Dominant effects of the Huntington’s disease HTT CAG repeat length are captured in gene-expression data sets by a continuous analysis mathematical modeling strategy

Jong-Min Lee1,∗, Ekaterina I. Galkina1, Rachel M. Levantovsky1, Elisa Fossale1, Mary Anne Anderson1, Tammy Gillis1, Jayalakshmi Srinidhi Mysore1, Kathryn R. Coser2, Toshi Shioda2, Bin Zhang3, Matthew D. Furia3, Jonathan Derry3, Isaac S. Kohane4,5,6, Ihn Sik Seong1, Vanessa C. Wheeler1, James F. Gusella1 and Marcy E. MacDonald1

1Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA, 2Massachusetts General Hospital Cancer Center, Charlestown, MA 02129, USA, 3Sage Bionetworks, Seattle, WA 98109, USA, 4Children’s Hospital Informatics program, Children’s Hospital, Boston, MA 02115, USA, 5Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02114, USA and 6i2b2 National Center for Biomedical Computing, Boston, MA 02115, USA

Received January 11, 2013; Revised April 2, 2013; Accepted April 9, 2013

In Huntington’s disease (HD), the size of the expanded HTT CAG repeat mutation is the primary driver of the processes that determine age at onset of motor symptoms. However, correlation of cellular biochemical parameters also extends across the normal repeat range, supporting the view that the CAG repeat represents a functional polymorphism with dominant effects determined by the longer allele. A central challenge to defining the functional consequences of this single polymorphism is the difficulty of distinguishing its subtle effects from the multitude of other sources of biological variation. We demonstrate that an analytical approach based upon continuous correlation with CAG size was able to capture the modest (∼21%) contribution of the repeat to the variation in genome-wide gene expression in 107 lymphoblastoid cell lines, with alleles ranging from 15 to 92 CAGs. Furthermore, a mathematical model from an iterative strategy yielded predicted CAG repeat lengths that were significantly positively correlated with true CAG allele size and negatively correlated with age at onset of motor symptoms. Genes negatively correlated with repeat size were also enriched in a set of genes whose expression were CAG-correlated in human HD cerebellum. These findings both reveal the relatively small, but detectable impact of variation in the CAG allele in global data in these peripheral cells and provide a strategy for building multi-dimensional data-driven models of the biological network that drives the HD disease process by continuous analysis across allelic panels of neuronal cells vulnerable to the dominant effects of the HTT CAG repeat.

INTRODUCTION

Huntington’s disease (HD) is a dominantly inherited disorder that in all cases is caused by an expanded allele of a CAG trinucleotide repeat polymorphism located in the 4p16.3 huntingtin gene (HTT, previously HD) (1). HTT CAG allele lengths of about 40 or more units reliably predict the degeneration of neurons, especially in the striatum, and the onset of characteristic motor, cognitive and often psychiatric signs of HD (2,3). Shorter alleles in the 35–39 repeat range, extending perhaps through the high-end normal range (27–34), exhibit increasingly reduced penetrance, while alleles in the 6–26 repeat range have not been associated with confirmed HD cases (4–6).
The striking inverse correlation of CAG repeat length with age at diagnosed HD motor symptoms shows that CAG repeat size is the primary determinant of the rate of the underlying pathologic process that leads to the onset of motor clinical manifestations (1,7–10). In contrast, the dosage of the expanded CAG allele is not an important determinant, as the age at onset of motor symptoms for HD homozygote individuals is determined by the longer CAG repeat allele, just as it is for typical HD heterozygotes (11).

From genotype–phenotype studies in HTT CAG allelic series designed to test the full spectrum of normal and HD-associated allele sizes in human lymphoblastoid cell lines, it has emerged that the continuous relationship between CAG size and dominant biochemical measures outcome extends below the expanded CAG HD range (12). The view of the HTT CAG repeat as a functional polymorphism with dominant quantitative biological consequences that become overwhelming at HD disease-producing CAG lengths is also consistent with the demonstrated continuous impact of the CAG repeat encoded polyglutamine tract on the activity of huntingtin protein (13), a ubiquitously expressed ancient HEAT domain protein (14,15) shown to be involved in a wide variety of cellular processes.

The knowledge that the HTT CAG repeat is the root cause of all cases of HD has provided a starting point for discovering, from unbiased genome-wide data, the biological network of CAG allele length-dependent molecular effects that lead eventually to the onset of diagnostic motor symptoms. However, most studies in the field, including gene-expression profiling studies (14–22), performed a standard dichotomous analysis to discover the effects of the repeat by comparing HD expanded allele samples with normal control samples, with each group comprising samples with many different CAG lengths. This approach fails to take CAG length into account. Moreover, since all observed differences are then ascribed to the mutation, it makes an assumption, likely to be misleading, that despite the CAG repeat representing a small DNA sequence difference at a single locus, its effects must account for the entirety of the biological variation between the two groups.

