Integrated genomics and proteomics define huntingtin CAG length–dependent networks in mice

Peter Langfelder¹, Jeffrey P Cantle²–⁴, Doxa Chatzopoulou², Nan Wang²–⁴, Fuying Gao², Ismael Al-Ramahi⁵,⁶, Xiao-Hong Lu²–⁴, Eliana Marisa Ramos²,³, Karla El-Zein⁵,⁶, Yining Zhao², Sandeep Deverasetty², Andreas Tebbe⁷, Christoph Schaab⁷, Daniel J Lavery⁸, David Howland⁸, Seung Kwak⁸, Juan Botas⁵,⁶, Jeffrey S Aaronson⁸, Jim Rosinski⁸, Giovanni Coppola²–⁴,⁹, Steve Horvath¹,¹⁰ & X William Yang²–⁴

To gain insight into how mutant huntingtin (mHTT) CAG repeat length modifies Huntington's disease (HD) pathogenesis, we profiled mRNA in over 600 brain and peripheral tissue samples from HD knock-in mice with increasing CAG repeat lengths. We found repeat length–dependent transcriptional signatures to be prominent in the striatum, less so in cortex, and minimal in the liver. Coexpression network analyses revealed 13 striatal and 5 cortical modules that correlated highly with CAG length and age, and that were preserved in HD models and sometimes in patients. Top striatal modules implicated mHTT CAG length and age in graded impairment in the expression of identity genes for striatal medium spiny neurons and in dysregulation of cyclic AMP signaling, cell death and protocadherin genes. We used proteomics to confirm 790 genes and 5 striatal modules with CAG length–dependent dysregulation at the protein level, and validated 22 striatal module genes as modifiers of mHTT toxicities in vivo.

Huntington's disease (HD) is a dominantly inherited neurodegenerative disease characterized by movement disorder, cognitive and psychiatric symptoms¹. The hallmark of HD neuropathology is selective degeneration of striatal medium spiny neurons (MSNs) and, to a lesser extent, cortical pyramidal neurons (CPNs)². Motor symptom onset for HD has a broad range but is usually in middle age. HD is relentlessly progressive, and patients die from complications of the disease about 10–20 years after onset. Currently there are no therapies to prevent the onset or slow the progression of HD³.

HD is caused by a CAG trinucleotide repeat expansion encoding an elongated polyglutamine (polyQ) stretch of the huntingtin (HTT) gene⁴. Unaffected individuals have fewer than 36 repeats, whereas affected individuals have 36 to 250 CAG repeats. HD belongs to a group of nine neurodegenerative disorders caused by CAG repeat expansion in distinct polyQ proteins⁴. Despite broad expression of these proteins, polyQ disorders present different clinical and pathological phenotypes that are attributed to distinct polyQ protein contexts.

A pivotal human genetic clue to pathogenesis of all polyQ disorders is that the length of the CAG repeat is inversely correlated with the age of disease onset⁴,⁵. Patients with HTT CAG lengths in the 40s often have motor onset in the fourth decade of life, while repeat lengths greater than 60 lead to juvenile onset⁴. In contrast to its influence on age of onset, the influence of CAG length on disease progression is much more modest¹,⁴, suggesting an important effect of CAG length early in disease pathogenesis⁴. Recent imaging studies of patients suggest that CAG length correlates with caudate atrophy⁷ and that combined CAG length and age is a useful predictor of many clinical outcomes in HD¹. Overall, HD patient studies underscore a critical role of CAG length in the early stages of pathogenesis.

The central role of CAG length on the age of HD motor symptom onset led to the polyQ molecular trigger hypothesis, which suggests that polyQ expansion in HTT leads to repeat length–dependent, dominantly acting pathogenic changes in the vulnerable neurons to initiate the disease⁵. The search for mutant HTT (mHTT) CAG length–sensitive molecular pathogenic changes has so far been limited to studies in cultured cells⁸,⁹, and very little is known about genes and pathways that are dysregulated by the expanded CAG mutant protein in the HD-vulnerable and pathogenically relevant brain regions, namely the striatum and cortex¹⁰. Here we undertook a multistep, integrative systems biology study of both the transcriptome and proteome of HD allelic series knock-in mice to identify murine mHTT CAG length– and age-dependent molecular networks, which is complemented by functional validation studies in a Drosophila model expressing a mHTT fragment. We also created a new interactive online resource (https://www.hdinhhd.org/) to disseminate our data.

¹Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, California, USA. ²Center for Neurobehavioral Genetics, Semel Institute for Neuroscience & Human Behavior, University of California, Los Angeles, Los Angeles, California, USA. ³Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, California, USA. ⁴UCLA Brain Research Institute, University of California, Los Angeles, Los Angeles, California, USA. ⁵Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA. ⁶Jan and Dan Duncan Neurological Research Institute, Texas Children’s Hospital, Houston, Texas, USA. ⁷Evotec Munich GmbH, Martinsried, Germany. ⁸CHDI Foundation/CHDI Management Inc., Princeton, New Jersey, USA. ⁹Department of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, California, USA. ¹⁰Department of Biostatistics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, California, USA.

Correspondence should be addressed to X.W.Y. (xwyang@mednet.ucla.edu) or S.H. (shorvath@mednet.ucla.edu).

Received 16 September 2015; accepted 28 January 2016; published online 22 February 2016; doi:10.1038/nn.4256
RESULTS

Longitudinal RNA-sequencing analyses of mHtt allelic series knock-in mice

We reasoned that transcriptomic study of HD-vulnerable brain regions (striatum and cortex) and a relatively disease-resistant peripheral tissue (liver) from an allelic series of HD knock-in mice at distinct ages (Fig. 1a) could provide crucial systems-level insight into CAG length–dependent selective pathogenesis in HD. Using deep mRNA sequencing (RNA-seq), we profiled the striatum, cortex and liver of 2-, 6- and 10-month-old mice that express one wild-type endogenous Htt allele and a second Htt allele with knock-in of human mHtt exon 1 carrying one of six different CAG lengths (denoted Q; that is, Q20, Q80, Q92, Q111, Q140 and Q175). These Htt knock-in mice were chosen to include those that exhibit progressive disease phenotypes (Q140 and Q175), as well as those with no overt behavioral deficits but with molecular signature changes (Q80, Q92 and Q111). We sequenced 8 animals per genotype and age for each tissue, with a total of 432 RNA-seq samples. Differential gene expression analyses showed robust age- and CAG length–dependent increases in aberrantly expressed genes in the striatum and, to a lesser degree, in the cortex, while such CAG length–dependent gene expression was not a feature of the liver transcriptome (Fig. 1b and Supplementary Table 1).

