# Expression analysis of Huntington disease mouse models reveals robust striatum disease signatures

John C. Obenauer[1]*, Jian Chen[2], Viktoria Andreeva[1], Jeffrey S. Aaronson[2], Ramee Lee[2], Andrea Caricasole[3,4], Jim Rosinski[2]*

[1] Rancho BioSciences, San Diego, CA

[2] CHDI Foundation, Princeton, NJ, USA

[3] IRBM, Rome, Italy

[4] Current Address: Kedrion Biopharma, Barga, Italy

* Corresponding authors

E-mail: john.obenauer@ranchobiosciences.com (JCO)

E-mail: jim.rosinski@chdifoundation.org (JR)

Short title: Analysis of HD mice reveals striatum disease signatures

**Abstract**

Huntington's disease is caused by expanded trinucleotide repeats in the huntingtin gene (HTT), and a higher number of repeats is associated with an earlier age of disease onset. Although the causative gene has been identified, its connections to the observed disease phenotypes is still unclear. Mouse models engineered to contain increasing numbers of trinucleotide repeats were sacrificed at different ages to detect RNA-level and protein-level changes specific to each repeat length and age in order to examine the transcriptional and translational characteristics of the disease. RNA-seq and quantitative proteomics data were collected on 14 types of tissues at up to 8 repeat lengths and in up to 3 different ages, and differential gene and protein expression were examined. A modified method of imputing missing proteomics data was employed and is described here. The most dysregulated tissue at both the RNA and protein levels was the striatum, and a strong gender effect was observed in all of the liver experiments. The full differential expression results are available to the research community on the HDinHD.org website. The results of multiple expression tests in the striatum were combined to generate an RNA disease signature and a protein disease signature, and the signatures were validated in external data sets. These signatures represent molecular readouts of disease progression in HD mice, which further characterizes their HD-related phenotype and can be useful in the preclinical evaluation of candidate therapeutic interventions.

**Author Summary**

Mouse models of Huntington's disease were engineered to allow a detailed examination of how the disease causes changes in gene activity in a variety of tissues. Among the 14 tissues

studied, the one most affected by the disease in our experiments was the striatum, a brain region involved in voluntary movement. The liver results showed large differences in gene activity between the male and female mice. In our analysis, we propose a minor change in how proteomics data is typically analyzed in order to improve the ranking of significant results. Using the striatum data in this study and in others, we identified robust genetic signatures of disease at both the RNA and protein levels.

47 **Introduction**

48 Expanded repeats of the DNA trinucleotide CAG in the huntingtin (HTT) gene cause
49 Huntington's disease (HD), a fatal neurological disorder characterized by progressive motor,
50 cognitive, and behavioral impairments (1). The CAG repeats in the DNA are translated to a
51 polyglutamine (polyQ) region in the HTT protein, but the mechanisms connecting the polyQ
52 mutation to the disease phenotypes are still being explored. Marcy MacDonald, Vanessa
53 Wheeler, Scott Zeitlin, and their respective collaborators engineered mice with a human exon 1
54 of HTT knocked into the endogenous locus. From that, they engineered a series of HTT alleles
55 having different CAG repeat lengths, which we will refer to as the mouse allelic series (2). The
56 range includes wild type (WT) mice containing the natural Q7 repeat length, as well as a Q20
57 knock-in that has been characterized as being behaviorally (2) and transcriptionally (3) similar
58 to WT. The disease alleles Q50, Q80, Q92, Q111, Q140, and Q175 were compared to the Q20
59 knock-in as a control whenever possible, in order to remove possible effects of the knock-in
60 construct. In the experiments lacking Q20 samples, the WT mice were used as controls. All the
61 mice are heterozygous for the knock-in allele. The mice were sacrificed at 2, 6, or 10 months of
62 age, and tissues extracted from the mice were studied using RNA-seq and liquid
63 chromatography tandem mass spectrometry (LC/MS/MS) label-free proteomics.

64

65 Table 1 summarizes all of the RNA-seq and proteomics data sets that were analyzed and lists
66 their GEO, SRA, and PRIDE accession numbers. Most of the RNA-seq experiments used a broad
67 range of HTT alleles (WT, Q20, Q80, Q92, Q111, Q140, and Q175) and a range of ages (2, 6, and
68 10 months), and these are called the "Full Series" in Table 1 and in the text. The ones labeled

69 “Miniseries” had fewer alleles, namely WT, Q20, Q50, Q92, and Q140, and only two ages, 6 and

70 10 months. The “Tissue Survey” experiment (GSE65775) had only Q175 and WT mice aged 6

71 months. The wild type Q length was Q7. All of the proteomics experiments used the alleles in

72 the RNA-seq “Full Series” plus Q50. Some of the proteomics experiments lacked a Q50 sample

73 at the 2-month age, as noted in Table 1.

74

75 Table 1. All of the mouse allelic series experiments analyzed in this study.

| Experiment ID | Molecule | Tissue | Description | Exceptions |
|---|---|---|---|---|
| GSE65770 (PRJNA274989) | mRNA | Cortex | Full Series | No Q50 |
| GSE65772 (PRJNA274987) | mRNA | Liver | Full Series | No Q50 |
| GSE65774 (PRJNA274985) | mRNA | Striatum | Full Series | No Q50, Extra housed-with samples |
| GSE65775 (PRJNA274984) | mRNA | Tissue Survey | 11 tissues | Only 6M, only Q175 and WT |
| GSE73468 (PRJNA297063) | mRNA | Cerebellum | Full Series | No Q50 |
| GSE73503 (PRJNA297190) | mRNA | Hippocampus | Full Series | No Q50 |
| GSE76752 (PRJNA308528) | mRNA | White Adipose Near Gonad | Full Series | Only 6M, no Q50 |

| Experiment ID | Molecule | Tissue | Description | Exceptions |
|---|---|---|---|---|
| GSE78270 (PRJNA313072) | mRNA | Cerebellum | Miniseries | No 2M, no Q80/Q111 /Q175, Extra housed-with samples |
| GSE78272 (PRJNA313070) | mRNA | Cortex | Miniseries | No 2M, no Q80/Q111 /Q175, Extra housed-with samples |
| GSE78273 (PRJNA313076) | mRNA | Liver | Miniseries | No 2M, no Q80/Q111 /Q175, Extra housed-with samples |
| GSE78274 (PRJNA313075) | mRNA | Striatum | Miniseries | No 2M, no Q80/Q111 /Q175, Extra housed-with samples |
| PXD005485 | Protein | Cortex | Full Series with Q50 | No Q50 at 2M |
| PXD005526 | Protein | Cerebellum | Full Series with Q50 | No Q50 at 2M |
| PXD005538 | Protein | Hippocampus | Full Series with Q50 | No Q50 at 2M |
| PXD005641 | Protein | Liver | Full Series with Q50 | |
| PXD006302 | Protein | Striatum | Full Series with Q50 | No Q50 at 2M |
| PXD010958 | Protein | Heart | Full Series with Q50 | |

| Experiment ID | Molecule | Tissue | Description | Exceptions |
|---|---|---|---|---|
| PXD010957 | Protein | Muscle | Full Series with Q50 | No Q140 at 2M |

76

77 Langfelder et al. published an analysis of the striatum, cortex, and liver RNA-seq data and part
78 of the striatum proteomics data (3) and identified correlated expression modules using
79 weighted gene coexpression network analysis (WGCNA). We extended Langfelder's
80 differential expression analysis of the striatum, cortex, and liver RNA to include the cerebellum,
81 hippocampus, and white adipose tissue near gonads, and added full proteomics analyses of the
82 striatum, cortex, cerebellum, hippocampus, liver, heart, and muscle. Our proteomics analysis
83 includes imputation of missing values in selected cases, as is done in the Bioconductor package
84 DEP (4), except that in our case the imputed values were calculated deterministically to make
85 the ranking of significant results more consistent. A striatal RNA disease signature is presented
86 here and contains genes overlapping two of the WGCNA modules in Langfelder's study. A
87 striatal protein disease signature is also presented here. These RNA and protein signatures will
88 be useful in understanding perturbations to transcription and translation in HD and, as
89 molecular biomarkers, can aid in evaluating the potential of candidate therapeutics to revert
90 the HD-related phenotype in mouse models.

