

Project A: End-to-End Pilot Combining Data, Building Models, and Querying Them

Andrew Kasarskis & Ilya Kupershmidt

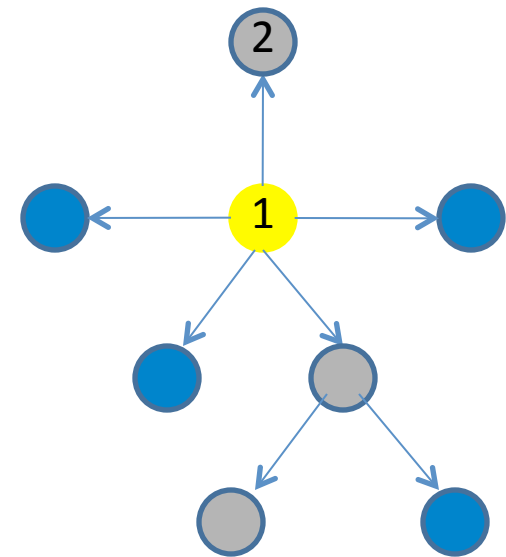
April 23, 2010

What we wanted from this pilot

- Global Coherent Data Sets
- Network Models
- An attempt at standardization
- Data and models accessible to tools
- Learnings

What do these network models look like?

- The core info
 - Node 1 (e.g. HMGCR)
 - Node 2 (e.g. LDLR)
 - Numbers that describe the association between nodes 1 & 2 (e.g. r^2)
- Meta-data about the network
 - Code and parameters used
 - Input data



Sage Commons System Concepts

Network Models

Global Coherent Data Sets

Source Code

**Emerging Global Coherent Data
Sets**

Perturbation Experiment Results

Global Coherent Data Sets

A data set containing genome-wide DNA variation and intermediate trait, as well as physiological phenotype data across a population of individuals large enough to power association or linkage studies, typically 50 or more individuals. To be coherent, the data needs to be matched with consistent identifiers. Intermediate traits are typically gene expression, but may also include proteomic, metabolomic, and other molecular data.

Status

Definition

Sage - Available

Dataset available from Sage website

Sage - Transition

Dataset in process of being made available

Requires Release

Dataset with known or anticipated legal release requirements prior to posting on Sage website

In progress

Dataset not yet complete

GCDs are current state of knowledge and subject to change as more information becomes available to Sage

Sage Data Sets – Available/In transition

Available	Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
	Mouse_CVD_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	CVD	Jake Lusis	UCLA	Sage - Available	334
	Human_Cancer_HCC_HKU	HCC	Human	Cancer	John Luk	HKU	Sage - Available	250
	Human_CVD_Liver_Vanderbilt/Pittsburgh/StJudes	Liver	Human	CVD	Guengrich/Strom/Schuetz	Vanderbilt/Pittsburgh/StJudes	Sage - Available	517

Transition	Dataset Name	Tumor/Tissue Type	Species	Disease	Investigator	Institution	Status	Approximate Number of Individuals
	Human_Cancer_Breast_BCCA	Breast	Human	Cancer	Aparicio/Caldas	BCCA/Cambridge	Sage - Available	1,500
	Human_Cancer_Glioblastoma_TCGA	Glioblastoma	Human	Cancer	TCGA	TCGA	Sage - Available-subset	413
	Human_Neurodegenerative_Brain:Prefrontal cortex_Visual Cortex_Cerebellum_HBTRC	Brain:Prefrontal cortex_Visual Cortex_Cerebellum	Human	Neuro-degenerative	Francine Benes	HBTRC	Sage - Transition	700
	Mouse_Metabolic_Liver_UCLA	Liver	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	111
	Human_Cancer_AML(pediatric)_FHCRC	AML(pediatric)	Human	Cancer	Sohail Meschini	FHCRC	Sage - Transition	200
	Mouse_Metabolic_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	442
	Mouse_Metabolic_Adipose_Liver_Brain_Muscle_UCLA	Adipose_Liver_Brain_Muscle	Mouse	Metabolic	Jake Lusis	UCLA	Sage - Transition	309

See <http://sage.fhcrc.org/downloads/index.php> for available data.

Network Models in Sage Commons Pilot

- Human B cell interactome – Mariano Alvarez
- Mouse embryonic stem cell differentiation – Sheng Zhong
- Yeast Genetic Interaction Map - Gregory Hannum
- TCGA Human Glioblastoma - Mariano Alvarez
- TCGA Human Glioblastoma – Sage
- Human luminal A breast cancer – Sage, with data and consult from S. Aparicio, C. Caldas, S. Shah
- Mouse BxH ApoE -/- liver, brain, adipose, and muscle – Sage & J. Lulis
- Human Liver Cohort – Sage (Guengerich, Strom, & Shuetz data)
- Hepatocellular carcinoma and adjacent normal tissue – Sage & J. Luk
- All available at <http://sage.fhcrc.org/downloads/index.php>

Network Modeling User Group Presentations

- Mariano Alvarez, "Functional mining of transcriptional networks"
- Gang Fang, "Subgroup-specific Differential Coexpression Networks: An Illustrative Analysis on a Sage Commons Pilot Dataset"
- Sohrab Shah, "METABRIC : profiling the genomic landscape of breast cancer with integrated copy number and gene expression analysis"
- Gregory Hannum, "Genome-wide association data reveal a global map of genetic interactions amongst protein complexes"
- bonus discussion & synthesis of needs!

