Reconstruction and simulation of regulatory networks in the Htt allelic series using causal machine learning

Jeanne Latourelle[1], Raymond Yan[1], Michael Beste[1], Tun-Hsiang Yang[1], Boris Hayete[1], Iya Kahalil[1], Jeff Aaronson[2], James Rosinski[2]
[1]GNS Healthcare, Cambridge, MA 02139; [2]CHDI Foundation, Princeton, NJ 08540

Objectives

- Although the autosomal dominant mutation in the huntingtin gene (Htt) is well-characterized, the downstream molecular machinery that mediates disease pathogenesis is poorly understood.
- High resolution transcriptional and behavioral profiling across the murine Htt allelic series is designed to capture the earliest molecular effectors of CAG-repeat expansion across striatum, cortex, hippocampus, and cerebellum at various stages of disease progress.
- We have applied GNS' Reverse Engineering Forward Simulation (REFS™) machine learning platform to model the effects of change in CAG repeat length on the contextual transcriptional causal network in a data-driven fashion.
- Exhaustive *in silico* experiments applied across models identify key drivers of the Htt-contextual regulatory network including those affected by varying CAG repeat length, age or both.

Methods

Allelic Series Design and Profiling

To systematically distinguish early from late molecular HD phenotypes, CHDI has deeply profiled three cohorts of transgenic Htt mutants, comprising:

I. Mutant HTT knock-in in BL/6 background (n=208)
II. Cohorts (n=104/104 M/F) aged 2, 6, and 10 months
III. WT and mutant Q20, Q50, Q80, Q92, Q111, Q140, Q175 (n=8 each)
IV. Five tissues
  - Striatum
  - Cortex
  - Hippocampus
  - Cerebellum
  - Liver
V. RNAseq (~20k transcripts)
VI. LC/MS Proteomics (~6k targets)
VII. PsychoGenics Behavioral profiles

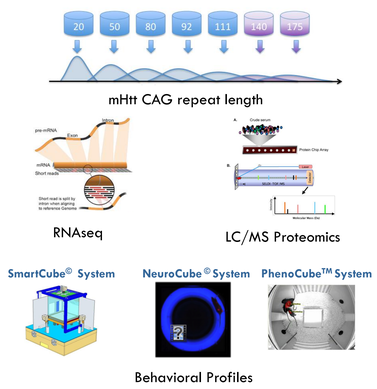


Figure 1. Experimental design and profiling platforms characterizing the mHtt allelic series.

Causal Inference via Reverse Engineering and Forward Simulation (REFS)

I. Bayesian networks are graphical models that encode structural relationships among variables of interest [1].
II. Structural models may encode causal relationships that reflect underlying mechanisms.
I. GNS' Reverse Engineering Forward Simulation (REFS) platform performs massively parallel inference of model structure at industrial scale [2-4].
II. REFS learns ensembles of model structures maximally supported by the data.

Figure 2. Graphical representation of a directed Bayesian network. Target nodes (children) are numerically predicted by their immediate upstream nodes (parents); e.g. $C = \alpha + \beta_1 B + \beta_2 H$

Multimodal Data
- Genomics
- Proteomics
- Phenotyping

→ Structure Learning: Ensemble sampling via massively parallel inference → Prediction: Forward simulation of targeted interventions

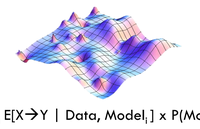
Numerical Sampling of Model Ensembles

I. In high dimensional domains (n < p), many models describe the data equally well.
II. Selection of a single network model underestimates prediction error.
III. Ensembles of network models - sampled from the posterior distribution P(Model | Data) - simultaneously capture parametric and structural uncertainty.
IV. A single ensemble naturally resolves high vs. low confidence structural relationships amongst variables of interest.

Bayesian Model Inference

$$P(\text{Model} \mid \text{Data}) \sim P(\text{Data} \mid \text{Model}) \times P(\text{Model})$$

Posterior Distribution | Marginal Likelihood | Prior Distribution



Posterior Landscape

$$E[X \rightarrow Y] = \Sigma_i E[X \rightarrow Y \mid \text{Data}, \text{Model}_i] \times P(\text{Model}_i \mid \text{Data})$$

Figure 3. REFS model construction via Monte Carlo sampling of the posterior model landscape. The confidence of a constituent edge X→Y is obtained by averaging its appearance over the ensemble of most probable models.