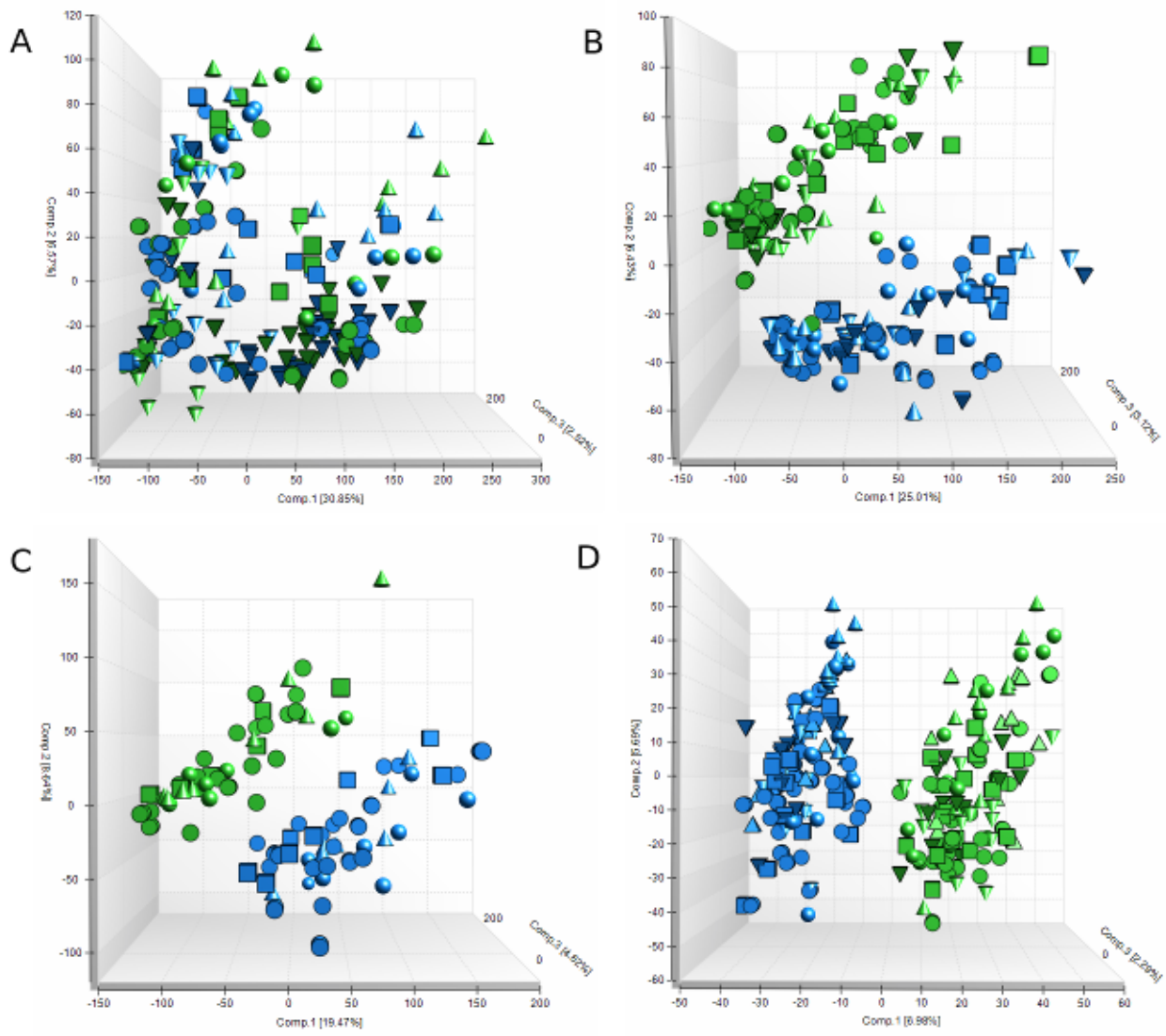
91

92 **Results**

93 *Gender effect in liver*

94 While examining PCA plots during outlier detection, a striking difference was seen between the
95 male and female samples in all the liver experiments, including the RNA Full Series, RNA

96 Miniseries, and protein Full Series. This gender effect was not seen in any other tissue. Figure
97 1 shows several PCA plots with female samples in blue and male samples in green, and with the
98 different Q lengths indicated by different shapes. Figure 1A shows the striatum RNA Full Series,
99 with the green and blue genders intermingled, and this intermingling was typical of the other
100 tissues and experiments. Figures 1B, 1C, and 1D show the liver RNA Full Series, RNA Miniseries,
101 and protein Full Series respectively, and blue female samples are clearly separated from the
102 green male samples. Gender differences in liver gene expression have been previously
103 reported in mice (5) and the differences have been observed to begin at puberty, around days
104 30 to 35 of age. Interestingly, gender differences have been observed in human HD. While no
105 gender-specific differences in disease burden or age of onset are observed, the progression rate
106 in women appears to be faster (reviewed in (6)). Since the high-Q and wild type samples are
107 within each gender cluster, this means the gender effect in liver is larger than the disease
108 effect. Because of this large difference, all the liver data differential expression was analyzed
109 three ways: combined liver (LIV), female liver (LVF), and male liver (LVM).
110
111 Figure 1. PCA plots showing female samples in blue and male samples in green. A, striatum
112 RNA Full Series. B, liver RNA Full Series. C, liver RNA Miniseries. D, liver protein Full Series.




113

114 *Transcriptional and translational effects in each tissue*

115 The significant differential expression results from all comparisons have been combined into

116 Supplementary Table 1. To determine which tissues are more affected by HD, the significant

117 genes from all Q-length comparisons are summarized in Table 2 for tissues with RNA Full Series

118 or protein Full Series experiments and for tissues in the RNA tissue survey experiment. Figure 2

119 separates the significant genes by age, Q length, and positive or negative directions of change

120 for the RNA Full Series, and Figure 3 does the same for the protein Full Series.

121 Table 2. Counts of significant genes by tissue in all comparisons. Percentages refer to the total

122 counts in each column.

| RNA Full Series | | Protein Full Series | | RNA Tissue Survey | |
|---|---|---|---|---|---|
| Striatum | 23,165 (42%) | Striatum | 3,032 (25%) | White Adipose around Gonad | 3,102 (27%) |
| Liver Female | 9,603 (18%) | Liver Combined | 2,757 (22%) | Cerebellum | 2,565 (23%) |
| Cerebellum | 6,239 (11%) | Gastrocnemis Muscle | 1,838 (15%) | Skin | 1,986 (18%) |
| Hippocampus | 6,051 (11%) | Liver Female | 1,556 (13%) | Brown Adipose | 1,400 (12%) |
| Cortex | 4,020 (7%) | Cortex | 1,300 (11%) | Hippocampus | 1,122 (10%) |
| Liver Combined | 3,334 (6%) | Liver Male | 759 (6%) | Brainstem | 642 (6%) |
| Liver Male | 2,438 (4%) | Heart | 648 (5%) | Thalamus/ Hypothalamus | 440 (4%) |
| Total | 54,850 (100%) | Cerebellum | 328 (3%) | Heart | 49 (0%) |

| RNA Full Series | Protein Full Series | | RNA Tissue Survey | |
|---|---|---|---|---|
| | Hippocampus | 81 (1%) | Corpus Callosum | 14 (0%) |
| | Total | 12,299 (100%) | Gastrocnemius Muscle | 8 (0%) |
| | | | White Adipose near Intestine | 1 (0%) |
| | | | Total | 11,329 (100%) |

