RESEARCH ARTICLE

A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment [version 1; peer review: 1 approved, 1 not approved]

Eleni Mina¹, Willeke van Roon-Mom¹, Pernette Verschure², Peter A.C. 't Hoen¹, Mark Thompson¹, Rajaram Kaliyaperumal¹, Kristina Hettne¹, Erik Schultes¹, Barend Mons¹, Marco Roos¹

¹Department of Human Genetics, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands

²Synthetic Systems Biology and Nuclear Organization Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, 1098 XH, The Netherlands

V1 First published: 25 Oct 2017, 6:1888
<https://doi.org/10.12688/f1000research.9703.1>
 Latest published: 25 Oct 2017, 6:1888
<https://doi.org/10.12688/f1000research.9703.1>

Abstract

Background: Huntington's Disease (HD) is an incurable disease of the adult brain. Massive changes in gene expression are a prominent feature. Epigenetic effects have been reported to be implicated in HD, but the role of chromatin is not well understood. We tested if the chromatin state of dysregulated genes in HD is affected at a genome-wide scale and examined how epigenetic processes are associated with CpG-island-mediated gene expression.

Methods: Our general approach incorporates computational and functional analysis of public data before embarking on expensive wet-lab experiments. We compared the location in the genome of the genes that were deregulated in HD human brain, obtained from public gene expression data, to the location of particular chromatin marks in reference tissues using data from the ENCODE project.

Results: We found that differentially expressed genes were enriched in the active chromatin state, but not enriched in the silent state. In the caudate nucleus, the most highly affected brain region in HD, genes in the active state were associated with transcription, cell cycle, protein transport and modification, RNA splicing, histone post-translational modifications and RNA processing. Genes in the repressed state were linked with developmental processes and responses related to zinc and cadmium stimulus. We confirmed that genes within CpG-islands are enriched among HD dysregulated genes in human and mouse in HD. Epigenetic processes were associated more with genes that overlap with CpG-islands than genes that do not.

Open Peer Review

Approval Status

	1	2
version 1
25 Oct 2017	view	view

1. **Núria Queralt-Rosinach** , Scripps Research Institute, La Jolla, USA
2. **Andreas R. Pfenning**, Carnegie Mellon University, Pittsburgh, USA
Irene Kaplow, Carnegie Mellon University, Pittsburgh, USA

Any reports and responses or comments on the article can be found at the end of the article.

Conclusion: Our results suggest that massive transcriptional dysregulation in HD is not matched by large-scale relocation of gene activity, i.e. inactive chromatin regions are altered into actively expressed chromatin regions and vice versa. We expect that changes in epigenetic chromatin state might occur at the level of single genes (e.g. promoters, gene body) and scattered genomic sites (e.g. CTCF sites, enhancer regions) instead of large-scale genomic regions.

Keywords

Huntington's disease, gene expression, CpG-islands, caudate nucleus, chromatin regions

This article is included in the **Bioinformatics** gateway.

Corresponding authors: Eleni Mina (e.mina@lumc.nl), Marco Roos (m.roos@lumc.nl)

Competing interests: No competing interests were disclosed.

Grant information: The research leading to these results is supported by grants received from the Netherlands Bioinformatics Centre (NBIC) under the BioAssist program, the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements No. 305444 (RD-Connect; HEALTH.2012.2.1.1-1-C) and No. 270129 (Wf4Ever; ICT-2009.4.1), and the Innovative Medicines Initiative Joint Undertaking project Open PHACTS (grant agreement No. 115191).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Mina E *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

How to cite this article: Mina E, van Roon-Mom W, Verschure P *et al.* **A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment** [version 1; peer review: 1 approved, 1 not approved] F1000Research 2017, 6:1888 <https://doi.org/10.12688/f1000research.9703.1>

First published: 25 Oct 2017, 6:1888 <https://doi.org/10.12688/f1000research.9703.1>

Introduction

Huntington's Disease (HD) is a complex disease of the brain associated with massive changes in gene expression. The genetic cause was identified in 1993¹, but no successful treatment has been found yet. HD is a dominantly inherited neurodegenerative disease that affects 1 – 10/100,000 individuals, making it the most common heritable neurodegenerative disorder².

Although HD is considered a monogenic disease³, extensive research since 1993 into the underlying pathology suggests that the disease mechanisms are more complex than originally considered. Transcriptional dysregulation is a widespread phenomenon in HD that can be observed well before the first clinical symptoms appear⁴. It suggests that mutant huntingtin causes a broad and complex cascade of downstream effects. There are several ways in which mutant huntingtin can interfere with the transcriptional machinery and alter gene expression^{5,6}. Given the extent of the transcriptional changes, the question arises if epigenetic mechanisms that operate at a genome-wide scale are also involved in HD. There is increasing evidence for epigenetic mechanisms playing a role in HD in human and model systems. For instance, earlier computational analysis of HD gene expression data showed that expression is deregulated in large genomic regions, indicative of a coordinated genome-wide mechanism⁷. These studies did not determine whether genome-wide alterations in gene expression are associated with changes in the composition of histone modifications.

More recently, H3K4me3 was shown to be enriched in 136 loci in an HD case/control study, which included genes that may affect the neuronal epigenome at large⁸. In HD mice, down-regulated genes were found to be associated with a selective decrease of H3K27ac and RNA Polymerase II⁹. Inhibitors of HDAC1, broad-range regulators of chromatin structure, have indeed been shown to be effective in HD mice¹⁰. The interplay of enzymes involved in post-translational histone modifications like histone deacetylases (HDACs), may make chromatin less accessible in many places and therefore alter gene expression patterns. In *Drosophila*, the homolog of htt was found to suppress position effect variegation (PEV), possibly by influencing PEV modifier genes¹¹.

A role for epigenetic mechanisms in HD is further corroborated by numerous neurodevelopmental and neurodegenerative disorders that have been associated with an altered chromatin structure^{12–14}. Neuroepigenetics has therefore become a prime topic of interest¹⁵, with researchers seeking to identify epigenomic signatures and how they can contribute to brain health or brain disease¹⁶.

Considering the growing body of evidence that epigenetic regulation is involved in neurological pathology, we asked if the massive transcriptional dysregulation in HD coincides with large scale changes in chromatin state across the genome. Under that assumption, we hypothesize that significant numbers of differentially expressed genes in HD will be found in regions that are not normally associated with active chromatin states.

Here we report on a computational test of our hypothesis before laboratory experiments using only publicly available data sets: HD gene expression from the Gene Expression Omnibus (GEO)¹⁷

and chromatin state from ENCODE¹⁸. We assessed the overlap between differentially expressed genes and chromatin state, and we applied literature-based concept profile analysis (CPA) to interpret our findings^{19,20}.

Our results indicate that the massive transcriptional dysregulation in HD is not matched by a significant largescale change of activity of genes that are part of inactive chromatin states in reference tissue. Our report includes a functional characterization of differentially expressed genes in HD in relation to huntingtin, chromatin state and CpG islands, based on a literature-based semantic analysis.

The analysis we performed to test our hypothesis is part of an interdisciplinary research approach where computational analysis is used to help steer laboratory experiments to increase the overall efficiency of a research laboratory²¹.

Methods

Concept profile analysis

In Concept Profile Analysis (CPA), the vector space model is used to associate two concepts mined from the literature with each other. Advantages of this model include efficient and transparent comparisons, and the possibility of attaching a weight to the association²⁰. The CPA algorithms have previously been used for a range of different gene expression data analysis purposes such as functional annotation²², comparison of studies²³, prediction of novel interactions²⁴, generation of gene sets^{19,25}, and association with chemical structures²⁶. The methodology has been described previously²⁷. In short: In our database, every concept is associated with PubMed records using the indexing engine Peregrine (<https://trac.nbic.nl/data-mining/>) which is equipped with an in-house thesaurus of biomedical and chemical concepts that have been prepared for text mining^{28,29}. For all concepts except genes and Gene Ontology (GO) terms the PubMed records are comprised of the texts in which the concept is mentioned. For genes, only a subset of PubMed records are used in order to limit the impact of ambiguous terms and distant homologs. GO terms are sometimes given as words or phrases that are infrequently found in the normal texts. To still provide broad coverage of GO terms, the PubMed records that were used as evidence for annotating genes with the GO term are added. For every concept in the thesaurus that is associated to at least five PubMed records, a vector containing all concepts related to the main concept (direct co-occurrence), weighted by the symmetric uncertainty coefficient is created. We call this a “Concept Profile”. Concept profiles are matched to identify similarities via their shared concepts (indirect relations). Any distance measure can be used for this matching such as the mutual information, inner product, cosine angle, Euclidean distance or Pearson's correlation. The CPA Web Services that we used for our analysis use an inner product measure³⁰. These web services can be found in the BioCatalogue web service registry <https://www.biocatalogue.org/services/3559>.