Given the genetic characteristics of the HTT CAG repeat and the fact that its biological consequences are likely to be subtle, as expanded alleles typically do not produce signs of HD until mid-life, we are interested in taking a different approach by phentypifying members of panels that sample the entire HTT CAG allelic spectrum and then distinguishing its effects by an analytic strategy based upon continuous analysis with CAG repeat size. We believe that this approach will provide a powerful genetically directed route to efficiently distinguish the subtle dominant consequences of this one functional variant from the manifold consequences of the other, far more numerous, sources of biological variation.

We first evaluated this idea with a small isogenic panel of heterozygous mouse Hdh CAG knock-in embryonic stem cell lines, comprising only four alleles drawn from the normal (CAG 18) and expanded (CAG 48, CAG 89, CAG 109) human ranges (23). The results confirmed that a continuous analytical strategy did efficiently distinguish genes whose level of expression was altered in a manner that was significantly correlated with CAG allele size (23). Notably, these transcriptional shifts delineated biological pathways/processes related to huntingtin functional pathways and demonstrated that the CAG-correlated genes formed a distinct and relatively small set of molecular responses that were not identified using the traditional dichotomous Htt CAG expanded allele versus Htt wild-type allele comparison (23). While encouraging, the identification of a CAG-correlated transcriptional response from this otherwise isogenic allelic cell panel could not reveal whether the approach would be sufficiently powerful to distinguish the effects of the CAG repeat in panels of human cells or tissues where, in addition to a variety of different environmental and chance factors, the genetic background is highly variable from person to person.

Here we describe a proof-of-concept study using a panel of 107 human lymphoblastoid cell lines designed to test the power of the continuous analytical approach as a strategy for discerning the dominant effects of the CAG repeat in genome-wide gene-expression data sets.

RESULTS

**HTT mRNA expression levels in human lymphoblastoid cell lines**

In order to test the power of our continuous analytical approach, we chose a set of 107 human lymphoblastoid cell lines (55 males and 52 females) for which the larger of the two HTT CAG alleles ranged from 15 to 92 repeats, spanning both the normal (41 lines) and the HD ranges (66 lines) (Supplementary Material, Fig. S1). Over all cell lines, the mean size of the longer and shorter allele was 40.9 and 17.6 CAGs, respectively. For each line, genome-wide gene expression was determined using the Affymetrix U133 Plus 2 array (54 675 probes) and the data sets were normalized and batch effect corrected (Gene Expression Omnibus accession number; GSE34721). Given the genetic parameters of HD (dependence on HTT CAG size rather than allele dosage), we are interested in identifying gene-expression signatures that reflect continuous effects of the CAG repeat, rather than effects that might instead reflect expression levels of HTT mRNA. Therefore, we first analyzed the data sets to determine whether there might be an association between HTT mRNA levels and CAG repeat length. Two of the microarray probes represented human HTT mRNA (RefSeq, NM_002111; Fig. 1A), with target sequences located within the alternate short (202389_s_at, red line) or long (202390_s_at, blue line) 3’ UTRs produced by the use of alternate polyadenylation sites, which yields two different transcripts that both encode the same huntingtin protein (1,24,25). The results of these analyses, summarized in Figure 1B–D, revealed that for both target probes, the signal intensity appeared to be independent of the longer CAG repeat, the shorter CAG repeat as well as the sum of longer and shorter CAG repeats. This was confirmed by statistical analyses, where HTT mRNA expression levels were modeled as a function of: (i) longer and shorter allele (HTT expression levels ~ longer CAG + shorter CAG) and (ii) sum of longer and shorter CAGs (HTT expression levels ~ sum of CAGs). As shown in Figure 1E and F, the HTT probe signal intensities could not be explained by the CAG repeat lengths, indicating that HTT mRNA level in lymphoblastoid cell lines is independent of CAG repeat length. Although the microarray probes representing the HTT mRNA cannot distinguish the expanded allele from the normal allele, the indirect evidence described earlier argues against a role for CAG repeat length in determining the steady-state HTT mRNA expression.
levels in lymphoblastoid cells. These observations also implied that gene-expression signatures identified in strategies using HTT CAG repeat size would be independent of the dosage of HTT mRNA.