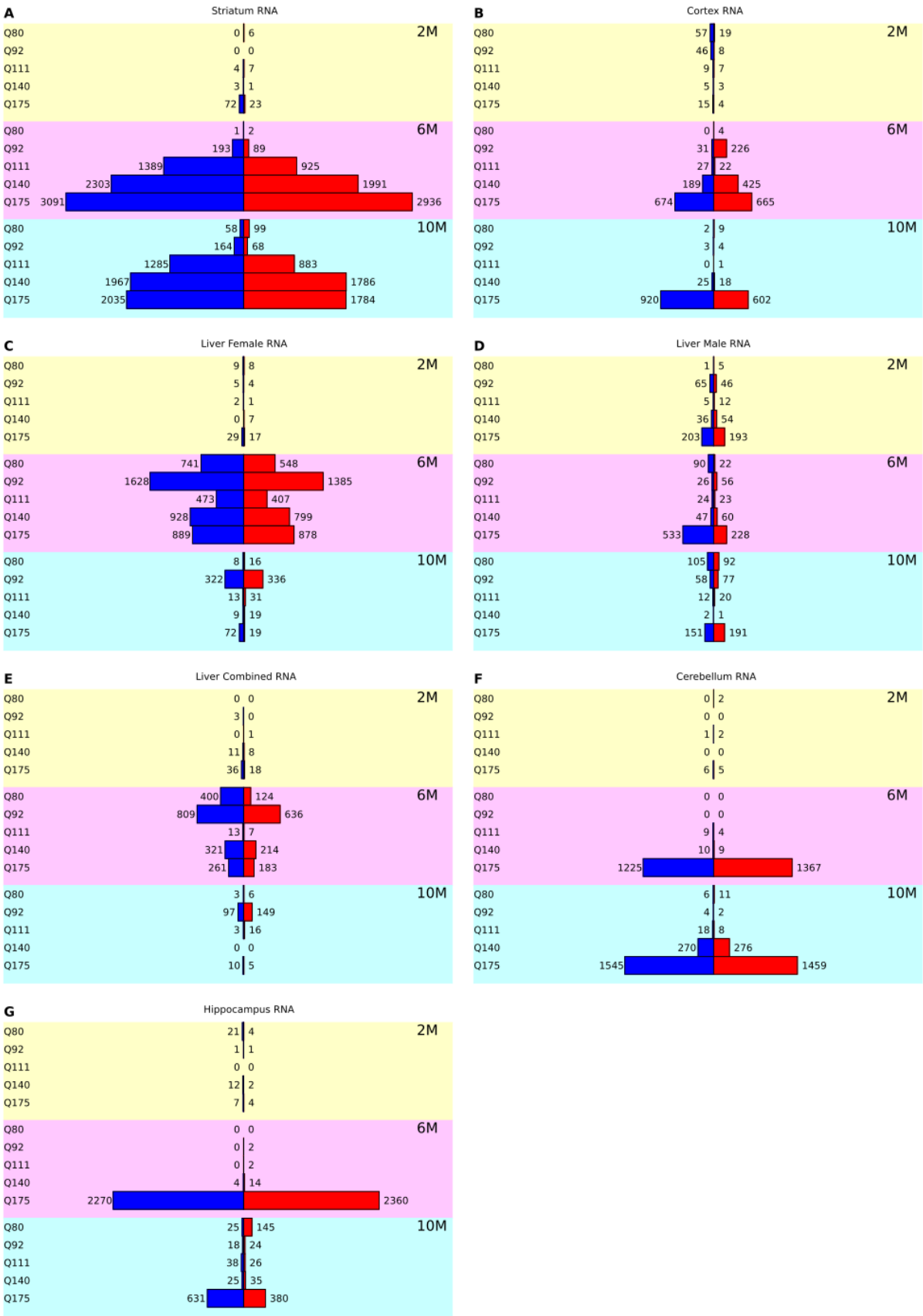
123

124 Figure 2. Counts of significant genes for the RNA Full Series tissues, separated by direction of

125 change. The tissues are (A) striatum, (b) cortex, (C) female liver, (D) male liver, (E) combined

126 liver, (F) cerebellum, and (G) hippocampus. All plots in this figure are drawn at the same scale

127 to allow comparisons between the tissues.

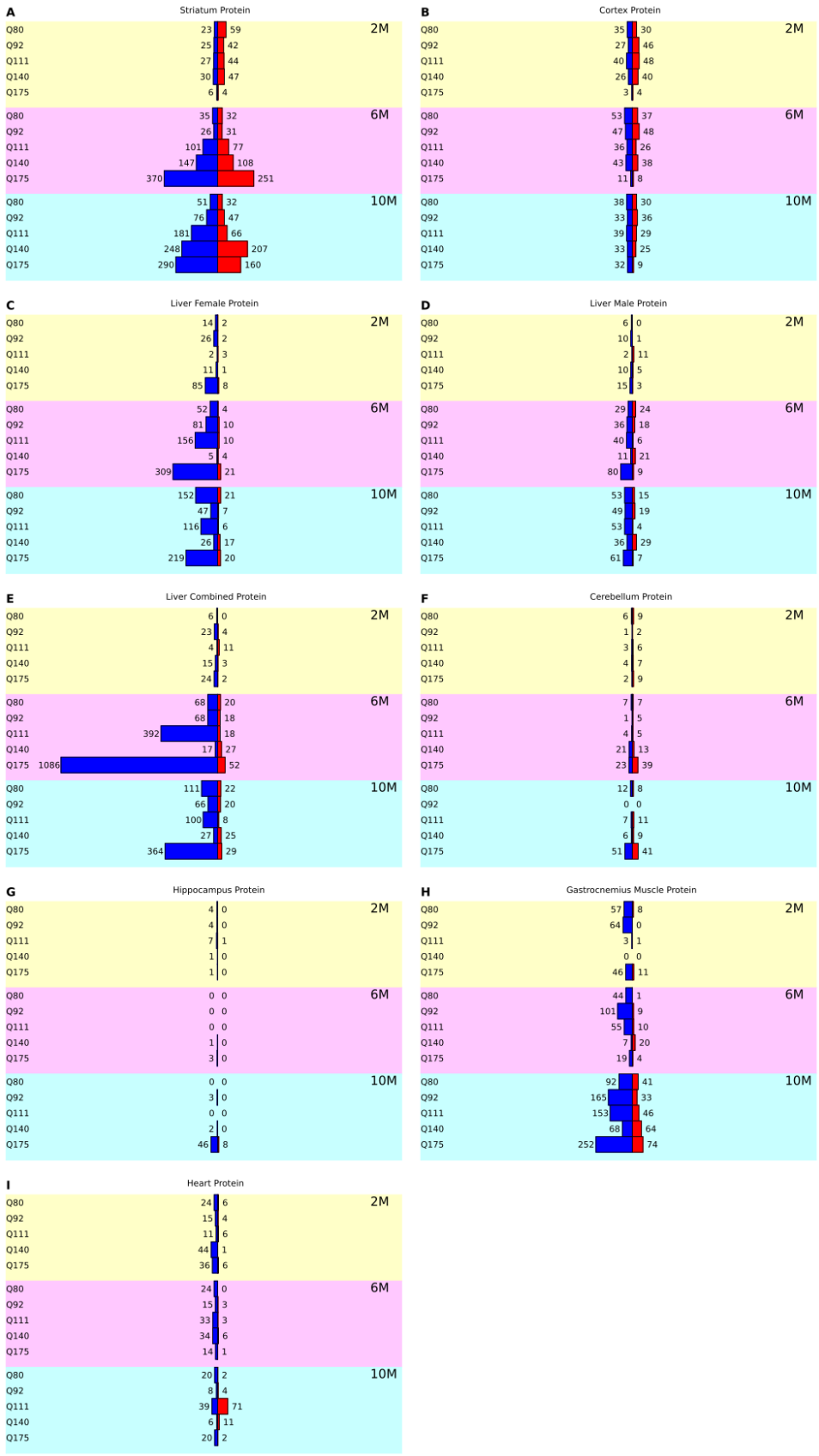

A
Striatum RNA
2M
Q80 0 6
Q92 0 0
Q111 4 7
Q140 3 1
Q175 72 23
6M
Q80 1 2
Q92 193 89
Q111 1389 925
Q140 2303 1991
Q175 3091 2936
10M
Q80 58 99
Q92 164 68
Q111 1285 883
Q140 1967 1786
Q175 2035 1784
B
Cortex RNA
2M
Q80 57 19
Q92 46 8
Q111 9 7
Q140 5 3
Q175 15 4
6M
Q80 0 4
Q92 31 226
Q111 27 22
Q140 189 425
Q175 674 665
10M
Q80 2 9
Q92 3 4
Q111 0 1
Q140 25 18
Q175 920 602
C
Liver Female RNA
2M
Q80 9 8
Q92 5 4
Q111 2 1
Q140 0 7
Q175 29 17
6M
Q80 741 548
Q92 1628 1385
Q111 473 407
Q140 928 799
Q175 889 878
10M
Q80 8 16
Q92 322 336
Q111 13 31
Q140 9 19
Q175 72 19
D
Liver Male RNA
2M
Q80 1 5
Q92 65 46
Q111 5 12
Q140 36 54
Q175 203 193
6M
Q80 90 22
Q92 26 56
Q111 24 23
Q140 47 60
Q175 533 228
10M
Q80 105 92
Q92 58 77
Q111 12 20
Q140 2 1
Q175 151 191
E
Liver Combined RNA
2M
Q80 0 0
Q92 3 0
Q111 0 1
Q140 11 8
Q175 36 18
6M
Q80 400 124
Q92 809 636
Q111 13 7
Q140 321 214
Q175 261 183
10M
Q80 3 6
Q92 97 149
Q111 3 16
Q140 0 0
Q175 10 5
F
Cerebellum RNA
2M
Q80 0 2
Q92 0 0
Q111 1 2
Q140 0 0
Q175 6 5
6M
Q80 0 0
Q92 0 0
Q111 9 4
Q140 10 9
Q175 1225 1367
10M
Q80 6 11
Q92 4 2
Q111 18 8
Q140 270 276
Q175 1545 1459
G
Hippocampus RNA
2M
Q80 21 4
Q92 1 1
Q111 0 0
Q140 12 2
Q175 7 4
6M
Q80 0 0
Q92 0 2
Q111 0 2
Q140 4 14
Q175 2270 2360
10M
Q80 25 145
Q92 18 24
Q111 38 26
Q140 25 35
Q175 631 380

