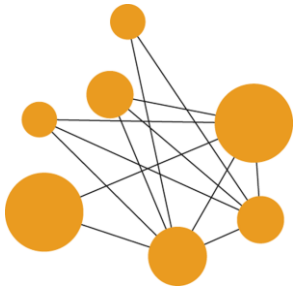
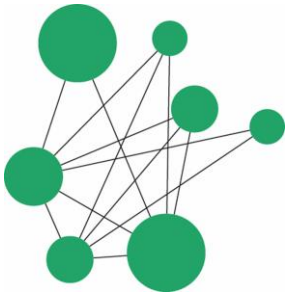
computational bioscience program

university of colorado school of medicine

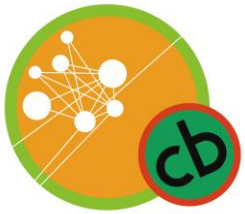
Building a Mind for Life

Lawrence Hunter, Ph.D.
Director, Computational Bioscience Program
University of Colorado School of Medicine



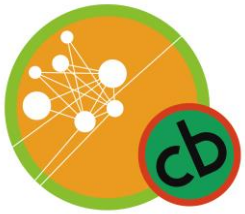


Larry.Hunter@ucdenver.edu
http://compbio.ucdenver.edu/Hunter
@ProfLHunter



Sometimes, two hard problems are easier to solve than one

- Understanding genomics is a superhuman task
- Human-like AI domain knowledge (even as just Bayesian priors) remains a distant goal



The first artificial mind will think about molecular biology

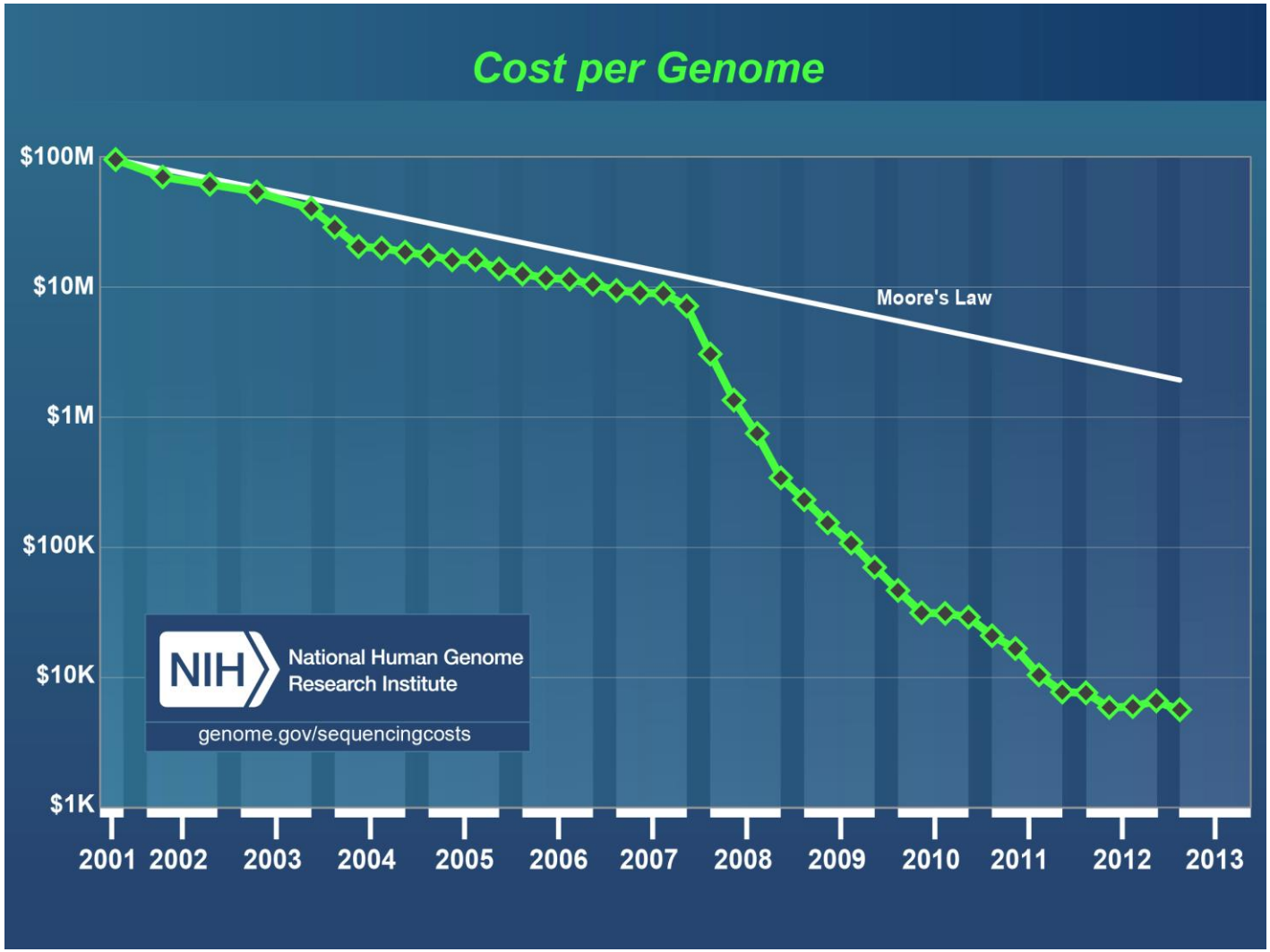
- “You can’t think about thinking without thinking about thinking about something.”