Brain and peripheral tissue transcriptomes in Q175 mice

To evaluate whether CAG length–dependent genes are also dysregulated in other brain regions and peripheral tissues, we generated a second set of RNA-seq data (173 samples) to survey the transcriptome of five more brain tissues and six peripheral tissues from Q175 and wild-type (WT) mice at 6 months of age (Fig. 1a and Supplementary Table 2). Although this tissue survey used WT controls carrying murine Htt (Q7) and not the Q20 controls used in the full series, we found very few if any differentially expressed genes between Q20 and WT mice (Supplementary Fig. 1). For consistency with the rest of the tissue survey, we also carried out differential analysis between striatum Q175 and WT samples (WT samples for liver and cortex were not available). Our tissue survey confirmed that the striatum, the brain region most vulnerable in HD, had the largest number of differentially expressed genes (Fig. 1c). Intriguingly, white gonadal fat, a tissue not previously implicated in HD but that is adjacent to testes, which do show degeneration in HD, showed the second highest number of differentially expressed genes in Q175 mice.

Figure 1 Workflow and differential expression analysis with respect to Htt CAG length.

(a) Overview of experiment design and analysis strategy. (b) Numbers of significantly (FDR < 0.05) differentially expressed genes in striatum, cortex and liver. Blue (red) bars represent genes significantly down- (up-) regulated with increasing CAG length (Q). (c) Numbers of differentially expressed genes in the 14 tissues for which we profiled Q175 samples and controls, consisting of striatum, cortex, liver and 11 more tissues. The screening in the cortex and liver corresponds exactly to that presented in (b); the numbers for striatum are slightly different because the analysis in (b) used Q20 controls whereas the one in (c) used WT controls. (d) Heat map showing correlation of differential expression Z statistics of individual genes across the 14 tissues. Correlations whose absolute value is at least 0.2 are shown explicitly. Ad., adipose; intest., intestinal; hypothal./thal., hypothalamus and thalamus.
differentially expressed genes. The tissue survey also replicated previously reported results, such as the lack of gene expression changes in the heart of HD mice.16

To gain insight into global tissue similarities and differences of transcriptomic response to polyQ expansion within Htt, we evaluated the concordance of differential expression among the brain and peripheral tissues (Fig. 1d). We observed that the brain tissues clustered relatively tightly together, while peripheral tissues form two looser clusters. None of the peripheral tissues correlated with brain tissue at or above the correlation coefficient threshold of r = 0.20, suggesting that Q175 mice exhibit brain-specific differential transcriptome changes at 6 months of age.

**Huntingtin CAG length– and age-dependent gene coexpression network**

We reasoned that a consensus weighted gene coexpression network analysis (WGCNA)17 across the three time-points would allow us to define modules of genes that are coexpressed at all three time-points and study their CAG repeat-dependent variation across all three ages. The three consensus analyses (one per tissue) identified 37 striatal, 35 cortical and 34 liver consensus modules (Supplementary Table 3) across a total of 140 striatum, 142 cortex and 141 liver samples. The modules ranged in size from 38 to nearly 1,900 genes and contained a total of 12,654 genes in the striatum, 11,325 genes in the cortex and 11,590 genes in the liver (each gene was allowed to belong to only one module or to be unassigned). In this manner, network analysis reduced thousands of genes across three ages to a relatively small number of coherent gene modules that represent distinct transcriptional responses to the varying CAG length alleles in the murine Htt locus. To quantify the overall relationship between a module and CAG length across all three ages, we used a meta-analysis of correlations of module eigengenes (summary expression profiles) with CAG length. We identified 13 striatal and 5 cortical modules that passed a meta-analysis significance Z statistic threshold of 5 (corresponding to P ≈ 10−7; Fig. 2 and Supplementary Table 3). None of the liver modules satisfied this criterion.

The network analyses provide several layers of information. First, the strength and significance of associations between modules and CAG length were strongest in the striatum, suggesting that CAG length affects entire coexpression modules more strongly in the striatum than in the cortex or liver. Second, the meta-analysis significance statistics allowed us to rank modules by their overall association with CAG length (Fig. 2c–f). Third, a major output of network analysis is a continuous (‘fuzzy’) measure of module membership for all genes in all modules (Supplementary Table 3). The module membership measures how similar the gene expression profile is to the eigengene (summary profile) of the module. Genes whose profiles are highly similar to the eigengene can be identified as intramodular...
Such genes are often useful for implicating relevant biological pathways and prioritizing genes for functional studies. To explore the biological implications of the CAG length–related modules, we carried out gene functional enrichment studies (Supplementary Tables 4–7). Striatal module M2 had the strongest negative association with CAG length (Fig. 2c) and was highly enriched with genes involved in CAMP signaling, postsynaptic density proteins and caudate (striatum) marker genes (Supplementary Tables 4 and 7). M11 and M52 appeared to be glia-related, and M25 was enriched for glutamate receptor signaling. Arguably, one of the most interesting modules was M34, which was downregulated in a CAG length–dependent manner even at 2 months and was highly enriched in transcription and chromatin factors.

Among the eight striatal modules that had a positive correlation with CAG length, M20 had the strongest correlation and was enriched for p53 signaling, cell division and protocadherin genes. M7, M39 and M46 were involved in stress responses, including cell death (M7), DNA damage repair (M39) and glucocorticoid signaling (M46). Intriguingly, DNA damage repair has been implicated in modifying pathogenesis in HD mice20 and age at onset in HD patients21. Other upregulated striatal modules revealed the impact of mutant CAG repeats on genes involving mitochondria (M9, M43) and proteostasis (M1, M10, M39), which have already been implicated in HD22,23.

Among the cortical modules correlated with CAG length (Fig. 2e,f), M4 had the most significant negative correlation and was enriched in genes related to calcium signaling, synapses and glutamatergic neurons (Supplementary Tables 4–6). These pathways are consistent with a pathogenic role of cortical neurons in HD10. Cortical M45 was a glial module involved in fibroblast growth factor signaling, ensheathment of neurons and fatty acid biosynthesis, and M6 and M7 were enriched in axon guidance genes. The latter suggests that certain neurodevelopmental pathways could be dysregulated by mHtt.

**Preservation of module–CAG length associations in independent HD mouse model and patient data sets**

We next studied whether genes in the 18 CAG length–dependent modules are dysregulated in other brain regions or peripheral tissues of HD mice. We found that average module–CAG length correlations were more preserved among the brain tissues; none of the modules strongly changed with genotype in the peripheral tissues in Q175 mice (Fig. 3a and Supplementary Table 8). Striatal M34 and M43 modules appear striatum specific, as they showed few correlations with genotype in other brain regions. Three modules enriched in glial genes (M11, M20 and M45) exhibited high correlation across multiple brain regions, suggesting that glial gene changes may be more widespread in HD mouse brains.

We next asked whether the CAG length dependence of these modules derived from allelic series knock-in mice are preserved in other HD mouse models (Fig. 3b; see Online Methods and Supplementary
We found that the majority of the module–genotype associations were preserved in R6/2 (ref. 27; 13 of 13) and BACHD-ΔN17 mice (12 of 13, Fig. 3b), both of which express neuropathogenic fragments of mHTT, and in two other full-length mHtt mouse models (Q150 knock-in and YAC128, a transgenic with a yeast artificial chromosome expressing Q128 mHtt)11.

