CHDI Workshop: How can RNA profiling best provide pathogenic insights and pharmacodynamic biomarkers for Huntington's disease (HD)?
May 4-5, 2011
Los Angeles, CA

Executive Summary

Although the workshop title suggests a focus on RNA-profiling data, much of the discussion by the group had more to do with systems biology. How can massive data sets provide a better understanding of HD? Data sets can now be analyzed in a way that reveals signatures, or "modules" that may reveal clues. Despite a seemingly simple trigger—a mutation in a single gene—HD manifests as a complex disease that affects many pathways and many sub-networks throughout the body's single, overall network. In order to apply a systems-biology approach to HD, it became clear during the workshop that the first step will be to gather available data into some central repository—including molecular and phenotypic—from various databases and individual laboratories throughout the HD research community. This data may require various types of quality control or re-analysis. Although it seems a daunting task, we must gather a comprehensive look at the available data set in order to move forward. Once the gaps in this information can be identified, future priorities can be set.

Understanding the perturbations that HD causes in the body's sub-networks will be key to pursuing early, disease-modifying therapies for HD. Despite the fact that the Htt gene was identified nearly 20 years ago, we don't yet have a clear understanding of the disease trajectory, or even if there is one single or multiple different trajectories among patients. The signatures generated from molecular and phenotypic data, from humans, animals, and cell models, can help to establish a clearer disease profile. One thing has emerged clearly: the CAG-repeat length serves as the critical biological determinant of disease severity, which lies on a quantitative continuum. This polymorphism is present in the entire population, making it imperative that data be collected from control subjects as well as from patients. CAG-repeat length in the poly-Q region of Htt appears important not only in understanding HD pathology but perhaps the function(s) of wild-type huntingtin as well. Another key question that emerged from the workshop addresses the reversibility of the effects of the mutant huntingtin protein and the disease process, which must be considered when pursuing effective therapies for HD patients.

Introduction

The technology and tools of inquiry into molecular biology have become increasingly powerful in recent years. Massive amounts of data can be generated relatively quickly and affordably, providing tremendous opportunities that should be undertaken selectively. As Martin McIntosh pointed out, "Data is not the same as information," while knowledge and understanding are even harder to mine from an experiment. One needs to organize those data in ways that allows one to understand the perturbation of the system, in this case by mutant huntingtin (mHtt). Analytical tools are now available that can reveal signatures or modules from the data.

Reflecting on the current workshop, Bob Hughes remembered a similar meeting a decade ago when participants thought that microarrays might revolutionize the understanding of HD through gene analysis. Despite the lack of a "Eureka!" moment, some good headway has been made through those tools. The situation might be similar with this next generation of tools, which could provide two key types of information. First, signatures could provide a valuable yet "agnostic" signature that could inform the disease process and serve as a biomarker of disease. Second, the tools could generate hypothesis-driven findings about the mechanism of mutant huntingtin-mediated pathology. Workshop participants agreed that those "signatures" lie at the core of a bigger-picture systems approach to HD. Whether comprised of molecular, phenotypic, or other types of data, analysis that generates common "modules" can reveal systemic perturbations and give rise to a better understanding of this complex, body-wide degenerative disease. An important point in assessing these perturbations is that huntingtin is expressed in every cell throughout the body—not just in the neurons. And it's expressed from conception until death.

CAG-repeat length critical to HD

Huntington's disease (HD) arises from a single mutation in the gene (Htt) for the huntingtin protein. The mutation takes the form of an expansion in the protein's polyglutamine repeat region, encoded by the nucleotide sequence CAG. The accepted view has been that disease arises with a repeat length around 37 or higher, but it has become clear as clinical data accrues from around world that the situation is more complex. The "normal range" is now restricted to 6-27 repeats, because no cases of HD have (yet) been identified in that range. A CAG-repeat length of 27-35 is now considered to carry increased risk with very low penetrance: cases have been identified in this range. With 36-39 repeats, penetrance increases dramatically, and if an individual has 40 or more repeats, overt clinical symptoms (including death) will inevitably ensue. With this deepening appreciation for the importance of the CAG repeat come several realizations: the CAG repeat is a polymorphism in everyone and represents a continuum. Moreover, CAG repeat length is directly tied to a phenotypic continuum. CAG allele size is correlated with neurological outcomes, psychiatric symptoms, onset of overt motor symptoms, and death.

