Supplementary Material

Approaches to Sequence the *HTT* CAG Repeat Expansion and Quantify Repeat Length Variation

Supplementary Table 1. MiSeq compatible PCR primers and corresponding TruSeq CD indexes for the sequencing of the *HTT* exon one trinucleotide repeat locus for up to 96 samples per MiSeq run. See Excel file.

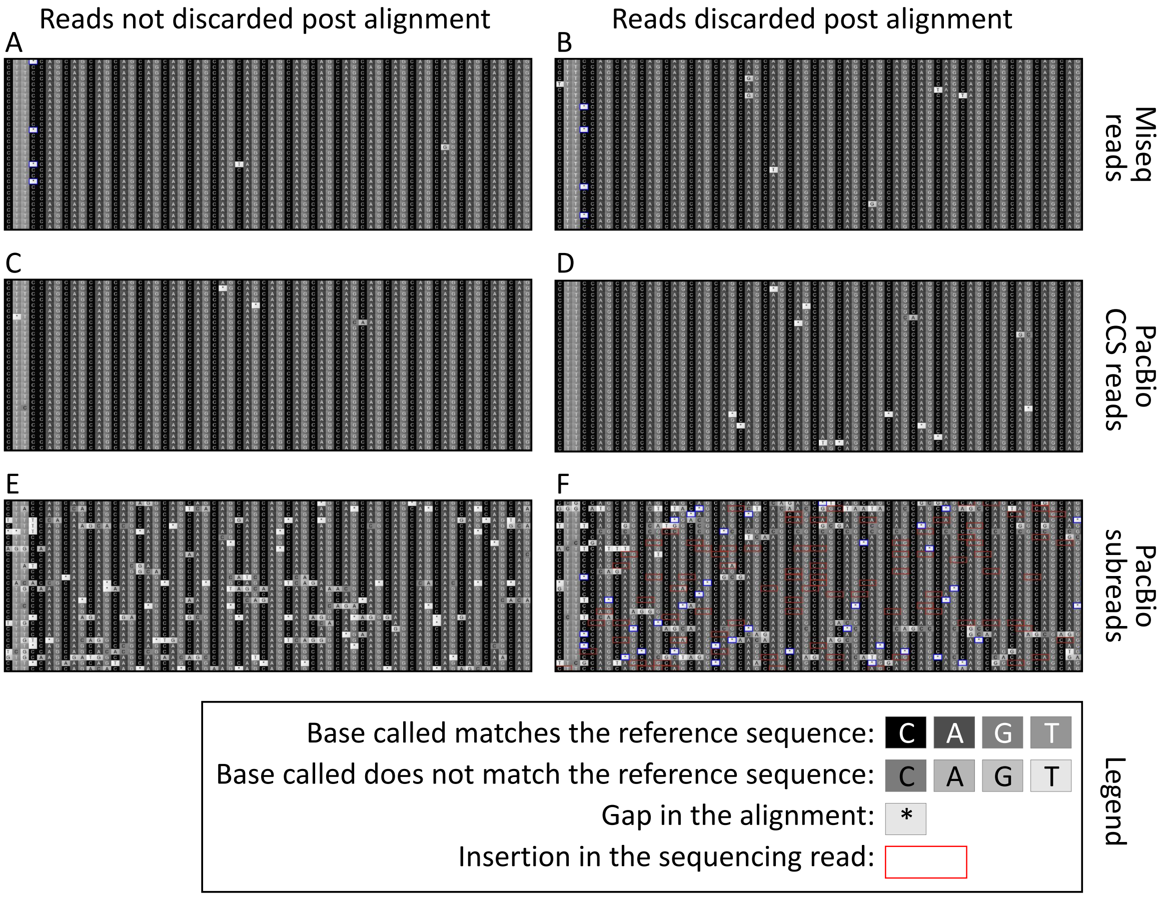
Supplementary Table 2. MiSeq compatible PCR primers and corresponding Nextera XT Index Kit v2 indexes for the sequencing of the *HTT* exon one trinucleotide repeat locus for up to 384 samples per MiSeq run. See Excel file.