  – Seymour Papert, 1974

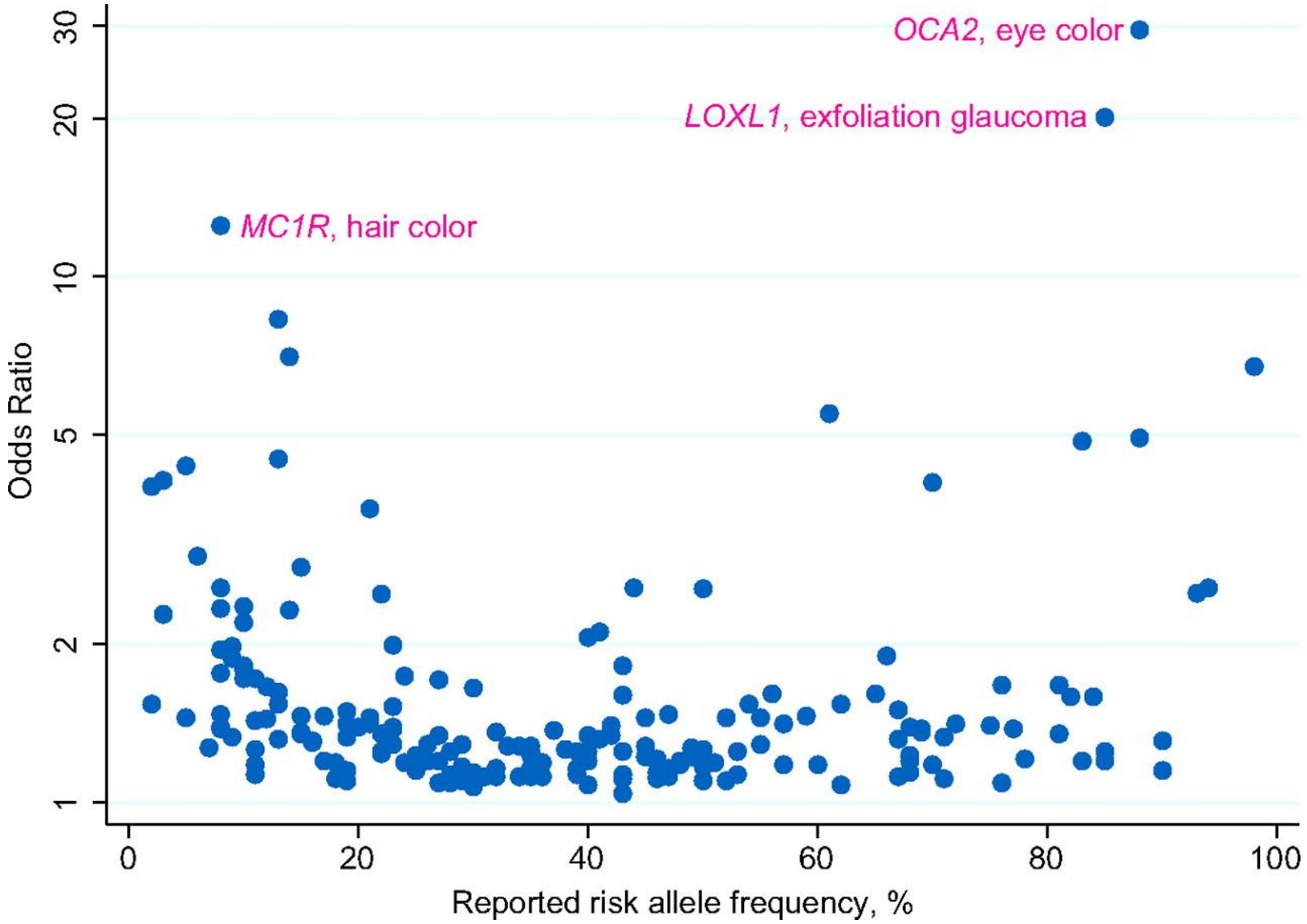
- “A thorough study of Human Physiology is, in itself, an education broader and more comprehensive than much that passes under that name. There is no side of the intellect which it does not call into play, no region of human knowledge into which either its roots, or its branches, do not extend.”