128

129 Figure 3. Counts of significant genes for the protein Full Series tissues: (A) striatum, (b) cortex,

130 (C) female liver, (D) male liver, (E) combined liver, (F) cerebellum, (G) hippocampus, (H)

131 gastrocnemius muscle, and (I) heart.


A
Striatum Protein
2M
Q80 23 59
Q92 25 42
Q111 27 44
Q140 30 47
Q175 6 4
6M
Q80 35 32
Q92 26 31
Q111 101 77
Q140 147 108
Q175 370 251
10M
Q80 51 32
Q92 76 47
Q111 181 66
Q140 248 207
Q175 290 160
B
Cortex Protein
2M
Q80 35 30
Q92 27 46
Q111 40 48
Q140 26 40
Q175 3 4
6M
Q80 53 37
Q92 47 48
Q111 36 26
Q140 43 38
Q175 11 8
10M
Q80 38 30
Q92 33 36
Q111 39 29
Q140 33 25
Q175 32 9
C
Liver Female Protein
2M
Q80 14 2
Q92 26 2
Q111 2 3
Q140 11 1
Q175 85 8
6M
Q80 52 4
Q92 81 10
Q111 156 10
Q140 5 4
Q175 309 21
10M
Q80 152 21
Q92 47 7
Q111 116 6
Q140 26 17
Q175 219 20
D
Liver Male Protein
2M
Q80 6 0
Q92 10 1
Q111 2 11
Q140 10 5
Q175 15 3
6M
Q80 29 24
Q92 36 18
Q111 40 6
Q140 11 21
Q175 80 9
10M
Q80 53 15
Q92 49 19
Q111 53 4
Q140 36 29
Q175 61 7
E
Liver Combined Protein
2M
Q80 6 0
Q92 23 4
Q111 4 11
Q140 15 3
Q175 24 2
6M
Q80 68 20
Q92 68 18
Q111 392 18
Q140 17 27
Q175 1086 52
10M
Q80 111 22
Q92 66 20
Q111 100 8
Q140 27 25
Q175 364 29
F
Cerebellum Protein
2M
Q80 6 9
Q92 1 2
Q111 3 6
Q140 4 7
Q175 2 9
6M
Q80 7 7
Q92 1 5
Q111 4 5
Q140 21 13
Q175 23 39
10M
Q80 12 8
Q92 0 0
Q111 7 11
Q140 6 9
Q175 51 41
G
Hippocampus Protein
2M
Q80 4 0
Q92 4 0
Q111 7 1
Q140 1 0
Q175 1 0
6M
Q80 0 0
Q92 0 0
Q111 0 0
Q140 1 0
Q175 3 0
10M
Q80 0 0
Q92 3 0
Q111 0 0
Q140 2 0
Q175 46 8
H
Gastrocnemius Muscle Protein
2M
Q80 57 8
Q92 64 0
Q111 3 1
Q140 0 0
Q175 46 11
6M
Q80 44 1
Q92 101 9
Q111 55 10
Q140 7 20
Q175 19 4
10M
Q80 92 41
Q92 165 33
Q111 153 46
Q140 68 64
Q175 252 74
I
Heart Protein
2M
Q80 24 6
Q92 15 4
Q111 11 6
Q140 44 1
Q175 36 6
6M
Q80 24 0
Q92 15 3
Q111 33 3
Q140 34 6
Q175 14 1
10M
Q80 20 2
Q92 8 4
Q111 39 71
Q140 6 11
Q175 20 2

132

133 The striatum is the most affected tissue, accounting for 42% of all differentially expressed genes
134 in the RNA Full Series and 25% of all significant genes in the protein Full Series. The numbers of
135 significant genes increase with Q length at both 6 months and 10 months, and at both the RNA
136 and protein levels. This trend is not observed for any other tissue. At the 2-month age, the
137 striatum shows more changes at the protein level than at the RNA level, and this is also true for
138 the cortex. The cerebellum and hippocampus are more affected (both 11%) than the cortex
139 (7%) at the RNA level, but less than the cortex at the protein level (cortex 11%, cerebellum 3%,
140 hippocampus 1%). The cerebellum shows similar numbers of changes in the Full Series RNA
141 (Q175 vs. Q20 at 6 months, 2,592 changes) and in the tissue survey (Q175 vs. WT at 6 months,
142 2,565 changes), but the hippocampus shows more changes in the Full Series (4,630) than in the
143 tissue survey (1,122). The cerebellum and hippocampus both show high counts at the Q175 Q
144 length in the RNA at 6 months and 10 months, but few counts at any other Q length and age.
145 These high Q175 counts for the cerebellum and hippocampus RNA are not seen in the
146 proteomics data.
147
148 The female liver is the second most affected tissue in the RNA Full Series experiments (18%),
149 while the male liver is the least affected (4%). The high counts in the female liver are primarily
150 restricted to the 6-month age, while the male liver has similar gene counts at all ages. At the
151 protein level, the female liver is again more affected than the male (13% vs. 6%), but the
152 combined liver has even more changes (22%), making it second only to the striatum among the
153 proteomics tissues. A likely explanation for the higher counts in the combined liver analysis is
154 that it has twice as many samples, increasing the statistical significance for some genes that

155 would not be significant with fewer samples. The most extreme result in the combined liver is
156 the Q175 vs. Q20 comparison at 6 months, with 1,138 significant genes, and this is the largest
157 count of significant changes among all the proteomics experiments. 1,084 of these 1,138
158 changes are negative. All of the proteomics results for the liver samples are dominated by
159 negative genes (down-regulated in the disease condition), while changes in both directions are
160 seen in the RNA.
161
162 The most affected tissue in the RNA tissue survey is white adipose around gonad, with 3,102
163 significant changes. This same tissue was investigated separately in experiment GSE76752 at a
164 range of Q lengths, and the Q175 vs. Q20 comparison still shows thousands of changes (6,802),
165 but all other Q lengths show much smaller changes (13 at Q140, 2 at Q111, 141 at Q92, and 22
166 at Q80). In contrast, the white adipose sample near intestine is the least affected tissue, with
167 only one significant gene change. Two types of muscle tissue were included in both the RNA
168 tissue survey and the proteomics Full Series, namely gastrocnemius (calf) muscle and heart.
169 The gastrocnemius muscle was more affected than the heart in the proteomics data (15% vs.
170 5%). The least affected brain tissue in the survey was the corpus callosum, with only 14
171 differentially expressed genes.
172
173 *Overlap between RNA and protein results*
174 The RNA and protein results for all the Full Series Q-length comparisons were examined to find
175 genes that were significant at both the RNA and protein levels. There were 1,503 comparisons
176 that showed the same gene significant in the RNA-seq and proteomics data. These are divided