Supplementary Table 3. Barcoded PCR primers and corresponding PacBio barcodes for PacBio RS II System in symmetric mode for the sequencing of the *HTT* exon one trinucleotide repeat locus on the PacBio RS II System. See Excel file.

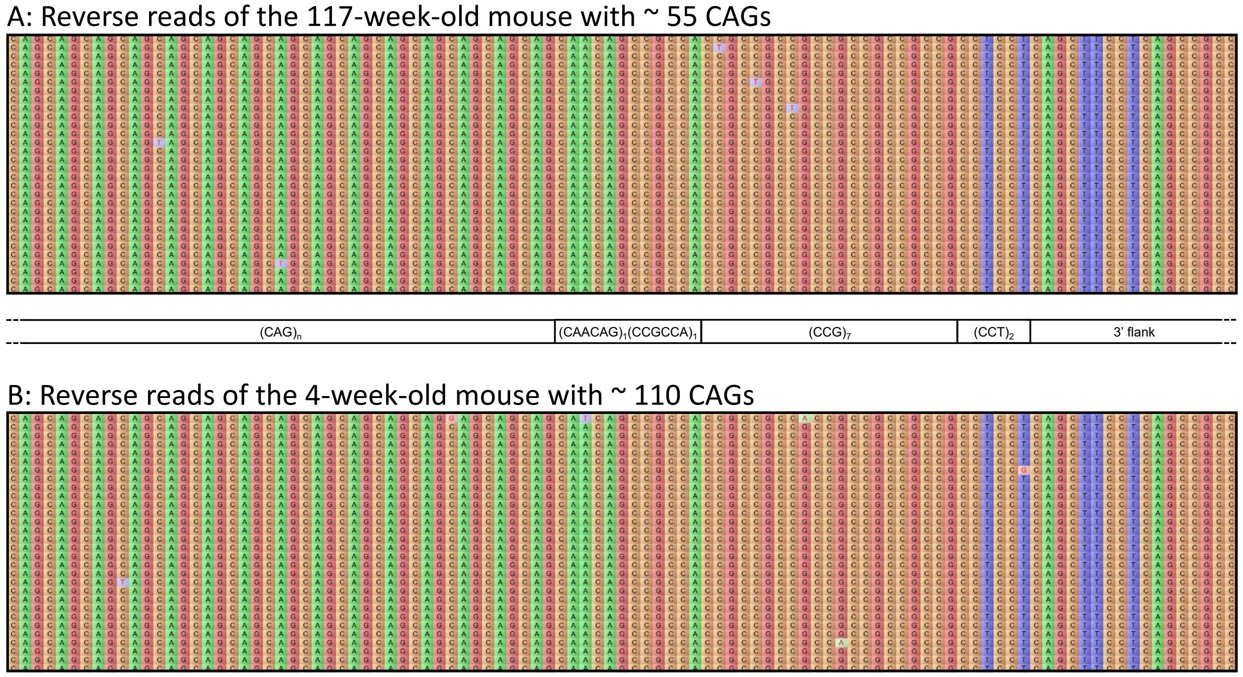
Supplementary File 1. Estimation of the percentage of on-target and full-length reads for each experiment. See PDF.

Supplementary file 2. Comparison between PacBio run 1 (PCR products generated from samples with ~55, ~110 and ~255 CAGs) and PacBio run 2 (PCR products generated from samples with ~255 and ~470 CAGs) by comparing the sample with ~255 CAGs in each run. See PDF.