Effects of number of probes and training samples on mathematical prediction model

Our primary hypothesis is that the continuous effects of HTT CAG repeat size are reflected in gene expression. To test this hypothesis, we determined whether or not CAG length can be predicted based upon gene expression by building and testing predictive mathematical models. First, to understand the effects of the model parameters on the performance of prediction models, we carried out pilot experiments. Our modeling procedures for the pilot experiments are described in Supplementary Material, Figure S2A. Briefly, to determine the optimal number of probes for prediction models, the discovery set (107 samples) was randomly split into a training set (97 samples) and a test set (10 samples), and the correlation between CAG repeat and expression levels was calculated for all probes using only the training samples. Then, for each split, fixed numbers of top correlated probes were used to construct partial least square regression (PLSR) models (i.e. models using top 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950 or 1000) (26) and the same model parameters were used to predict the CAGs of the unused test samples. These procedures were repeated 1000 times to understand overall effects of the number of probes on prediction accuracy. As shown in Figure 2A, the error rates of

Figure 1. CAG length-independent HTT mRNA expression levels in human lymphoblastoid cell lines. (A) Two microarray probes in the Affymetrix U133 plus 2 microarray (202389_s_at, red; 202390_s_at, blue) represent human HTT mRNA. The approximate locations of target sequences of those two microarray probes are shown relative to different transcripts (red and blue vertical lines). Genomic location was based on UCSC Genome Browser GRCh37/hg19 assembly. (B) HTT expression levels determined by two microarray probes (y-axis, log2 scale gcRMA values) were plotted against the length of longer CAG repeat lengths (x-axis). Red and blue circles represent 202389_s_at and 202390_s_at, respectively. (C) HTT expression levels determined by two microarray probes (y-axis, log2 scale) were plotted against the length of shorter CAG repeat lengths (x-axis). (D) HTT mRNA levels determined by microarray probes may represent the sum of mRNAs expressed from both alleles (an allele with longer CAG and the other allele with shorter allele). Therefore, HTT expression levels (y-axis, log2 scale) were plotted against the sum of longer and shorter CAG repeat lengths (x-axis). (E) To determine whether lengths of CAG repeats were significantly associated with the expression levels of HTT, HTT gene-expression levels determined by microarray probe 202389_s_at were modeled as a function of: (i) the length of longer CAG repeat and that of shorter CAG repeat and (ii) the sum of longer and shorter CAGs in a multiple regression model. (F) HTT gene-expression levels determined by microarray probe 20239090_s_at were modeled as a function of: (i) the length of longer CAG repeat and that of shorter CAG repeat and (ii) the sum of longer and shorter CAGs. Slope represents estimated slope of each variable. Summary statistics suggest that the lengths of CAG repeats are not significant determinants of HTT mRNA levels.
test samples gradually decreased when increased numbers of probes were used for the PLSR models (see red LOWESS trend line). However, the error rates of models using approximately the top 500 probes or more did not decrease even when more probes were introduced into the prediction models. These observations indicated that: (i) the performance of prediction models depends on the number of probes up to a certain point (e.g. 500 probes) and (ii) after passing that point, incorporating more information into the models did not result in significant improvement. It is also important to note that certain models using larger numbers of probes generated higher error rates than models using smaller numbers of probes for many instances (Fig. 2A). We reason that this phenomenon might be due to the quality of the training samples. Taken together, the findings from the pilot study suggested that the performance of prediction models is influenced both by the number of probes and by the manner in which the data are split. Since the optimal number of probes and the performance of a model could not be determined until a model has been tested using independent test samples, optimizing the modeling parameters (e.g. probe number) in a completely unbiased way is impossible. Therefore, to objectively test our hypothesis, instead of optimizing the model parameters, we constructed multiple models using different combinations of training and test samples for each data split in an iterative fashion. Although this approach is computationally intense and may not necessarily yield the optimal prediction model, this strategy offers an unbiased way to assess the prediction power of CAG-correlated gene-expression signatures without relying on a single model based upon a guess as to probe number or an arbitrary fixed selection of training and test samples.

Next, we investigated how the performance of the prediction models might be affected by the number of training samples. Prediction models using different numbers of training samples were constructed and the prediction power of the resultant models was evaluated in the unused test samples. In each iteration of
mathematical modeling, seven prediction models were built using 200, 250, 300, 350, 400, 450 or 500 probes, and the average predicted CAGs across the seven models was taken as the representative predicted CAGs of the test samples. As shown in Figure 2B and C, significant correlations between predicted and actual measured CAGs (P-value < 0.05) were observed for models using 35 or more training samples (B, red trace), and models using larger numbers of training samples tended to generate predicted CAGs of the test samples that were strongly correlated with the measured CAGs of the same samples (Fig. 2B, red trace). However, the correlation coefficient appeared to asymptote to ~0.45 starting at around 80 training samples, and adding more training samples into the models did not appear to significantly improve the accuracy of the prediction model (Fig. 2C, red). In contrast, models using permuted training sample CAG allele sizes produced predicted CAG sizes that were not significantly correlated with the true CAG allele sizes of the test samples, regardless of the number of training samples used in the models (Fig. 2B and C, blue traces). These observations suggested that: (ii) the performance of prediction models is also dependent on the number of training samples; (ii) the most stable correlation between measured and predicted CAG can be obtained from as few as ~80 independent lymphoblastoid cell lines; and (ii) although the predicted CAGs were significantly correlated with the experimentally determined allele sizes, it appeared that there was an upper limit to the correlation.

