Single-cell-resolution analysis of the HTT CAG repeat and genome-wide RNA expression

McCarroll Lab

Revision history:
January 16, 2025 Published version (Cell 2025)

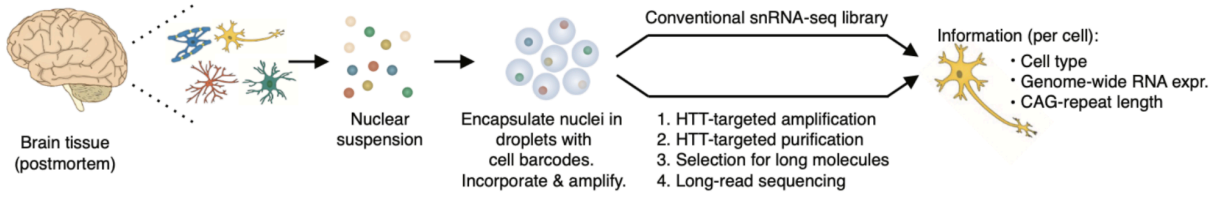
Contents

5. Single-cell HTT-CAG and RNA sequencing

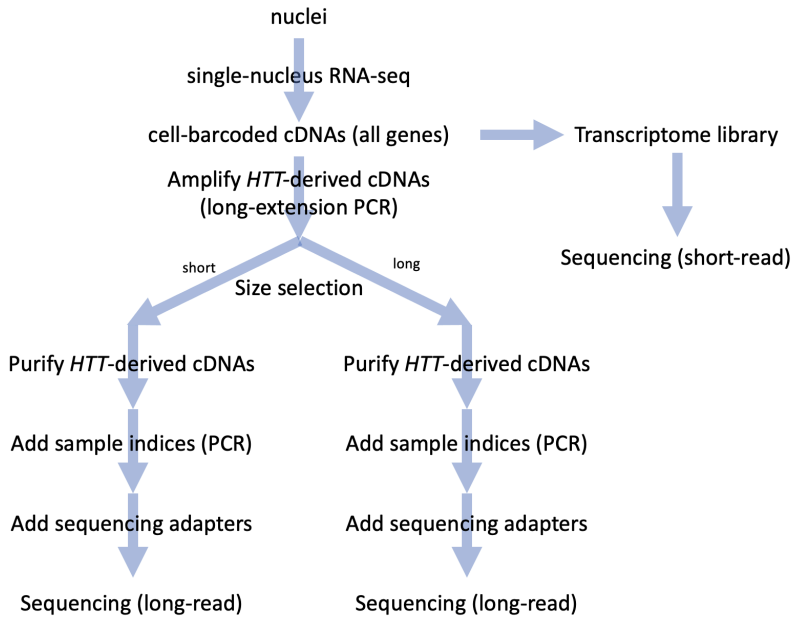
Single-cell-resolution analysis of repeats and genome-wide RNA expression: Constructing single-cell target-sequence (HTT-CAG) libraries.

5.1. Overview

The goal of this approach is to create, from cells or nuclei, libraries of gene transcripts (cDNAs) that (i) contain cellular and molecular barcodes identifying the cells and individual RNA molecules from which the cDNAs are derived, (ii) are greatly enriched for a target sequence of biological interest (the CAG repeat in exon 1 of the *Huntingtin* (*HTT*) gene), (iii) faithfully maintain information about key features of the target sequence (especially, the length of the CAG repeat sequence) throughout PCR amplification, and (iv) ascertain as many as possible of the repeat-containing RNA sequence molecules that were present in the original sample. We hope the following diagrams and discussions are useful in establishing this protocol and adapting it to other biological samples. We also suggest checking the McCarroll lab website for future protocol updates, improvements and discussions.



5.2. Summary of specific steps and their purpose



- Single-cell/nucleus encapsulation, transcript barcoding and reverse transcription
  This first step is shared in common with standard 3' snRNA-seq (e.g. steps 1.0 to 1.5 of the Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index) protocol) with a minor difference in preparing nuclei suspension (step 1.1): instead of water, we use 1%BSA in 1X PBS supplemented with RNase inhibitor (NxGen™ RNAse Inhibitor, 1U/µl). Cells or nuclei are encapsulated and lysed, and their RNA transcripts are primed for reverse transcription using primers (delivered by beads to each droplet) that have droplet-identifying "cell barcodes" and single-molecule-identifying Universal Molecular Identifiers (UMIs). These cell barcodes and UMIs will be used later in computational analysis to match molecules in the target-sequence (*HTT*-CAG) library to the wider molecular profiles of the cells from which they were obtained. Generally these initial steps include encapsulation of cells or nuclei in droplets, followed by (within-droplet) nuclear lysis, priming, and reverse transcription of the mRNAs into cDNAs.

- Transcriptome amplification
  In this step, barcoded cDNAs are amplified using PCR. This is also a standard step in the standard 3' snRNA-seq protocol e.g. from 10X Genomics. We modify this step by adding one or more "spike-in" primers we have designed to the target gene transcript (*HTT*) 5' of the target sequence (the *HTT* CAG repeat). The addition of spike-in primers increases yield of the target sequence by making successful amplification independent of the (only partially efficient) standard, template-agnostic "template switch" step. Since the template-switch step has only partial efficiency (with the result that many first-strand cDNAs are not carried forward into amplification), the addition of spike-in primers greatly increases the ascertainment of target molecules of interest.
  The product of this amplification – a complex mixture of cDNAs (from all genes), with cell barcodes and UMIs incorporated into the cDNA molecules, in which each founding molecule (with a distinct cell barcode and UMI) is now represented by many copies – is then split into fractions which are used respectively to prepare the conventional "transcriptome library" (for single-cell analysis of genome-wide RNA expression) and the DNA-repeat (*HTT*-CAG) library; the latter is prepared by the additional steps described here.

