GNS HEALTHCARE
Driving Intelligent Interventions

Data-Driven Reconstruction and Simulation of Transcriptional Regulatory Networks in the Htt Allelic Series

Michael Beste[1], Tun-Hsiang Yang[1], Jeanne Latourelle[1], Boris Hayete[1], Liliana Menalled[2], Dani Brunner[2], Vadim Alexandrov[2], Seung Kwak[3], David Howland[3], Jeff Aaronson[3], Iya Khalil[1], James Rosinski[3]

[1]GNS Healthcare, Cambridge, MA 02139; [2]PsychoGenics Inc., Tarrytown, NY 10591; [3]CHDI Foundation, Princeton, NJ 08540

CHDI FOUNDATION

Objectives

- Polyglutamine expansion within exon 1 of HTT is associated with transcriptional dysregulation contributing to disrupted neurotransmission and progressive loss of striatal medium spiny neurons.
- High resolution transcriptional and behavioral profiling across the murine Htt allelic series is designed to capture the most proximal effects of CAG expansion and resolve incipient molecular events across multiple tissues.
- We have applied GNS' Reverse Engineering Forward Simulation (REFS) machine learning platform to statistically model and orient CAG → transcriptional → behavioral pathways using the allelic series profiling compendium.
- Exhaustive interventional simulations across REFS graphical models naturally identifies high confidence upstream/downstream transcriptional influences relative to CAG expansion over time.

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. R6/2 transgenic HTT knock-in (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