Data analysis and interpretation

For data analysis and interpretation we implemented a series of workflows using the Taverna workbench (Taverna workbench version 2.4.0)^{31,32}. Taverna is an open source software for the development and execution of workflows. Our workflows are deposited online on the Zenodo repository (<https://doi.org/10.5281/zenodo.164201>)³³

and <https://doi.org/10.5281/zenodo.16419834>) and the myExperiment platform (<http://www.myexperiment.org/packs/553>).

Workflows for data analysis

The first Taverna workflow `load_data_identify_DE_genes_Array_A`, was implemented to examine differential gene expression between control and HD samples. Required workflow inputs were two data files with gene expression values and the phenotype information that describes the samples from the microarray experiment. Differential expression was computed using moderated *t* statistics with the package `limma`³⁵ (version 3.14.1), which is provided by the bioconductor project³⁶, R version <https://www.bioconductor.org/>. We analyzed each brain region separately, because previous analysis³⁷ revealed regional patterns in gene expression. The workflow maps the expression data from probes to entrez gene ids using the Affymetrix Human Genome U133 set annotation data, (packages `hgu133a` and `hgu133b`, version 2.8.0; <https://bioconductor.org/packages/hgu133a.db>; <https://bioconductor.org/packages/hgu133b.db>). When multiple probe names map to the same gene id, the ones exhibiting the most significant changes were used for further analysis. Final outcome of this workflow is a report. Each row is composed of a gene id, a fold change and its corresponding *P*-value indicating the significance of every change in gene expression, between HD and controls for each brain region. Adjusted *P*-values, generated by Benjamini and Hochberg's method for multiple testing correction, are also included³⁸. This workflow can be adjusted to compute differential gene expression between other variables such as male/female or the grade of disease pathology by editing the nested workflow “`compute_DE_limma`” within the R workflow component. An additional workflow `create_exprs_obj_download_files` is included in the myExperiment pack that was used to download data from the ArrayExpress repository. The workflow saves the gene expression data and the corresponding phenotype file in the directory indicated in the workflow input.

We note that this particular microarray experiment was composed of two microarrays, Human Genome U133A and U133B. For convenience we added the workflow: Get differentially expressed genes for Array B one brain region, but in principle the first workflow could be reused by adjusting the “libraries” component.

The second workflow `map_genes_on_chromosome` uses the output from the first workflow in order to map genes to their corresponding genomic location. The workflow uses the Biomart³⁹ service within R, to obtain information regarding the position of each gene at the chromosome, HGNC gene symbols⁴⁰, transcription start and end site and the transcription strand. The database that was used was the Ensembl genes 68 from Sanger institute and the *Homo Sapiens* dataset GRCh37.p8. The mouse assembly that was used to map genes to their chromosomal location was Dec 2011, GRCm38mm10.

The last workflow `get_promoter_region_calculate_overlaps`, first computes a promoter region for each gene and then operates on genomic intervals to compute gene promoters that overlap with a genomic region. The promoter region is computed for each gene, according to prespecified values, indicating the number of base pairs (bp) upstream and downstream of the transcriptional start site

(TSS); for the CpG island analysis 5000bp upstream and 2000bp downstream and for the chromatin states analysis 50bp upstream and 50bp downstream was used. The decision for the promoter size in each case was taken after discussing with the domain experts and from knowledge acquired from previous experiments using the ENCODE data from Ernst *et al.*⁴¹. Using those data we performed multiple runs with different input values for the “upstream” and “downstream” variables, and the overlap parameter (Supplementary File 1).

When genes had multiple transcription start sites, we computed a promoter region for every TSS. Next part of this workflow is to compute overlapping regions between the input datasets. It includes a two sample Kolmogorov – Smirnov test to compare the empirical cumulative distribution functions (ecdf) of the *P*-values between the gene promoters that overlap with a specific genomic region and the ones that do not. The null hypothesis tested here was that there is no difference between the two groups. The workflow returns two lists of genes, one for the genes that overlap with a particular genomic region and another that does not. Furthermore, the results of the statistical test are reported: the ks test statistic (maximum distance *D* between the ecdf of the two samples) and the *P*-value of the test. If *P*-value < 0.05 we reject the null hypothesis.

Workflows for data interpretation

The workflows that we implemented for gene interpretation and gene prioritization are based on the workflow pack at Zenodo, <https://doi.org/10.5281/zenodo.164198> (see also, at myExperiment: <http://www.myexperiment.org/packs/368>), and are implemented using CPA web services³⁰.

The CPA workflow `Annotate gene list with top ranking concepts` annotates a gene list with top ranking concepts by matching concept profiles of genes with for example in our case the concept set of Biological Processes. The web services part of this workflow query the Anni database²⁷ that stores the concept profiles for each concept of interest. The first web service `mapDatabaseIDListToConceptIDs` maps a list of concepts, in our case Entrez gene identifiers, to their corresponding concept profile ids. Necessary inputs are a concept list (gene list) in a comma separated file, and the database identifier of the gene list necessary for the mapping (EG for Entrez Gene, see here: https://www.biocatalogue.org/soap_operations/41197 for more details on database identifiers). The next web service `getSimilarConceptProfilesPredefined` matches our gene list with the predefined concept set “Biological Process” (ID = “5”), and gives the top scoring biological processes that describe our gene list. For a complete list of predefined concept sets, the workflow `List Concept Sets` (provided in the current pack) can be run to choose the ID of the predefined concept set of interest. The web service `getConceptName`, gives the complete (human readable) names of the top matching biological processes. Lastly, the workflow `Explain score between two concepts` can be included to the analysis to provide evidence for the association between each gene and the annotations of the biological processes. The evidence reported is a list of concepts that link one concept with another and the contribution to the overall strength of the association. In addition, the corresponding concept ids are reported.

The workflow Prioritize gene list can prioritize a set of genes with respect to their association with particular concepts, in our case the HTT concept and epigenetics (concept profile: “epigene”). In order to obtain the concept profile identifiers the workflow getConceptSuggestionsFromTerm needs to run first.

Data obtained from public sources

Human brain data. The HD human brain data that was used in this analysis was originally produced and analyzed by A. Hodges and co-workers³⁷. This experiment contains 44 HD positive cases and 36 age and sex matched controls. The processed data are available from the public repository NCBI Gene Expression Omnibus, entry GSE3790. Three brain regions were included and analyzed; the caudate nucleus, frontal cortex and cerebellum, with an Affymetrix Microarray GeneChip (Human Genome U133A and U133B). Furthermore, the HD positive cases were further classified based on whether symptoms were present or absent and according to Vonsattel grade of disease pathology (scale = 0 – 4). In our analysis we used the processed data and performed our own differential gene expression analysis (Dataset 1⁴²) with the workflow that performs differential gene expression analysis that was described previously in Methods.

Dataset 1. Gene expression data

<http://dx.doi.org/10.5256/f1000research.9703.d179468>

This folder contains the gene expression data for the three human brain regions and the three mouse brain regions.

CpG island data. CpG island information in the human genome was obtained from UCSC genome browser⁴³, hg19 assembly FEB 2009 1 (Dataset 2⁴⁴). Here, CpG islands are marked as the DNA regions where the following conditions hold:

- GC content of 50% or greater
- length greater than 200 bp
- ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs in the segment

The ratio observed/expected (Obs/Exp) CpG was calculated as follows:

$$\text{Obs/Exp CpG} = \frac{\text{number of CpG}}{\text{number of C} \times \text{number of G}} \times N,$$

where N is the total amount of nucleotides in the sequence that is being analyzed. For CpG island information of the mouse genome the assembly dec 2011 (GRCm38/mm10) was used.

Dataset 2. CpG islands

<http://dx.doi.org/10.5256/f1000research.9703.d179469>

This folder contains the CpG island information both for human and the mouse data.

Chromatin states data. The chromatin marks were obtained from the encode project⁴⁵. The chromatin states were part of an integrative analysis of 111 reference human epigenomes profiled for histone modification patterns based on DNA accessibility, DNA methylation and RNA expression. We used the two cell types that were more suitable for our analysis; the anterior caudate and dorsolateral prefrontal cortex (Dataset 3⁴⁶ and Dataset 4⁴⁷, respectively). The chromatin states we used for the current analysis were: active transcription start site proximal promoter (TssA), bivalent regulatory region (TssBiv), heterochromatin (Het) and repressed Polycomb (ReprPC).

Dataset 3. Chromatin states for the anterior caudate

<http://dx.doi.org/10.5256/f1000research.9703.d179470>

This folder contains the four chromatin state data for the anterior caudate. Active TSS proximal promoter: TssA, bivalent regulatory region: TssBiv, heterochromatin: Het, repressed Polycomb: ReprPC.

Dataset 4. Chromatin states for the prefrontal cortex

<http://dx.doi.org/10.5256/f1000research.9703.d179471>

This folder contains the four chromatin state data for the prefrontal cortex. Active TSS proximal promoter: TssA, bivalent regulatory region: TssBiv, heterochromatin: Het, repressed Polycomb: ReprPC.