- Target-sequence enrichment ("CAG Amp")
  In this step, we start with the purified output of the transcriptome-amplification step, which is also an intermediate created in the process of generating the 10X 3' snRNA-seq library. Making the standard transcriptome library generally uses only some (15 µl) of this intermediate, and we use part of the rest to make the target-sequence library. (We don't use all of it, in case something goes wrong with either library and necessitates a re-do. The key thing is to use sufficient volume that almost all UMI-tagged cDNAs amplified in the previous PCR are sampled at least once. For *HTT*-CAG libraries, we estimate that an input of 4ul generally accomplishes this, and that use of more sample yields more-incremental increases in the number of UMIs ascertained. We use a biotinylated gene-specific primer (designed 5' of the target sequence, and 3' of any

spike-in primers) and another primer designed to the 10X bead sequence, to selectively amplify molecules that contain the target sequence (the CAG repeat within exon 1 of the *HTT* gene).
The number of PCR cycles should be carefully calibrated to the sample, with the PCR ended while in exponential phase (sometimes also called “log phase”), to prevent late cycles in which incompletely replicated molecules then act as primers in subsequent PCR cycles; this priming by incompletely replicated molecules causes cell barcodes and UMIs to appear in association with the wrong cDNAs. We describe a way to calibrate the number of PCR cycles, at this and a subsequent step, using real-time PCR.

- Size separation
  The CAG Amp step has created molecular libraries that are somewhat enriched for the target sequence. In most applications, we now separate each library into two libraries, based on molecular size – creating “short” and “long” libraries. This size separation helps protect the longer molecules (e.g. those cDNAs containing very-long CAG repeats) from being out-competed by shorter molecules during subsequent PCR and (later) in reaching target sites on the PacBio sequencing flow cell. This is key for efficiently ascertaining the *HTT* transcripts with long CAG expansions (150+), which we have found to be the molecules of greatest interest for many biological inquiries. However, for some specific applications, such as scenarios in which CAG-length distributions are within a lower range (6-100) or have a smaller variance, this step is unnecessary, and it is simpler to go forward with a single library.
  (Note that the order of the size-separation w.r.t. earlier/subsequent steps could in principle also be reversed.)

- Streptavidin purification
  This step purifies the target molecules away from the rest of the library. Since the gene-targeting (5’ HTT) primer we used in the CAG amp step is biotinylated, the streptavidin beads will bind only the target molecules. The bead-associated molecules are the molecules we elute and carry forward into downstream steps.

- Indexing and further amplification
  This step further amplifies the target molecules (the cDNAs that contain *HTT* CAG-repeat sequences) and adds molecular indexes so that libraries from multiple samples can be pooled for sequencing. As with the target-sequence enrichment step, the number of PCR cycles should be carefully calibrated to the sample, with the PCR ended while in exponential (“log”) phase, to prevent late cycles with incompletely replicated molecules that then act as primers in subsequent PCR cycles, causing cell barcodes and UMIs to appear in association with the wrong cDNAs. We describe how to do this using real-time PCR.

- Final purification (1X SPRI)

This step removes primers and buffers, yielding a library that can be immediately sequenced on the Illumina platform or further prepared for sequencing on the PacBio platform.

- Sequencing and analysis
  Analysis by Illumina sequencing is useful for assessing the overall properties of the library (complexity, numbers of transcripts ascertained, numbers of cells from which these transcripts are derived, proportions of normal and HD alleles). Analysis by long-read sequencing is essential for measuring long CAG-repeat expansions that are beyond the length limits of Illumina sequence reads, and in general for ascertaining molecules that are longer than those the Illumina platform (e.g. with its cluster-generation step and optimal/recommended cluster-size ranges) is designed to ascertain effectively.

5.3. Materials

- 10X Genomics Chromium Next GEM Single Cell 3' Kit v3.1
- 10X Genomics Chromium Next GEM Chip G Single Cell Kit
- Qiagen UltraRun® LongRange PCR Kit
- 10X Genomics Dual Index Kit TT Set A
- Bio-Rad iQ SYBR Green Supermix (2X)
- Pacific Biosciences SMRTbell Express Template Prep Kit 2.0
- Primers

  HTT spike-in primers (working stocks: 100 µM):
  - Spike-in A (+std, Forward): HTT.5prime.pos56.CAG-112 (21bp)
    - 5’CCCAGAGCCCCATTCATTGCC
  - Spike-in B (+std, Forward): HTT.5prime.pos148.CAG-27 (22bp)
    - 5’GGCGACCCTGGAAAAGCTGATG

  CAG Amp primers (working stocks: 10 µM):
  - Primer1 (+std, Forward): Biotin IllumR-HTT-C (57bp)
    - 5’-/5BioagGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTTCGAGTCCCTCAAGTCCTTC
  - Primer2 (-std, Reverse): Illum_f_10X_3p_B (55bp)
    - 5’TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACACGACGCTCTTCCGATCT

Nextera indexing primers (working stocks: 10 μM):

