Pathological Short Tandem Repeats Analysis by Long-Read Sequencing in Affected Individuals

Hagar Mor-Shaked[1,2], Shira Silverstein[1], Chaggai Rosenbluh[1], Orly Elpeleg[1,2] and Tamar Harel[1,2]

1 Department of Genetics, Hadassah Medical Organization, Israel

2 Faculty of Medicine, Hebrew University of Jerusalem, Israel

Introduction: Pathological short tandem repeats (STRs) are repetitive DNA sequences found throughout the genome that can expand and contract in length, resulting in various genetic diseases. These diseases include Huntington's disease, Fragile X syndrome, and myotonic dystrophy, among others. Traditional short-read sequencing technologies have limitations in detecting and accurately characterizing STRs due to their inherent instability and repetitive nature. Long-read sequencing (LRS), on the other hand, can provide a potential solution for accurate and comprehensive STR analysis.

Methods: 9 affected individuals with known pathogenic expansions (previously characterized by repeat-primed PCR) in 9 different loci were sequenced on PromethION (Oxford Nanopore, ONT), to obtain long-read whole genome sequencing (LR-WGS). Short tandem repeat genotyping was performed using the straglr software as bundled within Oxford Nanopore Technologies wf-human-variation workflow using default parameters and with the `--str` flag to specify STR genotyping. Structural variations interpretation was performed on the Geneyx Analysis software.

Result summary: All known repeat expansions were detected by the straglr software.

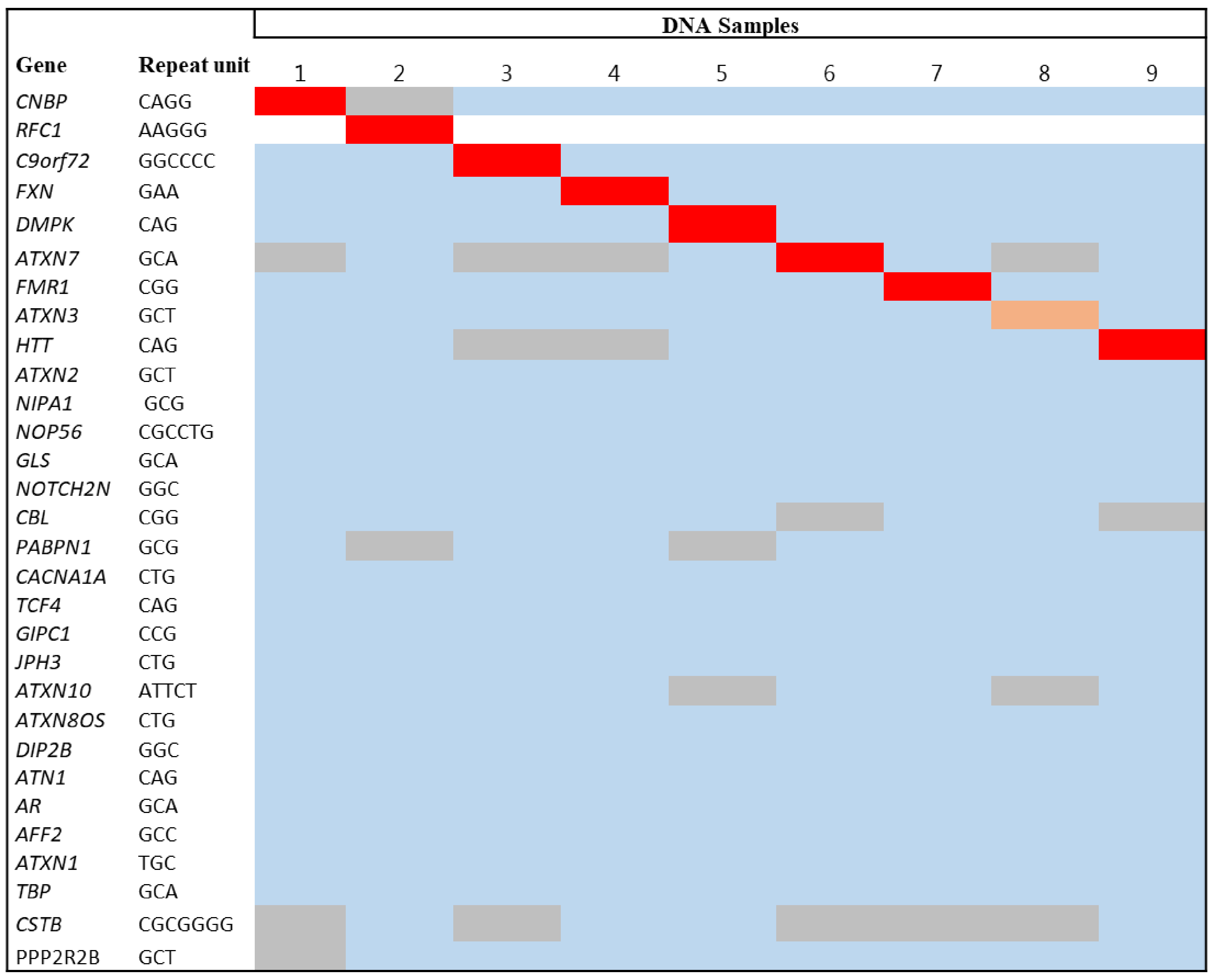
Table and Figure 1. Affected samples summary and results (A) and VCF summary of all samples (B)