Model Construction & Simulation

- REFS ensembles orient profiling measures into directed graphical networks composed of local structural models - i.e. generalized linear regressions – between upstream drivers and downstream effectors.
- Importantly, REFS distinguishes co-expression (correlation) from co-regulation (conditional independence) as most co-expression does not imply direct regulation.
- Conditional independence relations effectively prune network structure for parsimonious regulatory models.
- Comparisons of model reconstruction of Q50 striatum profiles (excluded from the training model) show good recovery of expression profiles (**Figure 4**).
- After model construction, exhaustive interventional simulations can then be computed to predict downstream effects of a hypothetical perturbation.
- This allows identification of targets downstream of CAG, which are intermediaries towards relevant phenotypes and pathways (**Figure 5**).
- Simulation network connectivity characteristics between different data modalities for the striatum models are summarized in **Table 1**.

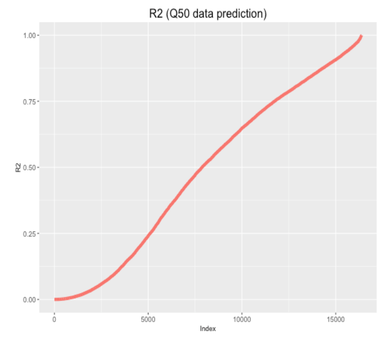


Figure 4. Prediction accuracy of Striatum Q50 Data show the coefficient of determination (R2) of predicted vs. actual values using network model.

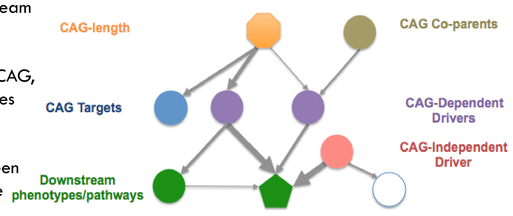


Figure 5. Simulations allow differentiation of drivers of downstream phenotypes or pathways of interest are dependent on CAG length (purple) from those that act independent of CAG (red)

Upstream	Downstream (Striatum)			
	mRNA	miRNA	Protein	Psych
AGE x CAG	10,287	161	28	3
mRNA	97,515	562	177	124
miRNA	129	1,086	2	11
Protein	292	4	2172	16

Table 1. Striatum model edges with nominal significance threshold of P<0.05 and absolute percent change of >1% over baseline across different data modalities shown.

Validation & Discovery

- The models that have been built are powerful, but remain mathematical constructs built upon a single large-scale experiment.
- To validate the ability to use the models in a predictive or prognostic mode, we have undertaken an equally large-scale validation project.
- In collaboration with William Yang, Steve Horvath and Giovanni Coppola at UCLA, a genetic perturbation project across the major nodes of this and other models created on the dataset is ongoing.
- Validation is based on RNA sequencing of the striatum of genetically or chemically perturbed mice from >100 of the most significant nodes of the network.
- Early results are promising with the perturbed mice confirming many of the assertions of the model.
  - Over one quarter of the genes altered by PDE10a inhibition (GSE89505)[5] were predicted by the model to be altered by changes in PDE10 expression levels ($p < 2x10^{-59}$) shown in **Figure 6.**
  - A heterozygous knock-out of Adcy5, a strong hub in the model, results in 406 genes rescued or exacerbated compared to control which totals around 20% of the overall predicted Adcy5 downstream genes in the model (**Figure 7**).

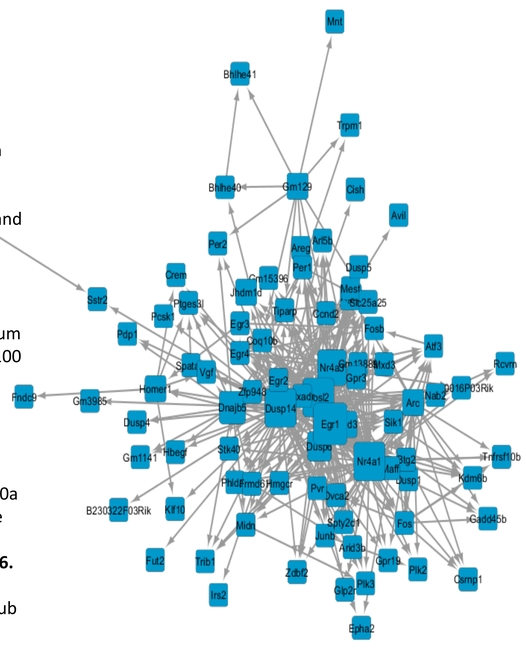


Figure 6. Genes changed by PDE10a inhibition in mice connected by the edges predicted by the model. This is strong evidence that PDE10a is a hub in the model, as expectation would have been few if any connections between the genes, not a rich network ($p < 2x10\text{-}59$).