This realization was at the core of a study recently published by McDonald, Gusella, and Jong-Min Lee (Jacobsen et al. HD CAG-correlated gene expression changes support a simple dominant gain of function. 2011 Hum Molec Gen). Marcy McDonald described the paper as a pilot to test the effects of CAG-repeat length on gene-expression changes. They started with a knock-in mouse expressing one allele of murine Htt with a very long CAG track of 111, and the other allele with the normal 7 repeats. They then used four more allele lengths of Q20, Q48, Q50, and Q98 to ask whether gene-expression data would reveal elements correlated with CAG repeat length. They found a series of genes whose expression varied in a quantifiable way with CAG length. The longer the allele, the more the expression went up or went down. This set of genes was completely different from those identified with the "standard" dichotomous approach of comparing HD- length (Q111) vs. non-HD-length (Q7). But many of the proteins that emerged in the two data sets participate in the same pathways, as might be expected. These CAG-length-dependent quantitative biological effects suggest, said McDonald, "it's not a qualitative change in the sense that Htt is non-toxic [at low CAG repeats] then becomes toxic [at a certain CAG repeat number]. It's a quantitative continuum." It's still unclear what happens to the protein biochemically with increasing CAG length, said McDonald, but "we would argue that CAG repeat length is *the* central metric, because it allows you to take anything you're measuring, whether it's RNA-seq data or splicing data or methylation or phenotyping, and tie the data to the CAG repeat size."

McDonald pointed out that the system provides a setting to next ask whether those CAG-correlated genes would overlap with those affected by having no Htt protein at all, as in the knock-out situation. This approach can start to address questions that are raised by the therapeutic strategy of knocking down huntingtin. Of course there are considerable questions of safety (as well as efficacy) with this therapeutic approach that will need to be addressed. But another key question is that of reversibility: if you reduce mutant Htt RNA (for example) as a therapy, what effects will it have on the system? In other words, what damage can be undone, and what effects are irreversible? Ruth Luthi-Carter suggested a simple inquiry: one could examine an animal model in which expression of mutant Htt is reversible; you would turn on the gene, let some moderate phenotype happen, turn it off, and then see whether the effects were also reversible. She sees the question as "important to know, and it's a feature independent of silencing of wildtype Htt."

Signatures reveal disease-related patterns

Gene-expression profiles in human brain regions are organized into robustly defined co-expression modules. Further, HD-related co-expression modules have been found in transcriptional studies of certain human brain regions. The network and pathway perturbations found in these modules, or signatures, provide more valuable information than examining changes in individual genes alone. What types of data can be incorporated into signatures? In addition to RNA-seq data looking at coding sequence, signatures can be revealed by other types of molecular data: non-

coding RNAs, micro-RNAs, and the epigenetic effects of DNA methylation among others. Magnetic resonance imaging (MRI) data can now be analyzed as a component of HD signatures. Nelson Freimer advocated going back to large-scale behavioral and cognitive studies and using that sort of phenotypic data as another component of a "signature."

Steve Horvath described how modules are created from data analysis. "I do what I refer to as biostatistical systems biology." Horvath developed statistical tools to identify targets and drivers of disease phenotypes using weighted gene co-expression networks and applying them to other types of data, like MRI and methylation data. He says that we now have a tremendous opportunity to mine the fantastic data sets generated over the last 20 years with treatments in mind, because it's now fairly easy to do using new software tools. When microarray data became heavily used, scientists struggled with how to normalize data from different arrays and data from different organisms, but "that has been solved now," and can be applied to these new technologies as well.

Do peripheral tissues reflect a mutant Htt signature?