Name	5’ common	Adaptor	3’ common
N701	CAAGCAGAAGACGGCATACGAGAT	TCGCCTTA	GTCTCGTGGGCTCGG
N702	CAAGCAGAAGACGGCATACGAGAT	CTAGTACG	GTCTCGTGGGCTCGG
N703	CAAGCAGAAGACGGCATACGAGAT	TTCTGCCT	GTCTCGTGGGCTCGG
N704	CAAGCAGAAGACGGCATACGAGAT	GCTCAGGA	GTCTCGTGGGCTCGG
N705	CAAGCAGAAGACGGCATACGAGAT	AGGAGTCC	GTCTCGTGGGCTCGG
N706	CAAGCAGAAGACGGCATACGAGAT	CATGCCTA	GTCTCGTGGGCTCGG
N707	CAAGCAGAAGACGGCATACGAGAT	GTAGAGAG	GTCTCGTGGGCTCGG
N708	CAAGCAGAAGACGGCATACGAGAT	CCTCTCTG	GTCTCGTGGGCTCGG
N709	CAAGCAGAAGACGGCATACGAGAT	AGCGTAGC	GTCTCGTGGGCTCGG
N710	CAAGCAGAAGACGGCATACGAGAT	CAGCCTCG	GTCTCGTGGGCTCGG
N711	CAAGCAGAAGACGGCATACGAGAT	TGCCTCTT	GTCTCGTGGGCTCGG
N712	CAAGCAGAAGACGGCATACGAGAT	TCCTCTAC	GTCTCGTGGGCTCGG
N501	AATGATACGGCGACCACCGAGATCTACAC	TAGATCGC	TCGTCGGCAGCGTC
N502	AATGATACGGCGACCACCGAGATCTACAC	CTCTCTAT	TCGTCGGCAGCGTC
N503	AATGATACGGCGACCACCGAGATCTACAC	TATCCTCT	TCGTCGGCAGCGTC
N504	AATGATACGGCGACCACCGAGATCTACAC	AGAGTAGA	TCGTCGGCAGCGTC
N505	AATGATACGGCGACCACCGAGATCTACAC	GTAAGGAG	TCGTCGGCAGCGTC
N506	AATGATACGGCGACCACCGAGATCTACAC	ACTGCATA	TCGTCGGCAGCGTC
N507	AATGATACGGCGACCACCGAGATCTACAC	AAGGAGTA	TCGTCGGCAGCGTC
N508	AATGATACGGCGACCACCGAGATCTACAC	CTAAGCCT	TCGTCGGCAGCGTC

5.4. Equipment

- Thermal cycler
- Thermal cycler with real-time system
- Bioanalyzer 2100 (Agilent) or TapeStation (Agilent)
- Tube rotator
- Benchtop mini centrifuge
- Magnetic rack separator

5.5. Step-by-step protocol

5.5.1. Transcriptome amplification

In this step, barcoded cDNAs are amplified using PCR. This is mostly a standard step in the standard 3' snRNA-seq protocol e.g. from 10X Genomics.
However, we have found that this step is usefully modified by adding one or more "spike-in" primers we have designed to the target gene transcript (*HTT*) 5' of the target sequence (the *HTT* CAG repeat). The addition of spike-in primers increases yield of the target sequence by making successful amplification independent of the (only partially efficient) standard, template-agnostic "template switch" step. Since the template-switch step has only partial efficiency (with the result that many first-strand cDNAs are not carried forward into amplification), the addition of spike-in primers increases the complexity (molecular ascertainment) of the resulting libraries.
When using spike-in primers, we add 2 µl of the 100 µM spike-in primers (1 µl per each primer, final concentration of each primer was 1 µM) in each reaction during the sep 2.2a of the standard 3' snRNA-seq protocol from 10X Genomics. The volume of the sample combined with the cDNA Amplification Reaction Mix is decreased accordingly to maintain 100 µL reaction volume.

5.5.2. Target-sequence amplification ("CAG Amp")

In this step, we start with the purified output of the transcriptome-amplification step, which is also an intermediate created in the process of generating the 10X 3' snRNA-seq library. Making the standard transcriptome library generally uses only some (15 ul) of this intermediate, and we use part of the rest. (We do not use all of it, in case something goes wrong with either library and necessitates a re-do. The key thing is to use sufficient volume that almost all UMI-tagged cDNAs amplified in the previous PCR are sampled at least once. For *HTT*-CAG libraries, we find that an input of 4ul generally accomplishes this.) We use a biotinylated gene-specific primer (designed 5' of the target sequence, and 3' of any spike-in primers) and another primer designed to the 10X bead sequence), to selectively amplify molecules that contain the target sequence (the CAG repeat within exon 1 of the *HTT* gene).

The number of PCR cycles should be calibrated to the sample, with the PCR ended while in log phase, to prevent late cycles with incompletely replicated molecules that then act as primers in subsequent PCR cycles (when this happens, it causes cell barcodes and UMIs to associate with the wrong cDNAs). In general, libraries with a larger number of founder molecules (e.g. due to more sample input, higher expression of *HTT*, or better RNA quality) or more-efficient amplification (e.g. due to smaller molecules with shorter CAG-repeat tracts) will not need as many PCR cycles at this stage.

5.5.3. Use of quantitative real-time PCR to optimize PCR programs and prevent chimerism

PCR chimerism arises when an incomplete PCR amplicon serves as a primer for successive PCR cycles. PCR chimerism causes a target sequence (CAG repeat sequence) to become associated with the wrong cell barcode (yielding an incorrect result). PCR chimerism can be particularly problematic when studying repeat expansions, as an incorrect molecule with a short repeat sequence may out-compete (during PCR) correct-but-inefficiently-amplifying molecules with a longer sequence and become the dominant sequence for that cell barcode and UMI.

As the incomplete PCR amplicons tend to be generated when PCR efficiency drops due to limited remaining concentrations of reagents (such as dNTPs, primers, and polymerases), one can substantially prevent PCR chimerism by terminating PCR reactions while they are still in "log phase" – i.e. by performing PCR up to the number of the cycles before the PCR efficiency drops substantially.

We have found that quantitative real-time PCR (qRT-PCR) can be used to optimize the number of PCR cycles to prevent PCR chimerism. The Cq values from the amplification curve in qRT-PCR can provide a good estimate for an optimal number of PCR cycles – one at which substantial product has been generated, but chimerism is minimized.