Conclusions

I. Large-scale Bayesian network inference provides a rigorous data-driven framework for transcriptional regulatory inference across the Htt allelic series.
II. Networks can be informed by biological priors to varying degrees to balance complexity and interpretability with novel data driven findings.
III. REFS forward simulations exhaustively enumerate the downstream effects of hypothetical network interventions and statistically quantify the magnitude and uncertainty of predicted effects.
IV. Simulation networks highlight a progressive expansion of CAG-mediated transcriptional dynamics, increasingly modulated by tissue-specific regulatory factors over time.
V. Validation results thus far are encouraging including:
  - more than one quarter of the genes altered by PDE10a inhibition (Beaumont, 2016 GSE89505) were predicted by the model as such ($p < 2x10\text{-}59$).
  - 20% of the genes predicted to be altered by Adcy5 changes are altered in an Adcy5 heterozygous knock-out.
VI. Many more perturbations are coming in and we will soon be able to validate and update the model based on *in vivo* results.

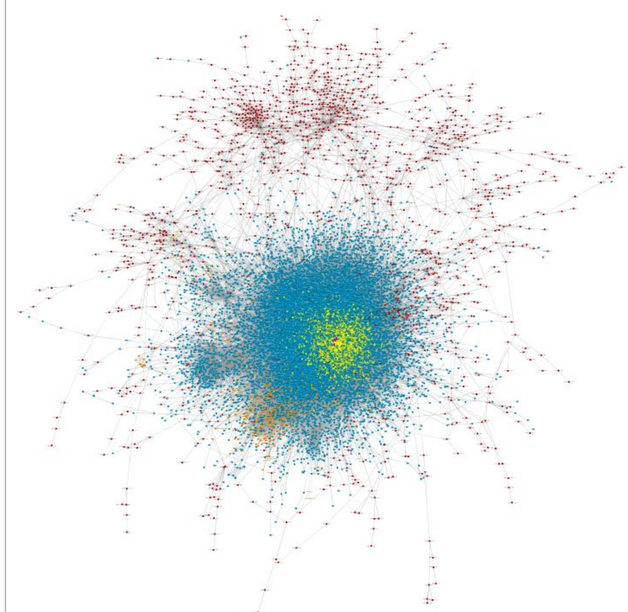


Figure 7. The entire project 5 model with RNA nodes in blue, miRNA nodes in orange, protein in red. The portion of the network under influence by Adcy5 is highlighted in yellow.

Resources For the HD Community

In conjunction with CHDI, GNS has prepared a suite of model files and annotations available via the HDinHD data portal:

I. Integrated and quality-controlled data frames for RNAseq, proteomics, and Psychogenics behavioral profiles from 15 tissue x age experiments.
II. Tabulated and annotated REFS simulation results from exhaustive pairwise interventional perturbations.
III. Cytoscape network files, including annotations and literature co-occurrence, for REFS simulation networks.
IV. OpenBEL namespaces and tissue-specific assertions for REFS simulations.

Acknowledgements

This work was generously supported by the CHDI Foundation.

References

1. Friedman N, Koller D. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*. 2003;50:95-125.
2. Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, Laramie JM, Markedian J, Piper BA, Willke RJ, Rublee DA. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records, *J Diabetes Sci Tech*. 2015; 10(1):6-18.
3. Steinberg GB, Church BW, McCall CJ, Scott AB, Kalis BP. Novel predictive models for metabolic syndrome risk: a "big data" analytic approach. *Am J Manag Care*. 2014;20(6): e221-e228.
4. Xing H, McDonagh PD, Bienkowska J, et al. Causal model-ing using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLOS Comp Biol*. 2011;7(3):e1001105.
5. Beaumont V, Zhong S, Lin H, Xu W et al. Phosphodiesterase 10A Inhibition Improves Cortico-Basal Ganglia Function in Huntington's Disease Models. *Neuron* 2016 Dec 21;92(6):1220-1237. PMID: 27916455