Considering that huntingtin is expressed throughout the body, peripheral tissues are likely to bear HD-related signatures as well. The disease takes 40 years to develop, and appears to affect very long-lived cells: adult neurons. But Martin McIntosh suggested that it might yet have an obvious signature with very subtle effects in seemingly "unaffected" peripheral cells that simply don't appear as a disease phenotype. The accessibility of peripheral tissues like blood from living patients makes this an important priority, particularly in the context of a disease biomarker.

Moshe Szyf stressed the importance of expanding the hunt for signatures to the periphery. He would like to see how well peripheral DNA methylation works as a behavioral disease readout. Although this approach is not aimed at understanding the disease mechanism in the brain, we need to pursue it as a tool. The approach requires that we appreciate the mind-body integration from a systems perspective. We are not a brain floating on the air, detached from the body, he said. He cautioned that tissue and even cell specificity might be crucial to finding reliable patterns. Indeed, some previous efforts to identify molecular signatures of HD from blood may have been thwarted by the mix of cell types present in whole blood.

Molecular techniques provide powerful tools

Clearly the technology available today provides ways of inquiring about molecular goings-on in the cell that might have sounded impossible even just a few years ago. But there is a danger in producing fruitless data if the techniques are not applied properly to your question. McIntosh cautioned that RNA-seq data has both its strengths and drawbacks. "While RNA sequencing is very powerful, it's also an artifact-finding machine. It's a way of looking at all the potential coding and non-

coding sequences in a tissue, but when we visually curate it we often find there's an artifact in alignment or processing or sample preparation."
One key observation lies at the root of the power of looking at gene transcription patterns in HD: mutant huntingtin (mHtt) appears to change the transcription profile in many different cells and systems. It remains unclear how or why this occurs, but it is a reminder of Htt's ubiquitous, constitutive expression as well as its cryptic functions. One possibility is that huntingtin itself may be an RNA-binding protein. Clotilde Lagier-Tourenne spoke about her work in Don Cleveland's lab with CLIP-seq data looking at RNA-interacting proteins involved in amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD), including the proteins TDP-43 and FUS/TLS. If mutant Htt—and specifically CAG repeat length—is really linked to system-wide gene expression changes, she said, then Htt may have a real role in transcriptional regulation. This might be investigated with experiments using ChIP-seq and CLIP-seq analysis. If Htt does not have a direct role in RNA processing, then what might lead to the changes observed in transcription profiles? She said that the poly-Q region could purportedly trap the RNA-processing proteins TDP-43 and FUS/TLS. It might seem very indirect and far away, but there's a real possibility that the CAG expansion can lead to RNA-processing changes by altering the function of other RNA-binding proteins. "We need to test if RNA-processing proteins are involved, and in particular the role of poly-Q in trapping them." This process is now seen in many genetic diseases, so it makes sense to investigate it in HD.

Systems approach requires organization, but how?
Keith Elliston, CHDI's new Vice President of Systems Biology, believes that when it comes to the massive amounts of data now available, "the first order of business is organization." When it came to *how* all that data might be organized and made available, participants varied in their opinions but contributed ideas throughout the course of the workshop. Elliston and others advocated for a centralized database. "I'd like to go to a single place and find all the data that has anything to do with HD on a molecular level," he said. Elliston finds raw data really useful, because everyone has a different method of analysis and there are many different platforms and techniques available today. He envisioned a database environment of raw data that would also contain analyzed data sets, linked to the expert who has processed the data in a usable way. Once these data are compiled, they can be analyzed in a way that might reveal useful signatures.

Jonathan Derry is in favor of analyzing existing data to come up with new information, but he cautions that in terms of building complex predictive models, it's very hard to take many different studies and piece them together. You need, he said, "some large coherent study to hang it off of." Like many others, Giovanni Coppola would like to have a comprehensive database of animal models that could serve to examine a particular gene, for example, across all models and patients. Peter Holmans agrees it would be nice to have a list of all available data that has been

quality-controlled, preferably in a standardized way, so that you know that you're dealing with high-quality data. In the long term, he would like to see data from samples of human and mice including phenotypic data, RNA-seq data, DNA methylation data, and importantly genotypic data as well, so that one can relate co-expression modules to genetic association or modulating effects. He also advocates that money be set aside to analyze the data—no small task in itself.

