nature biotechnology

Large-scale phenome analysis defines a behavioral signature for Huntington's disease genotype in mice

Vadim Alexandrov[1,4], Dani Brunner[1,4], Liliana B Menalled[1,4], Andrea Kudwa[1], Judy Watson-Johnson[1], Matthew Mazzella[1], Ian Russell[1], Melinda C Ruiz[1], Justin Torello[1], Emily Sabath[1], Ana Sanchez[1], Miguel Gomez[1], Igor Filipov[1], Kimberly Cox[1], Mei Kwan[1], Afshin Ghavami[1], Sylvie Ramboz[1], Brenda Lager[2], Vanessa C Wheeler[3], Jeff Aaronson[2], Jim Rosinski[2], James F Gusella[3], Marcy E MacDonald[3], David Howland[2] & Seung Kwak[2]

Rapid technological advances for the frequent monitoring of health parameters have raised the intriguing possibility that an individual's genotype could be predicted from phenotypic data alone. Here we used a machine learning approach to analyze the phenotypic effects of polymorphic mutations in a mouse model of Huntington's disease that determine disease presentation and age of onset. The resulting model correlated variation across 3,086 behavioral traits with seven different CAG-repeat lengths in the huntingtin gene (*Htt*). We selected behavioral signatures for age and CAG-repeat length that most robustly distinguished between mouse lines and validated the model by correctly predicting the repeat length of a blinded mouse line. Sufficient discriminatory power to accurately predict genotype required combined analysis of >200 phenotypic features. Our results suggest that autosomal dominant disease-causing mutations could be predicted through the use of subtle behavioral signatures that emerge in large-scale, combinatorial analyses. Our work provides an open data platform that we now share with the research community to aid efforts focused on understanding the pathways that link behavioral consequences to genetic variation in Huntington's disease.

Genome-wide association studies uncover correlations between large numbers of genetic variants and single diseases or traits. A related approach, the phenome-wide association study, is designed to discover associations between many diseases and traits and single genetic variants, facilitating the investigation of pleiotropy (the effect of genetic variants on multiple phenotypic outcomes)[1]. However, phenome-wide association studies typically examine genotype–phenotype associations in isolation, and thus provide a limited view of the broad phenotypic consequences that can arise from a genetic mutation. Here we present a combinatorial method for analyzing phenotypic data that characterizes the simultaneous and often subtle changes that occur in a large number of traits as a result of genetic variation.

We developed this approach in the context of Huntington's disease (HD), which is caused by a dominant CAG (encoding glutamine (Q)) expansion in the huntingtin (*HTT*) gene[2]. Early symptoms of the disease include cognitive and psychiatric deficits, and these progress to chorea, dystonia, bradykinesia, dementia and, eventually, death. It is well established that the onset of diagnostic motor signs of HD and ultimate patient survival are inversely correlated with both CAG-repeat length and age[3,4], and that dysfunction and degeneration of corticostriatal circuits are involved in the observed symptoms[5,6]. However, the complex effects of age and CAG-repeat length on the resulting behavioral traits have not been previously characterized.

Simple behavioral traits have been used in ENU mutagenesis projects[7] to identify fully penetrant mutations, and standard behavioral tests have been used to study genetic polymorphisms in a series of expanded repeats[8]. More recently, machine learning techniques have been used to advance the investigation of phenotype–genotype relationships in autism spectrum disorder[9]. However, to our knowledge no previous study has developed a comprehensive and unbiased computational analysis to define predictive behavioral signatures for a disease genotype.

We integrated molecular, cellular and behavioral data sets to understand the relationship between HD genotype and a broad range of phenotypes. Using custom-built computer vision software and machine learning algorithms, we measured 3,086 behavioral phenotypes generated by comprehensive high-throughput devices—SmartCube, NeuroCube and PhenoCube—that analyze different behavioral domains such as cognitive, motor, circadian, social, anxiety-like and gait[10]. We studied these phenotypes in a biological series that captured the broad phenotypic effects of progressively longer CAG repeats in *Htt*, allowing for influence by age, using a congenic series of heterozygous (HET) *Htt* CAG-knock-in (KI) mice. We used a computational method based on support vector machines (SVMs) to analyze the large-scale phenotypic information generated by the three systems and selected the phenotypes that best distinguished mice with CAG repeats of different lengths. The final model, which incorporated ~200 behavioral features, accurately predicted the CAG-repeat length of a blinded mouse line. Our results demonstrate the potential to predict underlying disease mutations by measuring subtle variations at the level of behavioral phenotypes.

[1]PsychoGenics Inc., Tarrytown, New York, USA. [2]CHDI Management/CHDI Foundation, Princeton, New Jersey, USA. [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to V.A. (vadim.alexandrov@psychogenics.com).

Received 11 March 2015; accepted 28 April 2016; published online 4 July 2016; doi:10.1038/nbt.3587

This work forms part of the HD Mouse Allelic Series Project, which is generating and integrating multidimensional data sets from mouse studies to construct a computational model for phenotypic changes that occur as a consequence of polymorphic expansions in CAG repeats. The transcriptomic and proteomic data sets accompanying this study are publicly available[11] to the wider research community (http://www.HDinHD.org), and additional data will follow as they are generated.

RESULTS

Generation of a systems biology data set

We generated a large, content-rich behavioral data set using a series of HET *Htt* CAG-repeat-KI mice with a range of CAG repeat lengths, assessed at different ages. In humans, the 40–55-CAG range is associated with the onset of HD motor symptoms in midlife. Longer CAG repeats are associated with onset during adolescence, and >110 CAG repeats result in early (juvenile) onset and severe signs of disease[12–14]. We used an allelic series consisting of three HET *Htt* CAG-KI lines expressing CAG-repeat lengths in the higher range (HdhQ^{80}, HdhQ^{92} and HdhQ^{111}, where Q indicates an expected average glutamine tract length expressed from CAG codons; Online Methods)[15] and extended the upper range with the HET CAG140 (ref. 16) and zQ175 KI[17] lines. Whereas the CAG140 and zQ175 HET mice showed robust pathological phenotypes within their first year[18], the HdhQ^{111} HET mice had a mild behavioral phenotype[17] late in life, and HET mouse models with lower CAG-repeat lengths did not display overt behavioral deficits. We also examined the HdhQ^{50} line, the CAG length of which is more representative of clinically relevant repeats in humans. To test our algorithm's predictive abilities, we also included a KI line with a CAG-repeat length blinded to the team responsible for testing, data analysis and data modeling.

Study design

Because HD pathology almost always occurs in a genetic context that includes wild-type HTT function, we focused on HET CAG-KI mice in which only one *Htt* allele carries an expanded CAG segment. We first tested three cohorts of HET mutant mice from six *Htt* CAG-KI C57BL/6J lines and corresponding wild-type littermates (Online Methods, **Supplementary Tables 1** and **2**) in the PhenoCube, SmartCube and NeuroCube systems over two consecutive weeks (Online Methods). We used a factorial design to test all lines at all ages to investigate the independent and combined effects of these two factors (gene mutation and age) on behavioral outcomes. In the 6-month-old cohort we included a hypomorphic HdhQ^{50neo} mouse line[19], which has a neomycin (neo) selection cassette oriented such that expression of the mutant *Htt* allele is substantially reduced compared to that in HdhQ^{50} mice, to probe the model's ability to differentiate between the different mouse lines. In the CAG140 and zQ175 mice, the neo cassette was inserted in reverse orientation to the *Htt* transcriptional unit and did not have a notable effect on the expression of *Htt* mRNA or protein in zQ175 C57BL/6J mice.

In a second study we tested HET mice from three KI lines, namely, HdhQ^{20}, HdhQ^{50} and CAG140, and an Hdh KI line with a CAG repeat length that was not disclosed to the computational modeling team. This study focused on 6- and 10-month-old mice and followed the same experimental procedures described above. In both studies, once testing was completed, tissue was collected from skin, skeletal muscle, liver, gonads, heart, pancreas, kidneys, subcutaneous brown and white fat, striatum, cortex, cerebellum, hippocampus, brain stem, thalamus/hypothalamus, corpus callosum, urine and plasma. Experimenters were blinded to genotype in both studies.

