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

GNS HEALTHCARE

Driving Intelligent Interventions

Michael Beste¹, Tun-Hsiang Yang¹, Jeanne Latourelle¹, Boris Hayete¹, Liliana Menalled², Dani Brunner², Vadim Alexandrov², Seung Kwak³, David Howland^{3,} Jeff Aaronson³, Iya Khalil¹, James Rosinski³

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



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.

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.

CONCLUSIONS

- 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.

- We have applied GNS' Reverse Engineering Forward Simulation (REFS) machine learning platform to statistically model and orient CAG \rightarrow transcriptional \rightarrow 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:

- R6/2 transgenic HTT knock-in (n=208) 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





Omic

Imaging/ Labs

Clinical

Prior

Posterior

Landscape





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.



- 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:

- Integrated and quality-controlled data frames for RNAseq, proteomics, and Psychogenics behavioral profiles from 15 tissue x age experiments.
- 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.

Liver **Behavioral Profiles** RNAseq (~20k transcripts) VI. LC/MS Proteomics (~6k targets) Figure 1. Experimental design and profiling VII. PsychoGenics Behavioral profiles platforms characterizing the mHtt allelic series.

Causal Inference via Reverse Engineering and Forward Simulation (REFS)

- Bayesian networks are graphical models that encode structural relationships among variables of interest [1].
- 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	Structure Learning	- Frediction

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 Theq



Ensemble Frequency

CAG

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

Figure 7. Co-regulators of the CAG transcriptional

response inferred via REFS were compared against

VALIDATION & DISCOVERY

• Importantly, REFS distinguishes co-

expression (correlation) from co-

• Most co-expression does not imply

• Conditional independence relations

parsimonious regulatory models.

effectively prune network structure for

direct regulation.

regulation (conditional independence).

• To validate and prioritize regulatory pathways inferred via REFS, simulation networks were subset for upstream regulators predicted to co-regulate CAG target genes.

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



REFSfs RStudio Simulation Interface

Using R Studio	Changing your password To change your password, enter	your user nan	ne, your old			
Login to Studio	R File Edit Code View Plots Session Build Debug Tools Help				chd	li2 Sign Ou
	🔍 🍳 + 🔄 🔒 🚔 i 🧥 Co to file/function				🔳 Proje	ct: (None) -
Within R Studio, your I	🕐 mij_demo.R × 🕐 mij_limR.R × 👘	Environment H	listory			
directory which contain	Source on Save 🧣 🖉 + 📃 🔅 🔿 Run 😂 🔿 Source -	Clabal Fastan	ort Dataset + 🛛 🔮 C	ilear 🚱	0	List
scripts. It also contains	66 # Find all parent nodes of a given node (at a given edge frequency cutof	Data				
you to share files with	67 # ensemble: a network ensemble. 68 # variables: a list of references to the output variables. May be either	🔘 striatum_10	49 c	bs. of 20880 variables		
You can upload files fr	69 # cutoff: the minimum proportion of networks a variable must be causal in	Ostriatum_2	47 0	bs. of 20876 variables		
Go to the "Files" tak	70 # maxpath: the maximum path length which will be counted as causal. The (71 upstream <- fsCausalVars(ensemble=ens_10, variables='Psych_Mobility.1', (Stridtum_6	53 0	ibs. of 20877 Variables		
Press the unload bu						
Select the file to unl	74 # Given a set of variables, returns all of the variables which are causa	Files Plots I	ackages Help	Viewer		
them on your local	75 # ensemple: a network ensemple. 76 # variables: a list of references to the input variables. May be either .	🤙 🧼 🔎 Zoo	m 🛛 🚈 Export 🕶	🧕 🥑 Clear All		¢
once they are uploa	77 # cutoff: the minimum proportion of networks a variable must be downstree 78 # maxpath: the maximum path length which will be counted as downstream. 79 downstream - foremetrangemark are seen the next for a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are its path forent in 79 downstream - foremetrangemark are seen to a variable are seen to a variable are its path forent in 70 downstream - foremetrangemark are seen to a variable are seen to a variab	2	nonths	6 months	10 mont	:hs
Similarly, you can dow	72:1 C (Untitled) ÷ R Script ÷	T F		1	1.4	
machine by:	Controls -/ ()	리고무	à l	리고문공	n -	
Go to the files tab		- 1- T	I - -	a-1코 물로곱	2	
Check the files you		ssion	<u> </u>	리 브립니	- AAIA	
Select "More " and		expre	白	9 .		'm_
		Ξ. Ξ.	4	n	n	Ψġ
		8		8	8	
		0	•	0	0 0	•
		• 1			ة 1	
		20	52 140	20 52 140	20 92	140