Even if raw data can be made available in such a centralized database, everyone agreed that it will likely require further quality-control measures. Horvath believes it would be valuable to go through and pre-process the individual data sets "one more time, to identify outliers or batch effects, which is also now completely routine." What once took a year to process in his lab can now be done in a day, using newly developed software tools. This fine-toothed sorting "makes a huge difference, especially with brain data," and can provide clearer results. Even some data sets in which batch effects weren't identified, "you can reconstruct and remove them." Horvath routinely makes available both software tools and tutorials or assistance in using them to achieve these tasks.

Lesley Jones encouraged everyone to keep their samples, because you don't know what they might yield in the future. This led the discussion to the suggestion of a sample registry. Rather than a physical repository of cells and tissues (which would be logistically very difficult), a registry would allow people to see what resources and samples might be available, and then to communicate with the lab through linked contact information. Human tissues, which are more rare and difficult to obtain, should also be included in such a resource registry. This might streamline the search process for samples, and might motivate individual labs to coherently store and organize their samples. Pacifici told participants that CHDI could usually provide the standard HD-related cell and tissue resources. Specialized tissues, for example with particular genetic manipulations or collected with laser capture micro-dissection, might be more difficult but could in some cases be provided. HD Crossroads was initially created with this sort of problem in mind, and could be expanded to include such a Samples Registry as well as a Data Registry of what people have available.

Robert Pacifici summed up the workshop by saying that CHDI is committed to pursuing this systems-biology approach, and that the creation of a central repository of data is "high on the list." As many people said, it's easier in theory than in practice, and he fears that we may be disappointed in the quality of the data available, and that it will need to be placed within some context. Beyond basic quality control, Pacifici is genuinely curious about the types of criteria people use to select subsets of signals to empanel: "How do you make that signature for tracking efficacy, for toxicity, etc?" The more we can learn from participants' thinking, he added, the more we benefit not only from the analysis, but a "meta-benefit" arises, in that we can incorporate that thinking in the future as we select priorities. He saw a

surprising unanimity in the support for the systems approach, and added, "There's a lot of fodder here for further collaboration."

EXPERIMENTS

At the conclusion of the workshop, Allan Tobin asked participants for suggested experiments as a first priority with respect to the technologies and tools we had been discussing. The participants' suggestions are summarized below.

Robert Pacifici outlined three "buckets," or "bins," into which experiments could be grouped; many participants characterized their suggestions according to these classifications:

1. -Omics surveys: detailed characterizations based on data collected from molecular techniques, including gene expression, modulation (e.g. methylation), proteomics, RNA-seq data, and so on.
2. Mechanism: Studies that aim to describe the still-unknown functions, deficits, and pathology associated with wild-type and mutant huntingtin.
3. Support of translational programs: investigations that will help support programs that will be entering clinical trials in the coming years, whether they measure safety, efficacy, or other parameters.

Participants came to a broad consensus that, particularly for "buckets" one and two, people would favor gathering data from humans rather than mice to start with.

Ethan Signer began by asking for a comprehensive comparison of the full-length Htt mouse models with the sort of data that Steve Horvath proposed—i.e. modules of molecular data that can provide a signature. He would ask, how do they compare with one another, and how do they compare with the human disease features? (During the workshop, participants also discussed the value in making comparisons of various animal models, and agreed they really are used as a research tool, and no animal serves as a reliable "model" of human disease. The utility in detailed characterization of these animal "tools" is in determining not the single best model (or tool) but the best tool *for your purpose*.

Signer suggested using a common black 6 mouse (C57BL/6) background to look at the molecular signatures of the YAC the BAC mice. This information would put people in a more powerful position to use the modules. When asked what specific parameters would be most of interest—for example brain region, time frame—Signer pointed out that there must be significant data already available about the molecular signatures of the various models. Although Ruth Luthi-Carter systematically looked at the existing data four years ago, more data from additional models would be available today. (E.g., the BAC model was not included in that survey.)