Characterization of the *Htt* CAG-KI lines

qPCR analysis with endogenous *Htt*-allele-specific primers showed that levels of endogenous wild-type *Htt* mRNA were 40–60% lower in cortex in all HET CAG-KI lines compared to the wild-type, as expected (**Supplementary Fig. 1**; for all comparisons $P < 0.0001$, analysis of variance (ANOVA) with Bonferroni *post hoc* correction). RNA-seq analysis demonstrated significant CAG-dependent (Q20 through Q175) changes in the steady-state levels of total *Htt* in HET mice across all ages tested (2, 6 and 10 months; **Supplementary Fig. 2**; for all comparisons $P < 0.003$; ANOVA with Bonferroni *post hoc* correction; HdhQ^{50} was compared using *t*-test, $P < 0.007$, at 10 months), consistent with previous reports of CAG length affecting *HTT* mRNA levels[20]. As CAG-dependent variation in steady-state *HTT* mRNA levels is independent of age, any CAG- and age-dependent behavioral deficits that are identified are likely to be the result of downstream effects on the function of the mutant HTT protein.

Two-class bioinformatics analysis

We first assessed the differences between the HET and wild-type mice using supervised machine learning algorithms to analyze the set of 3,086 features collected from the three Cubes. To identify the feature combination that best separated the two classes (HET versus wild-type), we constructed decorrelated (i.e., statistically independent) combinations of the original features that we then ranked according to their discrimination power (Online Methods). To quantify separability and build a 'discrimination index', we measured the rank-mediated overlap between the 2D Gaussian estimates approximating the two groups in the newly formed decorrelated/ranked feature space. The difference between the zQ175 HET and wild-type mice as determined using data from all Cubes is shown in **Figure 1a**, with

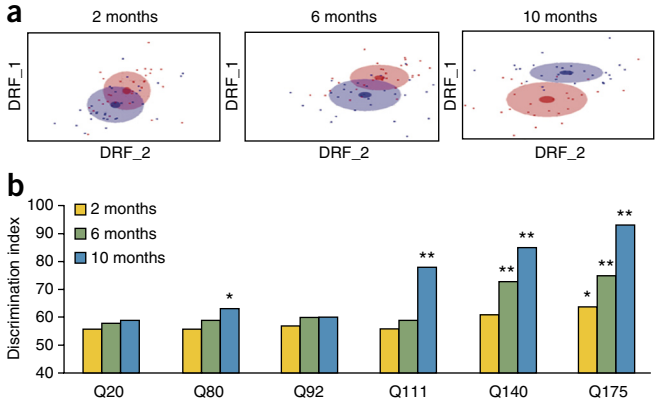


Figure 1 Discrimination between wild-type and HET mice. (**a**) Gaussian distribution of points corresponding to the HET mice from the zQ175 line (blue) as compared to wild-type littermate controls (red). The axes represent the two features that best separated the two samples. Data used were obtained with the three Cube platforms. The inner (darker) circles show the s.e., and the outer (lighter) circles show the s.d.; each individual point represents a single subject. DRF, decorrelated ranked feature. (**b**) Discrimination values for all CAG models against the corresponding WT controls, at all ages studied. Q20, HET mice from HdhQ^{20} line; Q80, HET mice from HdhQ^{80} line; Q92, HET mice from HdhQ^{92} line; Q111, HET mice from HdhQ^{111} line; Q140, HET mice from CAG140 KI line; Q175, HET mice from zQ175 line. The sample size for each group is presented in **Supplementary Table 2**. *P*-values (*$P < 0.05$; **$P < 0.0001$) were obtained from *t*-test calculations.

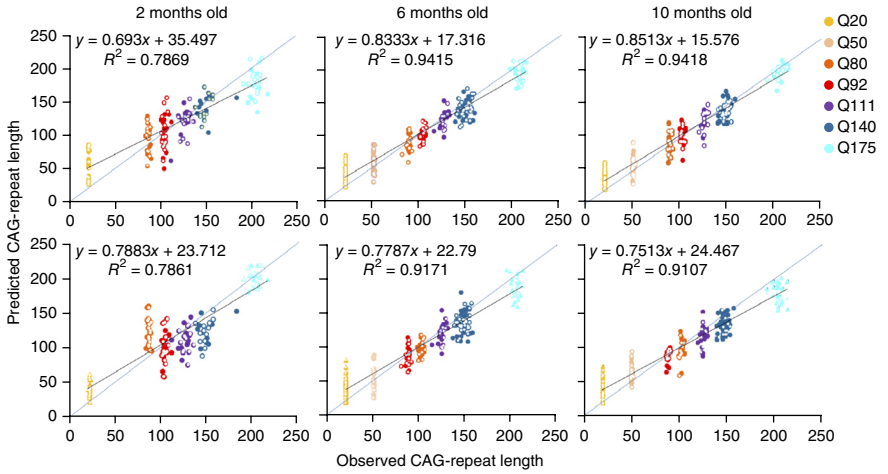
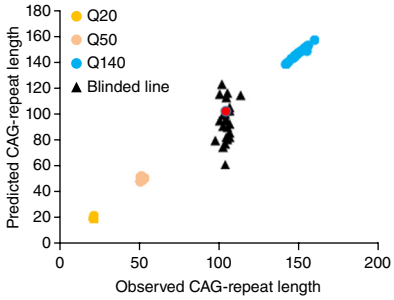


Figure 2 Performance of the CAG model during training and testing as assessed by regression on predicted versus observed CAG-repeat length. Open and closed symbols correspond to female and male mice, respectively. Top: the performance of the CAG models during prediction of CAG-repeat length for one example (LOOCV) not included in the training set was relatively good (both sexes combined). R^2 increased from 0.8 (in 2-month-old mice) to 0.9 (in 10-month-old mice), indicating increasing CAG signal strength with advancing age. The same was found for the observed-versus-predicted regression slope (regression lines shown in black, identity lines shown in blue). Performance of the CAG model built separately on male and female data sets required up to 50% fewer features to yield similar goodness of fit (R^2 ~ 0.9), suggesting that data sets for each sex were more uniform than the combined set. Bottom: to challenge the combined-sex model in a more stringent manner, we left out a whole line and trained with the remaining lines. The scattergrams show the observed-versus-predicted results for the predicted lines. The slopes and R^2 values remained comparable with those for the LOOCV example. Q20, HET mice from $HdhQ^{20}$ line; Q50, HET mice from $HdhQ^{50}$ line; Q80, HET mice from $HdhQ^{80}$ line; Q92, HET mice from $HdhQ^{92}$ line; Q111, HET mice from $HdhQ^{111}$ line; Q140, HET mice from CAG140 KI line; Q175, HET mice from zQ175 line.

the corresponding Gaussian distributions (clouds) projected onto the first two highest-ranked new feature axes. The overlap between clouds was significant at 2 months of age but decreased with advancing age, indicating (as expected) a progressive phenotype. The top features separating 6-month-old zQ175 HET mice from wild-type controls included a cluster that indicated decreased abrupt movements in response to startling stimuli and reduced startle response (clustered features in position 1 and 10 out of 62 clusters, respectively), and a cluster including measures of mobility, rearing, digging and complexity of trajectory (cluster 3 out of 62). Notably, the changes in these features were in the same direction in the 2- and 10-month-old mice. Indeed in the data set for 10-month-old zQ175 mice, reduced startle and associated measures ranked at the top, and the same cluster representing reduced mobility, rearing, digging and trajectory complexity ranked number 3. This suggests that there is a phenotype continuum consisting of similar deficits in the early and late disease stages in these lines, and that the machine learning algorithms were not simply finding arbitrary fluctuations in the data that happened to provide good class separation. We aimed to capture this continuum with the multiclass modeling described below.

The two-class discrimination index was calculated for all KI lines at all ages versus their corresponding wild-type controls. We also repeated the calculations using randomized labels to calculate the distribution of spurious discrimination indexes and a P value for the probability that was due to chance (Online Methods). CAG-repeat length and age were positively associated with better discrimination (**Fig. 1b**).



Multi-class bioinformatics analysis: CAG and age effects

We next sought to determine whether a continuum exists among CAG-repeat sizes in subjects of a given age. We analyzed all behavioral data (3,086 features) from the different CAG lines and from mice of all ages, using a probabilistic SVM (**Fig. 2**; Online Methods). This algorithm finds the vectors in hyperspace that best define separation between classes, in this case the different KI groups within each age class or the different age classes independent of CAG length (discussed below). We combined the SVM with a feature-ranking-and-selection algorithm to reduce dimensionality and to identify and more heavily weigh the top features that contributed most to the CAG signatures.