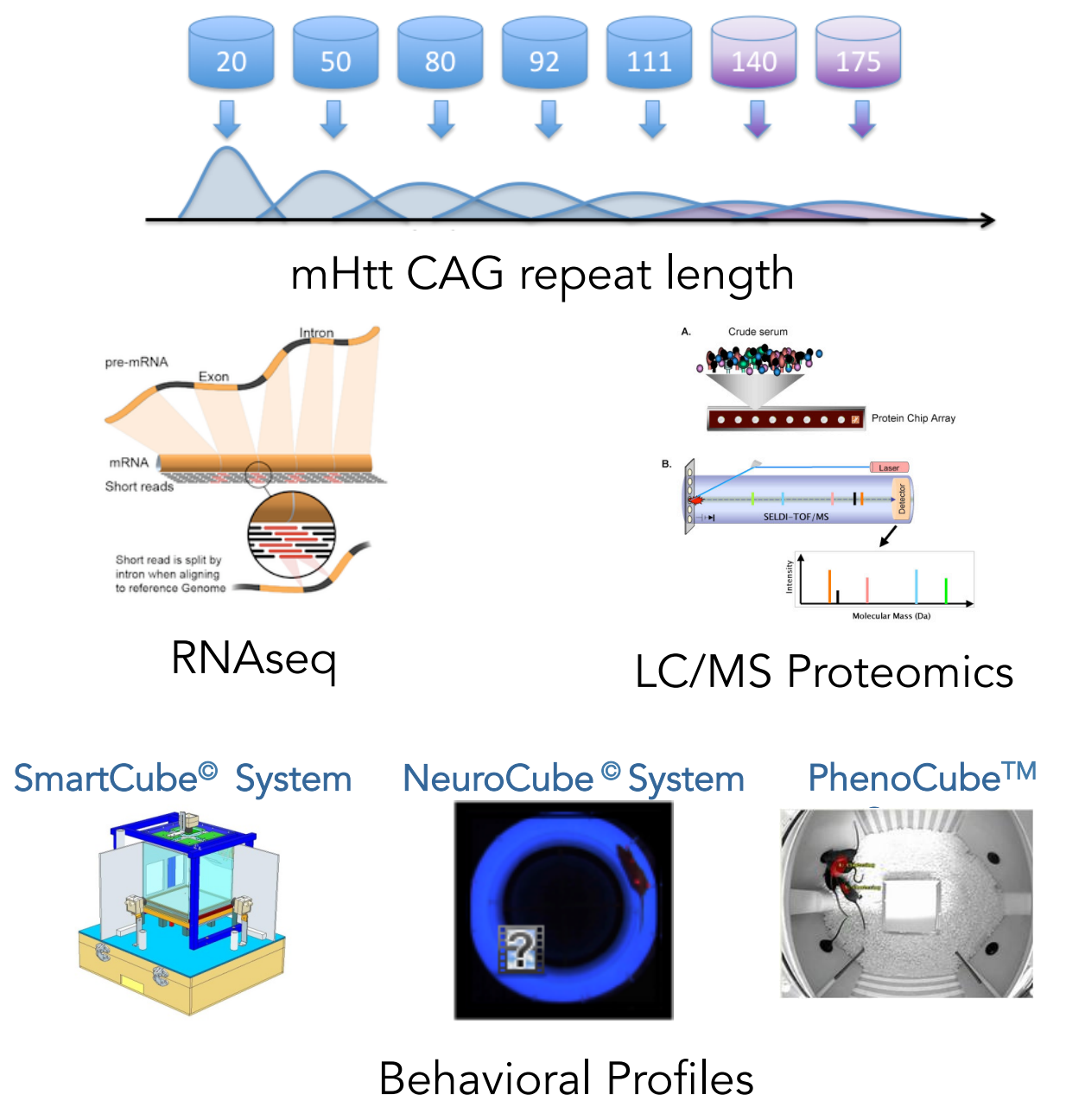


Figure 1. Experimental design and profiling platforms characterizing the mHtt allelic series.

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.

III. GNS' Reverse Engineering Forward Simulation (REFS) platform performs **massively parallel inference** of model structure at industrial scale [2-4].

IV. **REFS** learns ensembles of model structures maximally supported by the data.

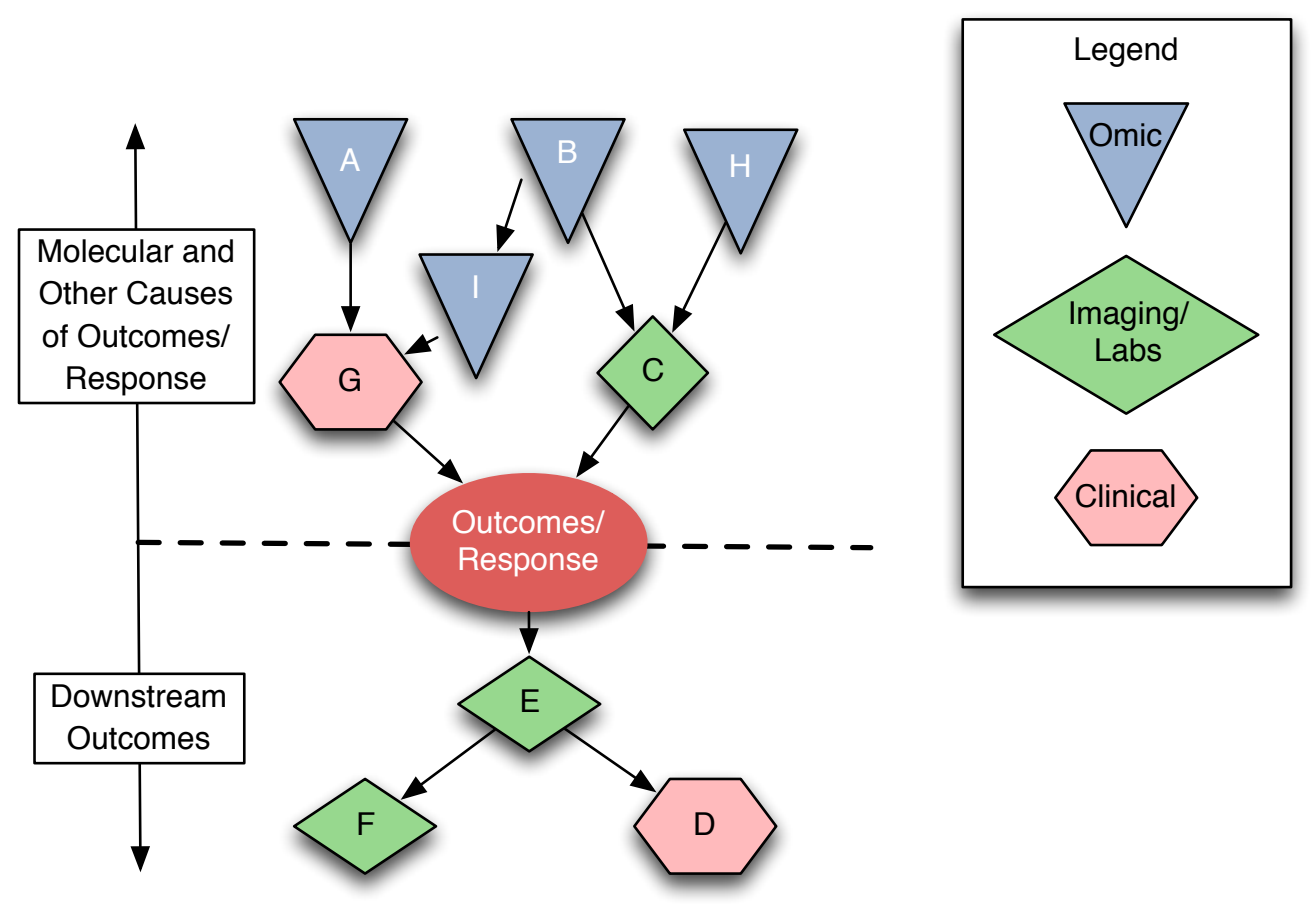
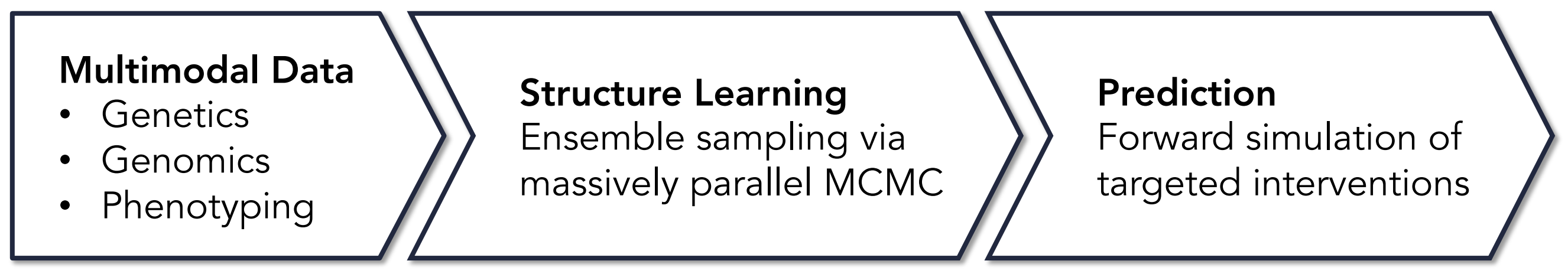


