal.pone.0135534

Wenger AE, Gomes J, Cashmore AJ, Lij, J, Kwreski D. 2014. Intermediate CAG repeat expansion in the ATXN2 gene is a unique genetic risk factor for ALS—a systematic review and meta-analysis of observational studies. PLoS One 9: e0135534. doi:10.1371/journal.pone.0135534

Wang, X., Zghoul F, Riccardi GR, Peters JF. 1987. The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. Science 237: 420–423. doi:10.1126/science.3630329

Welsh GA, McMillan C, Cullen CL, Boston SE, Turnbull J, Minassian BA. 2017. Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res 27: 697–708. doi:10.1101/gad.215095.116

Wenger AE, Peluso P, Rowell WJ, Chang P-C, Concepcion GT, Ehler J, Jungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 37: 1155–1162. doi:10.1038/s41587-019-0217-z

Weinberg ED, Altf RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, Baratz KH. 2012. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts fetal cerebral dysmyelination. PLoS One 7: e49083. doi:10.1371/journal.pone.0049083

Wijmenga C, Van der Tuin I, Clark LN, Dib C, Witteveen SP, Duyck R, De Vries HE, Gruter A-M, Hofker MH, Moerker P, Williamson R, et al. 1992. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. Nat Genet 2: 28–30. doi:10.1038/ng0992-28

Winnipeg CN, de Becker R, Ramsay J, Stroes D, Smit A, Fitzgerald DR, Kooij RF. 2007. CGG-Repeat expansion in the DMPK gene is associated with the fragile site FRA12A on chromosome 12q13.1. Am J Hum Genet 80: 221–231. doi:10.1086/510800

Wright GEB, Collins JA, Kay C, McDonald C, Dolzhenko E, Xia Q, Bec M, Kolesnikov A, Olson ND, et al. 2019. Length of interrupted CAG: independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. Am J Hum Genet 104: 1116–1126. doi:10.1016/j.ajhg.2019.04.007

Xu G-L, Bestor TH, Bourchis D, Hsieh C-L, Tommerup N, Bugge M, Hulten M, Xu Q, Russo JJ, Viegas-Péquignot E. 1999. Chromosome instability and tumour suppressor deficiency syndromes derived from a DNA methylation transference gene. Nature 402: 187–191. doi:10.1038/46052

Yeopong P, Pongpanich M, Srichomthong C, Assawapitaksak A, Shotelersuk V, Tantrirudnak J, Chunharas C, Supaphetkit P, Shotelersuk V. 2019. TTTCA repeat expansion in an intron of YEAT5 in benign adult familial myoclonic epilepsy type 4. Brain 142: 3360–3366. doi:10.1093/brain/awz267

Yeopong P, Chunharas C, Pongpanich M, Bennett MF, Srichomthong C, Chunharas C, Supaphetkit P, Shotelersuk V. 2019. TTTCA repeat expansion in an intron of YEAT5 in benign adult familial myoclonic epilepsy type 4. Brain 142: 3360–3366. doi:10.1093/brain/awz267

Yeopong P, Chunharas C, Pongpanich M, Bennett MF, Srichomthong C, Chunharas C, Supaphetkit P, Shotelersuk V, with the Human Epigenome Center. 2019. Length of interrupted CAG: independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. Am J Hum Genet 104: 1116–1126. doi:10.1016/j.ajhg.2019.04.007

Yuan Y, Liu Z, Hou X, Li W, Ni J, Huang L, Hu Y, Liu P, Hou X, Yue J, et al. 2020. Identification of CAG repeat expansions in the NOTCH2NL gene in amyotrophic lateral sclerosis. Neurology 95: e3394–e4305. doi:10.1212/WNL.0000000000010945

Zambonin JL, Bellomo A, Ben-Paz H, Everman DR, Frazer LM, Geraghty MT, Harper AD, Jones JR, Kamin B, Kernohan K, et al. 2017. Spinoocerebellar

26 Genome Research

www.genome.org

Genome Research
ataxia type 29 due to mutations in ITPR1: a case series and review of this emerging congenital ataxia. Orphanet J Rare Dis 12: 121. doi:10.1186/s13023-017-0672-7
Zhao X, Su L, Schaack S, Sadd BM, Sun C. 2018. Tandem repeats contribute to coding sequence variation in bumblebees (Hymenoptera: Apidae). Genome Biol Evol 10: 3176–3187. doi:10.1093/gbe/evy244
Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the α1A-voltage-dependent calcium channel. Nat Genet 15: 62–69. doi:10.1038/ng0197-62
Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguet A, Stone MD, Margolis J, Peterson M, Markowski TW, Ingram MA, et al. 2011. Non-ATG-initiated translation directed by microsatellite expansions. Proc Natl Acad Sci 108: 260–265. doi:10.1073/pnas.1015343108

Received July 29, 2020; accepted in revised form November 29, 2021.