Some mice (<5% of total) were not recognized by their own CAG classifier, and these intrinsically inconsistent samples[21] were excluded from all subsequent model training or cross-validation. We trained the algorithm on the complete data set comprising the six lines at the three ages tested, using the wild-type mice to identify and remove those features that reflected intertrial variability. To test the model's predictive power, we used the trained classifier to determine the CAG-repeat lengths of individual mice excluded from the training set (leave-one-out cross-validation (LOOCV[22])). Accuracy was still high (**Fig. 2**), indicating that the CAG behavioral model successfully captured CAG signatures. To assess predictive ability, we trained with five out of six lines and then predicted the CAG length of the sixth line left out after features correlation and ranking (Online Methods). We repeated this process for each of the six lines; although accuracy was lower at this step than for the LOOCV test set, the CAG model was still able to predict the CAG length, in particular for the 50–140-CAG range in the older mice (**Fig. 2**). These analyses show that the CAG

Figure 3 Prediction of the 'blinded line' by the SVR CAG model (10-month-old mice). The algorithm was trained on the features collected from all HET animals except for the blinded line. Selected features maximized cross-validation for performance. As expected, animals from all lines in the training set were perfectly recognized and thus lie along the identity line in the graph. The distribution of the predicted CAG-repeat lengths for the blinded line had a mean of 96.1 and s.e.m. of 3.1. It was revealed later to us that the line was $HdhQ^{92}$ with an actual average CAG-repeat length of 102.7 (red dot).

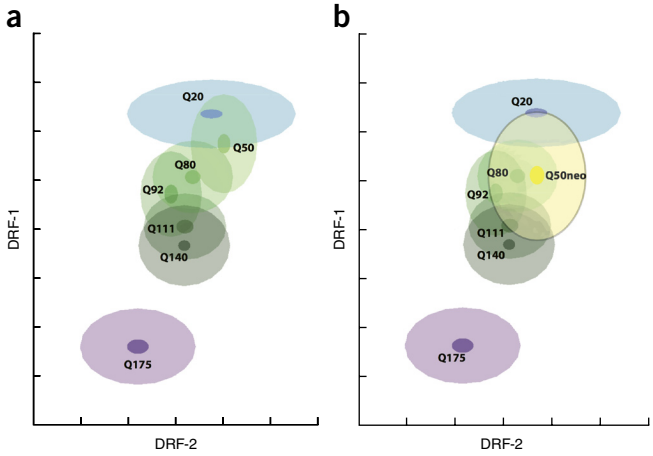


Figure 4 Projection of all Q lines onto the decorrelated ranked feature (DRF) plane defined by Q20 and Q175 lines at 6 months of age. The blue cloud represents the combined Q20 lines. The purple cloud represents the Q175 line. These two clouds define the DRF plane (the coordinate system formed from the original features in such a way that it separates the two clouds, purple and blue, the most). The rest of the data (from the other lines) are projected onto this Q20–Q175 DRF plane. The Q140 cloud is composed of data from the mice tested in both studies separated by over a year. (**a**) All Q lines perfectly line up in this coordinate system (including Q50 KI). (**b**) Q50neo, when projected (instead of Q50 KI) to this coordinate plane, does not line up with the other lines and occupies a much larger feature space compared to the other lines.

model successfully captured a CAG-length-dependent signal across the HET *Htt* CAG-KI mice. As expected, the predictive power of the CAG model was completely lost when the classifier was trained on samples with randomized labels.

We further validated our model by predicting the CAG-repeat length of a new mouse line with a blinded genotype. After analysis we determined a mean CAG length of 93.4 (with s.e. of 8.7) and 96.1 (s.e. of 3.1) for 6- and 10-month-old mice in this line, respectively. **Figure 3** shows the support vector regression (SVR) prediction for the blinded line at 10 months of age. This prediction was in close agreement with the actual mean CAG-repeat length, which was revealed to be 102.0 and 102.7 for the 6- and 10-month-old mice, respectively (**Supplementary Table 1**).

Finding features that predicted CAG length was not trivial, requiring 200–500 behavioral features (**Supplementary Fig. 3**). Using more than 700 features resulted in lower performance, probably as a result of overfitting and reliance on random feature variations that were poor predictors of the left-out example for the LOOCV task.

Testing HdhQ^{50} and HdhQ^{50neo}

We checked our approach for consistency by projecting all KI lines, together with the HdhQ^{50} line, onto the decorrelated ranked feature analysis plane (Online Methods) that best separated the zQ175 line from the first study and the pooled HdhQ^{20} control mice from both studies (to remove study-to-study variability). We monotonically aligned the HdhQ^{50} and other KI lines from lowest to highest (**Fig. 4a**).

When we used the HdhQ^{50neo} line instead, its projected feature values fell out of order (**Fig. 4b**; also note the large variability), indicating that this line differs substantially from the HdhQ^{50} and other lines.

Features that define the CAG signature

The top features selected by the feature-ranking algorithm for each of the three ages studied (2, 6 and 10 months) reflected the emerging phenotypic deficit. In 2-month-old mice, the top behavioral features were subtly different, particularly at low CAG-repeat lengths. For instance, both activity during the light phase and rearing were inversely proportional to CAG-repeat length (**Fig. 5a**). None of these changes alone was enough to distinguish the different lines; only the combined power of these features allowed discrimination at such an early age.

In 6-month-old mice, five of the top ten features describing the CAG-dependent signature were more progressively altered from low to high CAG-repeat length. Time huddling together was higher in the low-CAG-repeat than in the high-CAG-repeat lines, and a cognitive measure (win-shift) was lower in the high-CAG-repeat lines (PhenoCube). The temporal complexity of the locomotor trajectory, the same feature that ranked at the top in the two-class analysis, was lower in the high-CAG-repeat lines. The latency to stretch attend

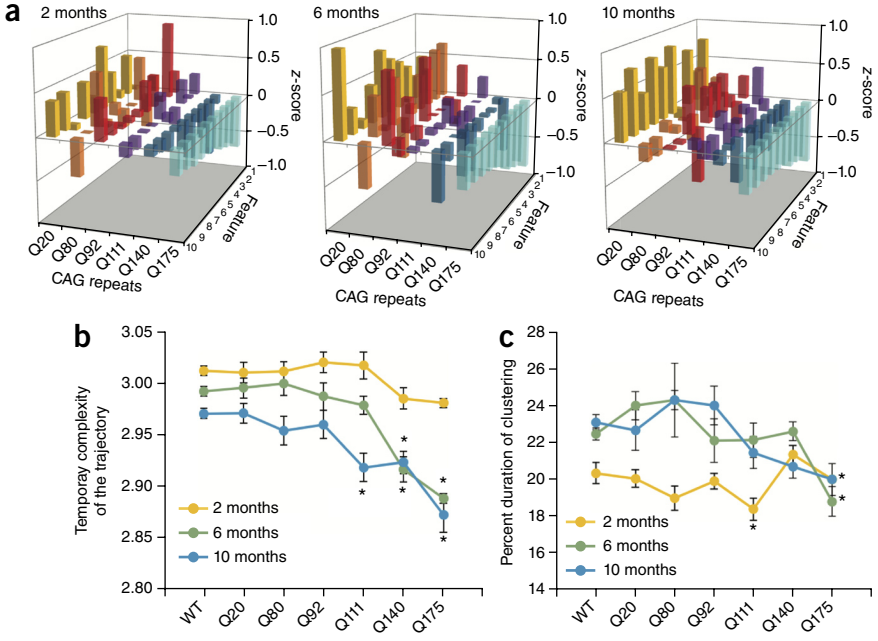


Figure 5 Top-feature score changes across different CAG-repeat lengths and ages. (**a**) Using Q175 as a guide, we chose 10 features with smaller Q175 *z*-scores from among the top 50 features for each age. The same ten features are shown for each of the CAG-repeat KI lines in each panel; from 1 to 10, trajectory complexity, clustering, rearing, huddling, cognitive (win-shift), gait (base width), stretch-attend, mobility, activity and transitions to activity. Very low Q175 *z*-scores tended to grow larger as the CAG-repeat length decreased; this pattern was more robust in older animals. (**b**,**c**) Complexity of the trajectory in time (**b**) and percent duration of clustering (**c**) for the three age groups as a function of CAG-repeat length. Asterisks denote significant differences as compared to the wild-type (WT) line (CAG main effect ANOVA, $P < 0.0001$; Bonferroni *post hoc* tests, $P < 0.0083$). Data are presented as group means ± s.e.m. The exact sample size for each group is presented in **Supplementary Table 2** ($n = 21$–32 in HET groups, $n = 183$–224 in wild-type groups).