Mouse data. The mouse brain data was taken from a published study¹⁰. There, total RNA was extracted from cortex, striatum, and cerebellum from WT and R62 transgenic mice. This study examines the effect of the HDACi 4b inhibitor on the disease phenotype. However only the data from animals treated with vehicle was used (Dataset 1⁴²). This study used Illumina Mouse Mouseref-8 Expression Beadchips v1. The raw data were analyzed using the Bioconductor packages and contrast analysis of differential expression was performed by using the LIMMA package. The differential expression values are available in the supplemental material of that publication.

Results

Chromatin state analysis and semantic interpretation

To test if and how differential gene expression in HD is associated with particular chromatin states we used publicly available datasets from the GEO and ENCODE public repositories. We selected HD gene expression data from three regions of the brain made available by Hodges and coworkers³⁷, and data from the ENCODE consortium carrying information about four chromatin states. These were: active TSS proximal promoter (TssA), bivalent regulatory region (TssBiv), heterochromatin (Het) and repressed Polycomb (ReprPC)⁴⁵. Briefly, the first state pertains to active genes, the second to repressed genes that are ready to be activated, and the latter two represent repressed genes. The chromatin states were part of an integrative analysis of 111 reference human epigenomes profiled for histone modification patterns, DNA accessibility, DNA methylation and RNA expression. The different chromatin states were defined by a computational model that is based on a multivariate Hidden Markov Model⁴⁸.

First we confirmed the results reported by the previous study³⁷: large and numerous changes in transcriptional activity in the caudate nucleus brain region and notably smaller changes in the frontal cortex and cerebellum. More specifically, the number of differentially expressed genes was 5219, 127 and 96 for each brain region respectively with a FDR of 0.05. These results confirm previous observations illustrating that specific HD-affected brain regions exhibit defined changes in gene expression, in line with observable physiological effects³⁷.

Secondly, we paired the data describing gene expression changes in the caudate nucleus and frontal cortex from the Hodges study to chromatin state data from the anterior caudate and dorsolateral prefrontal cortex from ENCODE. We considered these brain regions as the most comparable between the two studies. The cerebellum was excluded from this part of our analysis due to the absence of chromatin state data at ENCODE for the cerebellum. We determined the reference chromatin state of the genes from the gene expression study by determining the overlap of their promoter regions with the start and end positions of the chromatin state from the aforementioned brain regions used by ENCODE. The effect of a chromatin state on differential gene expression in HD was assessed by comparing the distribution of expression levels of genes overlapping with a particular chromatin state with the distribution of expression of non-overlapping genes. If chromatin state has substantial functional impact on the genes that are differentially expressed in HD, then these are expected to be significantly

different. More details on the promoter region calculation and the overlap between the genomic regions can be found in the Methods section.

Specifically, we compared the distribution of the *P*-values for differential expression between these two groups of genes and assessed the difference by a KolmogorovSmirnov test, in the two brain regions and for each chromatin state (Figure 1). In the caudate region, we found an enrichment of genes overlapping with the active TSS and a depletion of genes overlapping with ReprPC (Figure 1). Heterochromatic or bivalent chromatin were not significantly associated with a genomic location of differentially expressed genes. A similar, but less pronounced pattern was observed for the frontal cortex for both the active TSS and ReprPC. In addition, in the frontal cortex the number of genes overlapping with the bivalent state was reduced. The magnitude of the association of a brain region with each chromatin state is in line with the HD neurodegeneration pattern, where the caudate nucleus exhibits the largest gene expression changes while the frontal cortex exhibits an intermediate to low pathology. In summary, we found enrichment of differentially expressed genes in the active chromatin state of reference tissue, but no strong evidence for significant enrichment in chromatin regions that are associated with repressed gene expression activity.

To further interpret the results from the chromatin state analysis, we used literature information (CPA^{19,27}) to assess the biological

Figure 1. Chromatin states. For each brain region, the *P*-value distribution for differential expression in HD patients compared to controls was compared between genes overlapping with a specific chromatin state and all other genes, in the anterior caudate and dorsolateral prefrontal cortex cells. The plot displays the maximum absolute distance *D* between the cumulative distribution function of the *P*-values for each group of genes (overlapping and non overlapping), as reflected by the KS test statistic. Positive values correspond to an enrichment of differentially expressed genes in a chromatin state, negative values with a depletion. Stars indicate a significant enrichment/depletion of the KS test (*P*-value < 0.05). TssA= active TSS, TssBiv= bivalent Tss, Heterochr= Heterochromatin, ReprPC= repressed Polycomb state. For caudate: TssA *D*= 0.31, TssBiv *D*=0.04, Heterochr *D*=0.17, ReprPC *D*= 0.27, for frontal cortex TssA *D*=0.168, TssBiv *D*=0.097, Heterochr *D*=0.095, ReprPC *D*=0.147. The corresponding *p*-values were for the caudate : TssA *p*val < 2.2e-16, TssBiv *p*val= 0.59, Heterochr *p*val= 0.10 , ReprPC *p*val=4.5167e-12 , for frontal cortex TssA *p*val < 2.2e-16 , TssBiv *p*val= 0.0002, Heterochr *p*val= 0.872, ReprPC *p*val= 1.4596e-05.

processes within the lists of genes per chromatin state that were associated with gene deregulation (Dataset 5⁴⁹).

Using CPA, we found that for the caudate nucleus, genes in the active TSS state are mainly associated with processes related to transcription, cell cycle, protein transport and modification, RNA splicing, chromatin modifications and RNA processing. Genes in the repressed polycomb chromatin state are mainly associated with (brain) developmental processes and responses related to zinc and cadmium stimulus.

For the frontal cortex, genes in the active TSS state are associated with protein transport and modification, cell cycle, RNA splicing and signal transduction pathways (MAPK, notch, smoothened). Genes in the bivalent TSS are associated with brain development, neurogenesis, synapse and responses related to zinc and cadmium stimulus. We found that the genes that are part of the polycomb repressed group to be associated with similar functions as the genes in the bivalent state: (brain) developmental processes, responses related to zinc, cadmium and copper stimuli.

We note that at this point of our analysis, we filtered out genes of which the promoters were labelled with more than one state, using the criterion of at least 50 bps overlap in a promoter region of 100 bps. Interpretation of this set of genes would be ambivalent.

Dataset 5. Unique annotations of genes in each chromatin state with Biological Processes

<http://dx.doi.org/10.5256/f1000research.9703.d179472>

This file contains the results from the semantic analysis with Biological Processes of the genes that were overlapping with active TSS, bivalent TSS, heterochromatin and repressed polycomb chromatin state. The annotations that we present here are uniquely characterizing each gene list (as resulted from the CPA out of the top 50 annotations).

Overlap of HD deregulated genes with CpG islands and semantic interpretation

Because CpG island methylation is a known epigenetic regulatory mechanism that is ubiquitous in the human genome and may be a target for genome-wide regulation^{50,51}, we also applied our approach to test if genes within CpG islands are overrepresented among differentially expressed genes in HD. We measured this by similar KS test statistics as in the previous section. We found that genes overlapping with CpG-islands in their promoter region, were significantly enriched in the group of HD-deregulated genes in all three brain regions (Figure 2A; P -value < 0.05), which supports taking this mechanism into account to formulate hypotheses when studying gene deregulation in HD.

To further analyze and verify the association of gene deregulation in HD with the presence of CpG islands, we also analysed overrepresentation in data obtained from a HD transgenic mouse model¹⁰. Our results based on the mouse data corroborated those from the human data (Figure 2B). Comparable to the results for caudate nucleus in human data, the most highly affected mouse brain region, the striatum, showed the biggest overrepresentation of CpG islands among deregulated genes.

We next examined the biological processes that are associated with the differentially expressed genes that do or do not overlap with CpG islands by CPA. We inspected the top 50 annotations of each group. We identified many similarities between these two groups, but also annotations that were specific to each group (Dataset 6⁵²). For example, concepts related to chromatin alterations, such as chromatin remodelling and chromatin (dis)assembly, and histone post-translational modifications, such as (de)acetylation, and (de)methylation, were only found with a high rank in the list of genes containing CpG islands in their promoters.

Figure 2. CpG island effect in human and mouse data. Maximum distance between the cumulative distribution function of the P -values between genes containing a CpG island in their promoter and genes that do not. The maximum distance (D) for each brain region was plotted for both datasets (human (A) and mouse (B) datasets). The gradient adjacent to each plot indicates the extent of neurodegeneration in each brain region. Black represents severe and white mild neurodegeneration. **A:** The analysis performed in the human data. Caudate nucleus is exhibiting the largest differences with a distance D between the two distributions of 0.2808, frontal cortex follows with $D = 0.1482$ and cerebellum with $D = 0.147$. **B:** The analysis performed in the mouse data. The results are analogous to the human data with striatum showing the largest differences with $D = 0.118$, cortex following with $D = 0.0739$ and cerebellum with an insignificant difference (P -value = 0.596 >> 0.05) of $D = 0.0383$. * : depicts significance.

Conversely, lymphocyte activation, angiogenesis, antigen presentation and neurogenesis were only high ranking associations for the non-CpG containing genes.

Some annotations were found in both groups but in a different ranking order. For example, the rankings of gene silencing, RNA splicing and phosphorylation were increased for genes within CpG islands while transcriptional activation and mitotic cell cycle rankings were higher for non-CpG containing genes (Dataset 6⁵²).