Figure 2. Graphical representation of a directed Bayesian network. Target nodes (children) are numerically predicted by their immediate upstream nodes (parents); e.g. $C = \alpha + \beta_1 B + \beta_2 H$



Multimodal Data
- Genetics
- Genomics
- Phenotyping

Structure Learning
Ensemble sampling via massively parallel MCMC

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.

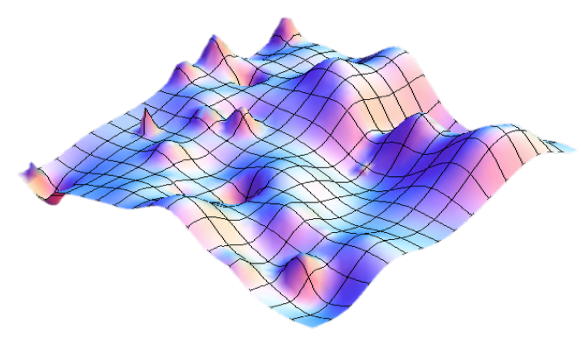
II. Selection of a single network model underestimates prediction error.

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

$$\underbrace{P(\text{Model} \mid \text{Data})}_{\text{Posterior Distribution}} \sim \underbrace{P(\text{Data} \mid \text{Model})}_{\text{Marginal Likelihood}} \times \underbrace{P(\text{Model})}_{\text{Prior Distribution}}$$



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

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 regulators and downstream effectors.
- Exhaustive interventional simulations – the numerical derivative of the underlying parametric models – are then computed to predict downstream effects of a hypothetical perturbation.