177 among all five tissues that have both RNA and protein Full Series data: the striatum (1,221),
178 cortex (43), cerebellum (72), hippocampus (29), and liver (73 female, 11 male, 54 combined).
179 90% (1,357) of the RNA and protein overlapping cases change in the same direction. 792 genes
180 are represented among these 1,503 overlapping comparisons. The table of matching RNA and
181 protein comparisons is Supplementary Table 2.
182
183 *Disease signatures*
184 The differential expression results for the striatum in Figures 2 and 3 show large transcriptional
185 and translational signatures at multiple Q lengths and ages. These multiple signatures suggest
186 that there might be a subset of overlapping genes that could be a reproducible indicator of
187 disease in the striatum. The other tissues don't show this same consistent trend. For example,
188 the cortex RNA shows changes in the Q92 and Q140 Q lengths at 6 months that seem to go
189 away at 10 months, and the cerebellum and hippocampus changes are mostly restricted to the
190 Q175 mice. The female liver RNA shows large transcriptional changes at 6 months that go away
191 at 10 months. Although the gender differences we observed in the liver are consistent with
192 other work (5), the large transcriptional signature at 6 months is not, and we authors are
193 concerned with the reproducibility of this 6-month liver data set, even though the data quality
194 itself is high. Based on the consistent changes seen in the striatum, we will focus on this tissue
195 to determine reproducible RNA and protein disease signatures.
196
197 The overlapping significance method described in Materials and Methods was used to
198 determine the RNA signature. Five comparisons from the allelic series striatum experiments

199 were combined with five other striatum data sets and were grouped by Q length and age: Q175
200 10 months (10M), Q140 10M, Q140 6M, and R6/2 3M. 37,633 genes that were defined in all 10
201 experiments were compared within these four groups and across all groups. The same
202 significance criteria were used for all 10 data sets: genes must have an adjusted p-value less
203 than 0.05 and a fold change of at least 20% in either direction. The counts of significant genes
204 in each experiment, genes shared within each group, and genes common to all four groups are
205 shown in Table 3.
206
207 Table 3. Counts of significant genes overlapping multiple striatum HD experiments.

<table>
<tr><th>Experiment</th><th>Significant Genes</th><th>Group Overlap</th><th>All Overlap</th></tr>
<tr><td>Full Series Q175 10M</td><td>1,725</td><td rowspan="3">1,088</td><td rowspan="9">287</td></tr>
<tr><td>Cohort1Time1 Q175 10M</td><td>3,056</td></tr>
<tr><td>Cohort1Time2 Q175 10M</td><td>2,379</td></tr>
<tr><td>Full Series Q140 10M</td><td>1,593</td><td rowspan="2">658</td></tr>
<tr><td>Miniseries Q140 10M</td><td>1,994</td></tr>
<tr><td>Full Series Q140 6M</td><td>1,625</td><td rowspan="3">439</td></tr>
<tr><td>Miniseries Q140 6M</td><td>984</td></tr>
<tr><td>Cohort2Time1 Q140 6M</td><td>2,563</td></tr>
<tr><td>HDAC R6/2 3M</td><td>6,089</td><td rowspan="2">4,551</td></tr>
<tr><td>KMO R6/2 3M</td><td>5,963</td></tr>
</table>

208

209 There were 287 genes that were significant in all 10 experiments and always changed in the
210 same direction. These include 21 predicted or uncharacterized genes, based on their naming
211 conventions: 11 gene models (like Gm10406), 8 Riken sequences (like A830036E02Rik), and 2
212 GenBank accessions (AW495222 and BC049352). These were removed from the list, leaving
213 266 well-characterized genes. We will refer to this 266-gene list of striatum RNA disease genes
214 as Str266R (for RNA). 71 (27%) of these genes increase expression in HD, while the other 195
215 (73%) decrease expression. The Str266R gene names, Ensembl IDs, and log2 fold changes are
216 shown in Supplementary Table 3.

217

218 To validate the Str266R signature, HD and WT mice from 4 additional experiments were used:
219 Cohort2Time2, Cohort3Time1, Cohort3Time2, and Cohort3Time3 (these data sets are described
220 in Materials and Methods). The fold changes and adjusted p-values for the Str266R genes in
221 these four experiments are included in Supplementary Table 3. 262 of the genes meet the
222 significance requirements in Cohort2Time2, where the 4 exceptions are Oscar, Pipox, Malat1,
223 and Rnf207. In the Cohort3Time1 data set, 262 genes again meet the significance
224 requirements, but the 4 exceptions are different: Tnip3, Psme1, Dpy19l3, and Cttnbp2. All 266
225 genes meet the significance criteria in both the Cohort3Time2 and Cohort3Time3 validation
226 sets. These validation results indicate Str266R is a robust and reproducible disease signature
227 for several HD model genotypes and ages.

228

229 The Str266R signature overlaps two of the striatum WGCNA modules published by Langfelder
230 (3). 180 (68%) of the Str266R genes are in module M2, and 54 others (20%) are in module M20.
231 In that study, modules M2 and M20 are the top two most CAG length-dependent modules,
232 where M2 correlated negatively with CAG length and M20 correlated positively. M2 had the
233 largest number of dysregulated genes, and M2's hub genes include the striatal medium spiny
234 neuron marker genes Ppp1r1b, Drd1, Drd2, and Gpr6, which are also in Str266R. M20 had the
235 strongest positive correlation with CAG length and was enriched for p53 signaling, cell division,
236 and protocadherin genes. The Str266R genes were tested for GO term enrichment, and the
237 most significant biological processes were responses to alkaloid and amphetamine, regulation
238 of glutamatergic synaptic transmission and membrane potential, synapse assembly, and protein
239 dephosphorylation. The full list of enriched GO terms is in Supplementary Table 4.
240
241 For cases where a smaller disease signature is desirable, the top 10 increasing and top 10
242 decreasing genes in Str266R were selected, except that the 8 genes that were not significant in
243 one of the four validation data sets were disqualified from this smaller signature. This small
244 bidirectional signature is called Str20R and is shown in Table 4.
245
246 Table 4. The Str20R signature, a subset of the larger Str266R signature.

| Gene | Direction | MinLog2FC | Gene | Direction | MinLog2FC |
|---|---|---|---|---|---|
| Wt1 | + | 3.8 | Sohlh1 | - | -2.8 |

| Gene | Direction | MinLog2FC | Gene | Direction | MinLog2FC |
|---|---|---|---|---|---|
| Onecut1 | + | 3.0 | Ifi27l2b | - | -2.7 |
| Sfmbt2 | + | 1.3 | Gpx6 | - | -2.1 |
| Rgs13 | + | 1.1 | Myo7b | - | -1.5 |
| Tnfrsf13c | + | 0.9 | Tmprss6 | - | -1.5 |
| Crnde | + | 0.9 | Wnt8b | - | -1.5 |
| Ifnlr1 | + | 0.9 | Ddit4l | - | -1.5 |
| Fgfr4 | + | 0.8 | Slc4a11 | - | -1.4 |
| Acy3 | + | 0.8 | Mafa | - | -1.4 |
| Smim24 | + | 0.8 | Odf4 | - | -1.3 |

247

248 The overlapping significance method was also used to find a cortex RNA signature. Ten cortex
249 experiments that had HD samples (either Q175, Q140, or R6/2) and WT samples at several ages
250 were tested for differential expression, assembled into similar groups (Q175, Q140 10M, Q140
251 6M, and R6/2), and the overlaps in their significant genes were calculated. The genes
252 significant in each experiment are listed in Supplementary Table 5 and the counts of
253 overlapping genes are summarized in Table 5. For the R6/2 mice, there is a strong cortex
254 signature that is reproducible at ages 6 weeks and 3 months, with 2,352 overlapping genes. For

255 the Q175 mice, there are 110 genes that overlap at ages 6 and 10 months. The Q140 overlaps
256 are very low at both 6 months and 10 months. In the two Q140 10M studies, only 9 genes
257 overlap: Il33, Slc45a3, Chdh, Gpr153, Enpp6, Gm5067, Apod, Flnc, and 5031410I06Rik. In the
258 three Q140 6M studies, only 2 genes overlap, Scn4b and Gm5067. Unlike the striatum analysis,
259 no genes pass significance tests in the cortex in all 10 experiments. It was not possible to
260 determine a consistent cortex disease signature using this method and these data sets. We
261 speculate that the active disease genes differ by age and by region within the cortex during the
262 progression of HD.