Table 1

	Gene	Disease	Molecular test	Repeat genotyping by LR-WGS
1	*CNBP*	Myotonic dystrophy 2	Expanded	1944,8
2	*RFC1*	Crohn's disease	Expanded	840 ,799
3	*C9orf72*	ALS	>145	626,2
4	*FXN*	Friedreich's ataxia	Expanded	514,9
5	*DMPK*	Myotonic dystrophy 1	200	277,20
6	*ATXN7*	SCA7	>100	120,10
7	*FMR1*	Fragile X syndrome	64	64,30
8	*ATXN3*	SCA3	64	62,19
9	*HTT*	Huntington's Disease	44	42,17

Repeat expansion results by a molecular test and by LR-WGS are summarized. All repeat expansions were captured precisely, and sequencing enabled sizing of large expanded repeats (Table 1). LR-WGS enabled detection of STR sizes in 30 loci, without false positive results (Figure 1). Occasionally, STRs were of insufficient coverage and were considered "no calls" (in grey).

Figure 1

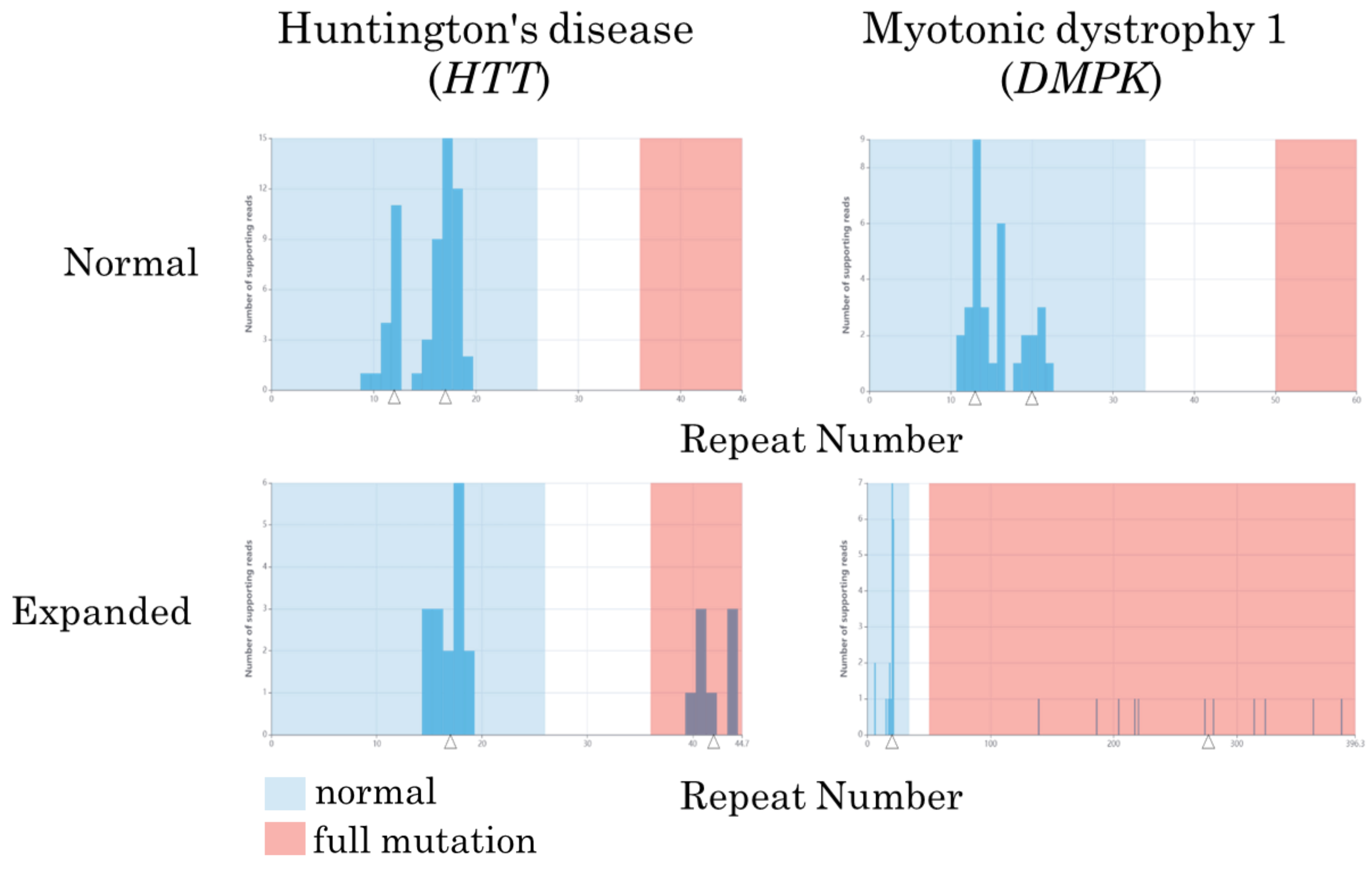


		DNA Samples								
Gene	Repeat unit	1	2	3	4	5	6	7	8	9
CNBP	CAGG	🟥	▨	🟦	🟦	🟦	🟦	🟦	🟦	🟦
RFC1	AAGGG		🟥							
C9orf72	GGCCCC	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦	🟦
FXN	GAA	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦	🟦
DMPK	CAG	🟦	🟦	🟦	🟦	🟥	🟦	🟦	🟦	🟦
ATXN7	GCA	▨	🟦	▨	▨	🟦	🟥	🟦	▨	🟦
FMR1	CGG	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟦	🟦
ATXN3	GCT	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟦
HTT	CAG	🟦	🟦	▨	▨	🟦	🟦	🟦	🟦	🟥
ATXN2	GCT	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
NIPA1	GCG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
NOP56	CGCCTG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
GLS	GCA	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
NOTCH2N	GGC	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
CBL	CGG	🟦	🟦	🟦	🟦	🟦	▨	🟦	🟦	▨
PABPN1	GCG	🟦	▨	🟦	🟦	▨	🟦	🟦	🟦	🟦
CACNA1A	CTG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
TCF4	CAG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
GIPC1	CCG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
JPH3	CTG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
ATXN10	ATTCT	🟦	🟦	🟦	🟦	▨	🟦	🟦	▨	🟦
ATXN8OS	CTG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