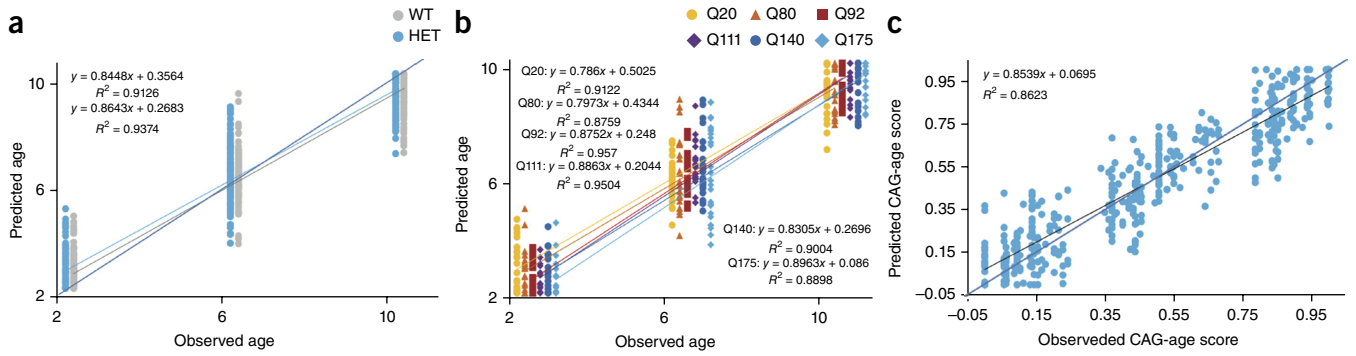


Figure 6 LOOCV performance of the age model during training and testing as assessed by regression on the predicted versus observed age. (**a**) The performance during the prediction of age for the examples not included in the training set for wild-type (WT) and HET mice was relatively good. The slopes of the regression lines were higher than 0.8, and R^2 was above 0.9 (regression lines shown in gray and sky-blue for WT and HET, respectively; identity line shown in dark blue). (**b**) Predicting age for each KI line separately yielded better regressions for the high-CAG-repeat lines, with slopes increasing from about 0.8 to 0.9 with increasing CAG-repeat length, and R^2 around 0.9. (**c**) Performance of the CAG–age model during LOOCV as assessed by regression on the predicted versus observed CAG–age combined score. The performance of the model during prediction of CAG–age for examples not included in the training set was relatively good. The slope of the regression and R^2 were both above 0.8 (regression line shown in black, identity line shown in blue).

(forward elongation of head and shoulders, followed by retraction to original position) was longer in the low-CAG-repeat lines (SmartCube) (**Fig. 5a**).

In 10-month-old mice, the top features showed a clear pattern of decreasing change from low to high CAG-repeat length, although not every individual feature precisely conformed to the pattern. Huddling time was higher for low-CAG-repeat lines than for high-CAG-repeat lines (PhenoCube), and rearing was robustly decreased in high-CAG-repeat versus low-CAG-repeat lines. Transitions between moving and scanning (horizontal head movements) were higher in low-CAG-repeat lines (SmartCube). Base width (side-to-side distance between the lines joining the two paws) was wider in the high-CAG-repeat lines, suggesting the appearance of a gait deficit (NeuroCube) (**Fig. 5a**).

Feature examples

We selected one top feature from the two- and multi-class analyses to plot individually as an example. The complexity of the trajectory is a measure of the convolutedness of the trajectory in time, that is, how much a mouse twists and turns over time. This feature showed a clear progression from low to high CAG-repeat length in the multiclass analysis (**Fig. 5b**, 6 months of age, feature 1), with complexity decreasing with both age and CAG-repeat length. These changes revealed significant effects of CAG-repeat length at each age ($P < 0.0067$; ANOVA with Bonferroni *post hoc* correction) mainly due to the high-CAG-repeat lines. This supports the view that a combination of more than 200 features, instead of a small subset, is required to capture the CAG-dependent behavioral effects for the low and middle CAG-repeat lengths.

The second example, a feature identified by the multi-class analysis as particularly important for the 6-month-old classification, is the duration of clustering (**Fig. 5c**), where clustering is registered when two or more mice are together for more than 1 s and less than 1 min. Although the data were noisier for 10-month-old mice and showed no difference for 2-month-old mice, mice with higher CAG-repeat lengths stayed together for a shorter time than mice with low CAG-repeat lengths did in the 6- and 10-month-old groups ($P < 0.02$; ANOVA with Bonferroni *post hoc* correction).

Age signature and defining features

We also trained our classifiers using age as the independent variable separately for the wild-type and HET mice (**Fig. 6a**,**b**). Accuracy, as measured by the coefficient of correlation of the regression line connecting the observed versus the predicted CAG-repeat lengths (R^2), was high for both models: 0.91 for HET and 0.94 for wild-type (**Fig. 6a**). **Figure 6b** shows the performance of the HET-only age model separately for each line. In general, age was predicted slightly more accurately in higher-CAG-repeat lines.

HD and aging

We performed several overlap analyses of features that best modeled age and CAG-repeat length. The overlap among features for wild-type-only and HET-only age models was about 50% (**Supplementary Fig. 4a**), suggesting that age-specific features were captured regardless of phenotype and that aging occurs differently in wild-type mice compared to HET mice, although many features contribute to both.

We also calculated overlaps among features modeling CAG-repeat length for each age and overlaps between these features and age features (**Supplementary Fig. 4b**). More than half (140–142 out of 250) of the best CAG-only modeling features for each age overlapped with age-only modeling features, indicating that (1) some features describe both age and CAG-repeat length and (2) there exist strictly CAG-only (i.e., totally age-unrelated) features for a given age. Notably, features best modeling CAG repeats for each particular age changed over time; that is, there was only ~20% overlap between each CAG feature for any two ages, and overlap between all three CAG models was very small (about 1%). Furthermore, we found no features that were shared simultaneously among all of the CAG and age models, suggesting that HD is not simply an accelerated aging process.

Modeling *Htt* CAG-repeat length and age simultaneously

As age and CAG-repeat length are factors with both common and independent features, we sought to evaluate the combined power of CAG-repeat length and age to enable simultaneous analysis of phenotypic changes in these two dimensions. Such a model could then capture changes in each of these two factors at once without missing their interaction, as single-factor models might. To predict both CAG-repeat length and age with our SVR model, we built a correspondence map by connecting the two-dimensional CAG–age pairs to a one-dimensional dependent variable (Online Methods). Cross-validation results are shown in **Figure 6c**. Unlike the age model (**Fig. 6a**,**b**) that almost perfectly discriminated animals by age, the

CAG–age model suggested that there is substantial overlap in behavioral patterns for older low-CAG-repeat and younger high-CAG-repeat mice. This algorithm allows for the incorporation of future data from other age × CAG model studies and will enable any given perturbation or treatment to be compared among all three models (CAG-dependent, age-dependent and simultaneous interactions). This could aid in the interpretation of behavioral signatures in terms of effects along age and CAG axes. For example, a treatment that shifts Q175 mice toward a younger age profile without affecting the CAG profile in the current model might be affecting compensatory networks underlying the aging process, which could be confirmed in the age-only model. Conversely, an anti-CAG treatment that shifts the signature toward a lower CAG-length profile but does not affect the age-only model might be affecting proximal HD causal networks.

DISCUSSION

We used a computational approach to describe an extended phenotype —complex changes in behavior—that is capable of distinguishing subtle differences in an underlying pathogenic repeat polymorphism. Our analysis provides a comprehensive behavioral characterization of the CAG-expanded HET KI lines in our allelic series[23–26]. Both the two-class and the multi-class models performed well in identifying CAG-repeat length. The combined CAG–age model seemed to capture the CAG signal even when age was integrated into the CAG dimension, suggesting that HD is not simply accelerated as a result of aging. The top features that captured aspects of the CAG signature in the two- and multi-class analyses included, not surprisingly, decreased mobility at higher CAG repeats. We found that mice with increasing CAG-repeat length spent less time huddling during the dark phase of the diurnal cycle and had decreased trajectory complexity (the extent to which the animal twists and turns). To our knowledge, the former observation has not been reported previously in a KI line, and it is tempting to speculate that this might reflect the social deficits observed in some cases of HD; it also suggests that this feature could be an indirect measure of dopamine depletion in the striatum. The latter observation is reminiscent of amphetamine action on increased locomotion complexity[27]. The model suggests that none of these features alone is sufficient to accurately separate mice according to either CAG-repeat length or age. Rather, it is the combination of more than 200 features that provided sufficient discriminatory power to accurately predict the CAG-repeat length of a blinded mouse line.

It is noteworthy that although *Htt* mRNA levels are comparatively decreased in higher-CAG-repeat KI mouse lines, independent of age, here behavioral CAG signatures emerged in an age-dependent manner. This suggests that the behavioral phenotype is driven by CAG-repeat-length effects that do not include reduced levels of *Htt* mRNA. We tested this hypothesis using a hypomorphic HdhQ^{50neo} HET KI mouse line that expresses approximately half the total amount of *Htt* mRNA as the wild type[19]. This line did not exhibit a behavioral signature consistent with the CAG-repeat-dependence signal captured by our CAG model, supporting the view that loss of function due to reduced HTT levels is not in itself sufficient to explain the behavioral signatures observed in this study. Previous work with a cellular allelic KI series revealed a CAG signature consisting of 73 CAG-length-dependent gene expression profiles involved in 172 CAG-length-correlated pathways[15]. Here we have extended that work to a CAG-correlative approach inclusive of downstream functional effects using a comprehensive behavioral battery and careful experimental design.