263 Table 5. Attempted cortex RNA disease signature using overlapping significance method.

| Experiment | Significant Genes | Group Overlap | All Overlap |
|---|---|---|---|
| Full Series Q175 10M | 631 | 110 | 0 |
| Full Series Q175 6M | 275 | | |
| Full Series Q140 10M | 22 | 9 | |
| Miniseries Q140 10M | 239 | | |
| Full Series Q140 6M | 221 | 2 | |
| Miniseries Q140 6M | 333 | | |
| Cohort2Time1 Q140 6M | 1,074 | | |
| HDAC R6/2 3M | 5,351 | 2,352 | |
| Cohort4 R6/2 6W | 3,981 | | |
| KMO R6/2 3M | 5,351 | | |

264

265 We next sought to determine a striatum protein disease signature but had access to only four

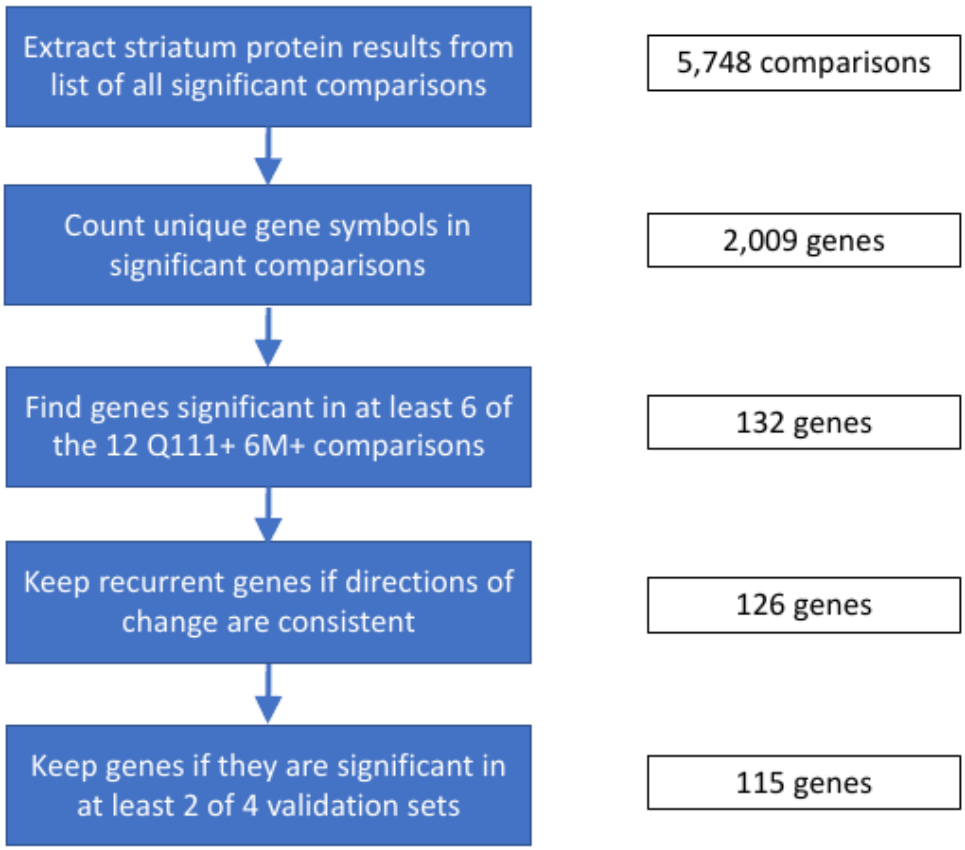
266 data sets outside of the allelic series (compared to nine for the RNA signature), so we used the

267 recurrence ranking method described in Materials and Methods and illustrated in Figure 4.

268

269 Figure 4. Recurrence ranking method used to determine the striatum protein signature,

270 Str115P.




271

272 12 of the allelic series striatum comparisons were selected to represent the disease signature

273 based on the differential expression seen in Figures 2 and 3: the 6 Q-length comparisons using

274 Q lengths of 111 or higher and ages 6 months or older, called Q111+ 6M+ (Q111 6M, Q111

275 10M, Q140 6M, Q140 10M, Q175 6M, and Q175 10M), and the 6 age comparisons that used

276 the same Q111+ 6M+ experimental groups but were compared to the Q-length-matched 2-
277 month controls. Starting with the 5,748 significant striatum protein comparisons in
278 Supplementary Table 1, we found that they were represented by 2,009 unique gene symbols.
279 The recurrence of significance among the Q111+ 6M+ comparisons was calculated for each
280 gene symbol, and UniProt IDs assigned to the same gene symbol were combined when testing
281 recurrence. 132 gene symbols were significant in 6 or more of the 12 target comparisons. 7 of
282 the 132 gene symbols were rejected because they changed in different directions in some
283 comparisons. The remaining 126 proteins were tested against the 4 proteomics validation data
284 sets. 115 proteins were significant in at least 2 of the 4 experiments, using adjusted p-values
285 less than 0.1 as the criterion (and no fold change threshold). We named these 115 validated
286 proteins the Str115P (for protein) signature. 17 of these proteins increase in disease while the
287 other 98 decrease. The Str115P signature and the steps used to determine it are documented
288 in Supplementary Table 6. These proteins were tested for GO term enrichment, and the most
289 significant biological processes were the regulation of synaptic plasticity, cognition, locomotory
290 behavior, learning or memory, visual behavior, and second-messenger-mediated signaling. The
291 full GO term enrichment results are included in Supplementary Table 4.
292
293 As we did with the RNA signature, we extracted a smaller protein signature by taking the top 10
294 increasing proteins and top 10 decreasing proteins from the Str115P and excluded proteins that
295 were not significant in all 4 validation sets. We call this smaller signature Str20P, shown in
296 Table 6. Note that the fold changes in Table 6 look smaller than the Str20R values in Table 4,
297 but the proteomics calculations are done in log10 intensities instead of log2.

298

299 Table 6. Str20P striatum disease signature.

| Gene | Direction | MinLog10FC | Gene | Direction | MinLog10FC |
|---|---|---|---|---|---|
| Chdh | + | 1.7 | Scn4b | - | -1.6 |
| Acy3 | + | 1.7 | Tcf20 | - | -1.6 |
| Macrod1 | + | 1.3 | Rasd2 | - | -1.5 |
| Armcx2 | + | 1.3 | Pde10a | - | -1.5 |
| Nagk | + | 1.3 | Rgs9 | - | -1.4 |
| Dus3l | + | 1.2 | Arc | - | -1.4 |
| Gprasp1 | + | 1.2 | Foxp1 | - | -1.4 |
| Lmnb2 | + | 1.2 | Sema7a | - | -1.4 |
| Psme1 | + | 1.2 | Rgs14 | - | -1.4 |
| Mri1 | + | 0.1 | Arpp19 | - | -1.4 |

300

301 The Str266R and Str115P signatures have 40 gene symbols in common, and all 40 change with

302 disease in the same direction at the RNA and protein levels. We define these 40 genes as a

303 combined RNA/protein signature, Str40RP. The 40 genes are shown in Table 7, ranked from

304 most positive to most negative RNA log2 fold change, and are included in Supplementary Table
305 6. Only the first four genes (Acy3, Chdh, Nagk, and Psme1) increase in disease; all the rest
306 decrease. This is not surprising because both the Str266R and Str115P signatures are
307 dominated by genes that decrease in disease. The most enriched biological processes among
308 these 40 genes are negative regulation of protein dephosphorylation, cAMP or cGMP metabolic
309 process, response to amphetamine, visual learning, visual behavior, and positive regulation of
310 long-term synaptic potentiation. The full GO term results for Str40RP are included in
311 Supplementary Table 4.
312
313 Table 7. The Str40RP RNA/protein signature, ranked from most positive to most negative RNA
314 log2 fold change.

| | | | | |
|---|---|---|---|---|
| Acy3 | Anks1b | Hpca | Adcy5 | Ppp1r16b |
| Chdh | Ppp1r1b | Itpr1 | Wipf3 | Homer1 |
| Nagk | Atp2b1 | Sh2d5 | Npl | Ppp4r4 |
| Psme1 | Camkk2 | Rgs9 | Drd1 | Pde10a |
| Cap1 | Sec14l1 | Ptpn5 | Pcp4 | Rgs14 |
| Baiap2 | Jcad | Rasgrp2 | Pde1b | Coch |
| Tbc1d8 | Rps6ka4 | Shank3 | Crocc | Arpp19 |

| Acy3 | Anks1b | Hpca | Adcy5 | Ppp1r16b |
|---|---|---|---|---|
| Bcr | Ano3 | Rgs7bp | Osbpl8 | Scn4b |

315

316

317 **Materials and Methods**

318 *RNA-seq and proteomics data sets*

319 All of the mouse experiments used in this study were downloaded from NCBI's Gene Expression
320 Omnibus (GEO) (7) and CHDI's Huntington's Disease in High Definition website (HDinHD) (3).
321 The raw FASTQ files are also available on NCBI's Sequence Read Archive (SRA) (8) and the mass
322 spectrometry data is on EBI's Proteomics Identifications Database (PRIDE) (9). The GEO data
323 sets were provided as read counts per gene, and the HDinHD data sets were provided as log10
324 label-free quantitation (LFQ) intensities. The RNA-seq read counts were produced from the raw
325 FASTQ files using OmicSoft ArrayStudio (10). The proteomics signal intensities were produced
326 from the mass spectrometry data by Evotec SE. The sample information for all experiments
327 was downloaded from HDinHD
328 (http://repository.hdinhd.org/data/allelic_series/Allelic_Series_Decoder_Ring-1.5.xlsx.gz). The
329 sample information file was edited to remove "LFQIntensity" from the sample names for the
330 liver PXD005641 proteomics samples in order to match the sample names in the intensities file.
331 Sample names were standardized throughout all experiments to indicate the Htt allele (a Q
332 length or wild type), age, tissue, gender, and replicate number, such as Q140_6M_STR_F_R1.