Predictive power of CAG-correlated gene-expression signatures

Taking the lessons learned from the pilot experiments into consideration, we next tested our formal hypothesis that a CAG-correlated gene-expression signature that reflects the continuous effects of HTT CAG repeat length may exist within the data sets and be discovered using a continuous analysis. A key prediction of this hypothesis is that CAG repeat length can be inferred from the CAG-correlated gene-expression signature. Model parameters were chosen based upon our observations concerning the influence of the numbers of probes and the training samples on model performance. Briefly, we: (i) split the entire discovery set (107 samples) into a training set (97 samples) and a test set (10 samples); (ii) identified HTT CAG repeat-correlated probes in the training set; (iii) built seven independent PLSR models using different number of probes (i.e. 200, 250, 300, 350, 400, 450 or 500 probes) in the training samples; (iv) used the same signature probes in the test samples and the same mathematical model to predict HTT CAG repeat lengths of the test samples; and, finally, (v) averaged the seven predicted CAGs for each sample to be compared with the experimentally determined CAG repeat lengths of test samples. These iterative procedures were repeated 10,000 times (Supplementary Material, Fig. S2B). This approach was chosen because, although it might not yield the optimal prediction models, it would test our key predictions in a completely unbiased way.

As shown in Figure 3A and B, the distribution of error rates of models using the correct CAG repeat lengths of the training samples (A; mean, 0.930) was significantly shifted to the left side of the error rate spectrum, compared with that of the models using permuted CAG repeat lengths for the training samples (B; mean, 1.65), indicating that the error rates of models using correct CAGs were significantly smaller than those of the models using permuted CAG allele sizes (Mann–Whitney U test P-value < 1.6E−16). In addition, the predicted CAG repeat sizes from the models using the correct training sample CAGs were significantly correlated with the measured CAG alleles (Fig. 3C; Pearson’s correlation coefficient, 0.4607, P-value, 5.94E−7; Spearman’s rho 0.3794, P-value, 5.563E−5). The variance in the predicted CAGs (y-axis) was best explained by a regression model with a slope of 0.229 (P-value, 5.94E−7), and this model accounted for 21.2% of the variance in the predicted CAG repeat lengths (Fig. 3C, red line; P-value, 2.02E−6). In contrast, the predicted CAG repeat sizes that emerged from the models using the permuted training sample CAGs were not correlated with measured CAGs (Fig. 3D; Pearson’s correlation coefficient, −0.0069; P-value, 0.9438), and permuted CAG could not explain any of the variance in the predicted CAG lengths (R², 0.0003725, P-value, 0.844). Although the current models did not perfectly predict the HTT CAG repeat lengths of the test samples, the results clearly demonstrated that gene-expression could capture a modest but significant amount of variance contributed by continuous CAG size, as 21% of the CAG allele effect was reflected in the form of gene expression. Our finding that a significant amount of variance in gene-expression levels can be attributed to the continuous effects of CAG repeats is consistent with observations of biochemical effects that are continuous across the CAG allele spectrum (12).

Optimized models

Our results showed that a significant amount of the variance in gene-expression values can be attributed to CAG repeat size. However, since an unbiased model does not necessarily represent the best prediction model, the strength of relationship between CAG repeats and gene expression in our models might be somewhat underestimated. Thus, we attempted to more precisely estimate the amount of variance in gene expression that was contributed by the CAG repeats by taking optimized modeling approaches. Briefly, we followed the same procedures of the unbiased modeling method described earlier to build and test prediction models except that the probes that generated the lowest error rate in the test samples were used for the final optimized model for any given data split. In contrast to unbiased models where averages of seven sets of predicted CAGs were considered as the predicted CAG sizes of the test samples, the best predicted CAG sizes of the test samples represented the predicted CAGs for the optimized models. As shown in Supplementary Material, Figure S3, the optimized model slightly improved the performance of the prediction models as 30.17% of variance in predicted CAG was explained by experimentally determined CAGs in the optimized model (Supplementary Material, Fig. S3) compared with 21.2% in the unbiased model (Fig. 3C). These results indicate that ~30% of variance in gene expression can be explained by CAG repeat length and that unbiased models can discover about 67% of this CAG-dependent variance.
Test of reproducibility of CAG-correlated gene-expression signature