meString=PsychoGenics Mouse Behavioral Feature Names			
yword=PGI			
mainString=SmartCube, NeuroCube, and PhenoCube featur	es		
eciesString=10090			
scriptionString=PsychoGenics (PGI) behavioral profil: proionString=20150219	# Document Properties Section	*******	
eatedDateTime=2015-02-19T12:00:00			
eryValueURL=http://www.psychogenics.com/proprietarydi	Is SET DOCUMENT Name = "CHDI Foundation Allelic Series: GNS Healthcare REPS simulation Assertions" SET DOCUMENT Description = "Strintum, Cortex, and Liver assertions. Approximately 559,000 statements"		
uthor]	SET DOCUMENT Version = "1.0"		
meString=GNS Healthcare	SET DOCUMENT Copyright = "Copyright (c) 2014-2015, CHDI Foundation, All r	ights reserved."	
pyrightString=Copyright (c) 2015, OpenBEL Project. Tř ntactInfoString=beste8gnshealthcare.com	SET DOCUMENT Authors = "CHDI Foundation and GNS Healthcare"	-,	
itation]	****	*****	
meString=PsycoGenics Inc., 765 Old Saw Mill River Roa	# Definitions Section		
scriptionString=PsychoGenics (PGI) behavioral profil:			
blishedVersionString=Thu, 19 Feb 2015 12:00:00	DEFINE NAMESPACE HGNC AS URL "http://resource.belframework.org/belframewo	rk/latest-release/namespace/hgnc-human-genes.belns"	
blishedDate=2015-02-19	DEFINE NAMESPACE MGI AS URL "http://resource.belframework.org/belframewor	k/latest-release/namespace/mgi-mouse-genes.belns"	
ferenceURL=http://www.psychogenics.com/proprietarydru	DEFINE NAMESPACE PGI AS "pgi-mouse-cubes.belns"		
'rocessing]	DEFINE ANNOTATION Anatomy AS URL "http://resource.belframework.org/belfr	amework/latest-release/annotation/anatomy.belanno"	
seSensitiveFlag=yes	DEFINE ANNOTATION Species AS URL "http://resource.belframework.org/belfr	amework/latest-release/annotation/species-taxonomy-id.belanno"	
limiterString=			
cheableFlag=yes	***************************************	******	
	A Chahamanha Cashian		
alues]	# Statements Section		
weight B			
ych_Alfa_rb B	SET STATEMENT_GROUP = "2 months, striatum"		
ych Alfa HI B			
web Alfa ND D	SET Evidence = "GNS Healthcare REFS simulation P-value <0.05"		
web ANGLE 1 B	SET Species = "10090"		
wch Angle 4. B	SET Anatomy = "striatum"		
wch Avg Spd B	SET Citation = {"Other", "GNS Healthcare Striatum Simulation Report", "GNS	Healthcare Striatum Simulation Report","","",""}	
wch Body Hor CV2 B			
ych Body Hor Mean B	measurement(PGI:Weight) increases measurement(PGI:PCWeight)		
ych Body Hor Move B	measurement(PGI:Weight) increases measurement(PGI:Psvch Body Hor Mean)		
ych_Body_Hor_Range B	measurement(PGI:Weight) decreases measurement(PGI:Psych PctTotalDurVisit.	D1.7AM)	
ych Body Hor Rel Move B	measurement(PGI:PCWeight) decreases measurement(PGI:Psych PctTotalDurVisi	t.D1.7AM)	
	measurement(PGI:PCWeight) increases measurement(PGI:Psych Body Hor Mean)	,	
	rnabhundance/MGT+2C158547 1) decreases rnabhundance/MGT+T16)		
	rnabundance(MCI:AC158547 1) decreases rnabundance(MCI:I]6)		
	minnahundende (MCI-Min2072) ingebesses mashundange (MCI-Tdus)		
	mines hunderes (NGT.Nix 2072) increases in abunderes (NGT.Thus)		
	milinaAbundance(MoliMiliSV/2) increases fhaAbundance(MoliIdua)		



Numerical Sampling of Model Ensembles

In high dimensional domains ($n \ll p$), many models describe the data equally well

Selection of a single network model 11. underestimates prediction error.

III. Ensembles of network models sampled from the posterior distribution P(Model | Data) - simultaneously capture parametric and structural uncertainty.



Posterior Marginal Likelihood Distribution Distribution $E[X \rightarrow Y] = \sum_{i} E[X \rightarrow Y \mid Data, Model_{i}] \times P(Model_{i} \mid Data)$ Figure 3. REFS model construction via Monte Carlo sampling of the posterior model landscape.

The confidence of a constituent edge $X \rightarrow Y$ is obtained by averaging its appearance over the ensemble of most probable models.

Bayesian Model Inference

P(Model | Data) ~ P(Data | Model) x P(Model)

 Predicted co-regulators were compared to those inferred from an Ingenuity regulatory analysis for the same CAG target gene set.

• Simulation networks recapitulate both canonical (HTT, CREB1, CREBBP, REST) literature inferred regulators for an identical set of and novel (ATN1, KMT2D) HD regulator CAG target genes. across multiple tissues.

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.

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.



Literature Inference

REFS Inference

CAG

co-regulators