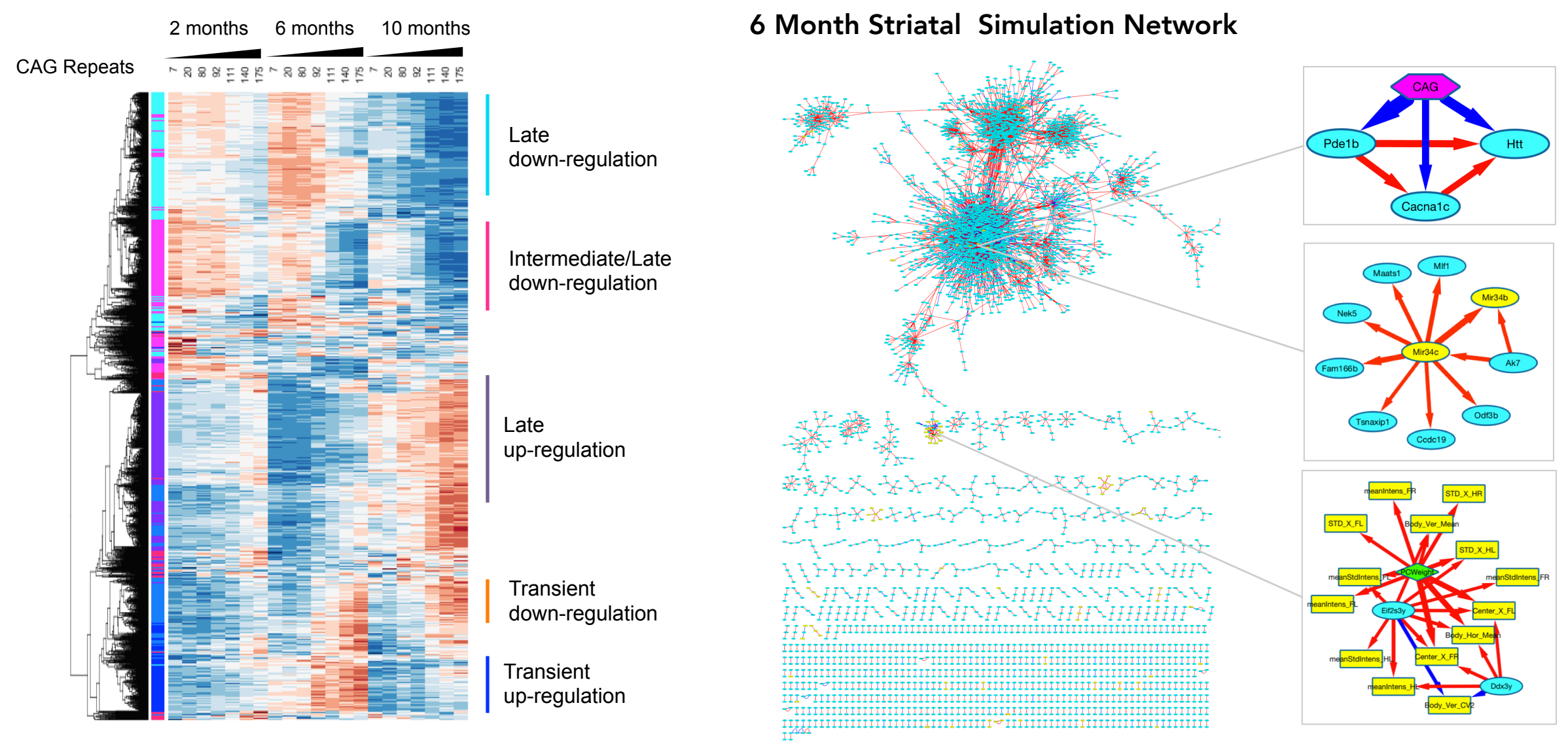


Figure 4. Normalized RNAseq expression profiles (left) are first oriented within a Bayesian network ensemble. A simulated response network (right) is subsequently derived from assessing significant pairwise perturbation effects across the ensemble.