Out of the 40 ul of the cDNA from 10X step 2, use 1.25 µl of 1:10 diluted cDNA (1/32 of cDNA that will be used for CAG amplification) for qRT-PCR.

Prepare the qRT-PCR mix

	1X (384 well), µl	2.2X (384 well), µl	1X (96 well), µl	2.2X (96 well), µl
SYBR	5	11	10	22
cDNA (1:10)	1.25	2.75	2.5	5.5
Water	2.75	6.05	5.5	12.1
CAG amp primers (F+R)	1	2.2	2	4.4
total	10	22	20	44

Make 10% larger volume to be enough. Technical duplication is recommended.

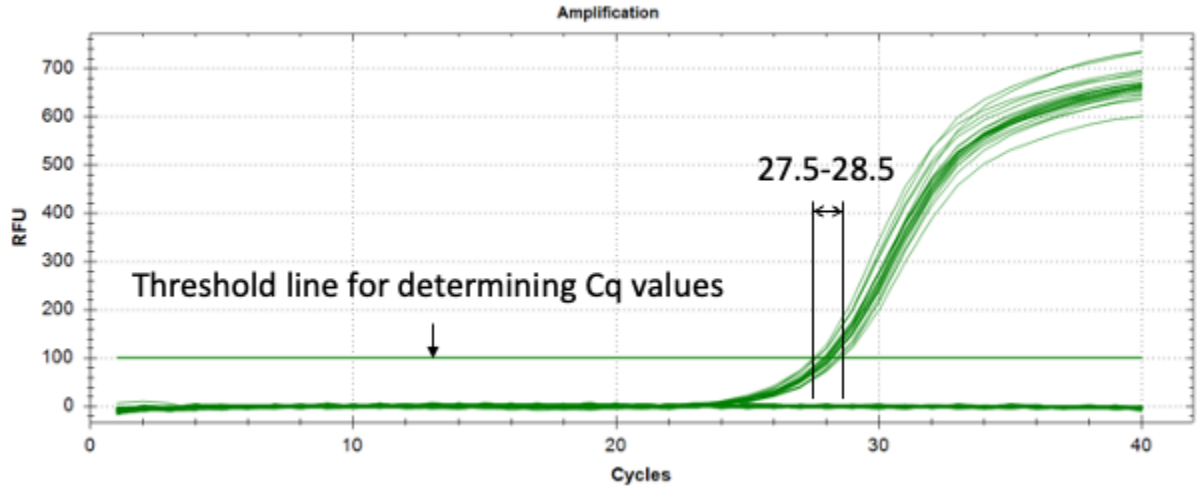
PCR program

1. 95C: 3 min
2. 95C: 15 sec
3. 60C: 6 minutes

Go to 2 and repeat 39X

5. Melt Curve 55.0 to 95.0 C, increment: 0.5 C, for 5 sec + Plate Read

Example of amplification curves



In this example, the Cq values are in the range of 27.5 to 28.5. Since this q-rtPCR pilot used only 1/32 of the cDNA that will be used for CAG amplification, the corresponding cycle numbers for actual CAG amplification will be reduced by about 5 (in the range of 22.5 to 23.5). Therefore, the amplification is not recommended to go beyond 22 cycles.

5.5.4. Amplification of barcoded cDNAs containing the CAG repeat ("CAG Amp" step)

Prepare the master mix

Component	**1x, Volume, µL**	**For 4 rxns, µL**	**For 8 rxns, µL**
Biotin IlumR-HTT-C, 10 µM	1	4.5	9
Illum_f_10X_3p_B, 10 µM	1	4.5	9
Solution Q (5X)	4	18.0	36
DNA polymerase (QRLR)	5	22.5	45
Water	5	22.5	45
cDNA (undiluted)	4	X	X
total	20	72 (aliquot 16 µL/rxn)	144 (aliquot 16 µL/rxn)

For each rxn, add 16 µL master mix to 4 µL undiluted cDNA and run PCR program below:

PCR program (~3h, Lid temp: 105C, vol: 20 µL)

1. 93C: 3 min
2. 93C: 30 sec

3. 60C: 30 sec
4. 68C: 6 minutes

Go to 2 and repeat for a number of cycles that is optimal for your sample (we used 18)

5. 72C: 10 min

5.5.5. Library size separation

In this step, we split the target-sequence library into two libraries based on molecular size (with some overlap). This is to ensure that longer molecules are not out-competed during subsequent steps.
Here we accomplish this with the use of SPRI beads (here at 0.4X concentration).

1. Add 30 µL water (to bring total volume up to 50 µL)
2. Add 20 µL SPRI to 50 µL CAG amp product (0.4x) and mix by pipetting
3. Incubate for 5 min
4. Place on magnet (High) and transfer 70 µL supernatant to new tubes

Bead pellet (“L”)

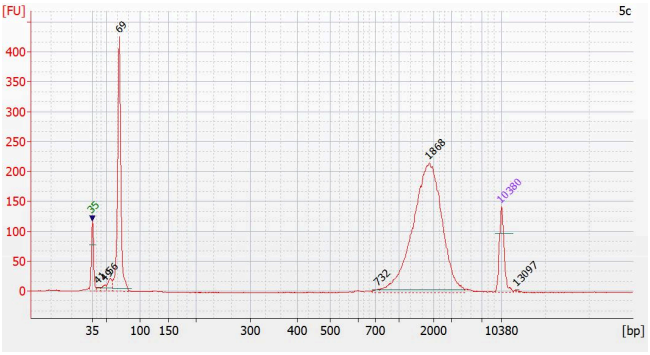
5. Add 200 µL of 80% ethanol to the beads and wait 30 seconds, repeat for a total of 2 washes
6. Spin down briefly, place on magnet (Low), and remove all ethanol
7. Remove from magnet and elute in 11 µL of water, incubate for 5 min
8. Place on magnet Low and transfer 10 µL of supernatant to new strip