  – Thomas Huxley,1893

The revolution in genomics



$5 billion worth of research



>1500 Hypertension genes

HuGE Navigator (version 2.0)
An integrated, searchable knowledge base of genetic associations and human genome epidemiology.

HuGE Navigator > Phenopedia (HuGEpedia)

Last data upload: 26 Apr 2013. (Total 2577 disease terms)

Phenopedia

Data collected since 2001

Home | About | Search Instructions | FAQs

Search Phenopedia **for** Enter a disease term Go Clear All

?

Download

Hypertension

Related Diseases

Summary

Total Publications
3769

Meta-Analyses
142

Genes
1533

GWAS
76

1537 genes have been reported with Hypertension

[Click to re-sort the table]

Group genes by KEGG | Display genes in UCSC

Gene	Total	Meta	GWAS	Gene-Env	Trend
ACE SNP	567	18	1	134	
AGT SNP	347	15	0	73	
NOS3 SNP	255	9	0	52	
AGTR1 SNP	236	5	0	61	
MTHFR SNP	169	13	1	51	
CYP11B2 SNP	156	7	0	28	
APOE SNP	152	5	1	30	
GNB3 SNP	141	6	0	29	

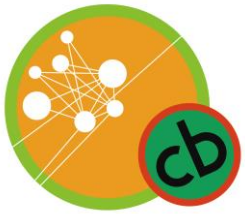
Analysis is the hard part

- “We are close to having a $1,000 genome sequence, but this may be accompanied by a $1,000,000 interpretation.”

  - Bruce Korf, president American College of Medical Genetics

- Not only is the cost of sequencing essentially free, but big computers and big storage are cheap, too. What will keep us busy for the next 50 years is understanding the data”