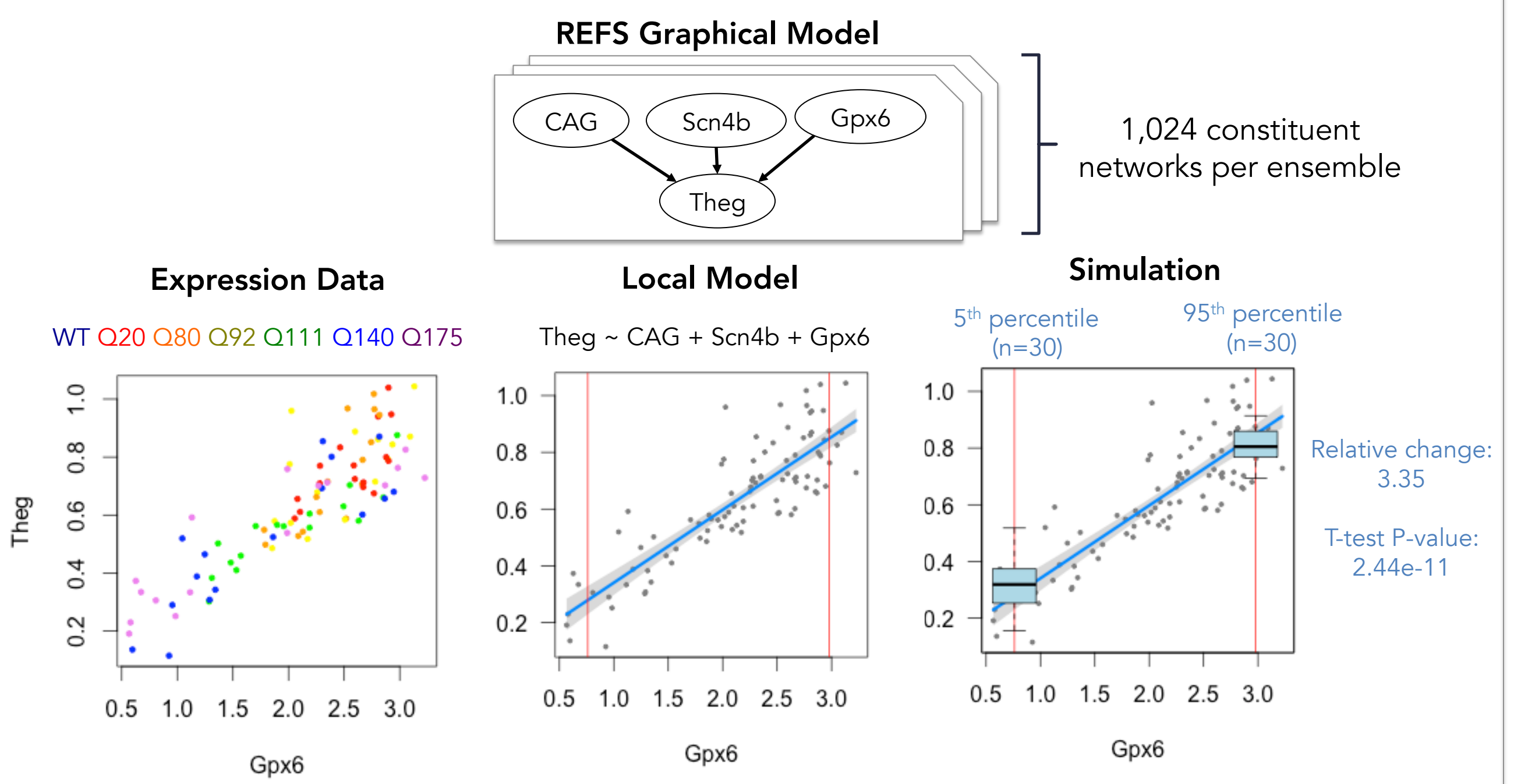


Figure 5. Exemplary REFS simulation for a constituent local model characterizing Theg regulation by CAG repeat length, Scn4b, and Gpx6 expression. Independent modulation of Gpx6 levels are predicted to positively regulate expression levels of Theg

- Importantly, REFS distinguishes co-expression (correlation) from co-regulation (conditional independence).
- Most co-expression does not imply direct regulation.
- Conditional independence relations effectively prune network structure for parsimonious regulatory models.

