Marissa B. Hirst ${ }^{1}$, John Obenauer ${ }^{1}$, Jeff Aaronson ${ }^{2}$, Mithra Mahmoudi ${ }^{2}$ and Douglas Macdonald ${ }^{2}$ ${ }^{1}$ Rancho BioSciences, San Diego, CA USA
${ }^{2}$ CHDI Management/CHDI Foundation, Los Angeles, CA USA

## Introduction

Huntington's Disease (HD) is a progressive, neurodegenerative disorder inherited in an autosomal dominant fashion and caused by a genetic mutation in the huntingtin (HTT) gene (1). The HTT protein is characterized by a polyglutamine (polyQ) repeat of variable length at the N -terminus of the protein (2). This variable domain is encoded by a CAG repeat in Exon 1 of the HTT gene. In the WT HTT gene, generally 10-28 CAG repeat lengths are observed and in a person with HD, the HTT gene contains 35 or more CAG repeat lengths, represented as an expanded polyglutamine (polyQ) tract in the mutant form of the HTT protein (mHTT) (2).
The variable domain length for the polyQ region in the HTT protein makes it difficult to specify a standard naming convention for the downstream amino acid residues towards the carboxy terminus. For example, Serine 421 is a highly conserved phosphorylation site in the HTT protein and across all vertebrates (3). In addition, Serine 421 has been shown to protect against the toxicity of the expanded polyQ HTT (4). With the variable length of the polyQ region in mHTT , is Serine 421 the $421^{\text {st }}$ amino acid residue in both WT HTT and mHTT?
The convention at CHDI Foundation is to number human HTT amino acids relative to the protein containing a Q23 polyQ repeat. For example, the plasmids Htt-Q23, mix, 1-90, human and Htt-Q73, mix, $1-90$, human are specified as containing 90 amino acids despite the latter having a polyQ region length of 73 . Alternatively, plasmids and proteins acquired from third parties have used a different nomenclature convention.
For example, CHDI Htt-Q23, mix, $1-573$, human has Q23 and a length of 571 amino acids in length. The plasmid from Novartis, Htt-Q25, GST, 1 - 548 , human has Q25 and 573 amino acids in length. These proteins, despite their apparent nomenclature length of 1-573 and1-548, are identical except the latter has a polyQ length of two additional glutamines. Despite the identical proteins, the nomenclature for these two HTT proteins are specified as two different HTT proteins.
CHDI has generated and collected 704 HTTDNA constructs that have been used in previous experiments, each with different lengths and mutations. The purpose of this work was to determine the consistency in the naming conventions for the HTT polyQ region across 129 sequence confirmed DNA constructs as well as further classify each HTT construct relative to the reference human HTT protein sequence (NM_002111.8).

## How were HTT protein sequences classified?

All DNA constructs were processed through a high throughput custom $R$ script for the analysis detailed below. All DNA constructs were translated from DNA to protein in six frames and the open reading frame (ORF) identified by the presence of a polyQ tract (all constructs contained Exon 1 of the HTT protein). Let's start with the example of $\mathrm{Htt}-\mathrm{Q} 16$, GST, 1-548, human (Novartis):

## Presence and location of epitope tags

1. Identify short tags with an exact sequence match
2. Identify long tags ( $>20 \mathrm{aa}$ ) with pairwise local alignment $>=75 \%$ match

MSPILGYWKIKGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIR YIADKHNMLGGCPKERAEISMLEGAVLDIRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNG DHVTHPDFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAWPLQGWQATFGGGDHP PKSDLEVLFQGPLGSMATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPPQA PKSDLEVLFQGPLGSMATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQA
QPLLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAME QPLLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENNIVAQSVRNSPEFQKLLGIAME
LFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKC LFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKC
RPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVS RPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVS
ICQHSRRTQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGS VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAASSGVSTPGSAG HDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAVPSDPAMDLNDGTQASSPISDSSQ TTTEGP

## Location of tag:

1. Index of tag location (ind)
2. Index of aa at the start of HTT protein (start)
3. Total length of HTT protein sequence (total)
ind $=1$ aa start $=232$ aa total $=564$ aa
|ind - start $\mid<$ ind - total| $=231<564$ aa $==$ amino-terminus GST tag
Length of polyQ and proline-rich tracts
4. Identify polyQ tract: 4 Q's in a row

QQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPPPG PAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSD DAESDVRM.....
2. Find proline-rich tract right after last $Q$

QQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPPPG PAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSD DAESDVRM.....
3. ID start and end of polyQ and proline-rich tract by index:
a. polyQ start: 18 aa, ends 33 aa
b. proline-rich tract start: 34 aa, ends 71 aa

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQ
PLLPQPQPPPPPPPPPP
4. Clip all sequence before and after the above indices and calculate length
a. polyQ: 16 aa
b. proline-rich tract: 38 aa

## References

1. A novel gene containing a trinucleotide repeat that is expanded unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell. 1993;72(6):971-983.
2. Gil JM, Rego AC. Mechanisms of neurodegeneration in Huntington's disease. Eur J Neurosci. 2008;27(11):2803-2820.

## Which polyQ length is referenced in the HTT protein sequence?

1. Identify and remove polyQ tract in each HTT protein sequence

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPP QAQPLLPQPQPPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENI VAQSVRNSPEFQKLLGIAMELFLLCSDDAESDVRM
2. Replace it with Q0, Q23 (NCBI-NIH reference sequence) or Q25 Q0
MATLEKLMKAFESLKSFPPPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPP
PPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAM ELFLLCSDDAESDVRM

Q23
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPP PQLPQPPPQAQPLLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVN HCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDDAESDVRM

## Q25

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPP PPPQLPQPPPQAQPLLPQPQPPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDR VNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDDAESDVRM
3. Calculate length of HTT protein: Is it equal to the original length?
a. PolyQ length of $Q 0=548$ aa!

Table 1 below reveals the total number of HTT protein sequences that referenced Q0, Q23 or Q25. The other 86 HTT protein sequences did not match any of the polyQ lengths. Many of these sequences assumed a static length (for example, a full-length HTT protein would be 3,144 aa with Q23), but the actual length of the HTT protein was never verified as 3,144 aa in length and was longer.

Table 1. Referenced polyQ lengths and the total number of HTT protein sequences that matched the given polyQ length out of a total of 129 HTT DNA constructs.

| Reference polyQ length | Number of Sequences |
| :--- | :--- |
| Q0 polyQ | 15 |
| Q23 polyQ | 24 |
| Q25 polyQ | 4 |
| Total | 43 |

## How do HTT protein sequences deviate from the reference?

1. Remove epitope tags
2. Align HTT protein sequence to Human HTT protein reference (NM_002111.8) using a local MUSCLE alignment
3. Create a string of 3,144 aa (length of reference sequence) using the following conventions:
a. Identical aa: "."
b. Mismatched aa: Use mismatched aa
c. Deletion: "-" for all locations
d. Additions: "+" at the aa index prior to any aa insertions (represents 1 < insertions)
e. Example: Htt-Q39, 1-90, K6A K9A K15A (IRBM)

## Htt-Q39, 1-90, K6A K9A K15A, human (IRBM)

MATLEALMAAFESLASFQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQPPPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPPPGPAVAEEPLHRP

## Human NP_002102

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPPQLPQ PPPQAQPLLPQPQPPPPPPPPPPGPAVAEEPLHRP

etc.

## Conclusions

1. A total of 57 HTT DNA constructs had amino- or carboxy-terminus tags
a. Amino-tags were a mix of GST, FLAG, SUMO and HIS
b. All carboxy-terminus tags were FLAGs
2. All HTT protein sequences had polyQ tracts ranging from Q10 - Q73
3. Majority of HTT protein sequences did not reference Q0, Q23 or Q25 (66.4\%)
a. $33.6 \%$ of HTT protein sequences referenced $\mathrm{Q} 0, \mathrm{Q} 23$ or Q 25
b. $\quad 18.8 \%$ of HTT protein sequences referenced Q23
c. Overall, no consistency of reference a polyQ length
4. What naming convention do you think should be used? Fill out our survey \& submit your suggestion! https://www.surveymonkey.com/r/ZB8ZJ27

2002;2(6):831-837.
4. Metzier $M$, et al. Phosphorylation of huntingtin at Ser421 in YAC128 neurons is associated with protection of YAC128 neurons from NMDA-mediated excitotoxicity and is modulated by PP1 and PP2A. J Neurosci. 2010:30(43): 14318-14329