Although tests utilizing the prediction models revealed a significant association between CAG repeat length and gene expression, the effect of CAG repeat length variation on gene expression was modest and there was not a one-to-one correspondence between CAG size and gene expression. Since gene expression can be influenced by many other factors, we were interested in testing whether the CAG-correlated signature is robust enough to be identified when other factors that potentially affect gene expression are present. We would expect that if a CAG-correlated gene-expression signature and the consequently predicted CAG size reflect the cellular state associated with CAG repeat length rather than other factors, then the gene-expression patterns of RNA samples from the same cell line are similar. Briefly, we randomly selected 20 cell lines from the initial set of 107 discovery lines, and for each cell line, we grew three independent frozen cell stocks and prepared two separate sets of RNA samples from each stock, generating a total of six sets of RNA preparations per cell line (so called replication samples). All replication samples were profiled using the Affymetrix U133 Plus 2 array (replication set). The 107 discovery data set and the 120 replication data set were then used to build and test prediction models. The following procedure was performed iteratively for each of the 20 chosen cell lines. In each iteration, the seven expression data sets from a given cell line (one in the discovery set and six in the replication set) were all excluded together from the training set, and all of the discovery sets and replication sets from the remaining samples were used to build the model. For each of the chosen 20 lines, seven independent prediction models were constructed using the top 200, 250, 300, 350, 400, 450 or 500 correlated probes, and each model was then used to predict the individual’s CAG allele length separately from each of the seven RNA preparations (one discovery and six replication)
that had been excluded from the training set. As shown in Figure 4, for these 20 individuals, the CAGs predicted in different replication data sets were significantly correlated (Pearson’s correlation, nominal P-value < 0.05), suggesting that gene-expression profiles of independently cultured cell lines were similar, thereby yielding correlated predicted CAG lengths. These results from the mathematical models, therefore, demonstrated that the CAG-correlated gene-expression signature was robust enough to be identified in spite of different frozen stocks and different culture conditions, supporting the power of a strategy of continuous analysis to discover dominant and continuous effects of the CAG repeat.

Genes and pathways in the CAG-correlated gene-expression signature

Next, to understand genes that contributed to the CAG-correlated gene-expression signature, we identified the model that yielded the smallest error rate in predicting CAGs of test samples in each iteration and extracted participating probes in that model. Briefly, we performed 10 000 iterations of optimized modeling procedure, and in each iteration we recorded the correlation P-value and the rank of correlation (e.g. the probe with the smallest correlation P-value would have the rank #1) for each probe that was a part of the optimized model. These procedures generated 10 000 sets of probes and subsequently, we summarized the results by calculating: (i) frequency (i.e. how many times the probe participated in the optimized models; a value that can theoretically range from 1 to 10 000); (ii) average correlation coefficient and (iii) average of correlation P-value for each microarray probe. A total of 10 639 unique probes were recorded from 10 000 optimized models, and summarized in Figure 5 and Supplementary Material, Table S1. As shown in Figure 5, mean ranks and frequencies varied greatly. For example, probes in the top left corner participated in the optimized models most of the time (y-axis frequency) and showed very high rankings (x-axis, mean rank). In contrast, many microarray probes at the bottom of the graph were parts of far smaller numbers of optimized models and showed varying levels of ranks. These observations re-inforce the notion that correlated genes and their correlation strength are largely determined by the fashion how the data set was split. Importantly, this analysis discovered consistent microarray probes that showed strong correlations with CAG repeat sizes regardless of the manner of data split, implying that those probes might be true CAG-correlated genes. Of note, two microarray probes represent the HTT gene in the

Figure 5. Genes consistently correlated with CAG repeat length. Ten thousand iterations of model optimization procedures have been performed to identify microarray probes participating in each of the 10 000 optimized prediction models. The analysis results were summarized so we could distinguish the probes consistently correlated with CAG regardless of the training samples (e.g. probes at the top left corner represent probes with a high frequency and a high mean rank) from the probes correlated in a training sample-dependent manner (e.g. probes at the bottom of the plot represent probes with a low frequency). Red circle indicates microarray probe 202389_s_at representing the HTT (frequency of 1, mean rank of 864, mean correlation coefficient of -0.2678 and mean correlation P-value of 0.008002).
Affymetrix U133 + 2 arrays, and one of them was a part of one optimized mode with a correlation rank of 863 (Fig. 5, red circle), indicating that the HTT gene-expression levels were not correlated with CAG repeat sizes in most of the data split.

As shown in Supplementary Material, Table S1, the consistently correlated genes (e.g. genes with frequency >4000) are involved in diverse cellular functions denoted as ribosomal process, translation, transcription, nucleic acid metabolism, adhesion, energy metabolism, hormone response, synaptic transmission and neurological process. We further analyzed the whole discovery set data to identify significantly correlated pathways using the sigPathway program (27). Pathways denoted as ribosomal process, energy metabolism, signaling, immune process, nucleotide metabolism and cell-death pathways were significantly correlated with CAG repeat length (Supplementary Material, Table S2). Signatures correlated with CAG repeat length, therefore, implicated polyglutamine modulated huntingtin function in numerous cellular processes, attesting to the multi-functional nature of huntingtin protein and the pleiotropic impact of the functional CAG polymorphism/mutation on the transcriptome.