Supernatant (“S”)

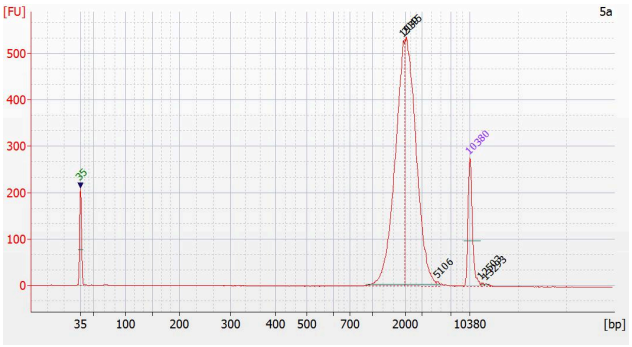
9. Add 30 µL SPRI to the transferred supernatant (1X SPRI)
10. Incubate for 5 min
11. Place on magnet (High) and discard supernatant
12. Add 200 µL of 80% ethanol to the beads and wait 30 seconds, repeat for a total of 2 washes
13. Spin down briefly, place on magnet (Low), and remove all ethanol
14. Remove from magnet and elute in 11 µL of water, incubate for 5 min
15. Place on magnet Low and transfer 10 µL of supernatant to new strip

An example of BioA traces before and after size separation

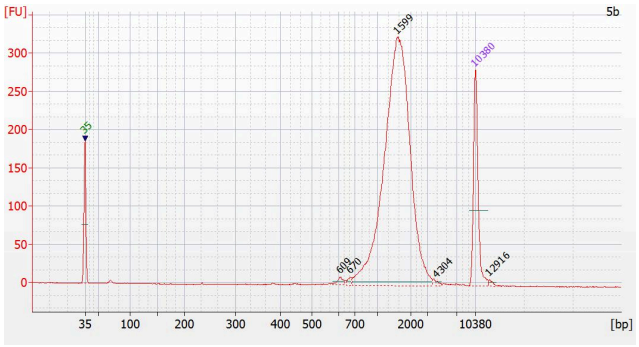
Before size separation



After size separation “L”



After size separation “S”



5.5.6. Purification of target molecules using streptavidin beads

This step purifies the target molecules away from the rest of the library. Since the gene-targeting (5’ HTT) primer we used in the CAG amp step is biotinylated, the streptavidin beads will bind only the target molecules. The bead-associated molecules are the molecules we elute and carry forward into downstream steps.

1. Make 2X W&B buffer (Make just before use)
   For 5 mL (enough for 8 samples):
   0.585 g NaCl
   10 µL 0.5M EDTA
   25 µL 2M Tris pH 7.5
   Bring up rest of volume with water

   Out of the 5 ml, keep 0.5 ml as 2X, and dilute 4.5 ml into 1X for washing.
2. Washing
   1) Resuspend beads (Take out 25 µl for 4 rxn)
   2) Wash with 1 mL of 1x W&B buffer and place on magnet for 1 min, remove supernatant. Repeat for a total of 3 washes

3) Resuspend beads in 2X W&B buffer at twice initial volume (10 µL for sample, 50 µL for 4 rxn)

3. Adding beads
    1) Add 10 µL washed beads to each reaction of purified CAGs and mix by pipetting
    2) Incubate on rotator for 30 min at RT
    3) Place on magnet Low. Remove supernatant and place on magnet High. Wash with 200 µL of 1X W&B for a total of 3 washes (1 min incubation for each wash. Spin down the beads and remove extra W&B buffer)
    4) Resuspend each sample in 10 µL of water

5.5.7. Optimizing the indexing/enrichment step

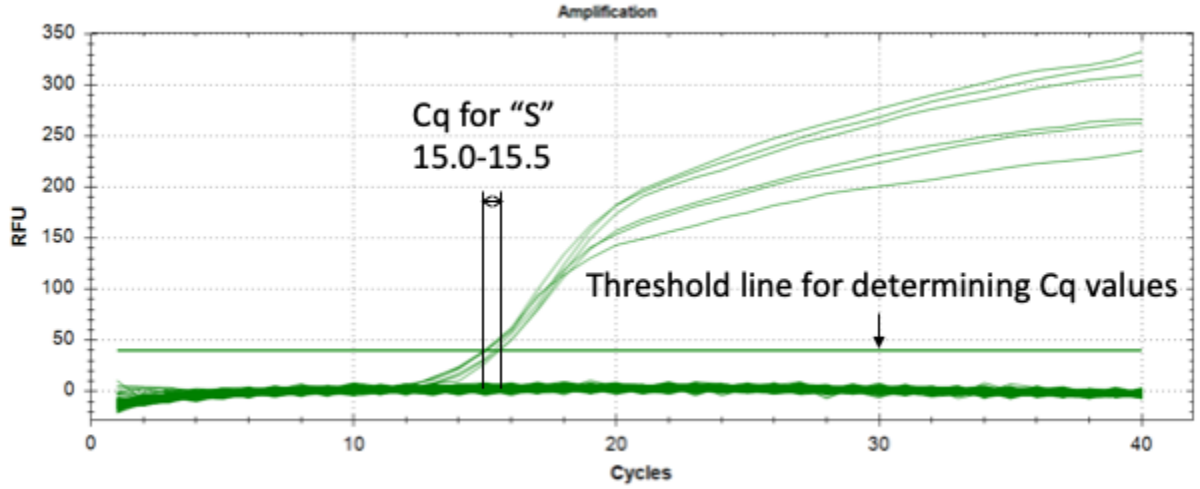
The next step will further amplify the target molecules (the cDNAs that contain *HTT* CAG-repeat sequences) while adding molecular indexes so that libraries from multiple samples can be pooled for sequencing. The “L” and “S” products from the same CAG library are indexed with the same indexes. As with the target-sequence enrichment step, the number of PCR cycles should be carefully calibrated to the sample, with the PCR ended while in log phase, to prevent late cycles with incompletely replicated molecules that then act as primers in subsequent PCR cycles, causing cell barcodes and UMIs to appear in association with the wrong cDNAs. We described how to do this using real-time PCR; after purification of “L” and “S” target molecules with streptavidin beads, the number of indexing PCR cycles can be tested by using 1/32 (3.13 ul of 1:10 diluted) of the streptavidin beads in each size.