The multiparametric phenotyping platform we describe has distinct advantages over other behavioral assessment tools. It generates high-content data sets comprising thousands of features that can be mined using machine learning algorithms, it is unbiased and captures every measurable behavior through computer vision algorithms, it eliminates human intervention and subjectivity, and it offers high-throughput drug screening for *in vivo* CNS activity[28] using a phenotypic approach[29]. Furthermore, this system could be used to estimate a drug's ability to rescue a disease phenotype while assessing its side effects[10], or to discover new and unexpected animal model phenotypes[30]. One limitation of the approach is that it excludes tests of higher cognitive functions (e.g., learning and memory) that require training in multiple sessions over time but may be important for translation to the clinic[31].

SVM and SVR approaches have been successfully used in the past to relate genotype and phenotype features in neurodevelopmental disorders[9]. Our SVM/SVR approach incorporates an inherent feature-ranking mechanism that selects the most relevant features for the algorithm and appropriately weights their specific contributions to the prediction of the dependent variable. In the analysis of high-content data sets where the total number of collected features far exceeds the number of samples—a classical systems biology problem—inclusion and/or equal weighting of all features' contributions usually makes machine learning models overfitted and unstable, whereas exclusion of tangentially relevant features inevitably leads to the loss of relevant information. Because our method does not require the inclusion of noisy or irrelevant features, it makes it possible to build good SVM models from data sets containing a virtually unlimited number of features of different origins.

Notably, we now have phenotypic measures to test moderate CAG-repeat-length models (such as Q50, Q80, Q92 and Q111) *in vivo*, which may offer new drug-testing strategies not dependent on aggressive, fast-progressing HD mouse models such as the R6/2 and zQ175 lines that may have narrow pharmacological sensitivity. Most of the studied pharmacological targets for HD have produced preclinical results in such mice that are difficult to reproduce[32–34] or have led to unsuccessful clinical trials[35–38]. Although some drug targets have been identified[39–43], they have been unsuccessful in slowing progression of the disease. Future studies could potentially use the CAG-dependent extended phenotype to test compounds in both mild and severe HD mouse models and interpret their effect across increasing genotoxic stress levels (i.e., CAG-repeat length). The computational models described here might also be used to quantify the effect of a therapy that may, for example, reduce the CAG-repeat profile from high to low or the age profile of a high-CAG-repeat model from old to young, or any interaction between these two axes.

The HD Mouse Allelic Series Project is providing coherent *in vivo* data sets that recapitulate some of the complexity of HD in an intact organism, including the dynamic effects of aging. We are developing a crowdsourcing platform (http://www.HDinHD.org) that the research community can use to mine and model these data sets and, ultimately, identify novel and robust drug-development opportunities for the treatment of HD.

METHODS

Methods and any associated references are available in the online version of the paper.

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

ACKNOWLEDGMENTS

We are grateful to E. Leahy, V. Rivera, J. Rivera, K. Cheng, D. Lignore and L. Homa for their assistance. We thank S. Noble, D. Baker and J.-M. Lee for their thoughtful comments. This work was supported by CHDI Foundation, Inc., a nonprofit

biomedical research organization exclusively dedicated to developing therapeutics that slow the progression of Huntington's disease. CHDI Foundation conducts research in a number of different ways; for the purposes of this manuscript, all research was conceptualized, planned and directed by all authors listed and conducted under a fee-for-service agreement at the contract research organization PsychoGenics, Inc.

AUTHOR CONTRIBUTIONS

V.A. developed and tested all machine learning models described in the paper. D.B., D.H., S.K., J.G., M.E.M., V.W. and L.M. designed the study. L.B.M. identified the spontaneous expansion giving origin to the Q175 line and managed further colony expansion. D.B. and L.B.M. managed all study performance and two-class data analysis. M.E.M. and V.W. created the $HdhQ^{20}$, $HdhQ^{80}$, $HdhQ^{92}$ and $HdhQ^{111}$ lines. A.K. managed tissue collection and some of the behavioral studies. J.W.-J. and M.C.R. managed animal care. M.M. monitored data collection and two-class analysis. M.M., J.T., E.S. and K.C. performed behavioral studies. I.R. developed the database used to track all information and helped manage data handling. A.S. and M.G. managed the health care and daily maintenance of the animals. I.F. developed the computer vision software. M.K. and A.G. performed and managed, respectively, RNA studies. S.R. provided general management. B.L. managed the breeding of all animals under study. J.A. and J.R. provided feedback on the application of the multi-class model to this study.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Denny, J.C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
2. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
3. Langbehn, D.R., Hayden, M.R. & Paulsen, J.S. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **153B**, 397–408 (2010).
4. Craufurd, D. & Dodge, A. Mutation size and age at onset in Huntington's disease. *J. Med. Genet.* **30**, 1008–1011 (1993).
5. Raymond, L.A. *et al.* Pathophysiology of Huntington's disease: time-dependent alterations in synaptic and receptor function. *Neuroscience* **198**, 252–273 (2011).
6. Kanazawa, I. *et al.* Studies on neurotransmitter markers and striatal neuronal cell density in Huntington's disease and dentatorubropallidoluysian atrophy. *J. Neurol. Sci.* **70**, 151–165 (1985).
7. Justice, M.J., Noveroske, J.K., Weber, J.S., Zheng, B. & Bradley, A. Mouse ENU mutagenesis. *Hum. Mol. Genet.* **8**, 1955–1963 (1999).
8. Suzuki, K. *et al.* DRPLA transgenic mouse substrains carrying single copy of full-length mutant human *DRPLA* gene with variable sizes of expanded CAG repeats exhibit CAG repeat length- and age-dependent changes in behavioral abnormalities and gene expression profiles. *Neurobiol. Dis.* **46**, 336–350 (2012).
9. Bruining, H. *et al.* Behavioral signatures related to genetic disorders in autism. *Mol. Autism* **5**, 11 (2014).
10. Alexandrov, V., Brunner, D., Hanania, T. & Leahy, E. High-throughput analysis of behavior for drug discovery. *Eur. J. Pharmacol.* **750**, 82–89 (2015).
11. Langfelder, P. *et al.* Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. *Nat. Neurosci.* **19**, 623–633 (2016).
12. Harper, P.S. *Huntington's Disease* (W.B. Saunders, London, 1996).
13. Vonsattel, J.P. & DiFiglia, M. Huntington disease. *J. Neuropathol. Exp. Neurol.* **57**, 369–384 (1998).
14. Wojaczyńska-Stanek, K., Adamek, D., Marszal, E. & Hoffman-Zacharska, D. Huntington disease in a 9-year-old boy: clinical course and neuropathologic examination. *J. Child Neurol.* **21**, 1068–1073 (2006).
15. 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).
16. Menalled, L.B., Sison, J.D., Dragatsis, I., Zeitlin, S. & Chesselet, M.F. Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. *J. Comp. Neurol.* **465**, 11–26 (2003).
17. Menalled, L.B. *et al.* Comprehensive behavioral and molecular characterization of a new knock-in mouse model of Huntington's disease: zQ175. *PLoS One* **7**, e49838 (2012).
18. Farrar, A.M. *et al.* Cognitive deficits in transgenic and knock-in HTT mice parallel those in Huntington's disease. *J. Huntingtons Dis.* **3**, 145–158 (2014).
19. White, J.K. *et al.* Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. *Nat. Genet.* **17**, 404–410 (1997).
20. Dragatsis, I. *et al.* CAG repeat lengths > or = 335 attenuate the phenotype in the R6/2 Huntington's disease transgenic mouse. *Neurobiol. Dis.* **33**, 315–330 (2009).
21. Meghanathan, N., Nagamalai, D. & Chaki, N. *Advances in Computing and Information Technology* Vol. 177 (Springer, 2013).
22. Geisser, S. *Predictive Inference: An Introduction* (Springer Science + Business Media, 1993).
23. Trueman, R.C., Jones, L., Dunnett, S.B. & Brooks, S.P. Early onset deficits on the delayed alternation task in the Hdh(Q92) knock-in mouse model of Huntington's disease. *Brain Res. Bull.* **88**, 156–162 (2012).
24. Trueman, R.C., Brooks, S.P., Jones, L. & Dunnett, S.B. Rule learning, visuospatial function and motor performance in the Hdh(Q92) knock-in mouse model of Huntington's disease. *Behav. Brain Res.* **203**, 215–222 (2009).
25. Trueman, R.C., Brooks, S.P., Jones, L. & Dunnett, S.B. Time course of choice reaction time deficits in the Hdh(Q92) knock-in mouse model of Huntington's disease in the operant serial implicit learning task (SILT). *Behav. Brain Res.* **189**, 317–324 (2008).
26. Brooks, S., Higgs, G., Jones, L. & Dunnett, S.B. Longitudinal analysis of the behavioural phenotype in Hdh(Q92) Huntington's disease knock-in mice. *Brain Res. Bull.* **88**, 148–155 (2010).
27. Geyer, M.A., Russo, P.V. & Masten, V.L. Multivariate assessment of locomotor behavior: pharmacological and behavioral analyses. *Pharmacol. Biochem. Behav.* **25**, 277–288 (1986).
28. Houghten, R.A. *et al.* Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing in vivo, and enhanced deconvolution by computational methods. *J. Comb. Chem.* **10**, 3–19 (2008).
29. Roberds, S.L., Filippov, I., Alexandrov, V., Hanania, T. & Brunner, D. Rapid, computer vision-enabled murine screening system identifies neuropharmacological potential of two new mechanisms. *Front. Neurosci.* **5**, 103 (2011).
30. Oakeshott, S. *et al.* Circadian abnormalities in motor activity in a BAC transgenic mouse model of Huntington's disease. *PLoS Curr.* **3**, RRN1225 (2011).
31. Oakeshott, S. *et al.* A mixed fixed ratio/progressive ratio procedure reveals an apathy phenotype in the BAC HD and the z_Q175 KI mouse models of Huntington's disease. *PLoS Curr.* **4**, e4f972cffe982c970 (2012).
32. Wood, N.I., Pallier, P.N., Wanderer, J. & Morton, A.J. Systemic administration of Congo red does not improve motor or cognitive function in R6/2 mice. *Neurobiol. Dis.* **25**, 342–353 (2007).
33. Brunner, D., Balci, F. & Ludvig, E.A. Comparative psychology and the grand challenge of drug discovery in psychiatry and neurodegeneration. *Behav. Processes* **89**, 187–195 (2012).
34. Menalled, L.B. *et al.* Comprehensive behavioral testing in the R6/2 mouse model of Huntington's disease shows no benefit from CoQ10 or minocycline. *PLoS One* **5**, e9793 (2010).
35. Huntington Study Group. A randomized, placebo-controlled trial of coenzyme Q10 and remacemide in Huntington's disease. *Neurology* **57**, 397–404 (2001).
36. Keene, C.D. *et al.* A patient with Huntington's disease and long-surviving fetal neural transplants that developed mass lesions. *Acta Neuropathol.* **117**, 329–338 (2009).
37. Bezprozvanny, I. The rise and fall of Dimebon. *Drug News Perspect.* **23**, 518–523 (2010).
38. Huntington Study Group TREND-HD Investigators. Randomized controlled trial of ethyl-eicosapentaenoic acid in Huntington disease: the TREND-HD study. *Arch. Neurol.* **65**, 1582–1589 (2008).
39. Subramaniam, S. & Snyder, S.H. Huntington's disease is a disorder of the corpus striatum: focus on Rhes (Ras homologue enriched in the striatum). *Neuropharmacology* **60**, 1187–1192 (2011).
40. Williams, A. *et al.* Novel targets for Huntington's disease in an mTOR-independent autophagy pathway. *Nat. Chem. Biol.* **4**, 295–305 (2008).
41. Giorgini, F., Guidetti, P., Nguyen, Q., Bennett, S.C. & Muchowski, P.J. A genomic screen in yeast implicates kynurenine 3-monooxygenase as a therapeutic target for Huntington disease. *Nat. Genet.* **37**, 526–531 (2005).
42. Stone, T.W. & Darlington, L.G. Endogenous kynurenines as targets for drug discovery and development. *Nat. Rev. Drug Discov.* **1**, 609–620 (2002).
43. Hu, G. & Agarwal, P. Human disease-drug network based on genomic expression profiles. *PLoS One* **4**, e6536 (2009).