333

334 *Outlier detection*

335 Outlier samples were detected by examining sample replicates by principal components
336 analysis (PCA) using OmicSoft Array Studio. For greater sensitivity, replicates of each sample
337 group (having the same experiment, tissue, age, and Q length) were examined independently
338 for outliers. Distances in the first principal component were given more weight due to its
339 higher percentage of total variability. In the case of the skin samples in the RNA tissue survey,
340 two additional samples were removed after seeing their behavior in the first round of
341 differential expression tests. The full list of outlier samples removed is included in
342 Supplementary File 11 in a file called Outlier_Samples_Removed.txt.

343

344 *RNA-seq differential expression*

345 Differential expression was performed using DESeq2 (11) in R (12) on the raw read counts. The
346 independent filtering option in DESeq2 was not used so that fold changes and p-values for all
347 input genes would be calculated. Genes were considered significantly changed if the adjusted
348 p-value after multiple test correction was below 0.05. No fold change threshold was used.
349 Each of the higher Q lengths (Q50, Q80, Q92, Q111, Q140, Q175) was compared to the Q20
350 control length, and these are called the Q-length comparisons in the text. For the tissue survey
351 experiment GSE65775, no Q20 samples were available, so the WT litter mate samples were
352 used as controls. All of the Q-length comparison results are in Supplementary File 8. Within
353 each Q length, the higher ages (6 and 10 months) were compared to the 2-month age, and we
354 call these the age comparisons. These age comparisons will include healthy aging genes in
355 addition to disease genes. The healthy aging genes were identified using the Q20 10-month vs.

356 Q20 2-month and Q20 6-month vs. Q20 2-month comparisons within each tissue. The presence
357 of healthy aging genes in the age comparison results could be a problem for some intended
358 uses of these gene lists, so the age comparisons are provided with the healthy aging genes
359 retained in Supplementary File 9 but with them removed in Supplementary File 10. Ensembl
360 identifiers in the differential expression results were converted to gene symbols using GFF3
361 annotation files from Ensembl (13), GRCm38 release 98.
362
363 *Proteomics differential expression*
364 Differential expression was performed using the limma package (14) in R on the normalized
365 log10 protein intensities. Proteins were considered differentially expressed if the adjusted p-
366 value was less than 0.1, a more permissive threshold than the 0.05 value used for RNA-seq. No
367 fold change threshold was used. UniProt identifiers in the differential expression results were
368 converted to gene symbols using protein annotations from UniProtKB (15) Swiss-Prot and
369 TrEMBL collections.
370
371 Most of the sample groups had 8 replicates. Proteins with fewer than 3 out of 8 measured
372 values in either group being compared were rejected from the analysis as insufficiently
373 reproducible. However, proteins with no measured values in one group and 3 or more
374 measured values in the other were given imputed values to allow a comparison. For example, if
375 a protein were absent in all the Q20 samples but had multiple measured values in the Q175
376 samples, it is biologically interesting but would not be detected by limma. To allow the
377 calculation of a fold change and p-value, a common practice is to impute missing values near

378 the experimental limit of detection, as is done in the Bioconductor (16) package DEP (4).
379 However, DEP uses random selection to impute values, resulting in different rankings of results
380 when the program is re-run. In addition, if 8 replicates all have missing values, DEP imputes
381 values for all 8 replicates, inflating the statistical significance. To avoid these problems, we
382 introduce here a deterministic imputation near the limit of detection. The lowest 2% of all
383 measured values in a proteomics experiment are extracted, and the average and standard
384 deviation of these low values is calculated. The average is used as the limit of detection (LOD),
385 and the standard deviation (SD) is used to calculate static replicate values that will be the same
386 for all imputed proteins. Rather than impute values for all 8 replicates, we imputed values for 3
387 replicates, leaving the others with missing values. The 3 values are LOD – ½ SD, LOD, and LOD +
388 ½ SD. The deterministic calculation means every imputed protein is compared to the same
389 reference values, and the low imputation count (n = 3) avoids raising all the imputed proteins
390 high in significance. This minor change to the common imputation practice gives reproducible
391 results and reasonable significance rankings. Example code for performing this imputation is
392 included in the proteomics analysis R script in Supplementary File S11.
393
394 *Disease signatures, overlapping significance method*
395 A simple way to identify a disease signature when many data sets are available is to examine
396 the genes that overlap each experiment's lists of significant genes and add the requirement
397 that their changes need to be in the same direction in every experiment. This was used to
398 determine the Str266R signature. In the selection of candidate genes from 10 RNA-seq data
399 sets, the significance criteria included both a fold change threshold (20% or higher) and an

400 adjusted p-value threshold (less than 0.05). When testing the candidate genes in the validation
401 data sets, only the adjusted p-value criterion was used.
402
403 *Disease signatures, recurrence ranking method*
404 When fewer data sets are available, a method based on ranking the recurring significance of
405 genes in multiple comparisons was used, and this was used to determine the Str115P signature.
406 Mice with Q lengths of Q111 or higher and ages 6 months or older (Q111+ 6M+) were
407 considered as having the disease phenotype base on the RNA and protein differential
408 expression results. The 6 Q-length comparisons and 6 age comparisons meeting these criteria
409 were used as the 12 comparisons to look for recurrent significance in. An adjusted p-value of
410 less than 0.1, with no fold change threshold, was used as the significance criterion. Genes were
411 ranked by their recurrence in the results. UniProt IDs assigned to the same gene symbol were
412 grouped together. For example, Anks1b is represented by four UniProt IDs in these
413 experiments (Q8BIZ1, Q8BIZ1-2, S4R1Q0, and A0A0R4J2A2), so Anks1b was considered
414 significant if any one of its UniProt IDs was significant. Proteins were required to change
415 consistently in the same direction in all of the significant experiments. Exceptions were made
416 for three proteins (Adcy5, Mri1, and Rap1gap), because in each case, one experiment changed
417 in the opposite direction, and this experiment had sparse data (5 of the 8 replicates in one
418 group had missing values, and all of the replicates in the second group had missing values).
419 Genes significant in 6 or more of the 12 comparisons and showing consistent directions of
420 change were used as candidates for validation. Significance in at least 2 of the 4 validation sets
421 was the requirement to be incorporated into the final Str115P signature.

422

423 *Validation data sets*

424 The striatum and cortex RNA signatures were validated using untreated HD and WT control
425 mice from other experiments, some of which have already been published and others which
426 are in pre-publication. Pre-publication data sets are referred to generically in the text as
427 "CohortX" until those data sets are public. Some cohorts have multiple time points, and those
428 will be identified as "CohortXTimeY". The public studies are HDAC (GSE104086), which studied
429 the effects of HDAC inhibitors, and KMO (GSE105158), which studied the effects of a KMO
430 inhibitor. All of the HD mice were either Q175 or Q140 mice aged 6 to 12 months
431 (Cohort1Time1, Cohort1Time2, Cohort2Time1, Cohort2Time2, Cohort3Time1, Cohort3Time2,
432 and Cohort3Time3) or R6/2 mice aged 6 weeks to 3 months (HDAC, Cohort4, and KMO). The
433 Cohort1, HDAC, Cohort4, KMO, and Cohort2 validation sets each had 7 to 10 HD mice and 7 to
434 10 WT mice. The Cohort3 validation sets each had 23 to 31 HD mice and 23 to 29 WT mice. The
435 Cohort1 and Cohort3 data sets had only striatum samples, while the HDAC, Cohort4, KMO,
436 Cohort2 sets had both striatum and cortex samples.