Dataset 6. Annotations of CpG containing genes and non CpG containing genes

<http://dx.doi.org/10.5256/f1000research.9703.d179473>

The top 50 annotations characterising each group of genes with CpG islands and without, resulted from the semantic analysis of those two groups of genes with Biological Processes.

Semantic analysis to identify proteins associated with HTT and epigenetics

CPA can also be used to prioritize findings by a specific biological interest. In our case, we aimed to prioritise the list of genes overlapping with CpG islands that were also differentially expressed in HD caudate nucleus by their association with HTT and epigenetics. Here, we prioritised 100 proteins based on their association with HTT and epigenetics and absence of a direct relation with HTT by CPA (Dataset 7⁵³). Such relations are typically novel, or lost in tables or figures. If a novel relation is found (i.e. the relation is not found in our database of relations in MedLine abstracts), then CPA also provides intermediate concepts that link the two concepts.

In Table 1, we present the top 5 novel proteins that have the strongest association, and the intermediate concepts that link each prioritized protein to HTT and epigenetics. The intermediate concepts are grouped under the semantic categories “General”, “Biological Processes”, “Disease or Syndrome”, “Homo Sapiens proteins” and “Molecular Functions”.

Grouping the evidence provides a more complete insight into the processes involved in gene deregulation in HD mediated by CpG islands. For example, for one of our candidates, the amyloid beta (A4) precursor protein-binding, family A, member 1 (*APBA1*), we found nerve tissue (General), endocytosis (Biological Process), Alzheimer’s Disease (disease or syndrome), *GRIN2B* (*H. sapiens* Genes) and kinesin activity (Molecular Function) as intermediate links with HTT. This suggests that mechanisms involving *APBA1* in HD share common components with the mechanisms involved in

endocytosis and Alzheimer’s Disease and those involving *GRIN2B* and kinesin activity. Intermediate links with epigenetics were respectively: epigene (General), methylation (Biological Process), hypomyelination and congenital cataract (disease or syndrome), *CDKN2A* (*H. sapiens* Genes) and *MGMT* - O6-alkylguanine DNA alkyltransferase - (Molecular Function). Accordingly, these concepts provide suggestions about the epigenetic role of *APBA1*, which can be taken into account when further studying the role of *APBA1* in HD.

Dataset 7. Gene prioritization

<http://dx.doi.org/10.5256/f1000research.9703.d179474>

Top 100 novel proteins, resulted from the semantic analysis associated with HTT and epigenetics.

Validation of concept profile gene prioritization

We next investigated whether the prioritized genes reflect valid biological knowledge. Figure 3 shows that CPA is able to prioritize true associations with huntingtin as measured by a gene expression experiment, but that combining experimental (differential expression) measurements and literature evidence enables to select even more specific HD signatures. We used our concept profile technology to match all genes in our database that have a concept profile (12,391 genes) to the “huntingtin” concept profile (black line). We then compared the distribution of those CPA scores to the CPA scores of genes that were found to be differentially expressed in the caudate nucleus (p value < 0.05). We included two gene lists in our analysis: the top 100 most differentially expressed genes (red line) and the top 1000 (green line). The shift in the distribution of CPA match scores between the differentially expressed genes (top 100 and top 1000) and the scores of all genes reflects the added value of CPA (CPA scores of top 100 and top 1000 can be found in Dataset 8⁵⁴). We found a significant shift in the scores when comparing all CPA scores from our database with the top 100 differentially expressed genes: $p = 2.67e-08$ and the top 1000 $p < 2.2e-16$.

Also the top 100 and top 1000 differ significantly ($p = 0.03184$), showing that it is useful to narrow down on the top ranks for follow-up research. In principle, more extreme p-values are associated with higher CPA scores. In addition, to show that our list of 100 prioritized gene-HTT CPA match scores would not have been found by chance, we assessed the percentile score of our list when compared to the frequency distribution of 100 match scores of randomly sampled gene-HTT concept pairs out of the 12,391 genes in our concept profile database (Dataset 9⁵⁵). All genes were in the top 95 percentile, except NTRK3 (55 percentile).

Table 1. Evidence table for the association between the top 5 proteins and the concepts of *HTT* and epigene, grouped in five semantic categories: “General”, “Biological Processes”, “Disease or Syndrome”, “Homo Sapiens Genes” and “Molecular Function”.

gene	concept	general	biological processes	disease or syndrome	h.sapiens gene	molecular functions
1.APBA1	HTT	nerve tissue trafficking	Endocytosis	Alzheimer's Disease	GRIN2B	kinesin activity
		mice,transgenic	protein transport	Friedreich Ataxia	CDK5	NMDA receptor
		PC12 Cells,	Pathogenesis	Neuropathogenesis	ITSN1	Protein Binding
		Caenorhabditis elegans	intracellular protein transport	Degenerative disorder	STX1A	membrane associated guanylate kinase
			RNA Interference	Malnutrition	DLGAP2	cyclin-dependent protein kinase activity
	epigene	Epigene	Methylation	hypomyelination and congenital cataract	CDKN2A	MGMT
		cpg islands	DNA Methylation	Werner Syndrome	APBA1	methyl-cpg binding
		Hypermethylation	Gene Silencing	marinesco-sjogren syndrome	MLH1	calcium channel activity
		Epigenetic Process	tumor suppressor activity	friedreich ataxia 1	DNMT3B	methylase activity
		microsatellite instability	cytosine methylation	Angelman Syndrome	APBA2	PDGFRA
2.KAT2A	HTT	histone acetylation	Transcription, Genetic	Neurodegenerative Disorders	EP300	Histone acetyltransferase activity
		ARID1A	Histone Acetylation	Huntington Disease	CREBBP	ubiquitin activity
		Transcription	histone modification	neu-laxova syndrome	TBP	acetyltransferase activity
		HDAC1	Transcriptional Activation	Recruitment Disease	KAT2B	cyclin-dependent protein kinase activity
			RNA Interference	Recruitment	PPARGC1A	Protein Binding
	epigene	histone epigene	Histone Acetylation	Recruitment	KAT2A	IGL
		chromatin location	Transcription	Adenovirus Infections	KAT2B	DNA Binding
		Transcription	histone modification	Neurodegenerative Disorders	HDAC1	histone deacetylase activity
		histone modification	chromatin remodeling	Infection Disease	EP300	transcription factor binding
			Methylation		DNMT1	histone binding

gene	concept	general	biological processes	disease or syndrome	h.sapiens gene	molecular functions
3.CARM1	HTT	Knock-in histone Nerve Tissue Ca150 PC12 Cells	Transcription histone methylation RNA Interference histone modification protein processing, post-translational	Recruitment muscular atrophy spinal muscular atrophy Malnutrition Disease	EP300 CREBBP ARID1A THOC4 DNM1L	nuclear hormone receptor activity acetyltransferase activity Protein Binding histone acetyltransferase activity DNA Binding
		epigene histone chromatin_immunoprecipitation chromatin location Protein Acetylation	Methylation histone modification histone methylation DNA Methylation Transcription	Recruitment Chimera Cholestasis Hyperhomocysteinemia Cerebrovascular accident	CARM1 EP300 EHMT2 PRMT1 PRMT5	methyltransferase activity DNA Binding methyltransferase 1 histone methyltransferase activity methylase activity
4.SLIT2	HTT	Nerve Tissue Tract Knock-in Caenorhabditis elegans Mutant	Neurogenesis Gene Silencing RNA Interference regulation of osteoblast differentiation central nervous system development	Adult disease Disease Degenerative disorder Malnutrition Kidney Diseases	HDAC5 HDAC6 ISL2 RBP1 PAX6	GTP Binding Protein Binding kinesin activity molecular function transcription factor binding
		epigene Hypermethylation Islands tumor suppressor genes Epigenetic Silencing	Methylation DNA Methylation Gene Silencing tumor suppressor activity Embryonic Development	hypomyelination and congenital cataract Adult disease Proteinuria Diabetic Nephropathy Asthma	SLIT2 SLIT3 SLIT1 DNMT1 CDKN2A	MGMT transcription factor binding deacetylase activity 1-phosphatidylinositol-3-kinase activity binding (molecular function)
5.BNIP3	HTT	caspase Mitochondria SETD2 FRAP1 EP300	Autophagy Cell Death Apoptosis RNA Interference Gene Silencing	Neurodegenerative Disorders dentatorubral-pallidoluysian atrophy muscular atrophy Ischemia Posttransfusion purpura	TGM2 CASP2 CREBBP CASP7 HTRA2	proteasome endopeptidase complex ubiquitin activity cytochrome c activity phenylalanine dehydrogenase activity Protein Binding
		epigene Hypermethylation 5-aza-2'-deoxycytidine Gene Silencing tumor suppressor genes	Methylation DNA Methylation Gene Silencing tumor suppressor activity Transcription, Genetic	NPC hypomyelination and congenital_cataract Ischemia Adenovirus Infections van der woude syndrome	BNIP3 HDAC1 MLH1 CDKN2A DmeICG3861	MGMT IGL DNA Binding cytochrome c activity ABL1

Figure 3. Combination of CPA with differential gene expression for effective gene prioritization. Cumulative distribution of match scores of the concept profiles (CP) between differentially expressed genes in HD with the concept profile of htt: match scores of all genes with a concept profile (black), match scores of the top 1000 differentially expressed genes (green), and the match scores of the top 100 of differentially expressed genes (red).