Doug Macdonald then pointed out that these models all contain extremely long CAG repeat lengths, and it might be beneficial to examine a mouse that expresses 20-50

CAG repeats. Although such a mouse would likely not display the overt phenotypic effects seen in mice with long repeats, this would not be a disadvantage for several reasons. The goal of such studies is to generate molecular data and not to assess behavioral outputs. These shorter repeats might more closely mimic (in cells throughout the body) the prodromal or pre-symptomatic phase of HD, which is becoming increasingly recognized as the more interesting disease phase in terms of developing disease-modifying therapeutics. And a heterozygous mouse would also be closer to the human disease and of more interest than would homozygotes. Marcy MacDonald requested that one also include a careful molecular characterization of the transgene, so that you know exactly what it looks like—for example, where the insertion has occurred, what rearrangement the gene has undergone, how many copies exist, etc. Such an analysis makes it "fair to the CAG repeat" in comparing different repeat lengths. Finally, Ruth Luthi-Carter proposed that "when we consider collecting new temporal data sets, we also we look at cerebral cortex," because striatum is not the only affected brain region, and moreover the cortex is not homogenous. Giovanni Coppola agreed with Luthi-Carter that he would like to see an analysis of mouse models with an emphasis on disease time-course analysis with frequent time points around onset of disease. He would like to see analysis not just of cortex or brain but of peripheral tissues as well, including blood, perhaps using cell-type specific profiling. In addition, Coppola suggested a human family-based analysis, that is looking at patients and controls from the same family to somewhat normalize genetic background.

Jim Gusella seconded these ideas, saying "it would make it easier for us to gather the sort of data we've been proposing," namely to home in on the consequences of CAG-repeat length at the molecular level. "We need a CAG-length-dependent comparison of CAG-length-dependent changes in the molecular signature of the animals." Gusella outlined the following screen: "I would take the mutant allele across from the null, and take the wild-type as two normals. Then take the wildtype over the null, then wildtype and the mutant together. Then do a comprehensive analysis of expression in the brains in those animals to see what changes are allele-dependent, are dosage-dependent, and I would take at least two or three allele sizes to see what the CAG repeat length does." Such a screen would also help to identify molecular pathways that might be of interest in pursuing the "Bin 3" support of program category, particularly in Htt-lowering strategies. Such studies could be very informative in lower organisms, even as simple as *Dictyostelium*.

Lesley Jones feels it's very important to obtain genetic data from people and correlate it with data from mice, so that you could sort out which aspects of what you see in animals really relate to patients.

Doug Macdonald said that CHDI is starting to form several *in vivo* experiments treating knock-in mice w/ siRNA, and it would be nice to share that tissue that with some of the participants. Also, he now realizes the importance of considering the

wildtype mice in any studies, with an emphasis on CAG-repeat length. Jim Wang would like to know, in terms of the current therapeutic program, what happens in mouse when both wild-type and mutant Htt are knocked down. Macdonald said those experiments have been done but not yet analyzed in a CAG150 mouse, collecting phenotypic and mRNA-level outcome measures. Jim added that in light of the importance of the CAG repeat length, it would be useful to go back and test a mouse with a more physiologic CAG length, like 40.

In terms of silencing the huntingtin gene, Ruth Luthi-Carter voiced concerns about the adverse consequences that might arise from silencing in the brains of HD patients. She suggests further study with a conditional knockout mouse. Doug Macdonald described a CHDI project with Artemis in which they created a mouse that inducibly expresses siRNA against wildtype Htt (Q7/Q7), thereby leading to whole-body knock-down of Htt. The problem was "it's leaky as all get-out," said Pacifici, even during development. This means that the siRNA is expressed even in the absence of doxycycline, the inducer. The consequences are "not good, if it's 90% knocked down." Plans are underway to remake the mouse.

In terms of experiments for "bins" one and two (the –omics survey and mechanistic studies), Andrew Kasarskis "would cheer on the Ruth/Steve contingent." In other words, look at human data and try to get other data from larger studies (including control subjects) to try to correlate CAG repeat length with molecular profiles from peripheral tissues. Additionally, he would like to further understand the normal function of wild-type Htt, where, he said, "there are a fair number of obvious experiments to be done." In terms of therapeutic support, he thinks that Ruth's line of research would be valuable to know how reversible and dynamic a phenotype is with expression of mutant Htt.