Prepare the qRT-PCR mix

	1X (384 well), ul	2.2X (384 well), ul	1X (96 well), ul	2.2X (96 well), ul
SYBR	5	11	10	22
Streptavidin beads (1:10)	3.13	6.89	6.25	13.75
Water	0.87	1.91	1.75	3.85
Nextera indexing primers (N7XX+N50X)	1	2.2	2	4.4
total	10	22	20	44

Make 10% larger volume to be enough. Technical duplication is recommended.
Use the same PCR program for the CAG amplification

Example of amplification curves



In this example, the Cq values are in the range of 15.0 to 15.5. Since this q-rtPCR pilot used only 1/32 of the streptavidin beads that will be used for indexing PCR, the corresponding cycle numbers for actual CAG amplification will be reduced by about 5 (in the range of 10.0 to 10.5). Therefore, the amplification is not recommended to go beyond 10 cycles.

5.5.9. Indexing and further amplification/enrichment PCR

Prepare master mix. Assuming you have done the size-separation step, you will now have twice as many reactions as you had for the CAG amp step because of the size-separation step, in which we split each library into two.

Component	**1x, µL**	**For 8 rxns (9x), µL**	**For 16 rxns (18x), µL**
Solution Q (5x)	8	72	144
Water	8	72	144
DNA polymerase (QRLR)	10	90	180
70X Index (10 µM)	2	X	X
50X Index (10 µM)	2	X	X
CAG amp product - with Beads	10	X	X
Total	40	234 (aliquot 26 µL/rxn)	468 (aliquot 26 µL/rxn)

Add 26 µL Master Mix to 10 µL CAG amp product
Add Index primers, mix by pipetting, and run PCR program below**:**
Index PCR_12 (~1.5 h, Lid temp: 105C, Volume = 40 µL)

1. 93C: 3 min
2. 93C: 30 sec
3. 60C: 30 sec
4. 68C: 6 min

Go to 2 and repeat 9X

5. 72C: 10 min

Remove streptavidin beads

1. Place indexed CAG product on magnet Low
2. **Transfer** supernatant to new strip

Cleanup

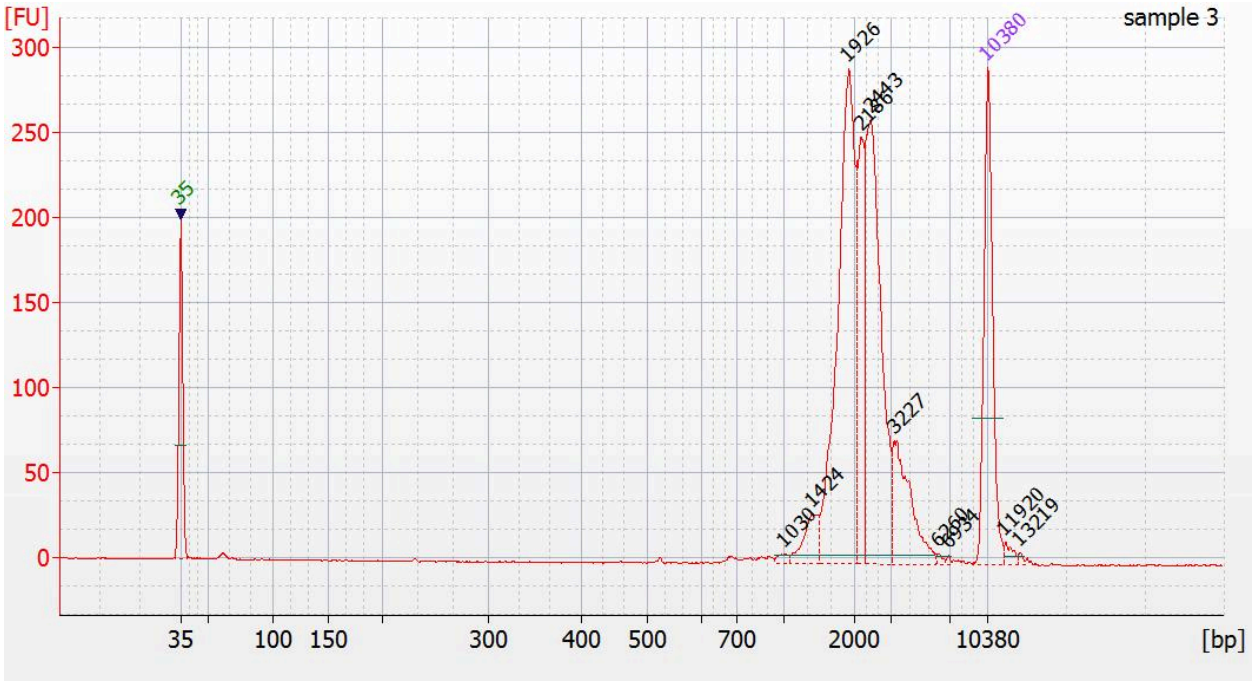
Here we use 1x SPRI to clean up the library