What did we learn?

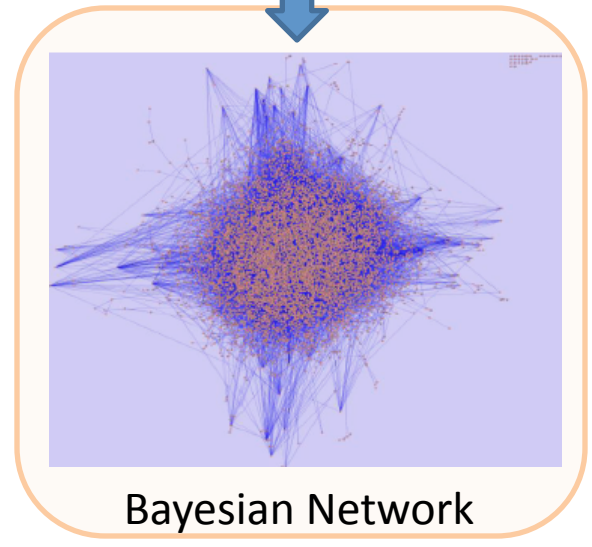
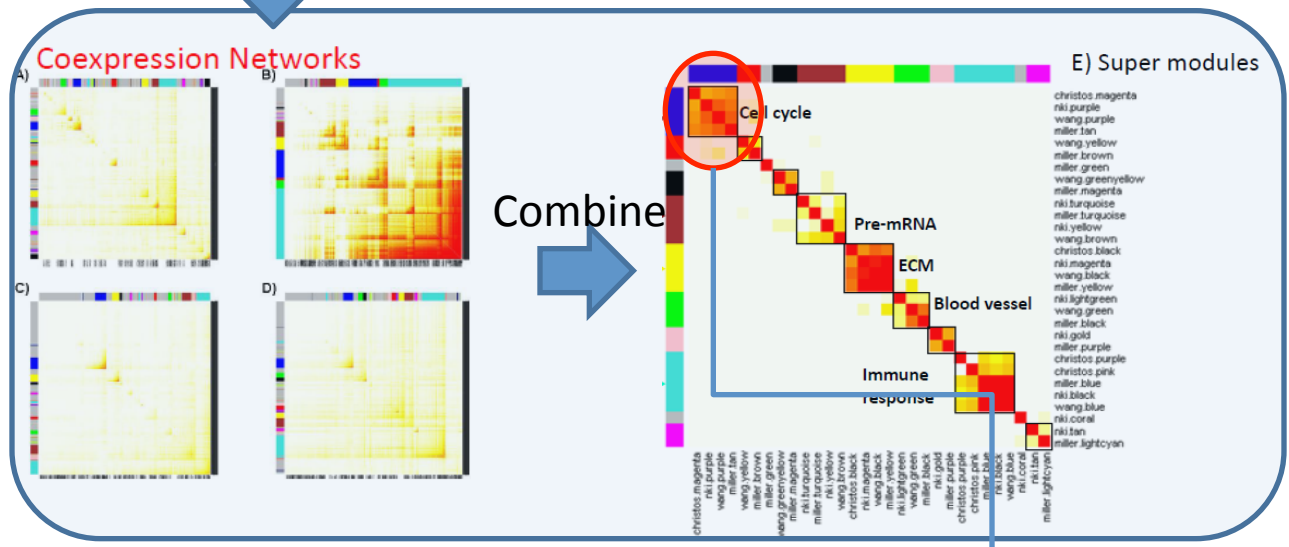
- All data is a lot of work
- Human data is *really* a lot of work
- Steep learning curve on leveraging ontologies
- Lots of applicable tools
- Network models are pretty easy to share –
new RDF model

Project A Activity During Congress

- Friday Discussion
 - Use cases
 - Challenges today
 - Gaps in what has been demonstrated in Friday AM Sessions
 - Other suggestions from the group on problems that need to be tackled
- Saturday
 - Issues ranked by cost/benefit of addressing
 - Commitments to near-term steps

Leveraging Breast Cancer Networks

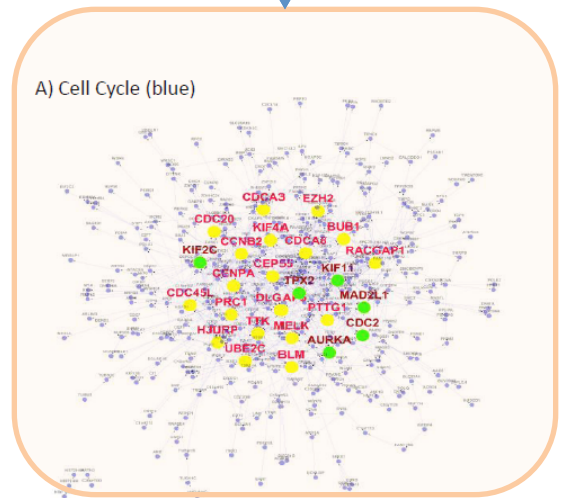
4 Published Breast Cancer Gene Expression Studies



Co-expression Network

Bayesian Network

Bayesian Sub-network with Global Drivers (yellow)



Data Mining