Dataset 8. Concept profile analysis (CPA) scores for the two gene lists

<http://dx.doi.org/10.5256/f1000research.9703.d179475>

This folder contains the CPA scores per gene list (top 100 differentially expressed genes in the caudate nucleus and top 1000 differentially expressed from the same brain region) that were obtained by matching the gene lists against the HTT concept profile.

Dataset 9. Percentile scores

<http://dx.doi.org/10.5256/f1000research.9703.d179476>

This document presents the percentile score of the prioritized gene list when compared to the frequency distribution of 100 match scores of randomly sampled gene-HTT concept pairs out of the 12,391 genes in our concept profile database.

Discussion

The computational analysis with public data that we present in this paper shows that there is no strong evidence for genome-wide

relocalization of gene activity to repressed chromatin states, at least not at a scale that could explain the massive transcriptional deregulation that we observe in HD. Most of the deregulated genes mapped to the active chromatin state of our reference tissue and were underrepresented in silenced states of chromatin (Figure 1). Previous reports supported the implication of large scale chromatin alterations in gene deregulation in HD. For instance, Anderson *et al.* reported that gene expression is deregulated in large genomic regions in blood and post mortem tissue of HD patients⁷. The authors inferred a relation with repressed and active chromatin, but did not use chromatin annotation data directly. Our results suggest that the association with genome clusters is mostly within the active state and does not extend to disruption of chromatin states at large scales.

Active chromatin is normally more prone to regulation and deregulation. Our results suggest that the epigenetic mechanisms that have been observed in HD are mainly bound to this fraction of chromatin. We speculate that the effects are more closely associated with transcription regulation of individual genes, than with large scale

higher-order rearrangements of chromatin structure⁵⁶. Our CPA analysis of deregulated genes with CpG islands corroborates this suggestion: chromatin-related concepts from our CPA ranked high in this set of genes, while CpG-islands are mostly known for their direct role in transcription initiation of the proximal gene. Nevertheless, CpG-mediated changes in gene expression are a common mechanism for many genes and deregulation of this mechanism is thus likely to have a genome-wide effect. It was unexpected that the differentially expressed genes are significantly depleted in the bivalent chromatin state in the prefrontal cortex. The bivalent state is expected to be associated with genes that are prone to become active. Our expectation was that deregulated genes in this brain region would be weakly enriched in this state, as we did find for the caudate nucleus. We currently lack a good explanation for this result.

We incorporated the computational analysis on public data in our research strategy to advise on next experiments. However, working with public data has its limitations. Here, we had to rely on reference tissue for the chromatin state in order to compare it with gene expression. In addition, the reference tissue chromatin state was measured in healthy individuals. Therefore, we cannot rule out that large scale chromatin effects will be observed when chromatin state and differential expression are measured together in new HD specific experiments. This could reveal new evidence for chromatin state deregulation in HD and give insight in the relationship with transcriptional deregulation, possibly at higher resolution than we could achieve in our current study. However, our results suggest that this is not very likely, leading us to advise fellow HD researchers to prioritise experiments that assess the role of epigenetic mechanisms in HD at the scale of individual genes or small clusters of genes, and within the active fraction of chromatin.

Our results can be used to refine hypotheses about molecular mechanisms involved in HD. For instance, we surmise that the reported association of DNA methylation and chromatin organisation, and the effects of HDAC on the HD phenotype in mice⁵⁷ are bound to the active fraction of chromatin. Furthermore, it appears that CpG islands located within the promoter region of a gene increase the probability that genes in such genomic regions are deregulated in HD. This is in accordance with a study where changes in DNA methylation were observed in cells expressing mutant huntingtin⁵⁸. This in turn suggests that DNA methylation in promoters is implicated in alterations in the brain, which is in accordance with a study that noted changes in DNA methylation in cells expressing mutant huntingtin⁵⁸. Based on our semantic analysis chromatin remodelling, chromatin (dis)assembly, and histone modification were associated with altered gene expression profiles. In contrast, non-CpG containing genes are more likely involved in immune response and neurogenesis, which represents functionally linked processes⁵⁹.

Furthermore, we used CPA as a means to prioritise genes for our hypotheses. For instance, we prioritised genes overlapping with CpG islands by their association with HTT, assuming that CPA rank scores are higher for genes that are higher-up in the cascade of events caused by the mutant protein. We recently showed that this is a fair assumption for CPA, although literature bias cannot be

completely mitigated⁶⁰. Similarly, we further prioritized candidate genes in terms of their association with epigenetics. Genes such as *APBA1*, *KAT2A*, *CARM1*, *SLIT2* and *BNIP3* came forward as the most likely candidates to play a role in HD in the context of downstream effects of HTT involving epigenetic mechanisms. These candidates were filtered on potential novelty: only those genes were reported for which there was no direct association found in our database of PubMed abstracts. This does not exclude associations that were reported in tables and supplemental material that are much harder to mine for technical and legal reasons. Our study also retrieved several well established associations consistent with earlier studies.

In summary, our results show how literature information in combination with data analysis present useful tools for exploration of hypotheses for possible future experiments.

Conclusions

Our methodology offers support for hypothesis generation to elucidate missing links in mechanisms involved in a complex disease such as HD. We have shown how the analysis of microarray data and the integration of publicly available datasets and literature information enables prioritization of associations, such as proteins and mechanisms, that are likely to be involved in HD. In addition, we were able to focus on mechanisms that are associated with epigenetic regulation that may regulate changes that are part of the disease pathology. We argue that such a methodology can be of great value to the scientific community for narrowing down the amount of possible associations but also providing evidence to support a particular hypothesis.

Data availability

All workflows are deposited in Zenodo (<https://doi.org/10.5281/zenodo.16420133>) and <https://doi.org/10.5281/zenodo.16419834>), and the myExperiment platform (<http://www.myexperiment.org/packs/553>).

Dataset 1: Gene expression data. This folder contains the gene expression data for the three human brain regions and the three mouse brain regions. doi, [10.5256/f1000research.9703.d179468](https://doi.org/10.5256/f1000research.9703.d179468)⁴²

Dataset 2: CpG islands. This folder contains the CpG island information both for human and the mouse data. doi, [10.5256/f1000research.9703.d179469](https://doi.org/10.5256/f1000research.9703.d179469)⁴⁴

Dataset 3: Chromatin states for the anterior caudate. This folder contains the four chromatin state data for the anterior caudate. Active TSS proximal promoter: TssA, bivalent regulatory region: TssBiv, heterochromatin: Het, repressed Polycomb: ReprPC. doi, [10.5256/f1000research.9703.d179470](https://doi.org/10.5256/f1000research.9703.d179470)⁴⁶

Dataset 4: Chromatin states for the prefrontal cortex. This folder contains the four chromatin state data for the prefrontal cortex. Active TSS proximal promoter: TssA, bivalent regulatory region: TssBiv, heterochromatin: Het, repressed Polycomb: ReprPC. doi, [10.5256/f1000research.9703.d179471](https://doi.org/10.5256/f1000research.9703.d179471)⁴⁷

Dataset 5: Unique annotations of genes in each chromatin state with Biological Processes. This file contains the results from the semantic analysis with Biological Processes of the genes that were overlapping with active TSS, bivalent TSS, heterochromatin and repressed polycomb chromatin state. The annotations that we present here are uniquely characterizing each gene list (as resulted from the CPA out of the top 50 annotations). doi, [10.5256/f1000research.9703.d179472](https://doi.org/10.5256/f1000research.9703.d179472)⁴⁹

Dataset 6: Annotations of CpG containing genes and non CpG containing genes. The top 50 annotations characterising each group of genes with CpG islands and without, resulted from the semantic analysis of those two groups of genes with Biological Processes. doi, [10.5256/f1000research.9703.d179473](https://doi.org/10.5256/f1000research.9703.d179473)⁵²

Dataset 7: Gene prioritization. Top 100 novel proteins, resulted from the semantic analysis associated with HTT and epigenetics. doi, [10.5256/f1000research.9703.d179474](https://doi.org/10.5256/f1000research.9703.d179474)⁵³

Dataset 8: Concept profile analysis (CPA) scores for the two gene lists. This folder contains the CPA scores per gene list (top 100 differentially expressed genes in the caudate nucleus and top 1000 differentially expressed from the same brain region) that were obtained by matching the gene lists against the HTT concept profile. doi, [10.5256/f1000research.9703.d179475](https://doi.org/10.5256/f1000research.9703.d179475)⁵⁴

Dataset 9: Percentile scores. This document presents the percentile score of the prioritized gene list when compared to the frequency distribution of 100 match scores of randomly sampled gene-HTT concept pairs out of the 12,391 genes in our concept profile database. doi, [10.5256/f1000research.9703.d179476](https://doi.org/10.5256/f1000research.9703.d179476)⁵⁵

Supplementary material

Supplementary File 1: Decision about the promoter region of the chromatin states analysis and CpG islands. This document is included as a reference to describe the decisions that were made concerning the promoter size using an older version of epigenetic data from ENCODE. In this file we present the additional runs of the workflow compute_overlaps that were performed with different parameters in order to test and decide for the best promoter region and overlap parameters.