1. Add 40 µL SPRI to 40 µL of indexed CAG product (1x) and mix by pipetting. Incubate for 5 min
2. Place on magnet (High) and discard supernatant
3. Add 200 µL of 80% ethanol and wait 30 seconds, repeat for a total of 2 washes
4. Spin down briefly, place on magnet (Low), and remove all ethanol
5. Remove from magnet and elute in 11 µL of water, incubate for 5 min
6. Place on magnet Low and transfer 10 µL of supernatant to new strip

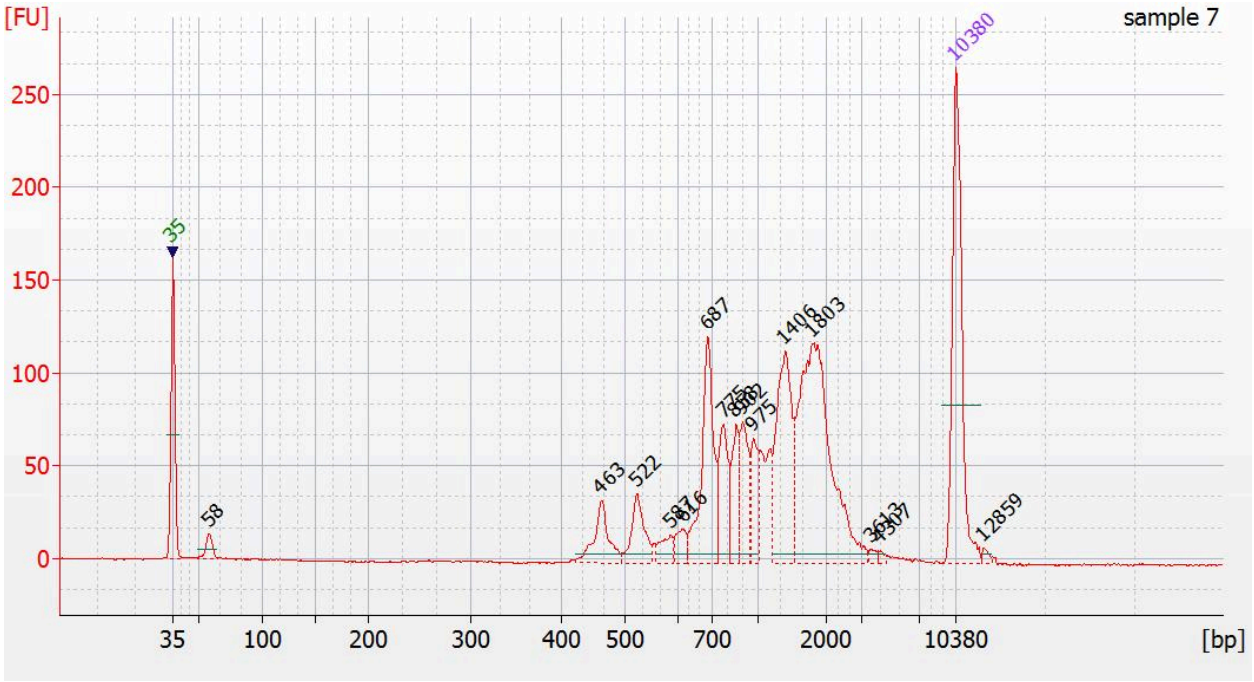
Quantification and assessment

1. Make a 1:10 dilution of final product
2. Run on bioA and record bp length/concentration

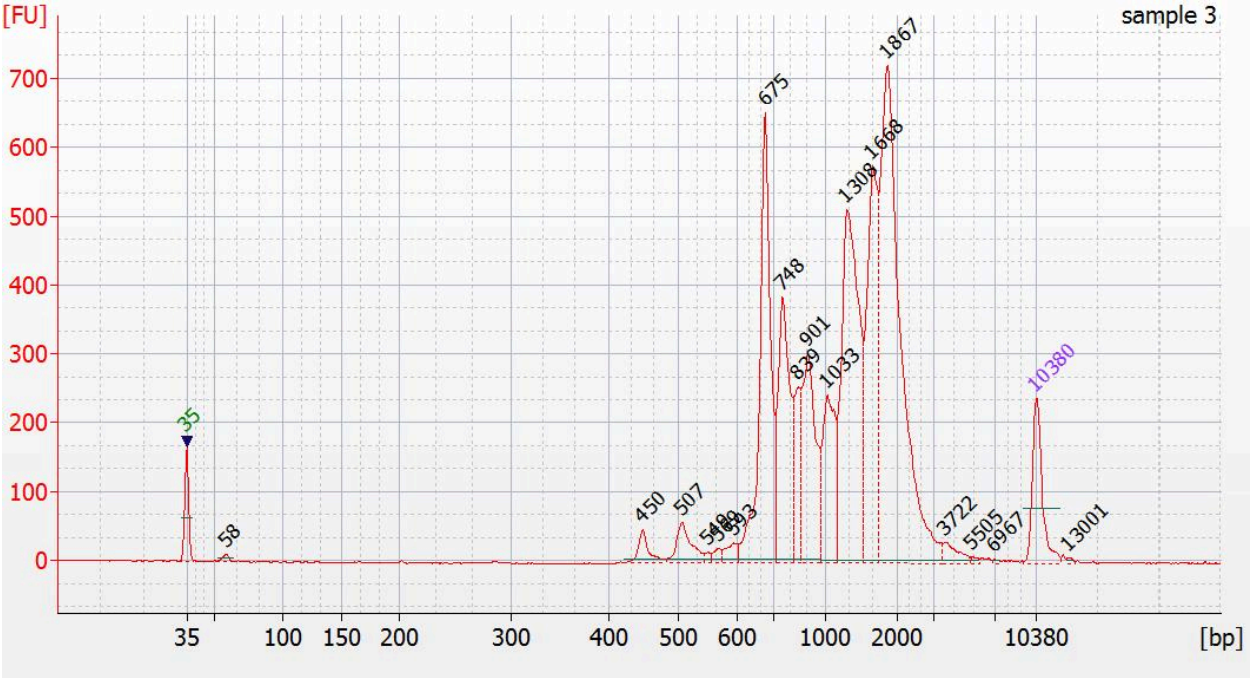
Example of bioA traces for indexed CAG of “L”



Example of bioA traces for indexed CAG of “S”



Example of bioA traces for indexed CAG of “L” with incomplete size separation



5.5.10. Preparation of *HTT*-CAG libraries for long-read sequencing

The “L” and “S” from the same indexed CAG library are separately pooled and prepared for PacBio library and sequencing. Pool the same molar amount of each sample in a total amount of 160-500 ng DNA in 47.4 ul volume.