ONLINE METHODS

Ethics statement. 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).

Subjects. We tested HET mutant mice from six KI lines and corresponding wild-type littermates. For each one of the six lines, male HET mice were crossed with C57BL/6J female mice at JAX. For each line, animals born within 3–4 d from litters having 4–8 pups were identified by ear tags, tail-sampled for genotyping and weaned at around 3 weeks of age, and re-genotyped after tissue collection. At 5 weeks of age, the selected experimental mice were re-housed in groups of ten according to the procedure described below. All lines were backcrossed over ten generations to the C57BL/6J strain.

Subject selection. Experimental animals were selected according to the following guidelines: No more than one animal per sex per genotype was selected from each litter. Animals originating from litters that could contribute to the experimental group with four animals were preferred over litters contributing three, which in turn were preferred over litters contributing two animals. HET mice were selected on the basis of CAG repeats to allow a Gaussian distribution of CAG repeats in the experimental cohort, to avoid skewed distributions. Best Gaussian fit was judged by eye (**Supplementary Table 1**). Experimental animals had to weigh more than 11 g (females) or more than 13 g (males) by 5 weeks of age (the re-housing week). Animals presenting any anomaly at the time of re-housing were excluded. Unacceptable anomalies were cataracts, malocclusion, missing/small eye, ear infection, and unreadable or missing tags. Sample size was determined according to our previous studies, where we established that 8+ animals per group were enough to detect differences among groups with 95% confidence (if such differences in fact existed). We used 16 animals per genotype per sex per line as a starting point to account for expected attrition due to aging-related problems, aggression, unexpected death, mix-up of animals during re-housing, experimental failures (for example, failure to lick during PhenoCube testing) and removal of outliers.

Husbandry. Final experimental cages housed 8–10 animals in rat Opticages (Animal Care Systems, Inc.)—about half HET and half wild-type (same sex), originating from ten different litters. Animals were housed at JAX with disposable nestlets and Shepherd Shacks (Shepherd Specialty Papers) as enrichment until shipping (at 6 weeks of age). Mice were fed 5001 rodent diet (Harlan-Teklad). The shipped Shepherd Shacks and a handful of bedding from the shipping crates were introduced in the new cage to reduce stress and aggression. In addition, cages were enriched with the standard PsychoGenics (PGI) enrichment: two play tunnels, a plastic bone and enviro-dri (Shepherd Specialty Papers). For each line and CAG-KI line, every week for 6 weeks we received 78–80 wild-type and littermate HET mice (half males and half females) in random order from JAX. On the week of arrival, one tail snip was collected for genotype confirmation, and electronic transponders (Data Mars) were implanted. One week after arrival, mice were handled twice for about 1 min each. The first cage change was scheduled around 10 d after arrival to minimize disturbance of the cages that could trigger fighting. From the first cage change onward, only the standard PGI enrichment was provided to the animals, and cage changes occurred weekly.

Extra animals (up to 2) were removed from the cage 2 weeks after arrival once re-genotyping results were received. The extra mice were euthanized. The goal of the removal of the extras was to create a final experimental cage containing four HET and four wild-type mice. Mice were removed from the study for differing reasons, such as failure to lick in PhenoCube, aggression, dermatitis and other causes. For the multi-class analysis (see below), we included only mice for which a complete data set existed for all time points and Cube technologies. This resulted in about 20–30% of mice excluded for this analysis only. **Supplementary Table 2** presents the number of animals per age/line/genotype/sex available for testing, tested in PhenoCube, included in the two-class analysis and in modeling (multi-class analysis).

Behavioral high-throughput systems: the Cubes. PGI's comprehensive high-throughput systems—the PhenoCube, NeuroCube and SmartCube systems—capture different domains of behavior, namely, cognitive, motor, circadian, social, anxiety-like, gait and others, using custom-built computer vision software and machine learning algorithms[10].

PhenoCube is a high-throughput platform that assesses circadian, cognitive, social and motor behavior exhibited by group-housed mice. Experiments are conducted using extensively modified Intellicage units (New Behavior AG), each with a camera mounted on top of the cage for computer vision analysis. Intellicages have four corners with small doors, containing antennas to pick up the ID data from the electronic chips in the mice. Inside the corners, two small gates give access to water bottles and allow measurement of nose-poking and cognitive performance[30,44].

We conducted PhenoCube experiments using eight units. We added intra-maze spatial cues to the environment by placing laminated paper with green and white stripes outside the long sides of the cage, with the stripes being horizontal on one side and vertical on the other. Additionally, two climbing rods were located along one of the long sides of the cage with an additional climbing structure positioned in the center of one short cage wall and a rectangular object in the center of the cage.