[Click here to access the data](#)

Author contributions

EM designed and executed the experiments, analyzed the data and wrote the manuscript; WvRM helped design and interpret the experiments as the Huntington's Disease expert and reviewed the manuscript; PH helped design and interpret the experiments as bioinformatics expert and reviewed the manuscript; PJV reviewed the manuscript; RK, MT provided technical support for the web services; KMH helped with the CPA and reviewed the manuscript; EAS reviewed the manuscript; MR helped design and interpret the experiments, reviewed the manuscript, general supervision; BM: senior advice.

Competing interests

No competing interests were disclosed.

Grant information

The research leading to these results is supported by grants received from the Netherlands Bioinformatics Centre (NBIC) under the BioAssist program, the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements No. 305444 (RD-Connect; HEALTH.2012.2.1.1-1-C) and No. 270129 (Wf4Ever; ICT-2009.4.1), and the Innovative Medicines Initiative Joint Undertaking project Open PHACTS (grant agreement No. 115191).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We gratefully acknowledge Herman van Haagen for methodological input, Jelle J. Goeman and Maarten van Iterson for statistical advice and Silvere van der Maarel for advice on the role of epigenetics in disease.

References

- MacDonald ME, Ambrose CM, Duyao MP, *et al.*: **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.** *Cell*. 1993; **72**(6): 971–983.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Landles C, Bates GP: **Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series.** *EMBO Rep*. 2004; **5**(10): 958–963.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Arning L: **The search for modifier genes in huntington disease - multifactorial aspects of a monogenic disorder.** *Mol Cell Probes*. 2016; **30**(6): 404–409.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Cha JH: **Transcriptional dysregulation in Huntington's disease.** *Trends Neurosci*. 2000; **23**(9): 387–392.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Luthi-Carter R, Cha JH: **Mechanisms of transcriptional dysregulation in huntington's disease.** *Clin Neurosci Res*. 2003; **3**(3): 165–177.
[Publisher Full Text](#)

6. Valor LM: **Transcription, epigenetics and ameliorative strategies in huntington's disease: a genome-wide perspective.** *Mol Neurobiol.* 2015; 51(1): 406–423.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Anderson AN, Roncaroli F, Hodges A, *et al.*: **Chromosomal profiles of gene expression in huntington's disease.** *Brain.* 2008; 131(pt 2): 381–388.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Bai G, Cheung I, Shulha HP, *et al.*: **Epigenetic dysregulation of hairy and enhancer of split 4 (HES4) is associated with striatal degeneration in postmortem huntington brains.** *Hum Mol Genet.* 2015; 24(5): 1441–1456.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Achour M, Le Gras S, Keime C, *et al.*: **Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of huntington's disease mice.** *Hum Mol Genet.* 2015; 24(12): 3481–3496.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Thomas EA, Coppola G, Desplats PA, *et al.*: **The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice.** *Proc Natl Acad Sci U S A.* 2008; 105(40): 15564–15569.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Dietz KN, Di Stefano L, Maher RC, *et al.*: **The *Drosophila* Huntington's disease gene ortholog dhft influences chromatin regulation during development.** *Hum Mol Genet.* 2015; 24(2): 330–345.
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Urdinguio RG, Sanchez-Mut JV, Esteller M: **Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies.** *Lancet Neurol.* 2009; 8(11): 1056–1072.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Jakovcevski M, Akbarian S: **Epigenetic mechanisms in neurological disease.** *Nat Med.* 2012; 18(8): 1194–1204.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. He F, Todd PK: **Epigenetics in nucleotide repeat expansion disorders.** *Semin Neurol.* 2011; 31(5): 470–483.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Shin J, Ming GL, Song H: **Seeking a roadmap toward neuroepigenetics.** *Neuron.* 2015; 86(1): 12–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Mo A, Mukamel EA, Davis FP, *et al.*: **Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain.** *Neuron.* 2015; 86(6): 1369–1384.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res.* 2002; 30(1): 207–210.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; 489(7414): 57–74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Jelier R, Goeman JJ, Hettne KM, *et al.*: **Literature-aided interpretation of gene expression data with the weighted global test.** *Brief Bioinform.* 2011; 12(5): 518–529.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Jelier R, Schuemie MJ, Roes PJ, *et al.*: **Literature-based concept profiles for gene annotation: the issue of weighting.** *Int J Med Inform.* 2008; 77(5): 354–362.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Mina E, Thompson M, Hettne KM, *et al.*: **Multidisciplinary collaboration to facilitate hypotheses generation in huntington's disease.** In: *IEEE 11th International Conference on e-Science (e-Science).* 2015; 118–125.
[Publisher Full Text](#)
22. Jelier R, Jenster G, Dorssers LC, *et al.*: **Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation.** *BMC Bioinformatics.* 2007; 8: 14.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Jelier R, 't Hoen PA, Sterrenburg E, *et al.*: **Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease.** *BMC Bioinformatics.* 2008; 9: 291.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. van Haagen HH, 't Hoen PA, de Morrée A, *et al.*: **In silico discovery and experimental validation of new protein-protein interactions.** *Proteomics.* 2011; 11(5): 843–853.
[PubMed Abstract](#) | [Publisher Full Text](#)
25. van Dartel DA, Pennings JL, Hendriksen PJ, *et al.*: **Early gene expression changes during embryonic stem cell differentiation into cardiomyocytes and their modulation by monobutyl phthalate.** *Reprod Toxicol.* 2009; 27(2): 93–102.
[PubMed Abstract](#) | [Publisher Full Text](#)
26. Hettne KM, Boorsma A, van Dartel DA, *et al.*: **Next-generation text-mining mediated generation of chemical response-specific gene sets for interpretation of gene expression data.** *BMC Med Genomics.* 2013; 6: 2.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Jelier R, Schuemie MJ, Veldhoven A, *et al.*: **Anni 2.0: a multipurpose text-mining tool for the life sciences.** *Genome Biol.* 2008; 9(6): R96.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Hettne KM, van Mulligen EM, Schuemie MJ, *et al.*: **Rewriting and suppressing UMLS terms for improved biomedical term identification.** *J Biomed Semantics.* 2010; 1(1): 5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Hettne KM, Stierum RH, Schuemie MJ, *et al.*: **A dictionary to identify small molecules and drugs in free text.** *Bioinformatics.* 2009; 25(22): 2983–2991.
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Hettne K, van Schouwen R, Mina E, *et al.*: **Explain your data by Concept Profile Analysis Web Services [version 1; referees: 2 approved with reservations].** *F1000Res.* 2014; 3: 173.
[Publisher Full Text](#)
31. Hull D, Wolstencroft K, Stevens R, *et al.*: **Taverna: a tool for building and running workflows of services.** *Nucleic Acids Res.* 2006; 34(Web Server issue): W729–732.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Wolstencroft K, Haines R, Fellows D, *et al.*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.** *Nucleic Acids Res.* 2013; 41(Web Server issue): W557–561.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Mina E: **HD data analysis workflows for paper: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment_v2.** *Zenodo.* 2016.
[Data Source](#)
34. Mina E: **HD data interpretation workflows for paper: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment_v2.** *Zenodo.* 2016.
[Data Source](#)
35. Smyth DK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol.* 2004; 3: Article3.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. **Bioconductor - home.**
[Reference Source](#)
37. Hodges A, Strand AD, Aragaki AK, *et al.*: **Regional and cellular gene expression changes in human Huntington's disease brain.** *Hum Mol Genet.* 2006; 15(6): 965–977.
[PubMed Abstract](#) | [Publisher Full Text](#)
38. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Roy Stat Soc B Met.* 1995; 57(1): 289–300.
[Reference Source](#)
39. Kasprzyk A: **BioMart: driving a paradigm change in biological data management.** *Database (Oxford).* 2011; 2011: bar049.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. **HUGO gene nomenclature committee home page | HUGO gene nomenclature committee.**
[Reference Source](#)
41. Ernst J, Kheradpour P, Mikkelsen TS, *et al.*: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature.* 2011; 473(7345): 43–49.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 1 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
43. **UCSC genome browser home.**
[Reference Source](#)
44. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 2 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
45. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, *et al.*: **Integrative analysis of 111 reference human epigenomes.** *Nature.* 2015; 518(7539): 317–30.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 3 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
47. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 4 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
48. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods.* 2012; 9(3): 215–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 5 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
50. Blackledge NP, Klose R: **CpG island chromatin: a platform for gene regulation.** *Epigenetics.* 2011; 6(2): 147–52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Teodoridis JM, Strathdee G, Brown R: **Epigenetic silencing mediated by CpG island methylation: potential as a therapeutic target and as a biomarker.**

Drug Resist Updat. 2004; 7(4–5): 267–78.