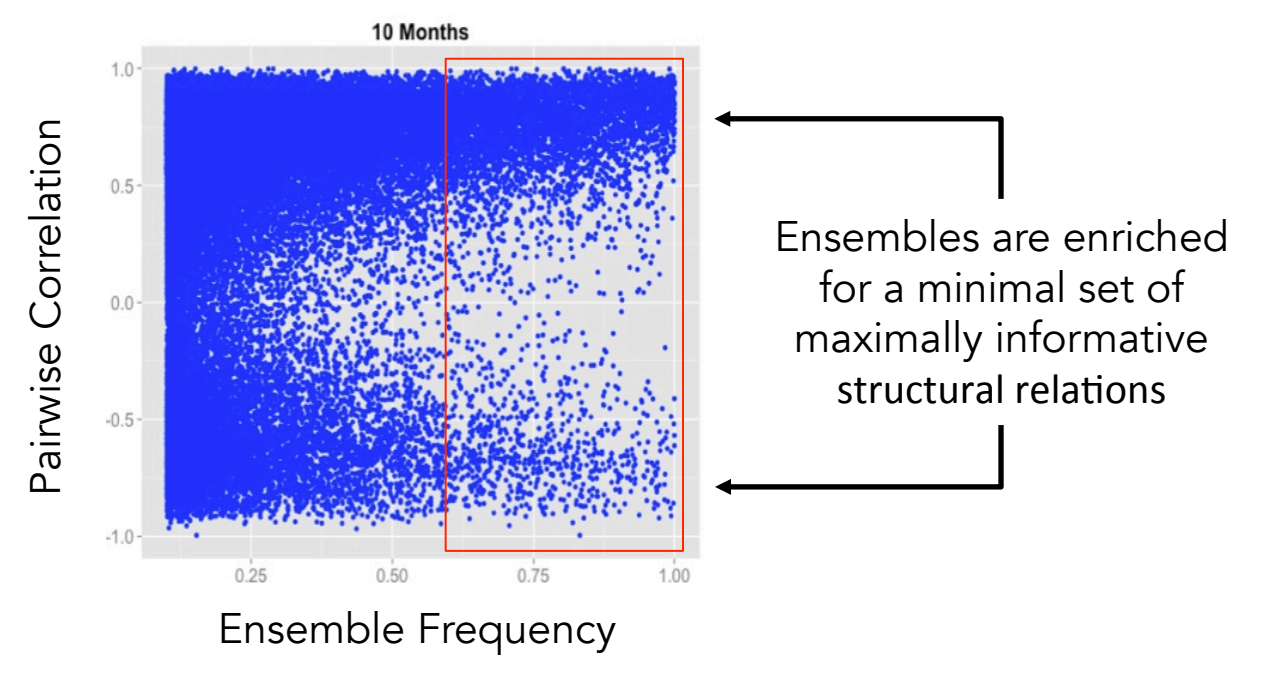


Figure 6. Bifurcation plot characterizing the relationship between model frequency (x-axis) and marginal correlation (y-axis) among all gene pairs.

Validation & Discovery

- To validate and prioritize regulatory pathways inferred via REFS, simulation networks were subset for upstream regulators predicted to co-regulate CAG target genes.
- Predicted co-regulators were compared to those inferred via an Ingenuity regulatory analysis for the same CAG target gene set.
- Simulation networks recapitulate both canonical (HTT, CREB1, CREBBP, REST) and novel (ATN1, KMT2D) HD regulator across multiple tissues.