Relevance of CAG-correlated gene-expression signature to HD

Next, we investigated the potential relevance of a CAG-correlated gene-expression signature that is predictive of CAG allele size in lymphoblastoid cells lines to the dominant consequences of the HTT CAG repeat in HD patients. First, we tested whether predicted CAG length would be at all associated with age at onset of motor symptoms, which is strongly correlated with true CAG repeat length (11). Age at onset of clinical motor symptoms data were available for 12 individuals whose cell lines were used in our modeling analysis. We observed a significant association between predicted CAG repeat length and age at onset of motor symptoms in those 12 samples (Fig. 6A; regression analysis, $R^2$, 0.54; $P$-value, 0.00584), supporting a functional connection between CAG-correlated gene-expression signature and an overt clinical disease outcome. Next, postulating similar fundamental effects of the HTT CAG repeat in different cell types, we tested whether CAG-correlated gene-expression signature in lymphoblastoid cell lines is also enriched in genes correlated with HTT CAG repeat length in human HD post-mortem brain. Among three brain regions tested (cerebellum, prefrontal cortex, visual cortex), the set of genes that were negatively correlated with CAG size in lymphoblastoid cell lines was significantly enriched in genes that were correlated with CAG repeat length in cerebellum (Fig. 6B; nominal $P$-value, 0.01691), suggesting some similarity between cellular responses of lymphoblastoid cells and cerebellum to the effects of expanded HTT CAG repeats.

DISCUSSION

The striking negative correlation between age at onset of clinically diagnosed HD symptoms and HTT CAG allele size strongly indicates the effects of a dominant CAG length-dependent mechanism in determining the rate of the underlying disease process that leads to overt manifestations. Integration of global molecular phenotyping using ‘omics’-based approaches, together with genome-wide DNA sequence data linked to clinical outcome measures (28–30), offers the promise of modeling the biological networks that drive this disease process. However, the ability to generate accurate mathematical models will require an efficient strategy for capturing the effects of the singular HTT CAG repeat from among the manifold effects due to the multitude of other sources of biological variation. Based upon the
hypothesis that the rate-limiting processes emanating from the disease mechanism are expected to be both dominant and progressive with CAG size, we have tested the utility of a continuous analytical strategy for generating CAG-dependent models.

Our results demonstrate the power of a continuous strategy to capture HTT CAG length-correlated gene-expression signatures conforming to the criteria expected for the effects of the mechanism that triggers the HD disease process. Our mathematical modeling for predicting HTT CAG allele length from quantitative data for 107 individuals with a full-range of CAG repeats revealed that CAG repeat size accounted for about 21% of the variance in cellular states defined by genome-wide gene expression measured in a microarray format. Knowledge of this analytically determined estimate is important because the relatively modest contribution of CAG allele size contradicts the commonly held notion that the expanded CAG would account for the vast majority of the variation in the gene-expression data, a key assumption that underlies the dichotomous HD versus normal control analytic approach now standard in the field, where HTT CAG repeat length is not taken into account (16,18–22,31). Although we reason that truly correlated genes would show significant differences in a dichotomous analysis, most studies using dichotomous analytical strategies have not been powered sufficiently to reveal correlated genes. Consistent with this view, in our study using 107 samples, we observed 100% concordance between the direction of correlation in continuous analysis and the direction of fold-change in dichotomous analysis for the probes significant in either continuous (718 probes, \( P\)-value < 0.01) or dichotomous analysis (464 probes, \( P\)-value < 0.01). However, only 154 probes were shared between them, suggesting that our sample size is still not sufficiently large to identify correlated genes through the dichotomous analysis strategy. Moreover, some of the genes showing significant differences in the dichotomous analysis are likely not to represent effects of the CAG repeat length, as their expression was not actually correlated with it.

Despite the modest contribution of CAG allele size to the overall variance, we demonstrated that mathematical models based upon continuous analysis of genome-wide expression values can predict the true CAG allele size. Furthermore, where clinical data were available, the predicted CAG size and age at onset of motor symptoms were significantly (inversely) correlated and the set of negatively correlated genes in lymphoblastoid cell lines was also significantly enriched in the top ranked CAG-correlated genes in HD brains. However, as expected for any approach attempting to capture the effects of sequence variation at a single locus, the prediction models were not perfect. The differences between actual CAG repeat size and gene-expression-predicted HTT CAG length seem likely to reflect a number of factors. First, HTT is a pleiotropic gene whose gene product (i.e. huntingtin protein) has been implicated in multiple processes, ranging from membrane trafficking to chromatin function. This implies that the relatively subtle impact of lengthening the huntingtin CAG encoded polyglutamine tract is unlikely to be completely circumscribed either by gene-expression values alone or indeed by any other type of unidimensional data. It seems reasonable, therefore, that model building to more completely capture the cellular state that is imparted by the CAG allele will require integration of different kinds of quantitative molecular and biochemical phenotypic data, measured across the members of the same allelic panel. It is noteworthy, however, that the biological processes that were highlighted by the CAG length-dependent expression signatures derived from the human lymphoblastoid cell line data sets were similar to those that emerged from continuous analysis of isogenic mouse knock-in HTT embryoic stem cell lines with increasing allele sizes (23). Both cell types yielded energy, ribosome and synaptic function, though in the former a novel ‘immune’ process was also significant, while in the latter a ‘developmental process’ was significant, implying the readout of some effects of the repeat may be cell-type specific.