[PubMed Abstract](#) | [Publisher Full Text](#)

52. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 6 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
53. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 7 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
54. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 8 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
55. Mina E, van Roon-Mom W, Verschure P, *et al.*: **Dataset 9 in: A putative role for genome-wide epigenetic regulatory mechanisms in Huntington's disease: A computational assessment.** *F1000Research.* 2017.
[Data Source](#)
56. Sadri-Vakili G, Cha JH: **Mechanisms of disease: Histone modifications in Huntington's disease.** *Nat Clin Pract Neurol.* 2006; 2(6): 330–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
57. Jia H, Morris CD, Williams RM, *et al.*: **HDAC inhibition imparts beneficial transgenerational effects in Huntington's disease mice via altered DNA and histone methylation.** *Proc Natl Acad Sci U S A.* 2015; 112(1): E56–E64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
58. Ng CW, Yildirim F, Yap YS, *et al.*: **Extensive changes in DNA methylation are associated with expression of mutant huntingtin.** *Proc Natl Acad Sci U S A.* 2013; 110(6): 2354–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
59. Kohman RA, Rhodes JS: **Neurogenesis, inflammation and behavior.** *Brain Behav Immun.* 2013; 27(1): 22–32.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
60. Hettne KM, Thompson M, van Haagen HH, *et al.*: **The Implicitome: A Resource for Rationalizing Gene-Disease Associations.** *PLoS One.* 2016; 11(2): e0149621.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 08 February 2018

<https://doi.org/10.5256/f1000research.10459.r30246>

© 2018 Pfenning A et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Andreas R. Pfenning

Carnegie Mellon University, Pittsburgh, PA, USA

Irene Kaplow

Carnegie Mellon University, Pittsburgh, PA, USA

Summary:

The purpose of this paper is to leverage publicly-available data to investigate the association between chromatin state and Huntington's Disease (HD). The authors do this by identifying genes that are differentially expressed in individuals with HD relative to healthy individuals and identifying the locations of these genes in the genome and the biological processes associated with these genes. They find that many of these genes' promoters are in the active chromatin state in healthy individuals and in CpG islands. They also find that many of these genes are related to biological processes related to HD and that some are in chromatin modification biological processes. Although this study suggests that there may be an association between chromatin state and HD, the nature of that association remains unclear.

Major Comments:

1. I appreciate how the authors integrated existing literature with differential gene expression results to prioritize biological processes, diseases, genes, and molecular functions. In addition, defining the similarity between concepts based on the number of shared concepts is similar to approaches that have been used for community detection in social networks (Blondel *et al.*, *Journal of Statistical Mechanics: Theory and Experiment*, 2008) and, more recently, for clustering cells based on protein expression (Levine *et al.*, *Cell*, 2015) (I do not think that the authors need to cite these papers), so I am not surprised that it worked well. I hope that the author's use of this approach will inspire others to use such methods for comparing biological concepts in literature and encourage future researchers to directly integrate literature with differential gene expression.
2. I found many of the results difficult to interpret because the authors seem to have done all of the analyses on the set of all differentially expressed genes. My expectations for up-regulated genes are different from those for down-regulated genes. In the Minor Comments, I point out specific analyses for which I think that separating the genes based

on the direction of the differential expression would be helpful. If the authors did use only down-regulated or only up-regulated genes, it would be great if they could make this clear in the Methods section and include what fold-change cutoff they used.

3. I thought that some of the claims in the Discussion section were not well-supported by the results. I have pointed out what these are in the Minor Comments. Most concerns come from the lack of separation between down-regulated and up-regulated differentially expressed genes in the analyses in this paper.
4. Although there is no chromatin state data from anywhere in the brain in HD individuals, there are H3K27ac and PolII datasets in the striatum of HD and control mice (Achour *et al.*, *Human Molecular Genetics*, 2015). This paper would be more convincing if it included a comparison between the differentially-expressed genes in mouse HD individuals versus controls and the differential H3K27ac regions from this dataset.
5. I found much of the Methods section difficult to follow. In the Minor Comments, I point out specific parts that I think should be re-ordered and specific details that I think should be added to make the Methods clearer. The authors should also include the exact version and settings that they used for every publicly available software package so that others can reproduce the results.

Minor Comments:

Introduction:

Page 3: Although the authors clearly describe literature suggesting that epigenetic mechanisms may be involved in HD, there is also some evidence against the role of epigenetics in HD. For example, a recent study profiled methylation in the cortex of HD individuals and controls using the Illumina HumanMethylation450K BeadChip array and found that there are no significantly differentially-methylated regions between HD individuals and controls (De Souza *et al.*, *Human Molecular Genetics*, 2016). The authors should cite this paper and explain why it does not demonstrate that epigenetics is not involved in HD (the assay used was not genome-wide, methylation is not the only component of transcriptional regulation, etc.).

Page 3: It is not clear to me why transcriptional dysregulation in HD would be associated with differentially expressed genes in regions that are not normally associated with active chromatin states. My understanding from the literature cited in the introduction is that many of the genes that are differentially expressed in HD individuals have *lower* expression in HD individuals than they do in control individuals. I would therefore expect that these genes would fall in regions that are normally associated with active chromatin states but may not be associated with active chromatin states in individuals with HD. It would be great if the authors could clarify the motivation behind this hypothesis.

Page 3: At the end of the introduction, the computational method is introduced as an approach to intelligently select which experiments to do. However, I was not sure from the introduction what types of experiments this method is designed to guide. It would be great if the authors could add a more detailed explanation of this earlier in the manuscript.

Methods:

Page 3: It would be easier to understand the advantages of the vector space model if they were listed after the description of the vector space model instead of before it.

Page 3: It would be helpful if the authors could describe how the subset of PubMed records are selected for genes or cite a previous paper that uses the same method that they used.

Page 3: It would be helpful if the authors could define "symmetric uncertainty coefficient."

Page 4: It would be helpful if the authors could list exactly what publicly available datasets were

used for each differential expression test before describing the differential expression test.

Page 4: It would be helpful if the authors could state what microarrays were used to generate the gene expression data before describing how the differential expression analysis was done.

Page 4: It seems like the authors did not account for potential confounding factors that were available, such as sex, age, and brain tissue, in the differential expression analysis. I am concerned that these confounding factors may affect the results.

Page 4: The authors state which human and mouse assemblies they used, but they had not previously stated that their analysis included data from mouse. It would be helpful if the authors could state exactly what species they are using for each part of their analysis earlier in the manuscript.

Page 4: The authors state that they used a Kolmogorov-Smirnov test to compare p-value distributions. It was not clear to me where these p-values come from. Are they the p-values for differential expression of the genes corresponding to the promoters? It would be helpful if the authors could clarify this.

Page 4: It was not clear how many Kolmogorov-Smirnov tests were done. The authors said that they rejected the null hypothesis if the p-value was < 0.05 . If they did more than one test, then they should do multiple hypothesis correction.

Page 4: It would be helpful if the authors could clarify the purpose of the concept ids.

Page 4: It would be helpful if the authors could provide a more detailed explanation of how the concept linking is done.

Page 5: It would be helpful if the authors could define the HTT concept and explain why they used it for prioritizing genes.

Page 5: It would be helpful if the authors could explain why they decided to prioritize genes based on the "epigene" concept. It seems like the authors are interested in genes that affect epigenetics, such as demethylases or histone modifiers. It is not clear to me how this selection relates to the hypothesis that was described in the introduction.

Page 5: It would be helpful if the authors could clarify exactly what differential expression tests were done with the human brain data and what the categories were for each test.

Page 5: It would be helpful if the authors could clarify whether the human brain data described here was the only human data used for differential expression analysis and, if it was not, what other data was used.

Page 5: It would be helpful if the authors could briefly describe how the CpG island data fits into the rest of the analysis.

Page 5: It would be helpful if the authors could explain why they selected the two cell types and four chromatin states that they used in the Methods section.

Page 5: I think that it might make sense to incorporate additional chromatin states, such as quiescent, weak repressed Polycomb, and enhancer, as strong repression is not always the cause of a promoter's inactivity.

Page 5: It would be helpful if the authors could clarify why they used only the mouse data from animals treated with the vehicle. My intuition is that it would make more sense to use the animals that did not receive the HDACi 4b inhibitor since the human subjects did not receive any kind of treatment. It is possible that I misunderstood the purpose of the mouse analysis.

Results:

Page 6: It is not clear to me why a difference in distribution of expression levels between genes overlapping a chromatin state and genes not overlapping that chromatin state implies that chromatin state has an effect on HD. I think that the authors mean that, if the *difference* in gene expression between individuals with and without HD is higher for genes overlapping a specific

chromatin state than overlapping other chromatin states, then there is an *association* between the chromatin state and HD.