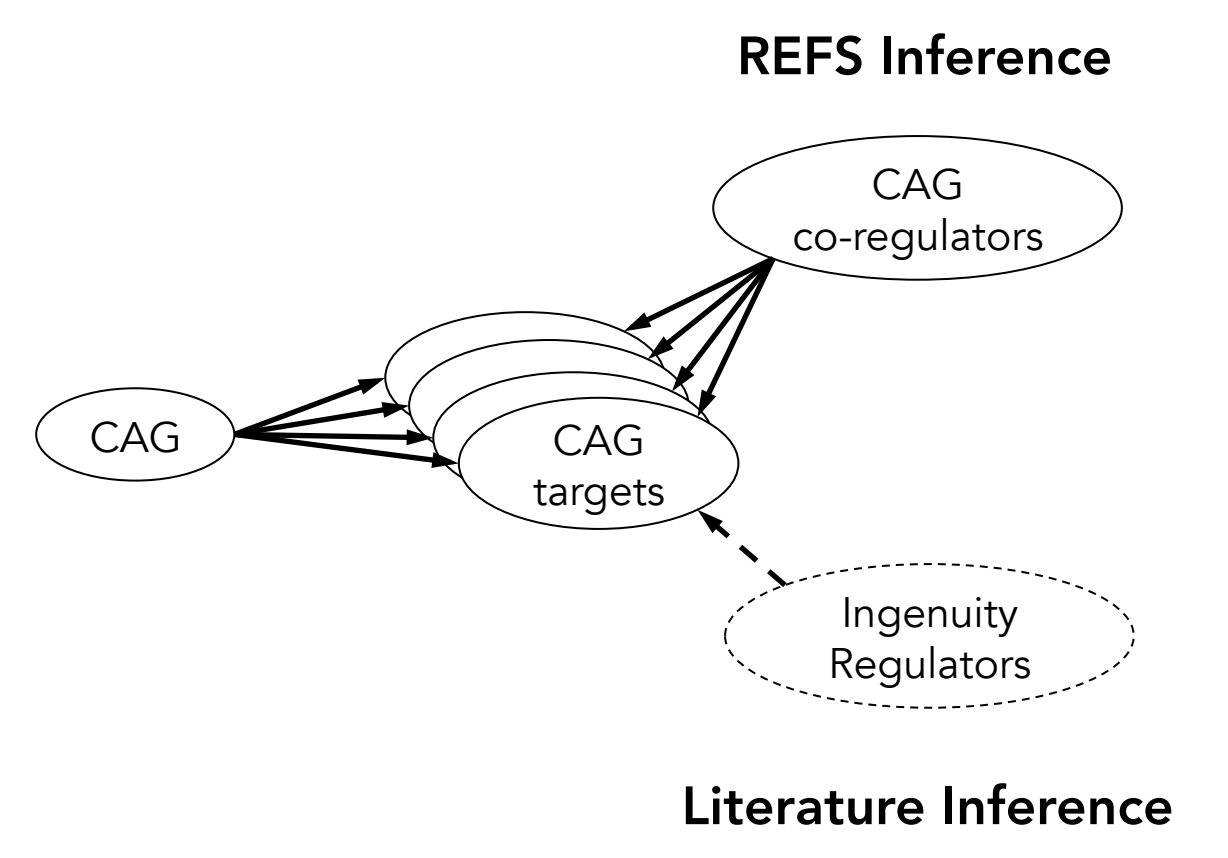


Figure 7. Co-regulators of the CAG transcriptional response inferred via REFS were compared against literature inferred regulators for an identical set of CAG target genes.

REFS Inferred Co-Regulator	Striatum	Cortex	Cerebellum
Htt	$<10^{-24}$	$<10^{-8}$	$<10^{-19}$
Crebbp	$<10^{-3}$	0.01	$<10^{-2}$
Rest	$<10^{-2}$	0.02	$<10^{-3}$
Creb1	$<10^{-5}$	$<10^{-2}$	$<10^{-11}$
Atn1	$<10^{-4}$	$<10^{-2}$	$<10^{-25}$
Kmt2d	$<10^{-2}$	0.04	$<10^{-2}$

Table 1. Statistical significance (P-value) for literature-based inference (Ingenuity) of CAG target co-regulators identified via REFS simulation networks.

Conclusions

I. Large-scale Bayesian network inference provides a rigorous data-driven framework for transcriptional regulatory inference across the Htt allelic series.

II. REFS forward simulations exhaustively enumerate the downstream effects of hypothetical network interventions and statistically quantify the magnitude and uncertainty of predicted effects.

III. Simulation networks highlight a progressive expansion of CAG-mediated transcriptional dynamics, increasingly modulated by tissue-specific regulatory factors over time.

IV. Independent validation of inferred co-regulators of the primary CAG response identified both canonical (Htt, Creb1, Crebbp, Rest) and novel (Atn1, Kmt2d) targets for further investigation.

V. Ongoing work aims to identify proximal sub-networks relevant to:
   - Investigational drug targets
   - Human age-of-onset modifier genes
   - Regulatory drivers of HTT somatic instability

Resources For the HD Community

In conjunction with CHDI, GNS has prepared a suite of model files and annotations for release 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.