Supplementary Figure 1. Representative sequence alignments of the 400 nt MiSeq reads (A and B), PacBio CCS reads (C and D) and PacBio subreads (E and F) aligned to a synthetic reference sequence with 115 CAGs. Alignments shown correspond to 30 sequencing reads obtained from the tail at weaning of the 20-week-old mouse with ~110 CAGs. The part of the alignment shown corresponds to the four nucleotides in the immediate 5'–flank of the *HTT* CAG repeat, followed by the first 20 CAGs. Both the reads that were not discarded (i.e., uniquely aligned to a single synthetic reference sequence, A, C and E) and discarded (B, D and F) post alignment are shown. Note that panels A, C and E are the same as panel A, C and E from Fig. 2 and are presented here to ease the comparison between the reads that were discarded or not post alignment.


Reads not discarded post alignment
Reads discarded post alignment
A
B
Miseq reads
C
D
PacBio CCS reads
E
F
PacBio subreads
Base called matches the reference sequence: C A G T
Base called does not match the reference sequence: C A G T
Gap in the alignment: *
Insertion in the sequencing read:
Legend

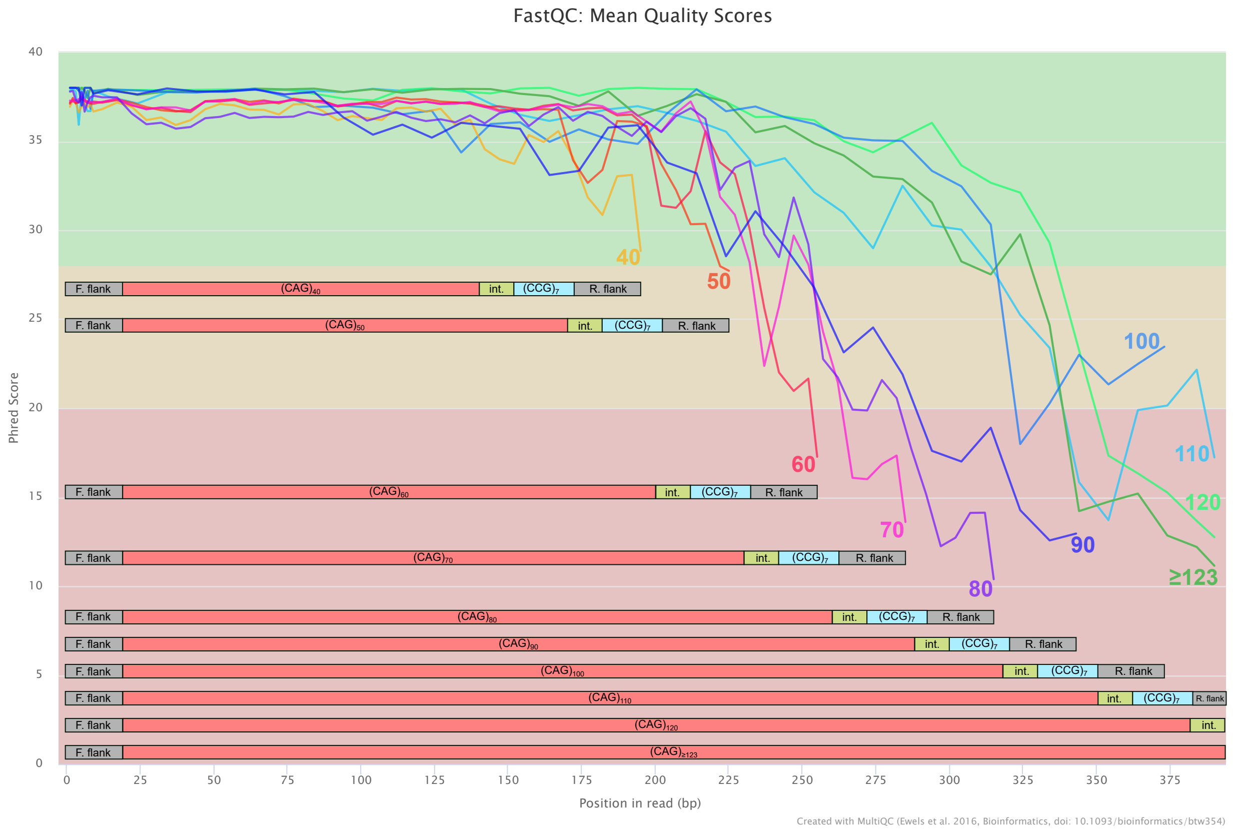
Supplementary Figure 2. Representative sequence alignments of the 200 nt MiSeq reverse reads aligned to the synthetic reference sequence $(CAG)_{100}(CAACAG)_1(CCGCCA)_1(CCG)_7(CCT)_2$-3'-flank. Alignments shown correspond to 30 sequencing reads obtained from the tail at weaning of the 117-week-old mouse with ~55 CAGs (panel A) and from the cerebellum of the 4-week-old mouse with ~110 CAGs (panel B). The part of the alignment shown corresponds to 15 CAGs, followed by the typical intervening sequence between the CAG and CCG repeats (i.e., $(CAACAG)_1(CCGCCA)_1$), $(CCG)_7$, $(CCT)_2$ and 17 nt of the 3'-flanking sequence. Reads were aligned using BWA-MEM with default parameters.