The PacBio library preparation workflow starts with the “DNA Damage Repair” steps in page 5 of the PacBio Protocol [Preparing Single-Cell Iso-Seq™ Libraries Using SMRTbell® Express Template Prep Kit 2.0 {Part Number 101-892-000 Version 01 (January 2020)}]. Each pooled CAG library (47.4 ul) undergoes “DNA Damage Repair”, “End Repair/A-Tailing”, “Overhang Adapter Ligation”, and “Cleanup SMRTbell Libraries” steps according to the protocol. The library is eluted in 12 uL of EB, and 1 ul of 1:10 diluted library was run on an Agilent Bioanalyzer to monitor the size, molarity, and concentration of DNA for PacBio sequencing.

The “L” and “S” CAG libraries are sequenced on different flow cells (this is to pre-empt a challenge Darren Monckton’s lab has noted, in which short molecules can occupy and fill productive positions on PacBio flow cells more quickly than longer molecules do, causing molecules with longer repeat expansions to be under-represented in the resulting data).

5.6. Computational analysis: decoding *HTT*-CAG reads

After sequencing the *HTT*-CAG libraries (both “L” and “S”), we perform basecalling and alignment using standard workflows for PacBio long reads (**STAR Methods**). The reads wareere then analyzed using a custom informatics pipeline to “decode” the reads to extract

information for downstream analysis. Each read is analyzed (“decoded”) based on the expected layout based on the library construction protocol (**Fig. SN5.1**). The decoding algorithm searches each read for a particular set of landmarks, identifying the landmark using a sensitive Smith-Waterman alignment algorithm designed to accommodate base-level errors and base insertions or deletions in the input reads. Based on the recognition of these landmarks, the read is divided into segments capturing features of the read used in downstream analysis (**STAR Methods**).

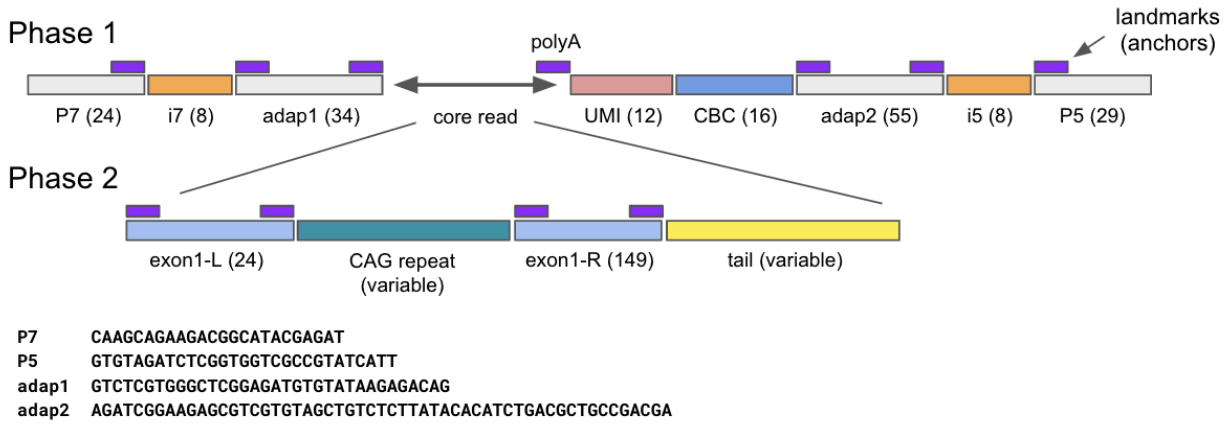


Figure SN5.1. Schematic illustration of long-read decoding in the *HTT* CAG repeat assay
The colored boxes are a schematic representation of the contents of one long read from our *HTT*-CAG library. The computational decoding of each long read was done in two phases. Phase 1 decoded the outer wrapper of the library construct, which included the Illumina sequencing adapters (P7 and P5), Illumina indexes (i7 and i5) which were used to multiplex different snRNA-seq libraries on the same PacBio flowcell, Illumina Nextera and Truseq adapters (adap1 and adap2) and the unique molecular identifier (UMI) and cell barcode (CBC) attached during snRNA-seq library construction. Phase 2 decoded the sequence derived from the host cell’s cDNA, including the initial portion of *HTT* exon 1 (exon1-L), the variable length CAG repeat region, the remaining portion of *HTT* exon 1 (exon1-R) and a variable length tail sequence, which may contain downstream exons of *HTT* or, if the source transcript was unprocessed, sequence from *HTT* intron 1. Numbers in parentheses indicate the expected length (in base pairs) of each read segment. Purple rectangles indicate fixed sequences which were recognized using fuzzy matching to known sequences to identify landmarks that were then used as anchors to divide each read into its constituent components.

5.6.1. Software availability

The software code we have developed for analyzing the HTT-CAG libraries is available at https://github.com/broadinstitute/HTT-CAG-Software