Finally, three striatal modules (M2, M25 and M7) were significantly preserved in two human post-mortem HD caudate data sets24,25, and one cortical module (M4) was preserved in two of four HD cortical data sets24,26 (Fig. 3c,d). These modules were mostly not preserved in cerebellar data sets, consistent with selective vulnerability in HD. Furthermore, we found 543 of our striatal CAG length–dependent

Table 9). We found that the majority of the module–genotype associations were preserved in R6/2 (ref. 27; 13 of 13) and BACHD-ΔN17 mice (12 of 13, Fig. 3b), both of which express neuropathogenic fragments of mHTT, and in two other full-length mHtt mouse models (Q150 knock-in and YAC128, a transgenic with a yeast artificial chromosome expressing Q128 mHtt)11.
module genes to be significantly altered (FDR < 0.05) in at least one of the publicly available HD caudate gene expression data sets\(^\text{24,25}\) and 431 of these genes were significantly altered in both patient data sets (Supplementary Table 10). Similarly, we found 25 CAG length–dependent cortical module genes that were significantly differentially expressed in at least one of the available HD patient cortical samples (Supplementary Table 10)\(^\text{24,26}\). As expected, these genes were significantly enriched in the three striatal modules (M2, M7 and M25) and one cortical module (M4), with M2 containing by far the largest number of such genes (Fig. 3e and Supplementary Table 10).

**M2 implicates mHtt CAG length in age-dependent impairment of striatal MSN identity gene expression in HD mice**

Our network analyses identified M2 as the module with the strongest association with CAG length and the largest number of dysregulated genes. Hub genes for module M2 (that is, those with highest correlation with the M2 eigengene; Fig. 4a) contained well-known striatal MSN marker genes (for example, Ppp1r1b, Drd1a, Drd2 and Grpr). Although striatal marker downregulation has been shown in HD mice and HD patients\(^\text{24,27}\), these studies could not rule out confounding effects such as neuronal loss (in postmortem brains), expression of only mHTT fragments, or transgene expression levels\(^\text{11}\). Our current study directly addresses whether striatal marker gene dysregulation is an early and CAG length–dependent pathogenic event in HD mice with endogenous levels of full-length mHtt expression.

To assess striatal marker gene dysregulation, we used a collection of striatum-specific marker genes ranked by the Allen Brain Atlas (ABA)\(^\text{28}\) that are also marker genes of striatal MSNs\(^\text{29}\). Remarkably, M2 contained 70 of the 88 top striatum-specific ABA marker genes that were present in our striatal transcriptome data sets (hypergeometric P-value \(3.3 \times 10^{-51}\); Fig. 4b and Supplementary Table 11). These striatal marker genes exhibited striking patterns of downregulation that were both CAG length– and age-dependent (Fig. 4b), which was not observed for ABA pan-neuronal marker genes in the striatum (Fig. 4c and Supplementary Table 11). This finding reveals, surprisingly, that Htt CAG length expansion selectively impairs age-dependent maintenance of striatal marker gene expression.

To further strengthen this observation, we asked specifically whether CAG expansion differentially impacts striatal direct-pathway MSNs (D1-MSNs) that are affected earlier in HD compared to indirect pathway MSNs (D2-MSNs) that are affected later in HD\(^\text{30}\). Indeed, the M2 module is significantly enriched with genes in two independently generated D2-MSN-specific gene sets while such enrichment was absent for D1-MSN-specific gene sets\(^\text{11,33}\) (Fig. 4e and Supplementary Table 11). This latter finding provides fresh evidence that mHtt CAG length selectively affects D2-MSN-enriched

---

**Figure 5** Cell death genes in striatum module M7. (a) Network plot of the top 50 hub genes in striatum module M7. Red, pink and white indicate significant (FDR < 0.1) and consistent differential expression between HD samples and controls in both, one or none of the human caudate nucleus data sets; blue color indicates genes implicated in cell death. (b) Graphical representation of the cell death pathway with dysregulated genes from striatum module M7 highlighted in red. (c–e) Association of cell death genes in striatum module M7 with genotype or HD status in other mouse striatum data (c), the tissue survey (d) and human postmortem data (e). Bars show the weighted average correlation of the cell death M7 genes with the relevant genotype or HD status; asterisks show the corresponding permutation test significance (*P < 0.05; **P < 0.01; ***P < 0.001). The correlations and P-values are listed in Supplementary Table 12. Ad., adipose; intest., intestinal; hypothal./thalam., hypothalamus and thalamus.
Figure 6  Protocadherin dysregulation across multiple modules. (a) Schematic of the clustering of mouse protocadherin gene. Each vertical bar represents one gene. Colored bars show protocadherins that are members of CAG length–dependent allelic series consensus modules. Variable: variable exons; cons.: constant exons. (b–e) Network plots of top 10 hub genes (inner ring) and other hub genes and protocadherins (outer ring) in striatal modules M20 (b), M34 (c), M39 (d) and M46 (e). Colored circles indicate clustered protocadherins (colors corresponding to schematic above: green, Pcdha; purple Pcdhb; blue, Pcdhg) and unclustered protocadherins (pink).

genes. Finally, M2 genes are also well preserved in laser-captured HD patient striatal MSNs24, suggesting that MSN-enriched genes may be similarly affected in HD patients (Fig. 4e and Supplementary Table 11).

Emerging concepts suggest that operationally defined neuronal identity genes are the genes whose expression are enriched in a group of neurons, stable throughout their lifetime and that subserve critical functions33,34. Our finding that the M2 module contains the most of the known ABA MSN marker genes implies that mHtt CAG expansion may impair age-dependent maintenance of striatal MSN identity gene expression. Notably, we found that mHtt CAG length may also modestly affect ABA cortical marker genes—that is, those in the M4 module (Fig. 4d and Supplementary Table 11)—suggesting that mHtt CAG expansion may impair age-dependent maintenance of striatal and likely cortical neuronal identity genes, but with the magnitude of dysregulation much greater in the striatum.

M7 module reveals upregulation of cell death signaling genes in HD mouse striatum

Annotation of the striatal M7 module unexpectedly revealed the CAG length– and age-dependent upregulation of 17 cell death–related genes in HD mouse striata (Fig. 5a, b). The cell death genes in M7 were differentially expressed in brain but not peripheral tissues in Q175 mice, and their differential expression was also preserved in other HD mouse and patient striata (Fig. 5c–e and Supplementary Table 12). This finding is surprising because heterozygous HD knock-in mice do not show striatal MSN cell loss up to 12 months of age13,14. We used unbiased stereology to confirm that Q175 heterozygous mice did not show changes in the striatal neuron and GFAP⁺ astrocyte cell numbers at 4.5 and 10 months of age (Supplementary Fig. 2). Thus, we conclude that the upregulation of cell death genes in M7, as well as downregulation of MSN marker genes in M2, is unlikely to be due to cellular composition changes (for example, MSN loss or proliferation of astrocytes); instead, they likely reflect CAG length–dependent increase in the vulnerability of MSNs to degeneration (akin to a prodegenerative state) in HD mice.