DIP2B	GGC	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
ATN1	CAG	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
AR	GCA	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
AFF2	GCC	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
ATXN1	TGC	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
TBP	GCA	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
CSTB	CGCGGGG	▨	🟦	▨	🟦	🟦	▨	▨	▨	🟦
PPP2R2B	GCT	▨	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

▨ no call 🟦 normal 🟧 pre-mutation 🟥 full mutation

Figure 2. Examples of plots produced by the STR module in *HTT* and *DMPK* genes (affected vs. normal)

Figure 2



Straglr software output includes an image graph for each STR locus.

Two loci are presented as examples: the *HTT* gene, which causes Huntington's disease and the *DMPK* gene, which causes myotonic dystrophy 1 (DM1) (Figure 2). The light blue and pink shading designates normal and full mutation STR ranges, respectively.

The pathogenic repeat expansion in the *HTT* gene is typically characterized by more than 40 CAG repeats. In contrast, the pathogenic repeat expansion in the *DMPK* gene can range from hundreds to thousands of CTG repeats. Larger repeat expansions are also characterized by somatic repeat instability, which is shown in the expanded samples of *DMPK*.

Both repeat sizes were captured precisely.

Conclusions: As the cost of genome sequencing continues to drop and new technologies emerge that can detect all types of genetic variations in one test, it is likely that clinical STR detection will eventually shift towards LR-WGS. This is further supported by the ability to analyze DNA methylation from ONT data, which is relevant and important for several pathological STRs. Before implementing LR-WGS for clinical STR detection, further validation is necessary, which includes testing multiple affected individuals for each genetic locus, and ensuring sufficient depth of coverage at each locus.

Note that Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.