437

438 The proteomics signature had 4 validation data sets: R6/2 mice aged 2 months (R6/2 2M) or 3
439 months (R6/2 3M) from PRIDE experiment PXD013771; R6/2 mice aged 6 weeks from a JNK3
440 knockout experiment (JNK3 R6/2 6W); and Q175 mice aged 10 months from Cohort5. The R6/2
441 2M, R6/2 3M, and JNK3 R6/2 6W validation sets had 10 to 12 R6/2 mice and 10 WT mice. The
442 Cohort5 Q175 10M set had 19 Q175 mice and 20 WT mice.

443

444 **Discussion**

445 Previous work has explored the RNA consequences of HD in mouse knock-in models with
446 repeat lengths of Q80, Q92, Q111, Q140, and Q175 through RNA sequencing of the striatum,
447 cortex, and liver, comparing each of these Q lengths to Q20 control mice (3). We extended that
448 RNA-seq analysis to include additional tissues (cerebellum, hippocampus, and white adipose
449 tissue near gonads) and added proteomics analysis of 7 tissues (striatum, cortex, liver,
450 cerebellum, hippocampus, skeletal muscle, and heart). These allelic series experiments, already
451 publicly available as counts and intensities per gene, will now be available with fold changes
452 and adjusted p-values from a collective analysis using uniform methods, thresholds, sample
453 names, and data files. The significant differential expression results for all genes and proteins in
454 these experiments have been combined into Supplementary Table 1 and are also available on
455 HDinHD.org. The programs and data files needed to reproduce the differential expression tests
456 are provided in Supplementary File 11.

457

458 The numbers of dysregulated genes in the striatum increased consistently with both Q length
459 and age, a behavior not seen in the other tissues. We combined these consistent changes with
460 other RNA-seq and proteomics data sets to generate robust striatum RNA and protein disease
461 signatures. Both signatures were validated using data sets that included a different HD mouse
462 model, the R6/2 mice that overexpress a fragment of human exon 1 of HTT (17). We further
463 derived smaller bidirectional versions of these signatures for experimental platforms that target
464 fewer genes, and also derived a small signature that works for both RNA and protein
465 experiments. These signatures will be valuable as molecular readouts of pathology progression

466 for researchers investigating the efficacy of experimental therapeutics and disease
467 mechanisms.
468 Future work could investigate the role of genes highlighted in these signatures. The gene
469 showing the highest change in the Str266R and Str20R signatures is WT1 (or Wt1 in mice). WT1
470 is a transcription factor that can act as either an activator or repressor, by binding to DNA at its
471 GC-rich consensus site, by modifying DNA methylation states, or by modifying chromatin
472 accessibility (18). Its consensus site, GCGGGGGCG, occurs in the promoter and exon 1 of
473 human and mouse HTT. The observed majority of genes that decrease in disease in Str266R
474 and Str115P could be due to negative regulation by the highly overexpressed WT1 in the HD
475 striatum RNA. In addition, further exploration of disease signatures in other tissues like the
476 liver may yield genes detectable in easily accessible tissues like blood and plasma, having
477 clinical biomarker potential.
478

479 **Acknowledgements**

480 This research was supported by CHDI Foundation, Inc. We thank Marcy MacDonald, Vanessa
481 Wheeler, and Scott Zeitlin for designing and engineering the expanded mouse lines in the allelic
482 series. We thank PsychoGenics for breeding the knock-in allelic series and dissecting the tissues
483 as part of a contract research agreement with CHDI. We thank Massachusetts General Hospital,
484 David Howland, and Seung Kwak for conceiving and executing the mouse allelic series study.
485 We thank Q2 Solutions Expression Analysis for generating RNAseq data and Evotec for
486 generating proteomics data.
487

488 **Supporting Information**

489 **NOTE: Because of their large file sizes, Supplementary Files 7, 8, 9, 10, and 11 will not be**

490 **uploaded to bioRxiv.org. They will remain available on HDinHD.org.**

491 **Supplementary Table 1. Combined table of all significant differential expression results,**

492 **Supp1_all_significant_results.txt.**

493 **Supplementary Table 2. Overlap in RNA and protein differential expression results,**

494 **Supp2_RNA_protein_comparison.xlsx**

495 **Supplementary Table 3. Str266R signature and validation results,**

496 **Supp3_Str266R_signature.xlsx.**

497 **Supplementary Table 4. GO terms for Str266R, Str115P, and Str40RP signatures,**

498 **Supp4_Signatures_GO_Terms.xlsx.**

499 **Supplementary Table 5. Cortex signature attempt, Supp5_cortex_signature.xlsx.**

500 **Supplementary Table 6. Str115P signature and validation results,**

501 **Supp6_Str115P_signature.xlsx.**

502 **Supplementary File 7. Full differential expression results for all genes, whether significant or**

503 **not, Supp7_Full_DEG_Results.zip (462 MB).**

504 **Supplementary File 8. Significant Q-length comparisons,**

505 **Supp8_Significant_Qlength_Results.zip (7 MB).**

506 **Supplementary File 9. Significant age comparisons with healthy aging genes included,**

507 **Supp9_Significant_Age_Results.zip (14 MB).**

508 **Supplementary File 10. Significant age comparisons with healthy aging genes removed,**

509 **Supp10_Significant_Age_Results_NoHealthy.zip (17 MB).**

510 **Supplementary File 11. All programs and input files needed to reproduce the RNA-seq and**

511 **proteomics differential expression results, Supp11_Programs_and_Data.zip (133 MB).**

512

513 **References**

514


515 1. Bates GP, Dorsey R, Gusella JF et al. Huntington disease. Nat Rev Dis Primers.
516 2015;1:15005.
517 2. Alexandrov V, Brunner D, Menalled LB et al. Large-scale phenome analysis defines a
518 behavioral signature for Huntington's disease genotype in mice. Nat Biotechnol.
519 2016;34:838-844.
520 3. Langfelder P, Cantle JP, Chatzopoulou D et al. Integrated genomics and proteomics define
521 huntingtin CAG length-dependent networks in mice. Nat Neurosci. 2016;19:623-633.
522 4. Zhang X, Smits AH, van Tilburg GB, Ovaa H, Huber W, Vermeulen M. Proteome-wide
523 identification of ubiquitin interactions using UbIA-MS. Nat Protoc. 2018;13:530-550.
524 5. Mode A, Gustafsson JA. Sex and the liver - a journey through five decades. Drug Metab
525 Rev. 2006;38:197-207.
526 6. Zielonka D, Stawinska-Witoszynska B. Gender Differences in Non-sex Linked Disorders:
527 Insights From Huntington's Disease. Front Neurol. 2020;11:571.
528 7. Barrett T, Wilhite SE, Ledoux P et al. NCBI GEO: archive for functional genomics data
529 sets--update. Nucleic Acids Res. 2013;41:D991-5.
530 8. Wheeler DL, Barrett T, Benson DA et al. Database resources of the National Center for
531 Biotechnology Information. Nucleic Acids Res. 2008;36:D13-21.
532 9. Perez-Riverol Y, Csordas A, Bai J et al. The PRIDE database and related tools and
533 resources in 2019: improving support for quantification data. Nucleic Acids Res.
534 2019;47:D442-D450.
535 10. Li J, Hu J, Newman M, Liu K, Ge H. RNA-Seq Analysis Pipeline Based on Oshell
536 Environment. IEEE/ACM Trans Comput Biol Bioinform. 2014;11:973-978.
537 11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
538 RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
539 12. Team RC. R: a language and environment for statistical computing. Vienna, Austria: 2018
540 13. Yates AD, Achuthan P, Akanni W et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682-
541 D688.
542 14. Ritchie ME, Phipson B, Wu D et al. limma powers differential expression analyses for
543 RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
544 15. UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res.
545 2019;47:D506-D515.
546 16. Huber W, Carey VJ, Gentleman R et al. Orchestrating high-throughput genomic analysis
547 with Bioconductor. Nat Methods. 2015;12:115-121.
548 17. Mangiarini L, Sathasivam K, Seller M et al. Exon 1 of the HD gene with an expanded CAG
549 repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. Cell.
550 1996;87:493-506.

551 18. Hastie ND. Wilms' tumour 1 (WT1) in development, homeostasis and disease.
552 Development. 2017;144:2862-2872.
553