A second important contributory factor is the genetic background of the individual, which may vary considerably across the cell lines used in our study, and is likely to influence the steady-state level of gene expression (and other measures) to different extents for different genes in different individuals, even in cells with the same HTT CAG repeat allele. The influence of genetic background is expected because the age at onset of HD motor symptoms is highly variable and, while ~66% of the variance can be explained by the HTT CAG repeat length (11), the remainder of the variance may be explained by a combination of genetic factors, interactions with environmental factors and/or simply chance events (32,33). The eventual availability of genome-wide genotyping data for each of the individual samples that comprise our lymphoblastoid cell line panel, and for other HTT CAG allelic cell or tissue panels, will permit genetic background to become an integral contributor to mathematical modeling of the CAG-size correlated biological networks that drive the disease process. Certainly, knowledge of DNA sequence variants at loci elsewhere in the genome that modulate processes continuously correlated with CAG allele size would become top candidates for potentially modifying HD pathogenesis.

In summary, we have evaluated an analytical discovery approach that is grounded in a genetic hypothesis, based upon observations of the effects of the HTT CAG repeat in human studies. This posits that the CAG repeat acts as a functional polymorphism across the allele distribution but whose effects are sufficiently extreme at the high end of the allele distribution to lead to the overt clinical manifestations of HD in a fully penetrant manner. Consistent with this hypothesis, continuous analysis across a full spectrum of CAG sizes identified CAG-correlated gene-expression signatures that were reproducible and somewhat predictive of true CAG allele size (and age at onset), despite the ‘noise’ contributed to the expression data sets by a host of other more prevalent factors. The success of our proof-of-concept study, with a single type of molecular phenotypic data, in a cell type for which extensive collections exist but which is not the most vulnerable to the disease process, supports the use of modeling based on continuous analysis for discovery of the HD disease process. Our results should spur such model building using integrated comprehensive multi-dimensional data sets in allelic collections of more disease-relevant cell types, such as those now enabled by the advent of human-induced pluripotent stem cell and human neuronal progenitor cell (NPC) technologies. Assuming a similar effect size of R-square value ~0.21, 30 samples are required to detect a significant correlation between HTT CAG repeat length and gene expression (95% significance level and 80% power). It is conceivable that in cells of neuronal character,
such as induced pluripotent stem cell-derived NPC or differentiated neurons, the R-square value (i.e. effect size) may be higher, increasing power if the CAG has a greater proportional effect on gene expression. In addition, generation of isogenic NPC lines from HD patients will significantly decrease the noise contributed by the variable genetic background in patient samples and will therefore further facilitate refinement of the relationship between HTT CAG repeats and biological changes.

**MATERIALS AND METHODS**

**Microarray gene-expression profiling**

To test the predictive power of CAG-correlated gene-expression signature, 107 EBV-transformed human lymphoblastoid cell lines (discovery set) from 105 independent subjects (normal control and HD subjects) were grown in RPMI-1640 (37°C, 5% CO₂, 10% fetal bovine serum, 100 IU/ml penicillin and 100 IU/ml streptomycin) for the microarray gene-expression experiments. The HTT CAG repeat length was determined by a method described previously (34,35). For microarray experiments, total RNA from 107 samples were extracted using TRIzol reagent (Invitrogen) and further cleaned up by RNeasy kit (Qiagen). All RNA samples passed quality controls including 260/280 ratio and 28S/18S ratio (Bioanalyzer assay). Subsequently, total RNA (5 μg) was converted into cDNA using SuperScript II reverse transcriptase (Invitrogen) for cRNA synthesis, and labeled probe (25 μg) was hybridized to Affymetrix U133 + 2 arrays as recommended by the manufacturer.

To test the reproducibility of the CAG-correlated gene-expression signature, we chose 20 cell lines from the initial discovery set. Three independent frozen cell stocks for each cell line were used as discussed earlier to obtain six sets of RNA specimens per cell line from six different culture experiments (two RNA samples/frozen stock). Total RNA from the 6 × 20 replication samples was prepared and tested for quality control as discussed earlier. To synthesize labeled probes from total RNA, Single-Round RNA Amplification and Biotin Labeling System (Enzo) was used, and labeled probe (25 μg) was hybridized to Affymetrix U133 + 2 arrays as recommended by the manufacturer. All microarray data including the discovery set (107 samples) and replication set (120 samples) were processed together for background correction and normalization using geRMA (R, 2.11.1; gcrma, 2.20.0) followed by batch effect correction (36,37).