Page 6: It would be helpful to split Figure 1 into two parts, one for genes that have higher expression in people with HD and another for genes that have lower expression for people with HD. My intuition is that most of the differences in p-value distribution are coming from the second category because, since the chromatin state data comes from people without HD, I would expect that genes in an active chromatin state would have higher expression in healthy individuals. Adding onto that, regions of closed chromatin cannot decrease because the genes are not expressed. Regions of open chromatin could either increase or decrease, potentially leading to more variability.

Page 6: It would be helpful to have a supplemental figure with all chromatin states because it is not clear from Figure 1 if the differences occur for TSS's in all active chromatin states (including inactive genes that are acting as enhancers for other genes) or only from genes that are transcribed.

Page 6: It would be helpful if the authors could clarify if the overlaps in Figure 1 are done using the entire gene, only the TSS, or the gene's promoter.

Page 7: For the biological process analyses, I think that using a tool for differential enrichment between the two groups of genes would provide more interpretable results than comparing the top hits from CPA because such a tool looks for terms that are significantly enriched in one gene set relative to another. An example of such a tool is CompGO (Waardenberg *et al.*, *BMC Bioinformatics*, 2015).

Page 8: It would be helpful if the authors clarified what they mean by "top novel protein." Does novel mean that the gene had not been associated with HD in a previous paper?

Page 8: It was not clear why Figure 3 shows that CPA is able to prioritize true associations with huntington as measured by a gene expression experiment and why combining differential expression measurements and literature evidence enables the selection of even more specific HD signatures. It would be great if the authors could clarify this.

Page 8: It would be helpful if the authors could include the direction of the CPA score shifts for the different groups of differentially expressed genes.

Page 8: The authors say that "the top 100 and top 1000 differ significantly." It would be helpful if they stated the way in which these gene sets differ.

Page 11: It would be helpful if the authors could clarify what x is in Figure 3.

Discussion:

Page 11: I am not sure if the paper provides a lack of evidence for genome-wide re-localization of gene activity to repressed chromatin states. The paper combined all of the up-regulated and down-regulated genes instead of separating them. If the paper had shown that the genes that are up-regulated in people with HD are not found in repressive chromatin states in healthy individuals, then I would be more convinced of this lack of re-localization. However, I would not be fully convinced because changes in chromatin state do not always cause changes in gene expression. For example, a previous study showed that most single nucleotide polymorphisms associated with histone modifications are not associated with transcription, suggesting that histone modification differences between individuals do not always correspond to gene expression differences (Grubert *et al.*, *Cell*, 2015). Thus, it is possible that there are chromatin state differences between HD individuals and controls in parts of the genome where there are no differentially-expressed genes.

Page 11: The authors suggest that HD is not associated with the disruption of chromatin states at a large scale. To investigate the association of HD with chromatin state using existing data, the

authors would need determine if genes that are up-regulated in people with HD tend to fall in repressive chromatin states and if those that are down-regulated in people with HD tend to fall in active chromatin states. Because there do not seem to be separate evaluations of up-regulated and down-regulated genes, I do not think that the results in this paper can be used to evaluate the relationship between chromatin state disruptions and HD.

Page 12: I think that CPA's high ranking of chromatin-related concepts for differentially expressed genes suggests an association between chromatin reorganization and HD. If differentially expressed genes near CpG islands include genes involved in chromatin structure, that suggests that there is *cis*-regulatory change in the regulation of those genes, which could have a downstream effect on chromatin organization.

Page 12: Although the paper shows that there are more differentially expressed genes in the active chromatin state in healthy individuals, I am not sure that there is sufficient evidence to conclude that most important changes in HD are occurring in the active chromatin state. For example, if the majority of differentially expressed genes are down-regulated in individuals with HD, then the findings in this paper would match my expectations, even if the most important differentially-expressed genes are up-regulated and are not found in the active chromatin state in healthy individuals.

Supplemental Datasets

Dataset 1: Some of the line breaks seem to be missing.

Dataset 2: The column breaks seem to be missing.

Dataset 8: The column breaks seem to be missing for the top 100 differentially expressed genes.

Supplementary File 1: The first word in the first figure caption seems like it should be "Illustration."

Supplementary File 1: It would be great if the authors could clarify what they mean by "x2."

Supplementary File 1: It would be great if the authors could explain why they are using the HMEC and NHEK cell lines.

Is the work clearly and accurately presented and does it cite currently literature?

The authors seem to clearly describe what they do and cite most of the relevant literature.

However, as I mentioned in the fifth major comment, I found parts of the Methods section difficult to follow.

Is the study design appropriate and is the work technically sound?

As I mentioned in the first major comment, I do not think that the authors can test their hypothesis with their study design because they combine the up-regulated and down-regulated genes.

Are sufficient details of methods and analysis provided to allow replication by others?

The authors provide publicly available workflows for almost everything they did. However, as I mentioned in my fifth major comment, the lack of clarity in parts of the Methods section might make reproducing some of the results difficult.

If applicable, is the statistical analysis and its interpretation appropriate?

Most of the statistical analysis seems appropriate, but most of the interpretation does not make sense because the up-regulated and down-regulated genes were combined.

Are all the source data underlying the results available to ensure full reproducibility?

Yes.

Are the conclusions drawn adequately supported by the results?

As I mentioned in my second major comment, I think that many of the conclusions are not supported by the results.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neurobiology, epigenetics, computational biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 24 November 2017

<https://doi.org/10.5256/f1000research.10459.r27321>

© 2017 Queralt-Rosinach N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Núria Queralt-Rosinach

Department of Integrative Structural and Computational Biology, Scripps Research Institute, La Jolla, CA, USA

Summary: This is a computational study devoted to investigate the hypothesis that epigenetic mechanisms dysregulate transcription at genome-wide scale in Huntington's disease (HD). The

authors designed their experiments to evaluate two regulatory mechanisms: 1) change of the chromatin state, and 2) CpG island methylation. By means of statistical analysis of experimental data they evaluated the association of dysregulated genes in HD with these two processes, and they provided results and a literature-based semantic analysis for functional interpretation. Furthermore, by application of a semantic analysis they provided a list of prioritized proteins based on their newly predicted association with HTT and epigenetics. Lastly, they evaluated their semantic analysis for gene prioritization. The main conclusions are that their findings do not support the hypothesis of a massive transcriptional dysregulation in HD is linked to large-scale relocation of gene activity, thus the authors speculate that epigenetic effects might be more closely related to dysregulation of individual genes. Finally, the authors claim that their methodology for hypothesis generation can be of great value for the scientific community as it helps in narrowing down the key associations and the evidence underlying them.

Reviewer notes:

- This is an interesting work on the possible epigenetic mechanisms that contribute to transcription dysregulation in Huntington Disease (HD). It is very well written in a clear and accurate manner. They based their hypothesis on a very cited current literature. I have only detected one typo in the whole manuscript: limma vs LIMMA, the authors should be consistent in the format.
- Introduction: I would like to highlight that this study is the first time that is done, so is novel and relevant. Good literature citation.
- Methods: The experiments are properly designed and the methods used are well established and based on state-of-the-art approaches. All is very FAIR: data, workflows and webservice used. The study is based on public and open resources (data and software). The methodology is well described but I still missed some information: lack of the statement about the R version, and lack of the specific parameters used in the statistical analyses that other scientists would need to replicate the experiments. Why the authors chose these statistical tests? Regarding CPA- co-occurrence can come from refuting evidence of association, is this taken into account in the concept profile? if does, it is that reflected in the evidence graph downweighting some edges? state the version of CPA database of relations used.
- Design: They tested two regulatory mechanisms: 1) via changes in the chromatin state, 2) via methylation in CpG island content. Both using the promotor gene region to 1) overlap with the chromatin state, 2) to overlap with the CpG content of deregulated genes. They assessed the association via KS statistic test, why this test? As the authors said, there are more regulatory regions in the DNA that could be target of epigenetic regulation, could a cumulative dysregulation in all these regions derive to a large-scale?
- Results: All the resulted data is available for reproducibility check. Highlight they reproduce previous published results. I am wondering if they had issues accessing, pre-processing the data to adapt it for their analysis workflows. An explanation of these issues and if their workflows help the community on this regard facilitating to overcome these issues in a systematic, reproducible and traceable manner would be of importance.

- Importantly, their analyses integrate experimental data and knowledge and evidence from the literature. Regarding the text mined noisy/literature-biased knowledge that may come from their CPA approach, inclusion of ontologies could be beneficial by leveraging the intrinsic knowledge using automated logical reasoning. Have the authors any plans on this regard?
- Discussion: I agree with the authors that the combination of experimental data analyses with Literature-based functional interpretation of the results (CPA) is relevant and add value to the results. In the second paragraph, can the authors suggest next steps to try to explain the unexpected results?
- Conclusions: The conclusions are supported by their results. In the conclusions section I missed the conclusion about the title of the paper is investigating, which is the plausibility of a genome-wide scale epigenetic dysregulation of transcription in HD although are stated in the abstract and discussion.

I would emphasize the relevance of their approach performing synergistic research work between computational and wet lab researchers. This interdisciplinary research approach seems to me the way to go in an innovative and efficient big data driven research.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research