Jamshid Arjomand is looking forward to seeing "systems biology at work to come up with mechanisms of action to generate hypotheses." Like many others, his priority would be to start with human tissue rather than mouse. If one could see a signature that's consistent, one could generate stem cells and even generate hypotheses from looking at those stem cells.

Megan Mulligan reiterated that a tremendous amount of data has been generated for the large family of BXD inbred mouse strains, using mRNA-expression, behavioral, and physiological measures. The BXD set of recombinant inbred strains is derived from a cross between C57BL/6 (B) and DBA/2J (D) inbred mouse strains. The BXD population contains over 80 strains and another 80 strains are currently being bred. The BXD strains have been densely genotyped, and she and her colleagues have characterized roughly five million B vs. D single-nucleotide polymorphisms (SNPs) segregating these strains. In addition to brain expression profiles from all major platforms, they can also provide traditional array expression profiles from multiple peripheral tissues. This legacy data could be useful in creating

molecular signatures of HD, which could be used to ask questions about HD disease progression and about genetic-modifier effects in the BXD family of strains.

Marcy MacDonald agreed with others and added that, in terms of "bucket 3, given that you have modules and readouts of broad, global networks…it comes down to the test of whether or not the true dominance in this disorder reflects something that you can reverse if you remove the mutant protein." That question is testable if you have a system set up with the networks that exist in human cells with human endogenous alleles that represent the wild-type state, and then longer and longer states—both in terms of CAG length and longitudinal progression. The question of reversibility must be answered.

Martin McIntosh: With the expertise gathered here, we are in a position to generate data and have ten people analyze it in 20 different ways. McIntosh suggests that we "generate the low-hanging-fruit data set to look at each in our own way, then call us back a year later…and see what to pursue next." Although human samples are harder to come by, one could use banked samples of brain and other tissues to generate data series. Specifically, McIntosh would go to the mouse models, look at a variety tissue types including brain, blood, and muscle, and generate data sets on potential regulatory modifications. He proposes funding five postdocs for six months, then calling the group back together to present the findings. As he points out, "information is more useful than data."

With the last word, Keith Elliston said, "it's very clear to me that we have a grand challenge here to develop a deep biological understanding of HD." This data-driven challenge requires a direct focus, and the first step is to figure out what we already have, and then do a good gap analysis and figure out what else we need. He seconded McIntosh's idea of coming back together and sharing not only data but also information and ultimately knowledge, to "make this work as a community rather than as individuals."

COLLECTED RESOURCES

Jonathan Derry at Sage Bionetworks, a non-profit in Seattle, is trying to facilitate data sharing of large genomic data sets in easily usable formats to drive development of molecular models of disease to inform biology and drug development. They provide a platform to share data, and then people can use it in community-based efforts. Although they have not focused on HD, a Merck data set is available that contains a substantial HD data set that represents 200 HD patients from three brain regions, including genome-wide genotyping and transcriptional profiling. HD patients were used as a "reasonable control" for this study of Alzheimer's disease (AD). The data focuses on building network models of disease from gene expression data. Some sets of genes were strongly correlated with CAG repeat length. Available at Sage Bionetworks Repository; click on "Harvard Brain Tissue Resource Center."

Jim Wang and Mike Pallozolo have put together a database of 600 genes shown to impact HD models. It is available the Crossroads portal CHDI set up, and has a pod focused on datasets curated in various models. Crossroads is meant to include bidirectional communications.

Ed Lein works at the Allen Institute, a non-profit neuroscience research organization. They have created large-scale gene-expression atlases and are moving into more complex brain analysis, focusing on profiling gene-transcript distribution across species and across development, including microarray profiling and expression profiles from mouse, non-human primate, and human brain. Information and data-mining tools are freely available at the Allen Institute for Brain Science portal (www.brain-map.org).

Martin McIntosh has a project dealing with visual curation of RNA-seq data called GBM Genome browser; its purpose is to build the isoform- or splicing machinery-viewer into it for better visualization of RNA-seq data.