A: Reverse reads of the 117-week-old mouse with ~ 55 CAGs

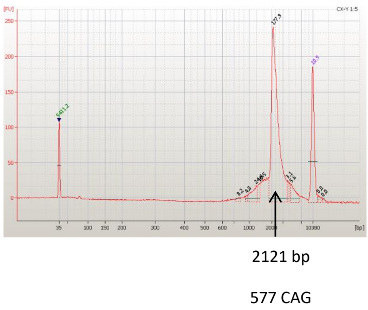
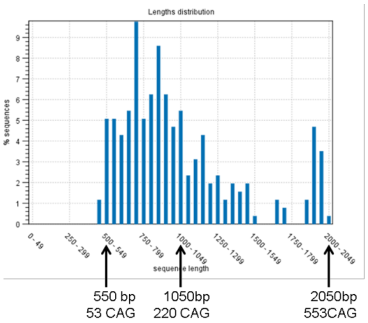
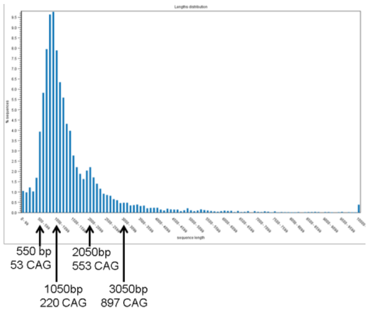
B: Reverse reads of the 4-week-old mouse with ~ 110 CAGs

Supplementary Figure 3. Representative plots of mean quality scores (representing the relationship between the average PHRED score and the position of the base called in the sequencing read) for sequencing reads containing different numbers of CAG repeats. The reads containing different numbers of CAG repeats were extracted from the BAM files produced for the genotyping of the striatum of the 117-week-old mouse with ~55 CAGs and the 20-week-old mouse with ~110 CAGs. Reads with ≤ 90 CAGs were generated for the striatum of the 117-week-old mouse with ~55 CAGs and reads with ≥100 CAGs were generated for the striatum of the 20-week-old mouse with ~110 CAGs. The position of the PHRED score drop-off in the reads is CAG-length dependent, i.e., the position of the drop-off is shifted towards the end of the read with increasing CAG length. This is likely caused by intra-cluster DNA polymerase slippage on the MiSeq flow cell during cluster generation.