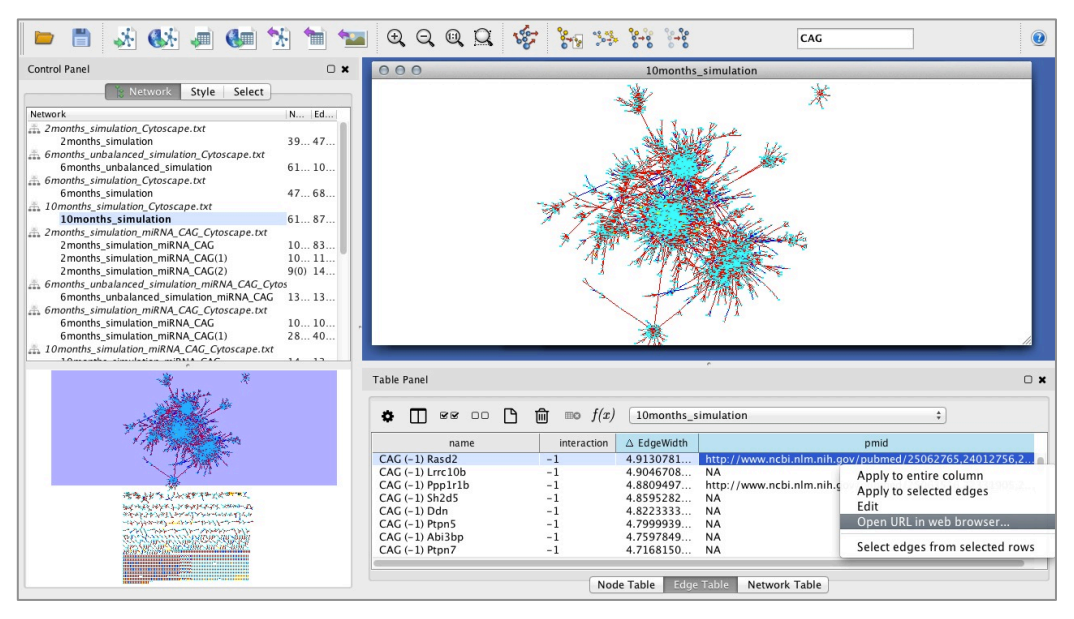
II. Tabulated and annotated REFS simulation results from exhaustive pairwise interventional perturbations.

III. Cytoscape network files, including annotations and literature co-occurrence, for REFS simulation networks.

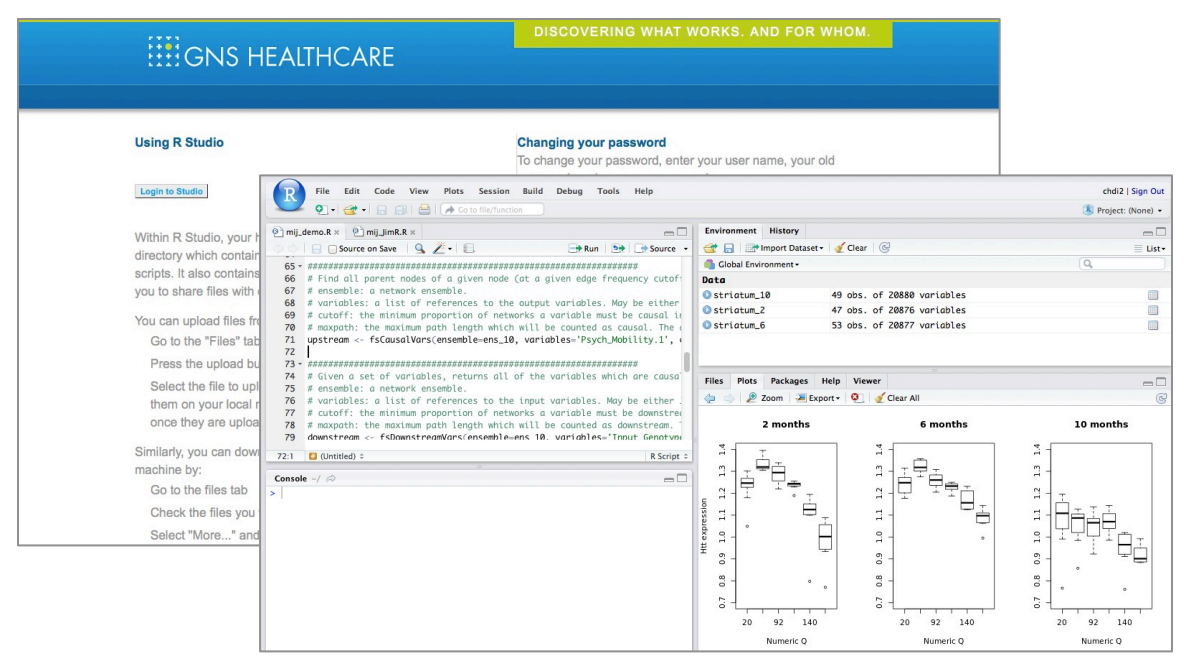
IV. OpenBEL namespaces and tissue-specific assertions for REFS simulations.

V. Hosted Rstudio access to GNS' REFSfs R simulation package for custom ensemble topology queries and model simulations.

Cytoscape Simulation Networks



REFSfs RStudio Simulation Interface



OpenBEL Namespaces and Assertions

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.

Acknowledgements

This work was generously supported by the CHDI Foundation.