Dysregulation of protocadherin clusters in striatal modules

Our results raise the question of how relatively subtle changes in the mutation (that is, CAG repeat expansion) in Htt lead to graded transcriptional changes in allelic-series HD mouse brains. A potential clue to this question is that 4 of the 13 striatal modules (M20, M34, M39 and M46) are highly enriched in cadherin and protocadherin (Pcdh) genes (Fig. 6). Cadherins and protocadherins function in neurodevelopment, intercellular signaling, adhesion, synaptic function and neuronal survival35, with a combinatory code of protocadherins thought to underlie single neuron identity in vitro and in vivo36. Prior studies have unraveled critical roles of the transcriptional and chromatin factors Ctcf, Rad21 and Rest in regulating the expression of Pcdh genes in neurons37,38. We found striking dysregulation of 37 out of 58 clustered Pcdh genes (Fig. 6). Intriguingly, among the known Pcdh transcriptional regulators, REST is already known to be dysregulated by mHTT in vitro and in vivo39, and both Ctcf and Rad21 were present in prominent CAG length–dependent modules (M34 and M7, respectively). Bioinformatic analyses revealed that the two highest Htt CAG length–dependent modules (M2 and M20) were significantly enriched in known Ctcf binding sites in the brain (parametric permutation P = 4.1 × 10⁻⁷ (Z = 4.9) for M2 and P = 7.6 × 10⁻⁷ (Z = 4.8) for M20; Supplementary Fig. 3). Together, our findings suggest that the study of Pcdh gene clusters may be a tractable route to dissecting the upstream regulatory mechanisms underlying mHtt CAG length–dependent molecular networks in HD mouse brains.
verified more than a dozen genes in the M2 module that are dysregulated in Q175 mice, more relevant evidence for validation of CAG length-dependent transcriptome changes is unbiased validation at the proteome level. To this end, we used striatal tissues from the same 6-month mouse brains used for RNA-seq to perform quantitative proteomic analyses with mass spectrometry (Fig. 7a). Prior studies have shown that such proteomic studies allow the identification and quantification of a substantial proportion of all proteins expressed in the investigated cells. Furthermore such studies can be used to build unbiased WGCNA protein networks. We quantified relative protein abundance in 45 of the 48 striatal samples using an accurate label-free quantitation method (MaxLFQ; Supplementary Table 13) that permits application of standard statistical tests. In total, 254,543 unique peptides and 10,047 proteins were identified according to an accepted FDR of less than 0.01 on the protein and peptide level after removing contaminants (36) and reverse hits (250). Not all proteins could be quantified in every sample, but on average 7,774 proteins were identified and quantified per sample.

Across the 45 common striatal samples and 7,039 protein–mRNA pairs present in both filtered protein data (Online Methods) and mRNA network analysis, the correlation of protein LFQ intensities and variance-stabilized mRNA data was 0.42. This agrees with other high-throughput proteomic studies, which have found similar correlations.

Using the quantitative protein readouts from MaxLFQ, we found CAG length–dependent increases in dysregulated proteins in Q111, Q140 and Q175 mice, with a total of 1,370 proteins found to be CAG length correlated. Moreover, we confirmed 790 proteins with continuous CAG length–dependent changes at both RNA and protein levels in the striatum, with 133 proteins in Q111 mice, 301 proteins in Q140 mice and 533 proteins in Q175 mice (Fig. 7b and Supplementary Tables 14 and 15). We have thus validated a substantial number of genes that are altered at both the RNA and protein level in a CAG length–dependent manner, hence presenting a list of candidate genes for future functional studies in HD model organisms.

Intriguingly, although the majority of proteins that were significantly correlated with CAG length were positively correlated with their mRNA (of the 1,370 CAG length–correlated proteins identified at an FDR level of 0.1, 1,023, or 75%, had a positive correlation with their mRNA (of the 1,370 CAG length–correlated proteins identified at an FDR level of 0.1, 1,023, or 75%, had a positive correlation with their mRNA), a few proteins (29) with significant CAG length
correlation had a negative and nominally significant ($P < 0.05$) correlation with their mRNA levels (Fig. 7c and Supplementary Table 15). This subset of genes may reflect the impact of mHtt CAG expansion on post-transcriptional regulation of a subset of proteins (for example, translation, protein stability or clearance). Among these proteins, F8a (or Hap40) is a known HTT interactor that mediates vesicular trafficking$^{43}$. With increasing CAG length, F8a was downregulated at the protein level but upregulated at the RNA level. Our data are consistent with a previously observed reduction of Hap40 in the synaptosome of Q175 mice$^{46}$, suggesting the need to study how mHtt may regulate Hap40, as well as other proteins (for example, Ogt and Golgb1; Fig. 7c) with opposing variation in RNA and protein levels, and whether such regulation could be pathologically relevant.

The advantage of performing high-throughput proteomic validation instead of validating individual proteins is the ability to examine CAG length–dependent molecular networks at both RNA and protein levels. We first used enrichment analyses to show that eight CAG length–dependent transcriptome modules were significantly enriched with differentially expressed proteins in HD mice, with the M2 and M7 modules having most such proteins (Fig. 7d and Supplementary Table 14). Moreover, by performing WGCNA directly on the proteomic data set using label-free quantitative protein inputs$^{12,43}$, we defined protein network modules (pMs), with several modules showing association with CAG length (Supplementary Table 15). Summary profiles (eigenproteins) of the five protein modules that passed a stringent significance threshold ($Z > 4.5 \times 10^{-5}$) are shown in Figure 7e-i. These independently created CAG length–dependent protein modules showed remarkable overlap with the respective transcriptome modules (pM2 with M2, pM7 with M7; Fig. 7) and Supplementary Table 16), demonstrating the robustness of these CAG length–dependent molecular (RNA and protein) networks and strengthening confidence in selecting hub genes from such modules for further functional studies.

**Modifiers of mHtt toxicity in Drosophila**

An important step for exploring the functional significance of the CAG length–dependent modules is to genetically perturb top hub genes in an established HD animal model system. To this end, we tested 49 striatal module hub genes (Supplementary Table 17) as genetic modifiers in a Drosophila model expressing a mHtt fragment, using age-dependent climbing deficits as a sensitive readout of mHtt-induced motor dysfunction$^{42,47}$. We found 22 modifier genes using such an in vivo assay, 11 of which were in the M2 module, including the top hub genes Arpp21, Camk2b and Chn1 (Fig. 8a,b, Supplementary Fig. 5 and Supplementary Tables 18 and 19). Four M7 hub genes and hub genes from M20, M34, M43 and M9 were also modifiers in this assay. In support of our hypothesis that the hub genes of M34, such as Ctcf, may be critical regulators of mHtt CAG length–dependent transcriptional deficits and neuronal toxicities, we found that two independent heterozygous loss-of-function alleles of the Drosophila Ctcf gene could substantially, albeit partially, ameliorate climbing deficits in this Drosophila model (Fig. 8a,b and Supplementary Tables 18 and 19). Moreover, 11 of our validated modifier genes were those shown to be CAG length dependent at both RNA and protein levels (Supplementary Table 18), providing particularly strong evidence that these genes may be relevant for future study in mammalian HD models.