  - Russ Altman, chair of Biomedical Engineering at

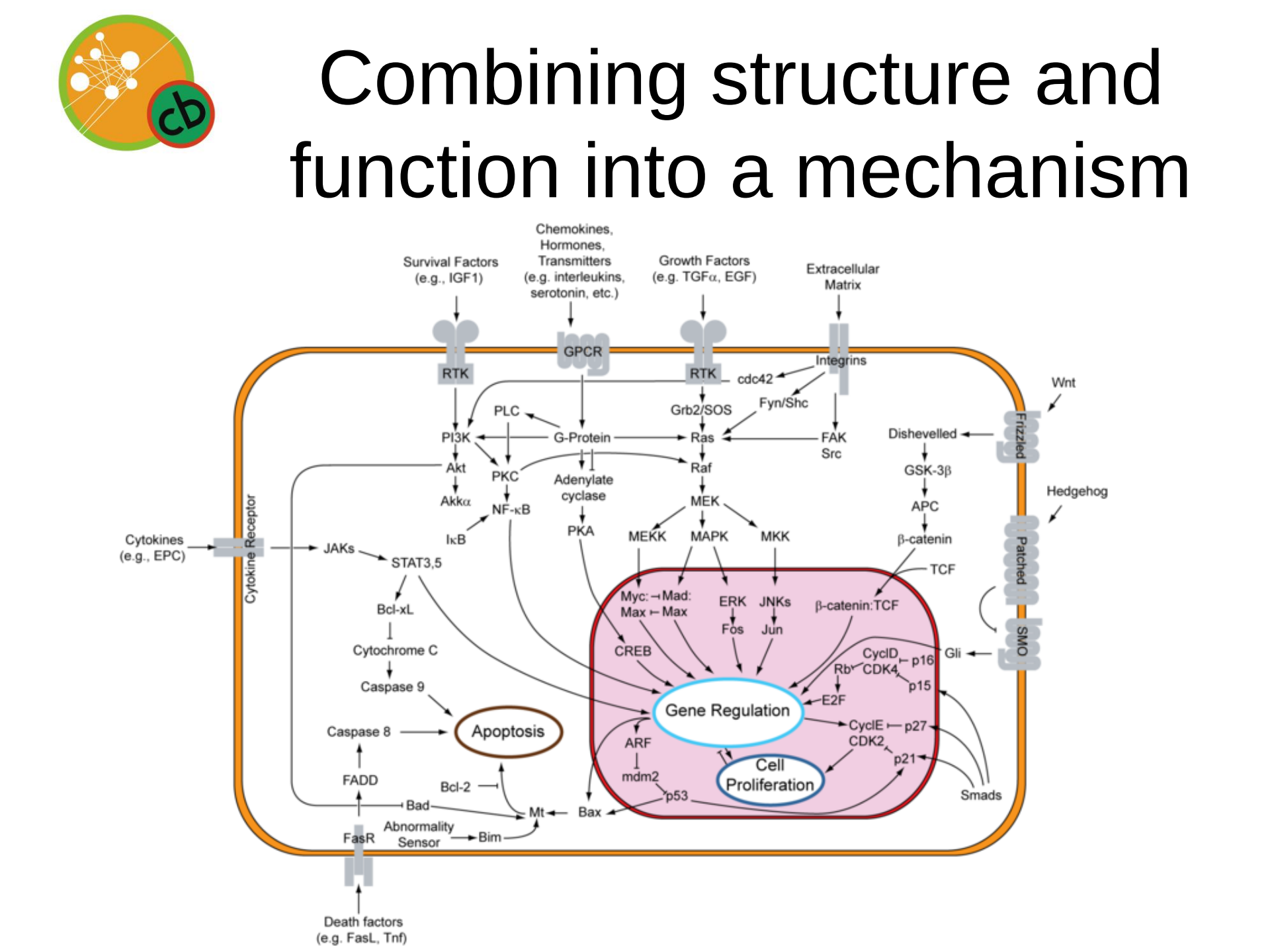


One Motivating Use Case

- Given a large set of genes (or the like) experimentally implicated in a phenomenon under study…
- Produce:
  - An explanation of the reasons that those genes are (or are not) relevant to the phenotype
  - Evidence to support the explanation(s)
  - Alternative explanations
  - Reasons to prefer one explanation over another

Biomedical Explanation

- Explanation is *creative thinking about causality*
- Biological explanations are *mechanistic,* involving
  - Structures (e.g. specific molecules, organs)
  - Processes (e.g. manage energy, synthesize biochemicals, sense the environment)
  - Evolution (e.g. selection, common origins, adaptation)
- Explanations are combinatoric
  - Complex interactions among many components
    - Physical interactions
    - Multi-layered regulation of production & activity
    - Signaling & responsiveness to stimuli
  - Multi-scale dynamics through time and space

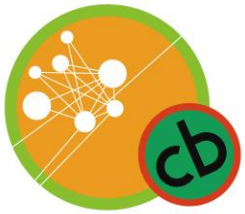


Combining structure and function into a mechanism

Abduction, and AI

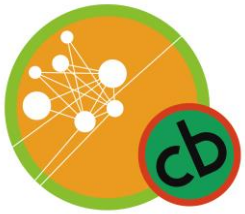
- Charles Sanders Pierce,1931:
  “However man may have acquired his faculty of divining the ways of Nature, it has certainly not been by a self-controlled and critical logic. Even now he cannot give any exact reason for his best guesses…. For though it goes wrong oftener than right, yet the relative frequency with which it is right is on the whole the most wonderful thing in our constitution.”