Animals were evaluated in 72-h test sessions, being placed in the Intellicage environment after a 16-h water-deprivation period in the home cage. The cages were maintained on a 12:12 light/dark cycle, with white light during the day and red light during the night, and a low subjective light level maintained for the subjects during the night period. While mice were inside the cage, water was available only from within the Intellicage corners, whereas food was freely available on the cage floor at all times. When possible, mice were left undisturbed during the course of experimental sessions.

The test animals initially received magazine training through a simple 'habituation' protocol, which allowed them to freely retrieve water from the Intellicage corners. Prior to lights-out on day 1, after the mice had spent 6 h in the cage, we switched the protocol to a training protocol described as 'alternation', which required the animals to visit specific locations to retrieve water and to alternate between potentially reinforced locations (**Supplementary Fig. 5**).

Habituation. In the habituation phase used at the start of an experiment, all four of the Intellicage corners were open, with both doors to water opening as soon as any mouse entered and remaining open until the mouse left the corner. Measures included visits to corners, nose-poking frequency and alternations.

Alternation. Mice were required to visit two of the four Intellicage corners in alternation in order to gain access to water. For each subject, two adjacent (active) corners were contingently rewarded, and the other (exploratory) two were never rewarded. The alternation protocol was set up to train the animals to switch between the two active corners, with reinforcement only for alternating visits. For example, if corners 1 and 2 were active, an initial visit to corner 1 would be considered a correct visit and would be rewarded. To obtain further reinforcement, the mouse would then be required to visit corner 2; repeat visits to corner 1 would be classified as incorrect, and mice would not receive a reward for them. After the mouse visited corner 2, the corners would switch again such that reinforcement would be available only in corner 1, and so on. The only event leading to a switch in the correct-corner identity was a visit to the correct corner at that time, in which reinforcement was available.

Alternation data were calculated within an interval of leaving an active corner such that only a visit to the correct corner 113 s or less after exiting an active corner counted as a correct. Alternation with any visit to the incorrect corner counted as incorrect; visits to the exploratory corners were irrelevant. Each corner contained two nose-poke recesses used to deliver water reinforcement during correct visits. Only the left-hand side provided a reward in active corner 1, and only the right-hand side provided a reward in active corner 2. Reward consisted of 8 s of access to the liquid reinforcement. No penalty was imposed for initially nose-poking on the incorrect side. Data collected included frequency of alternations, repeat visits to all corners, total visits, number of alternations after obtaining or failing to collect reward, nose pokes to the correct side and the percentage of correct initial nose pokes in each visit.

Computer Vision data. In addition to collecting data through the Intellicage apparatus, we collected general activity data through PGI's proprietary computer vision automated video scoring system. The primary measures used for this study were measures of distance traveled, time in locomotion, time immobile (in isolation), time on an object, time climbing, time and frequency of rearing, and time huddling or in occlusion (two or more mice together). For the bioinformatics analysis, data were summarized in 12-h bins.

NeuroCube is a platform that uses computer vision to detect stance characteristics, gait geometry and dynamics in rodent models of disorders.

Mice were allowed to acclimate in the experimental room for 1 h before testing. After acclimation, mice were placed in the NeuroCube and allowed to walk in the apparatus for 5 min. Subjects were returned to their colony room after testing. Data collected included gait features (stride length, step length, base width, stride duration, stand duration and swing duration), speed, 'paw features' (area of contact of the paw, perimeter of the paw print, minimal and maximal diameter of the paw image, intensity of the paw image), body motion (range of change of body dimensions, variability in body dimensions, range of body motion, variability in body motion), coordination (correlation of gait signals between pairs of paws) and paw positioning (paw position relative to body center, angles defined by every possible three-paw position).

SmartCube is a platform that uses computer vision and mechanical actuators to detect changes in body geometry, posture and spontaneous behavior and reactions to particular challenges[29]. Mice were taken in their home cage to the SmartCube suite of experimental rooms, where they remained until they were placed in the apparatus. A standard SmartCube protocol ran for a single session lasting 45 min. After the session, mice were placed back into to their home cage and were returned to the colony room. Any abnormal behaviors before and after the session, such as seizures or tremors, were noted. Mouse behavior was captured by digital video using purposely designed hardware that presents multiple challenges in a single test session and was analyzed with computer algorithms. Digital videos of the subjects were processed with computer vision algorithms to extract more than 1,400 dependent measures, including frequency and duration of behavioral states such as grooming, rearing, locomotor trajectories, posture, abrupt movements, stretched attend posture and startle.

Quantitative assessment of disease phenotype. The outcome of all of our Cube analyses was a vector of hundreds of features (behavioral parameters) that could be used for various other analyses (for example, one run through SmartCube produces more than 2,500 features, whereas PhenoCube and NeuroCube result in ~400 and ~100 feature values, respectively). Many of these features were correlated (for example, rearing counts and supported rearing counts). Therefore, using a proprietary decorrelation algorithm, similar in ideology to semi-blind independent component analysis[45], we formed feature-rank-weighted statistically independent combinations of the original features (referred to herein as decorrelated features) that discriminated between the two groups more effectively. Each decorrelated feature extracted information from the whole cluster of the original features so that the new feature space had lower effective dimensionality, for two reasons. First, the clustered features formed statistically independent combinations, and thus the number of such combinations was limited (these combinations were essentially eigenvectors of the original features with combination coefficients additionally modified by feature ranks). Second, the ranks (weights) of the new features decayed a lot faster than the ranks of the original features, so only a handful of the new features meaningfully contributed to the discrimination. Typically, we were able to obtain a faithful 2D representation (for visual assessment) regarding the similarity between the two groups, although weighted contributions from all new features were accounted for in the calculation of the actual quantitative similarity measure. The rank of the newly formed features decayed so rapidly that we refer to it as 'reduced' dimensionality, even though contributions beyond dimension 5 rarely matter for any practical purposes. We typically applied a decorrelated ranked feature analysis approach to get an overall qualitative picture of a data set and assess the discriminability of its components (for example, Q lines), whereas we used original (non-transformed) features for building the SVM/SVR models that we used for actual CAG and/or age predictions, as SVM-type methods can gracefully handle virtually unlimited numbers of features, highly correlated or not.

Next we applied a proprietary feature-ranking algorithm to score each feature's discrimination power (i.e., its ability to separate two groups, such as control and disease). In brief, the more sensitive the decision line (defined by support vectors) is with respect to the change of a particular feature value, the higher the rank (weight, relevance) that will be assigned to that feature. Ranking is an important part of our analyses because it weighs each feature change by its relevance: if there is a significant change in some irrelevant feature measured for a particular phenotype, the low rank of that feature will automatically reduce the effect of such a change (and effective dimensionality) in our analyses, so we don not have to resort to the conventional 'feature selection' approach and discard information buried in the less informative features. The ranking algorithm can be applied to either original or new features to gain insight about the key differences between disease and control states (**Supplementary Fig. 6**).

In the new feature space, the overlap between the 'clouds' (Gaussian distributions approximating the groups of mice in the ranked decorrelated-features space) served as a quantitative measure of separability ('distinguishability') between the two groups (**Supplementary Fig. 7**). For visualization purposes we plotted each cloud with its semi-axes equal to 1 s.d. along the corresponding dimensions. Note, however, that although the overlap between any two Gaussian distributions is always nonzero, it might not necessarily be seen at the '1-sigma' level.

Discrimination significance (generalized *P* value, i.e., the confidence measure of the discrimination probability) was calculated in the following way. First, each labeled set of candidates was randomly split with a 1:3 ratio, and larger groups from each set were used to calculate the discrimination probability using previously described methods. This procedure was repeated multiple times with different random splitting for each iteration to build a distribution of the 'true' discrimination probability p_{true} (**Supplementary Fig. 8**) (step 1). The number of iterations was limited by a fraction of the total number of split combinations available. Next, all candidates from both groups were combined together without individual class labels (step 2). Similar to the previous step, this set was split randomly multiple times. The larger group of candidates for each split was randomly divided into two 'classes' and used to calculate the discrimination probability. After many iterations, the distribution of 'random' discrimination probability p_{random} was built (**Supplementary Fig. 8**) (step 2). Both distributions were normalized, and their mutual weighted overlap was calculated. The resulting value represented a generalized quantity of what is well known as the *P* value indicating statistical significance. Basically, our method is a bootstrap procedure for determining how label randomization influences a classifier's accuracy.