Although fly modifier results may not simply translate to those in mammalian models, we reason that loss-of-function mutants may mimic genes in modules with negative CAG length correlation (for example, M2) and overexpression mutants may mimic the positively correlated modules (for example, M7 and M20). Nine out of 11 modifiers in the M2 module were loss-of-function suppressors, suggesting that their downregulation in HD mice may represent compensatory changes to mitigate mHtt toxicities. By contrast, two genes (Camkv and Nagk) showed CAG length–dependent changes that may result in enhanced toxicities. Future genetic analyses of CAG...
length–dependent module genes in both fly and mouse models of HD will help to systematically validate genes that are modifiers of HD pathogenesis in vivo.

**HDinHD: an online resource for HD research**

Raw and processed data, as well as network analysis plots and phenotypic data related to the allelic series HD mice, are hosted on a dedicated website and server at https://www.hdinhd.org/ (see Online Methods). Currently, HDinHD comprises three sections: (i) a data repository and gene expression browser, allowing access to genomic and proteomic data from the allelic series, including raw and normalized RNA-seq counts and relative protein expression values (MaxLFQ) from mass spectrometry; (ii) a network browser for online mining of the WGCNA analysis reported here; and (iii) a suite of analysis tools that allow the user to save, share and annotate gene lists, which can be compared with existing gene sets, either published or generated by other HDinHD users. A long-term goal of HDinHD is to compile a structured catalog of the HD-related genomic, proteomic and phenotypic data sets. We expect that HDinHD will help access the large-scale HD allelic-series RNA-seq and proteomic data and consensus network analysis results and will provide a collaborative environment for HD researchers to access and share both published and unpublished data sets. Online analysis tools will allow the user to annotate gene lists and to test hypotheses in silico in order to facilitate mechanistic and therapeutic research for HD.

**DISCUSSION**

We applied extensive genome-wide transcriptomics with deep RNA-seq using mice with Htt CAG repeat expansion in the endogenous gene, followed by strategic proteomics-based validation studies, to elucidate the mHtt CAG length–dependent molecular networks in disease-relevant brain regions. One finding was the number of CAG length–dependent genes and modules appearing to correlate with known vulnerabilities in HD (striatum more than cortex and none in liver), hence supporting the hypothesis that CAG length–dependent molecular changes are likely to be relevant to the underlying vulnerability in the disease.

We identified 13 striatal and 5 cortical gene coexpression modules that were strongly associated with Htt CAG length. These modules highlight several biological pathways affected by Htt CAG repeat expansion in the striatum and cortex. A continuous measure of module membership identifies module hub genes that, at least from a statistical point of view, deserve prioritization for genetic perturbation studies. The HD relevance of these CAG length–dependent modules is strengthened by the preservation of their association with mHtt expression in multiple HD mouse and HD patient data sets. Moreover, our proteomics data provide high-confidence validation of about 790 proteins as being Htt CAG length dependent at both RNA and protein levels in the striatum of HD mice. We found that five independently generated, CAG length–dependent protein modules (including M2 and M7) showed strong overlap with transcriptional modules also preserved in other HD models and patients. Hence, the hub genes in such modules should be prioritized for further target validation studies in HD mice.

One insight from this study is that CAG repeat expansion in endogenous Htt impairs the age-dependent maintenance of striatal MSN identity gene expression. While previous studies have shown that a subset of MSN marker genes are dysregulated in HD mice and patients, these studies were not designed to assess the impact of endogenous mHtt CAG length and evaluated only a subset of MSN identity genes. Our comprehensive transcriptomic analyses of Htt knock-in allelic series mice allowed us to uncover the graded effect of mHtt CAG length and age in selective downregulation of most MSN identity genes, but not general neuronal marker genes, in the striatum. Remarkably, module M2 also is enriched in genes selectively expressed in D2-MSNs, the most vulnerable neuronal cell type in HD. Our results are consistent with the emerging concept that neuronal cell type identity genes need to be actively maintained in adult brains by a specific set of transcription and chromatin factors. Disruption of such a program is likely to lead to progressive and selective loss of neuronal identity followed by neurodegeneration, which has been found in animal models. Our study highlights the importance of future studies unraveling the molecular program involved in maintaining striatal MSN identity gene expression and elucidating how mHtt CAG expansion may cause age-dependent interference with such a program. Ultimately, we need to test whether restoring such a program could prevent disease onset or progression in HD.

An important long-term goal of our systems-level study of HD pathogenesis is to identify the precise molecular mechanisms linking mHtt polyQ expansion to transcription and chromatin dysregulation. To this end, several clues have already emerged from our analyses. First, the dysregulation of many clustered Pdcd gene in four of our modules suggests that regulatory factors for Pdcd expression (namely, Ctf, Rad21 and Rest) may also be involved in mHtt-induced transcriptional dysregulation. Indeed, Rest is a known polyQ length–dependent interactor with mHtt that is involved in age-dependent neuronal transcription. Hence it would be worthwhile to study the function of Rest in Htt CAG length–dependent gene expression in vivo. Similarly, Ctf and Rad21 are critical chromatin loop regulators in two independent heterozygous null allelic of Ctf can ameliorate behavioral deficits in a Drosophila model of mHtt fragment toxicities, it would now be important to investigate whether Ctf is dysregulated in HD striatum and contributes to mHtt CAG length–dependent transcriptionopathy, including MSN identity gene downregulation. Finally, M34 is not only enriched in transcription and chromatin factors but is also altered in a CAG length–dependent manner at an early age (2 months). Hence, M34 module genes (for example, Ctf and Ezh2) should be prime candidates in studying early and causal transcriptional regulators that underlie CAG length– and age-dependent modules in HD mouse striatum.

In summary, our study provides a large-scale, comprehensive transcriptomic and proteomic characterization of the molecular pathogenic effects of CAG repeat expansion in endogenous murine Htt in vivo, identifying a consistent set of genes (both RNA and proteins) and networks that are dysregulated in a CAG length– and age-dependent manner in HD mouse brains. We provide integrative genomic evidence to show converging molecular networks that are perturbed in both HD mice and patients, and provide proof-of-concept that 22 genes from such networks modify mHtt fragment toxicities in Drosophila. Together, our integrative systems findings and online database (HDinHD) constitute a rich and novel resource to facilitate the discovery of mHtt CAG length–dependent pathogenic mechanisms and novel therapeutic targets for HD.

**METHODS**

Methods and any associated references are available in the online version of the paper.
Accession codes. Gene Expression Omnibus: GSE65776. Transcription data are also available through our online tool, HDinHD. 