- Judea Pearl, 1984:
  “The ability to interpret and generate such explanatory sentences, or to select the expression most appropriate for the context, is one of the most intriguing challenges of research in man-machine conversation.”



Why abductive AI has failed (so far)

- Causal knowledge is highly interdependent
- Not just about the connection between an explanation and the thing explained, but must also be “consonant” with other explanations.
  - “Complete enough” knowledge is key
  - Have to know many other explanations.
- Need “judgment” to compare the qualities of alternative explanations.

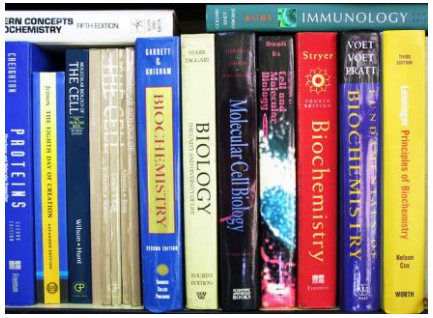


Interestingness functions

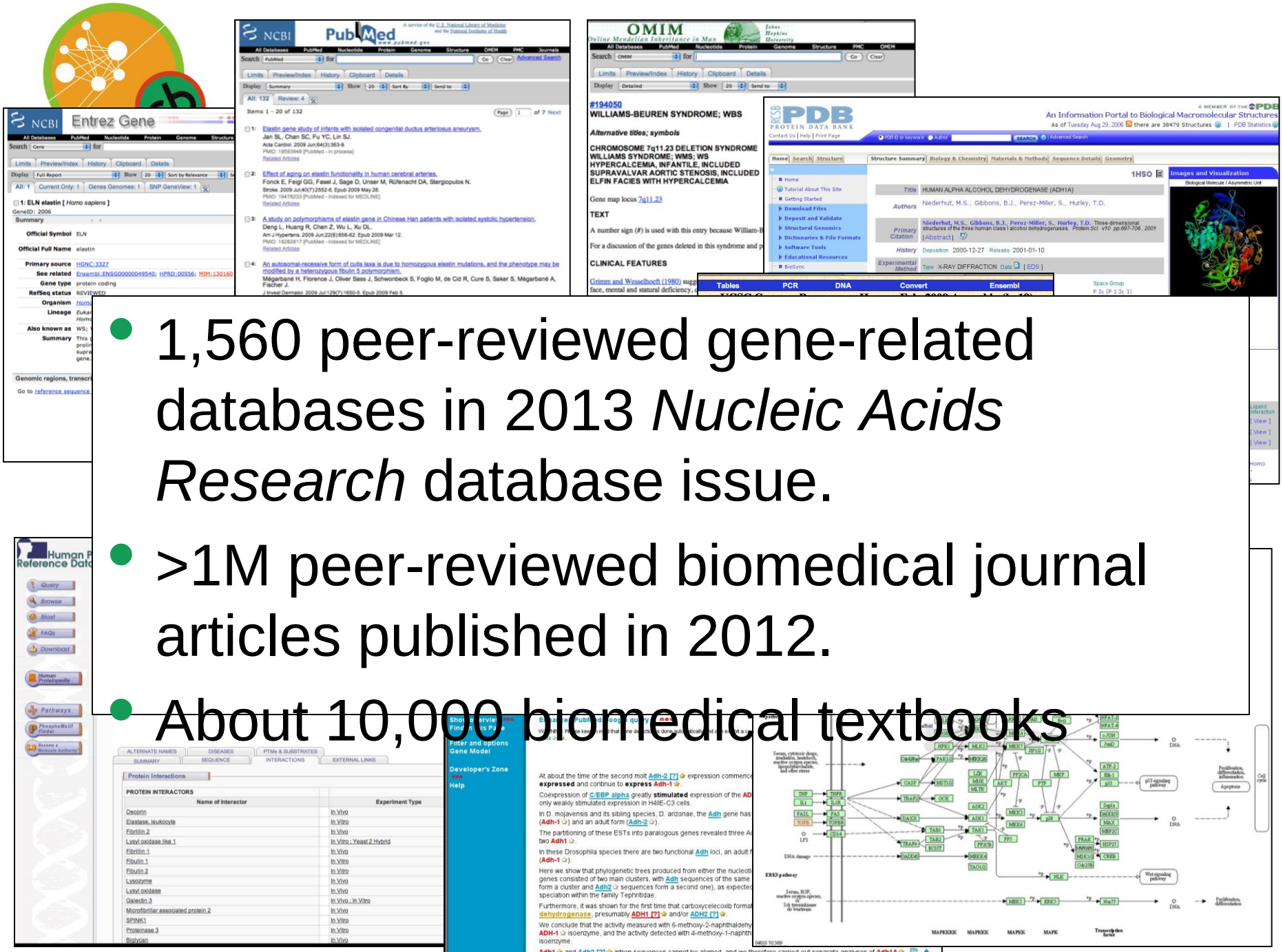
- Interestingness is a key judgment about an explanation (likelihood is the other)
- Virtuous cycle:
  - Judgments about explanations get better explanations
  - Explanations of judgments lead to better judgments
- The features of interestingness in our use case:
  - Open questions, state of the field, relationships to hypothesis that generated the data, background of the analyst, stage of analysis.
- Structure-based interestingness (a la Lenat 1980)