**Identification of CAG-correlated gene-expression signature, model construction and testing for predictive power**

Supplementary Material, Figure S2 describes our analysis flows for the pilot experiments and the formal testing. Briefly, to build prediction models, discovery set was randomly split into a training set (97 samples) and a test set (10 samples). For each split, only the training set was used to calculate correlation (Pearson’s correlation) between expression level and HTT CAG repeat length for each probe. Then, the training set was sorted based upon the correlation analysis results. We used the partial least squares algorithm in order to generate decreased numbers of variables that capture the covariance between the gene-expression levels and the CAG repeat length. We selected the top n number of correlated probes to construct 10 PLS variables and then used them as independent variables to model the HTT CAG repeat length in a multiple regression analysis context. The same model parameters were used for the test samples to predict CAG repeat length, and predicted CAG repeat lengths were compared with experimentally determined CAG repeat length. These procedures were performed iteratively in order to assess prediction error rates and the general model performance. For the formal statistical test (Supplementary Material, Fig. S2B), the discovery set was split into a training set (97 samples) and test set (10 samples) 10 000 times, and for each split, seven prediction models were constructed using top 200, 250, 300, 350, 400, 450 or 500 probes in the training set. The same model parameters were used to predict CAGs of the test samples, and the average of the seven predicted CAGs for each test sample was taken as the final predicted CAG repeat length of test samples. Variables (e.g. CAG and gene-expression data) were standardized for statistical modeling.

**Effects of the size of training samples**

In order to determine the effects of sample size in the training set on prediction accuracy of models, we split the discovery set into a training set and a test set using different ratios. For a given ratio, for example 10 training samples and 97 test samples, seven prediction models using different top probes (200, 250, 300, 350, 400, 450 or 500) were prepared and CAGs of the test samples were predicted. For a given split, these procedures were repeated 1000 times. The average of predicted CAG for a given ratio and split were compared with the true CAG of test samples to calculate correlation and significance of correlation. The same analysis was repeated using permuted CAGs of training samples.

**Test of reproducibility of CAG-correlated signature**

To test reproducibility of CAG-correlated gene-expression signature, we built prediction models and compared predicted CAGs of independent RNA samples from the replication samples. For each replication sample, we excluded corresponding seven expression data (one from discovery set and six from replication set) to build prediction model using 200, 250, 300, 350, 400, 450 or 500 top correlated probes. Then, the same model parameters were used to predict CAG lengths of seven unused data, and the average of the seven models were taken as the final predicted CAGs of test samples. These procedures were repeated to predict CAGs of all 20 replication samples. Pair-wise correlation was calculated to estimate the similarity of predicted CAGs of same cell lines.

**CAG-correlated genes and pathways**

To identify genes and pathways that are correlated with CAG repeat length, we used 107 discovery samples to perform correlation statistics. To identify genes consistently correlated with the CAG repeat length, we randomly split the discovery set into a training set (97 samples) and a test set (10 samples), and performed model optimization procedures as described in the Optimized Model result section. Once the optimized model has been identified in each iteration, we calculated correlation coefficients and P-values with CAG repeat length for all participating probes.
These procedures were repeated 10 000 times, and we summarized the results by calculating frequency, mean correlation coefficient, and mean correlation P-value for each participating probe.

For pathway analysis, the sigPathway program was used (27). Briefly, the entire discovery set was used to calculate correlation statistics and then the data set was permuted 10 000 times to construct null distributions of enrichment scores. In this analysis, enrichment score is the sum of correlation coefficients. Significantly correlated pathways were identified by gene-set permutation P-value.

Gene-set enrichment analysis of CAG-correlated gene sets in human HD brain regions

To test statistical enrichment of lymphoblast CAG-correlated genes in CAG-correlated genes in human HD brains, we compiled a positively correlated gene set (315 probes; Pearson’s correlation P-value <0.01) and a negatively correlated gene set (403 probes; Pearson’s correlation P-value <0.01). These gene sets were compared with the human HD brain gene-expression data from Sage Bionetworks (sagebase.org; cerebellum, prefrontal cortex and visual cortex) from samples for which we had previously determined HITT CAG allele size, using a similar algorithm as used in the sigPathway program (27). Briefly, gene accession numbers were used to match lymphoblast and human brain expression data, and correlation between gene expression and CAG repeat length in the human brain gene-expression data were calculated. Then, the sum of correlation P-values of genes in the gene sets were calculated (i.e. true gene set enrichment score) and null distributions of enrichment scores were constructed by permuting the brain gene-expression data 100 000 times followed by calculating gene-set enrichment score. Significance of enrichment was evaluated by one-tail empirical P-value.

Statistical analysis

All computational and statistical analyses were performed using R (version, 2.11.1).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by grants from CHDI Foundation, Inc., the National Institutes of Health, NINDS Massachusetts HD Center Without Walls (grant number NS16367 to J.F.G. and M.E.M. and NS32765 to M.E.M.) and NLM i2b2 (grant number LM008748-01 to I.S.K.) and a gift from an Anonymous Donor. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES