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository at PXD003442.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank PsychoGenics for help in breeding the knock-in allelic series and dissecting the tissues as part of a contract research agreement with CHDI. The research was supported by CHDI Foundation, Inc. HD research in the Yang laboratory is also supported by NINDS US National Institutes of Health grants (R01NS074312, R15NS049501 and R1IN084298). X.W.Y. is also supported by the David Well fund from Semel Institute, the Carol Moss Spivak Scholarship in Neuroscience from the Brain Research Institute at UCLA, and the Leslie Gehry Brenner Prize from the Hereditary Disease Foundation. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

AUTHOR CONTRIBUTIONS

X.W.Y., P.L., S.H., G.C., J.R. and J.S.A. designed and supervised the study. D.H. and S.K. supervised allelic series HD knock-in mouse tissue collection, RNA-seq and stereotactical counting of MSNs and astrocytes in Q175 mice. F.G. and G.C. performed RNA-seq data processing, PL. and S.H. performed WGCNA consensus module analyses, preservation studies and WGCNA analyses of proteomic data sets. J.P.C., N.W.Y., K.E.-Z. and X.W.Y. contributed to analyses and generation of data and graphs used in Figures 3–6 and Supplementary Table 4. J.P.C. performed studies for data shown in Supplementary Figure 4. A.R., K.E.-Z. and J.B. performed the mutant huntingtin Drosophila modifier study. D.C., Y.Z., S.D. and G.C. created the HDinHD database. E.M.R. and G.C. performed Ctf4 enrichment analyses. A.T.S. and D.J.L. performed striatal tissue proteomic studies for the Hit knock-in mice. X.W.Y., P.L., S.H. and G.C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.
Animal breeding and husbandry. We analyzed six heterozygous Htt knock-in lines expressing CAG repeats of length 20, 80, 92, 111 (ref. 51), 140 (ref. 52) and 175 (ref. 12). For each line, male heterozygous mice were crossed with C57BL/6 female mice at Jackson Laboratory (Bar Harbor, ME). Animals from litters having four to eight pups born within 3–4 d were identified by ear tags, tail sampled for genotyping and weaned at around 3 weeks of age. Heterozygous mice were selected according to CAG repeat length to allow a Gaussian distribution of CAG repeats in the experimental cohort and thereby avoid skewed distributions. Best Gaussian fit was judged by eye. Experimental animals had to weigh more than 11 g (females) or more than 13 g (males) by 5 weeks of age. Animals presenting any anomaly were excluded. Unacceptable anomalies were cataracts, malocclusion, missing or small eye, ear infection, unreadable or missing tag. Mice were housed in cages enriched with two play tunnels, a plastic bone and Enviro-Dri (Shepherd Specialty Papers). Animal cage changes occurred weekly. The cages were maintained on a 12:12 light/dark cycle. Water and food were freely available at all times.

This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals, NRC (2010). The protocols were approved by the Institutional Animal Care and Use Committee of PsychoGenics, Inc., an AAALAC International accredited institution (unit #001213).
Continuous measure of module membership. Module eigengenes lead to a natural measure of similarity (membership) of all individual genes to all modules. We define a continuous (’fuzzy’) measure of module membership $MM$ of gene $i$ in module $l$ as

$$MM_i^l = \text{cor}(x_i, E_i^l)$$

where $x_i$ is the expression profile of gene $i$ and $E_i^l$ is the eigengene of module $l$. This definition is applicable to every individual network (data set). The value of module membership lies between $-1$ and $1$. Higher $MM_i^l$ values indicate that the expression profile of gene $i$ is more similar to the summary profile of module $l$. Since we use signed networks here, we consider module membership near $-1$ low. The advantage of using correlation to quantify module membership is that the corresponding statistical significance ($P$-values) can be easily computed. Genes with highest module membership are called hub genes. Hub genes are centrally located inside the module and represent the expression profiles of the entire module. Some genes may have high continuous module membership in two or more modules and may, in this sense, be considered members of (or intermediate between) several modules.

Module membership in consensus modules. In a consensus module analysis, we calculate the fuzzy module membership $MM$ for each gene in each data set. Thus, for each consensus analysis of three data sets there are three values for the module membership of each gene in each module. We then use meta-analysis to summarize the three module memberships into a single meta-analysis $Z$ statistic. Genes with the highest module membership meta-analysis $Z$ statistics are called consensus hub genes. It has been shown that consensus hub genes can be useful in studying functional categories associated with clinical traits.

Meta-analysis. Our analysis methods make extensive use of meta-analysis since we often pool association and module membership statistics across the three time points. A simple yet powerful meta-analysis method relies on combining the $Z$ statistics from individual data sets. Specifically, for each gene $i$ and data set $a$, one obtains a $Z$ statistic $Z_{ia}$ for example, by the inverse normal transformation of the $P$-value. Next, a meta-analysis $Z_i$ statistic for each gene is calculated as

$$Z_i = \frac{1}{\sqrt{N_{\text{sets}}}} \sum_{a=1}^{N_{\text{sets}}} Z_{ia}$$

The meta-analysis statistic $Z_i$ is approximately normally distributed with mean 0 and variance 1; the corresponding $P$-value is then calculated using the normal distribution.

Matching of genes across data sets and organisms. To compare gene expression between different sets, we used the following gene matching procedure. We first transformed all expression data to gene-level measurements. For microarray data where several probes or probe sets may represent a single gene, the probe-level data were turned into gene-level data using the function collapseRows. For Affymetrix and Illumina microarrays, we used the default settings of collapseRows (this implies selecting the probe with the highest mean expression as the representative for each gene); for two-color Agilent microarray data, we selected the most variant probe. For comparisons between different species, we mapped genes using the gene homology mappings provided by the Mouse Genome Informatics website, The Jackson Laboratory, Bar Harbor, Maine (http://www.informatics.jax.org/homology.shtml), retrieved April 2014.

Conversion between gene symbols and Entrez identifiers. We consistently used Entrez identifiers to unambiguously identify genes. Since many external resources identify genes by symbols, we used the Bioconductor package org.Mm.eg.db to convert gene symbols to Entrez identifiers. Typically not all gene symbols can be unambiguously mapped to an Entrez ID; genes with such ambiguous mappings were discarded. For example, from the top 100 ABA striatum markers, only 88 could be mapped to an unambiguous Entrez ID.

Ctcf enrichment analysis. A list of Ctcf target genes was created by scanning the canonical promoter region (1,000 bp upstream of the transcription start site) of *Mus musculus* RefSeq genes (assembly mm9) for the presence of Ctcf peaks in the “Cortex Adult 8 weeks CTCF TFB cis-Seq Peaks from ENCODE/LICR” UCSC track. For each of the 18 consensus top modules, overlap with Ctcf target genes was determined, as well as the overlap of randomly generated RefSeq gene lists (with the same number of genes in the module) with the Ctcf target genes (10,000 permutations). $Z$-score was calculated as the number of s.d. our observed value was above the mean of the observed values within the empirical null distribution.

Preservation studies in independent human and mouse data. To quantify module association with genotype or disease status in independent data, we downloaded the following eight human and four mouse data sets (sample numbers below reflect our outlier removal):

- Durarenberger (2011) CN (GSE26927). 19 human postmortem caudate nucleus samples, assayed on the Illumina HumanRef-8 v2.0 expression beadchip.
- Hodges (2006) CN, BA4, B9, CB (GSE3790). Human postmortem data from HD patients and controls from four different brain regions, assayed on the Affymetrix U133A and B microarrays.
- Harvard Brain Tissue Resource Center PFC, VC, CB. The roughly 800 individuals contributing to these data sets consist of approximately 400 with Alzheimer’s disease (AD), 230 with Huntington’s disease and 170 controls matched for age, gender and postmortem interval. The tissue specimens for this study were provided by the Harvard Brain Tissue Resource Center (HBTRC). Three brain regions (cerebellum (CB), visual cortex (VC) and dorsolateral prefrontal cortex (PFC) were profiled on a custom-made Agilent 44K microarray of 39,280 DNA probes uniquely targeting 37,585 known and predicted genes, including splice variants, miRNAs and high-confidence noncoding RNA sequences. Clinical outcomes available include age at onset, age at death, Braak scores (AD), Vonsattel scores (HD) and regional brain enlargement/atrophy. The data can be accessed at http://www.synapse.org/ under access code syn4505. An analysis of AD samples and controls has been reported previously; here we used only the HD and control (non-AD) samples.
- BACHD-ΔN17 (GSE64386). Striatum samples from 4 wild-type and 4 BACHD-ΔN17 mice at each of 2, 7 and 11 months of age. For our analysis, we discarded the 2-month data and treated the BACHD-ΔN17 7-month and 11-month mice as the transgenic group, with the 7- and 11-month wild-type samples as controls.
- R6/2 (GSE9857). Striatum samples from 9 mice expressing a short N-terminal fragment of mutant huntingtin (R6/2) and 9 wild-type controls assayed on the Affymetrix Mouse Genome 430 2.0 Array.
- Q150 (GSE32417). Striatal samples from 4 Hdh Q150 mice and 4 wild-type littermates at each of 6, 12 and 18 months of age, assayed on the Affymetrix Mouse Gene 1.0 ST array. In our analysis of preservation of association with genotype, we used only the 12- and 18-month samples.
- YAC128 (GSE18551). Striatal samples from 9 transgenic mice (4 at 12 months and 5 at 24 months) expressing human huntingtin with 120 CAG repeats (YAC128) and 9 wild-type littermates, assayed on the Affymetrix Mouse Genome 430 2.0 Array.

Preprocessing of independent data. For Illumina and Agilent data, we started our preprocessing from the author-normalized data available online; for Affymetrix data sets, we started from the raw data contained in CEL files and applied the robust multi-chip average (RMA) normalization method. We then applied the following steps: (1) removal of low-expression probes whose expression indices are likely mostly noise; (2) log-transformation of the expression data (only for data that had not been log-transformed yet), which makes the data more amenable to standard statistical analysis; (3) removal of potential outliers identified using sample network methodology; (4) adjustment for technical covariates, including batch effects, and for gender and age where necessary; (5) quantile normalization where there was evidence of significant correlation of quantiles of individual samples. Finally, to facilitate cross-platform and cross-organism comparisons, the probe-level data were turned into gene-level data using the function collapseRows. Although our aim was to preprocess all data sets in a uniform manner, the preprocessing differs by necessity somewhat from data set to data set depending on platform, availability of technical and biological sample information, sample numbers and other factors.
An independent cohort of 6-month Q175 npg applying a linear gradient with increasing acetonitrile concentration from 7% to 30% (GE Healthcare). Peptides were eluted and separated at a flow rate of 1 mL per minute 200 × 4.6 mm analytical column (Waters) operated with the Äkta Explorer system liquid nitrogen and lyophilized.

Proteomic sample preparation and profiling. Striata were lysed in 100 µl lysis buffer (8 M urea, 50 mM Tris-HCl pH 8.2, 75 mM NaCl, 5 mM EDTA, 5 mM EGTA, 10 mM sodium pyrophosphate, protease inhibitor cocktail Complete Mini (Roche, Mannheim, Germany). Tissues were crushed using forceps and extracts were sonicated using a beaker resonator BR30 (Bandelin Electronic, Germany). Samples were centrifuged at >14,000g, 4°C for 20 min to remove tissue and cell debris, and the protein concentration of the supernatant was determined by Bradford assay.

For each sample, 200 µg of protein was reduced with 10 mM dithiothreitol for 30 min and alkylated with 55 mM iodoacetamide for 30 min in the dark. Subsequently, the endoproteinase Lys-C (Wako) was added at an enzyme-to-substrate ratio of 1:200 and incubated for 4 h at room temperature. Samples were thereafter diluted 1:4 with 20 mM Tris-HCl pH 8.2 before adding trypsin (Promega) at an enzyme-to-substrate ratio of 1:100 followed by overnight incubation. The resulting peptide mixtures were acidified by addition of TFA to a final concentration of 0.5% and subsequently desalted using C18 Sep-Pak columns (100 mg sorbent weight, Waters). Peptides were eluted with 50% acetonitrile, 0.5% acetic acid, samples split into two aliquots (100 µg each), snap-frozen in liquid nitrogen and lyophilized.

Of each sample 100 µg of peptides were subsequently fractionated by high pH reverse-phase chromatography. Briefly, hypophylizes were reconstituted in 20 mM ammonium formate, pH 10 (buffer A), and loaded onto an XBridge C18 200 × 4.6 mm analytical column (Waters) operated with the Äkta Explorer system (GE Healthcare). Peptides were eluted and separated at a flow rate of 1 mL per minute applying a linear gradient with increasing acetonitrile concentration from 7% to 30% buffer B (buffer A supplemented with 80% acetonitrile) over 15 min followed by an increase to 55% over 5 min, a washing phase for 5 min at 100% and a final equilibration phase at 0% B for 10 min. The collected 18 fractions of eluting peptides were combined in a concatenated way to generate 6 fractions for each individual sample. Samples were frozen in liquid nitrogen and lyophilized. After desalting via C18 Sep-Pak columns (100 mg sorbent weight, Waters), samples fraction were snap-frozen, lyophilized, and reconstituted in 0.1% formic acid for mass spectrometry (MS).

Samples were loaded onto a reverse phase analytical column packed in-house with 1.9 µm C18 beads (Dr. Maisch, HPLC GmbH, Ammerbuch, Germany) by an EASY nLC1000 UPLC system (Thermo Fisher Scientific) at flow of 400 nL/min at 95% buffer A (5% DMSO, 0.1% formic acid). Peptides were resolved by a linear gradient over 96 min from 10% to 30% buffer (5% DMSO, 80% acetonitrile) followed by an increase over 50% B in 13 min to 60% B in 6 min. Eluting peptides were electrosprayed via a nanoelectrospray ion source into a Q Exactive mass spectrometer (Thermo Fisher Scientific). The Q Exactive mass spectrometer was operated in the data-dependent acquisition mode acquiring full scans at a resolution of 17,500 in the Orbitrap mass analyzer.

Proteomic sample preparation and profiling. Striata were lysed in 100 µl lysis buffer (8 M urea, 50 mM Tris-HCl pH 8.2, 75 mM NaCl, 5 mM EDTA, 5 mM EGTA, 10 mM sodium pyrophosphate, protease inhibitor cocktail Complete Mini (Roche, Mannheim, Germany). Tissues were crushed using forceps and extracts were sonicated using a beaker resonator BR30 (Bandelin Electronic, Germany). Samples were centrifuged at >14,000g, 4°C for 20 min to remove tissue and cell debris, and the protein concentration of the supernatant was determined by Bradford assay.