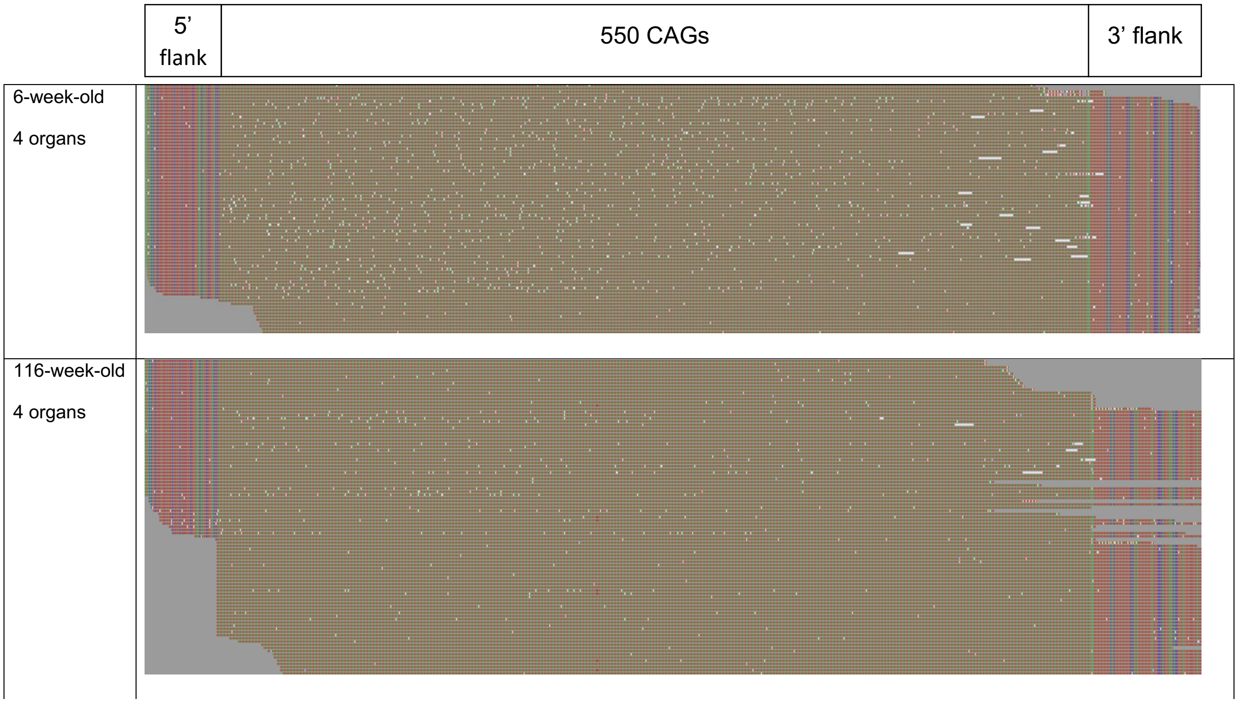


Created with MultiQC (Ewels et al. 2016, Bioinformatics, doi: 10.1093/bioinformatics/btw354)

Supplementary Figure 4. Strong bias toward the sequencing of short fragments when attempting to sequence *HTT* alleles with ~470 CAGs with PacBio sequencing: cortex of a 6-week-old R6/2 mouse with ~470 CAGs as an example. A: Bioanalyzer trace of the barcoded PCR products prepared for PacBio sequencing. B: Read length (bp) frequency distribution of the PacBio circular consensus sequencing reads. C: Read length (bp) frequency distribution of the PacBio subreads.

Size of the PCR products with ~470 CAGs submitted for PacBio sequencing (bp)	Read length (bp) of the PacBio circular consensus sequencing reads	Read length (bp) of the PacBio subreads
A 	B 	C 

Supplementary Figure 5. PacBio sequencing of the *HTT* repeat in R6/2 mice with ~470 CAGs. PacBio circular consensus sequencing (CCS) reads aligned to a synthetic reference sequence containing 550 CAGs. Alignment visualised with Tablet (Milne et al. (2013) *Briefings in Bioinformatics* **14**, 193-202).


5' flank
550 CAGs
3' flank
6-week-old
4 organs
116-week-old
4 organs