Multi-class analysis of independent CAG and age effects. We combed all behavioral data from the different CAG-KI lines and ages studied using a proprietary version of SVM/SVR. The SVM algorithm finds the landmarks (vectors) in hyperspace that best define separation between classes, in this case the different KI groups for each age class or the different age classes independent of CAG-repeat length. We used a feature-ranking selection algorithm to reduce dimensionality and to identify and more heavily weight the top features that contributed most to the CAG and age signatures. Our SVM/SVR approach is very similar to LIBSVM[46] with a modified Gaussian kernel with an individual exponent for each feature (modification of each exponent is proportional to the corresponding feature's rank). Cost value was optimized to achieve maximal LOOCV performance. Features were ranked by their sensitivity, i.e., the magnitude of decision line change as a response to each feature variation.

We used a feature-ranking curve to select the optimal number of top-ranked features that maximized the classifier predictive performance (for example, to predict CAG values for the HET lines). Furthermore, we analyzed the data pool for the wild-type mice in order to remove features that formed spurious feature patterns in the training set (which would disappear with larger sample sizes). To achieve this, we ranked all features according to their ability to discriminate between wild-type mice belonging to different KI-line cohorts, separately for each age group. We then removed the top-ranked features that separated the wild-type groups and made them undistinguishable, effectively removing any effects that were not CAG- or age-dependent.

doi:10.1038/nbt.3587

Code availability. The code that uses publicly available open source code to demonstrate the features of the described approach and reproduce major results presented in this paper is freely available from http://www.psychogenics.com/FileShare/BehavioralAllelicSeriesCode.zip and as **Supplementary Code**.

CAG signature modeling. We tested the ability of many machine learning methods to model CAG signatures. We found that the methods derived from minimization of Bayesian probability of misclassification—namely, Kernel Gaussian Process, Relevance Vector Machines/Regression and especially SVM/SVR—turned out to be especially well suited. SVM seems to be especially convenient for the problem at hand, not only because of its typically superior performance but also because proprietary PGI SVM-based models can be used immediately for feature ranking/selection, which often improves performance.

We judged the robustness of training by using an LOOCV approach[45]. For each subject we trained a classifier on all other subjects, excluding the one for which we made a prediction. After we had performed this procedure for all subjects, we calculated the coefficient of determination from predicted versus true values[47].

Age signature modeling. We used the modeling technique described above (namely, a proprietary version of SVR) to build the age-predicting classifier. In brief, ages 2, 6 and 10 months served as the one-dimensional 'response variable' to which machine learning regression (SVR) mapped the features collected from SmartCube, NeuroCube and PhenoCube.

Multi-class analysis of simultaneous CAG and age effects. To account for both CAG-dependent and age-dependent effects/features, we built an age model in which the dependent (response) variable was a combination of an animal's age and its CAG-repeat length. To predict both CAG-repeat length and age with our SVR model, we built a correspondence map by combining the two-dimensional CAG–age pairs with a one-dimensional dependent variable. Note that modeling of a multi-dimensional dependent variable (for example, CAG + age) can be easily reduced to the one-dimensional case via the so-called RGB-to-wavelength approach (**Supplementary Figs. 9** and **10**).

Statistical analysis of complexity of the trajectory and percent duration of clustering. We analyzed data using ANOVA. For *post hoc* tests we used a Bonferroni correction for the critical alpha value, which yielded $\alpha = 0.05/6 = 0.0083$, considering only six comparisons against the wild-type line.

Quantification of endogenous huntingtin by qPCR *Total mRNA extraction.* Cortical and cerebellar tissues were homogenized twice for 1 min each time at 25 Hz in 750 µl of QIAzol Lysis Reagent (79306; Qiagen, Valencia, CA) with TissueLyser (Qiagen, Valencia, CA) and 5-mm stainless steel beads (69989; Qiagen, Valencia, CA). Once tissues were disrupted, samples were allowed to incubate at room temperate for 5 min. For RNA extraction, we followed the manufacturer's protocol for the RNeasy 96 Universal Tissue Kit (74881; Qiagen, Valencia, CA) for RNA isolation. Briefly, 150 µl of chloroform (C2432; Sigma-Aldrich, St. Louis, MO) was added and samples were shaken vigorously for 15 s, after which they were incubated for 3 min at room temperature. The aqueous phase was separated from the organic phase by centrifugation at 6,000*g* (Beckman Coulter Avanti J-30I) and 4 °C for 15 min. The aqueous phase was then transferred to a new 96-well block, and total RNA was precipitated with an equal volume of 70% ethanol. The entire content was transferred to an RNeasy 96-well plate and then centrifuged at 6,000*g* (Beckman Coulter Avanti J-30I) at room temperate for 4 min. Total RNA bound to column membranes was treated with the RNase-Free DNase set (79254; Qiagen, Valencia, CA) for 30 min, and this was followed by three washing steps with RW1 and RPE buffers (provided with the RNeasy 96 Universal Tissue Kit). RNA samples were eluted with RNase-Free water (20 µl for striatum samples and cerebellum samples).

Total mRNA quantification and reverse transcription. Samples were quantified using a NanoDrop 8000 (Thermo Scientific). Total RNA (0.1 µg of RNA) was reverse-transcribed into cDNA with 3.2 µg of random hexamers (11034731001; Roche Applied Science, Indianapolis, IN) and 1 mM each dNTP (11814362001; Roche Applied Science, Indianapolis, IN), 20U Protector RNase Inhibitor (03335402001; Roche Applied Science, Indianapolis, IN), 1X Transcriptor Reverse Transcription reaction buffer and 10U Transcriptor Reverse Transcriptase (03531287001; Roche Applied Science, Indianapolis, IN) in a 20-µl total volume. The reactions were allowed to proceed at room temperature for 10 min and then at 55 °C for 30 min, and then they were inactivated at 85 °C for 5 min in a GeneAmp PCR Systems 9700 thermal cycler (Applied Biosystems, Foster City, CA). cDNA samples were diluted tenfold with RNase-Free water for qPCR analysis. For statistical analysis of the first study, we used pooled data from wild-type mice from all lines for the 6-month-old age group. Because of the high cost, we reduced the number of wild-type controls and used only those obtained from the Q20 line for the 2-month and 6-month age groups for the first study and for the Q50 analysis in the second study.

Quantitative PCR (qPCR). For all reactions using Universal Probe Library Probes, 5 µl of the diluted cDNA was amplified with 12.5 µl of 2× FastStart Universal Probe Master (Rox) (04914058001; Roche Applied Science, Indianapolis, IN), 0.5 µl of Universal Probe Library Probe (Roche Applied Science, Indianapolis, IN), 200 nM gene-specific primer (HPLC-purified; Sigma-Aldrich, St. Louis, MO) in a 25-µl reaction volume. For all reactions using hydrolysis probe (GAPDH), 5 µl of the diluted cDNA was amplified with 12.5 µL of 2× FastStart Universal Probe Master (Rox) (04914058001; Roche Applied Science, Indianapolis, IN), 300 nM TaqMan probe (FAM-labeled), 400 nM each gene-specific primer (HPLC-purified; Sigma-Aldrich, St. Louis, MO) in a 25-µL reaction volume. The reactions were run on the ABI 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA). qCPR conditions were 95 °C for 10 min for activation of FastStart Taq DNA Polymerase followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. For primers and Universal Probe Library used for qPCR, please refer to **Supplementary Table 3**.

qPCR data analysis. Small aliquots of z_Q175KI HET samples (tissue-matched) were pooled and used as a calibrator (the calibrator was diluted in the same way as the sample cDNA) to normalized plate-to-plate variations. Each cDNA sample (diluted 1:10) was assayed in triplicate, and the C_t values were averaged. Values that were greater than 0.5 s.d. of the average were discarded. The relative quantity of the PCR product (relative to the calibrator) was calculated as follows: Relative Quantity = (PCR Efficiency)(C_t calibrator − C_t sample). The geometric mean (GM) for the three housekeeping genes was then calculated and used to normalize the level of the target gene as follows: Normalized Quantity = Relative Quantity/GM. Finally, DNA levels were further normalized to the 5-week-old wild-type group.

Statistical analysis of qPCR results. Analyses were carried out using ANOVA at each age. For *post hoc* tests, a Bonferroni correction was used for comparison against the wild type only (endogenous *Htt*) yield $\alpha = 0.0062$. For multiple comparisons, $\alpha = 0.0025$.

44. Balci, F. *et al.* High-throughput automated phenotyping of two genetic mouse models of Huntington's disease. *PLoS Curr.* http://dx.doi.org/10.1371/currents.hd.124aa0d16753f88215776fba102ceb29 (2013).
45. Cichocki, A. & Amari, S. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications* (Wiley, 2006).
46. Chang, C.-C. & Lin, C.-J. in *LIBSVM: A Library for Support Vector Machines* Vol. 2 (ACM, 2011).
47. Steel, R.G.D. & Torrie, J.H. *Principles and Procedures of Statistics with Special Reference to the Biological Sciences* (McGraw-Hill, 1960).