For each sample, 200 µg of protein was reduced with 10 mM dithiothreitol for 30 min and alkylated with 55 mM iodoacetamide for 30 min in the dark. Subsequently, the endoproteinase Lys-C (Wako) was added at an enzyme-to-substrate ratio of 1:200 and incubated for 4 h at room temperature. Samples were thereafter diluted 1:4 with 20 mM Tris-HCl pH 8.2 before adding trypsin (Promega) at an enzyme-to-substrate ratio of 1:100 followed by overnight incubation. The resulting peptide mixtures were acidified by addition of TFA to a final concentration of 0.5% and subsequently desalted using C18 Sep-Pak columns (100 mg sorbent weight, Waters). Peptides were eluted with 50% acetonitrile, 0.5% acetic acid, samples split into two aliquots (100 µg each), snap-frozen in liquid nitrogen and lyophilized.

Of each sample 100 µg of peptides were subsequently fractionated by high pH reverse-phase chromatography. Briefly, hypophylizes were reconstituted in 20 mM ammonium formate, pH 10 (buffer A), and loaded onto an XBridge C18 200 × 4.6 mm analytical column (Waters) operated with the Äkta Explorer system (GE Healthcare). Peptides were eluted and separated at a flow rate of 1 mL per minute applying a linear gradient with increasing acetonitrile concentration from 7% to 30% buffer B (buffer A supplemented with 80% acetonitrile) over 15 min followed by an increase to 55% over 5 min, a washing phase for 5 min at 100% and a final equilibration phase at 0% B for 10 min. The collected 18 fractions of eluting peptides were combined in a concatenated way to generate 6 fractions for each individual sample. Samples were frozen in liquid nitrogen and lyophilized. After desalting via C18 Sep-Pak columns (100 mg sorbent weight, Waters), samples fraction were snap-frozen, lyophilized, and reconstituted in 0.1% formic acid for mass spectrometry (MS).

Samples were loaded onto a reverse phase analytical column packed in-house with 1.9 µm C18 beads (Dr. Maisch, HPLC GmbH, Ammerbuch, Germany) by an EASY nLC1000 UPLC system (Thermo Fisher Scientific) at flow of 400 nL/min at 95% buffer A (5% DMSO, 0.1% formic acid). Peptides were resolved by a linear gradient over 96 min from 10% to 30% buffer (5% DMSO, 80% acetonitrile) followed by an increase over 50% B in 13 min to 60% B in 6 min. Eluting peptides were electrosprayed via a nanoelectrospray ion source into a Q Exactive mass spectrometer (Thermo Fisher Scientific). The Q Exactive mass spectrometer was operated in the data-dependent acquisition mode acquiring full scans at a resolution of 17,5000 and fragmentation spectra (MS/MS mode) of the ten most abundant peptide ions at a resolution of 17,500 in the Orbitrap mass analyzer.

Processing of MS data. All raw files acquired were processed with the MaxQuant software suite (version 1.5.2.10) using the Andromeda search engine for peptide and protein identification and quantification. The experiments were collectively searched against a Uniprot mouse database (version November 2014). Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionine and N-terminal acetylation were set as variable modifications. The minimum required peptide length was seven amino acids and up to two missed cleavages were allowed. A false discovery rate (FDR) of 1% was selected for both protein and peptide identifications and a posterior probability less than or equal to 0.01 for each peptide-to-spectrum match was required. The match between runs option was enabled for a time window of 0.5 min. For protein quantification, the MaxLFQ feature of MaxQuant was enabled and used with default parameters. In particular, the minimum LFQ ratio count was set to 2, as suggested by Cox et al.
**Unbiased stereology.** Unbiased stereological counting of the total numbers of NeuN+ neurons and GFAP+ astrocytes in striatum at 4.5 and 10 month in WT and Q175 heterozygous mice (N = 6 per age per genotype) were performed by MBF Labs (Williston, VT) using the optical fractionator method. Serial 60-m-thick m-thick sections of the striatum were prepared and every sixth section was stained free-floating with anti-NeuN (1:100,000, Millipore MAB377) followed by goat anti-rabbit (1:250, Vector BA-1000) for cell counting.

**Screening in Drosophila.** Mutant and overexpression alleles were obtained from the Bloomington Drosophila Stock Center at Indiana University (http://flystocks.bio.indiana.edu). The inducible shRNAs were obtained from the Vienna Drosophila Resource Center (http://stockcenter.vdrc.at/control/main/). Nervous system expression was achieved using elav-Gal4 (C155). The NT-HTT128Q (F33A) line used for this study has been previously reported and expresses human HTTN231Q128. Motor performance tests were carried out in a second batch (41 genotypes) against common controls; using two replicates of 15 age-matched females per genotype and 10 trials per replicate. The NT-HTT128Q model of Huntington's disease with 140 CAG repeats.

**HDinHD database access.** Server, data sets and network browser for the allelic series transriptome study and consensus network analyses are available at https://www.hdinhd.org/.

**Supplementary Methods Checklist** is available.

---

51. Jacobsen, J.C. et al. HD CAG-correlated gene expression changes support a simple dominant gain of function. *Hum. Mol. Genet.* **20**, 2846–2860 (2011).

52. Menalled, L.B., Sison, J.D., Dragatsis, I., Zeitiin, S. & Chesselet, M.F. Time course of early, mid and late phases of neurodegeneration in a knock-in mouse model of Huntington's disease with 140 CAG repeats. *J. Comp. Neurol.* **465**, 11–26 (2003).

53. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

54. Gu, X. et al. Unbiased stereological counting of the total numbers of NeuN+ neurons and GFAP+ astrocytes in striatum at 4.5 and 10 month in WT and Q175 heterozygous mice (N = 6 per age per genotype) were performed by MBF Labs (Williston, VT) using the optical fractionator method. Serial 60-m-thick m-thick sections of the striatum were prepared and every sixth section was stained free-floating with anti-NeuN (1:100,000, Millipore MAB377) followed by goat anti-rabbit (1:250, Vector BA-1000) for cell counting.

Screening in Drosophila. Mutant and overexpression alleles were obtained from the Bloomington Drosophila Stock Center at Indiana University (http://flystocks.bio.indiana.edu). The inducible shRNAs were obtained from the Vienna Drosophila Resource Center (http://stockcenter.vdrc.at/control/main/). Nervous system expression was achieved using elav-Gal4 (C155). The NT-HTT128Q (F33A) line used for this study has been previously reported and expresses human HTTN231Q128. Motor performance tests were carried out in a second batch (41 genotypes) against common controls; using two replicates of 15 age-matched females per genotype and 10 trials per replicate. The NT-HTT128Q model of Huntington's disease with 140 CAG repeats.

**HDinHD database access.** Server, data sets and network browser for the allelic series transcriptome study and consensus network analyses are available at https://www.hdinhd.org/.

**Supplementary Methods Checklist** is available.

51. Jacobsen, J.C. et al. HD CAG-correlated gene expression changes support a simple dominant gain of function. *Hum. Mol. Genet.* **20**, 2846–2860 (2011).