People don't have implicit knowledge of molecular biology

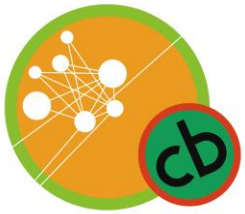
- Everything anyone knows about MolBio comes from some combination of:
  - Textbooks
  - Scientific publications
  - Databases (e.g. NCBI)



- There is no *elicitation barrier* to capturing <u>everything</u> known about molecular biology
- Formal representation of a biology textbook: http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html



- 1,560 peer-reviewed gene-related databases in 2013 *Nucleic Acids Research* database issue.
- >1M peer-reviewed biomedical journal articles published in 2012.
- About 10,000 biomedical textbooks

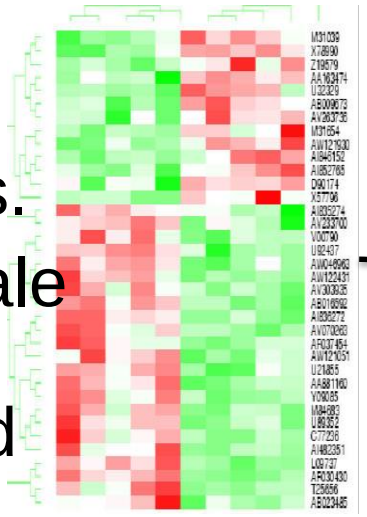


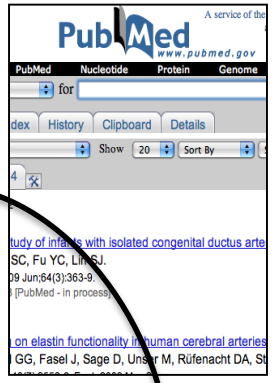
Practical advantages of a Life Mind AI research agenda

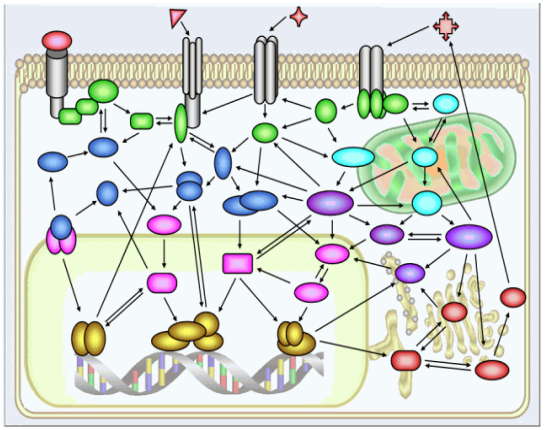
- User community desperate for help
  - Clear criteria for success (publication)
- Community-curated ontologies developed and used by molecular biologists (GO, BFO, etc.)
  - Fiducial, in that arguments among experts about meaning have been resolved.
  - OWL, but not yet much inference
- Biomedical language amenable to NLP
  - BioCreative, TREC genomics, etc. evaluations

Hanalyzer PoC

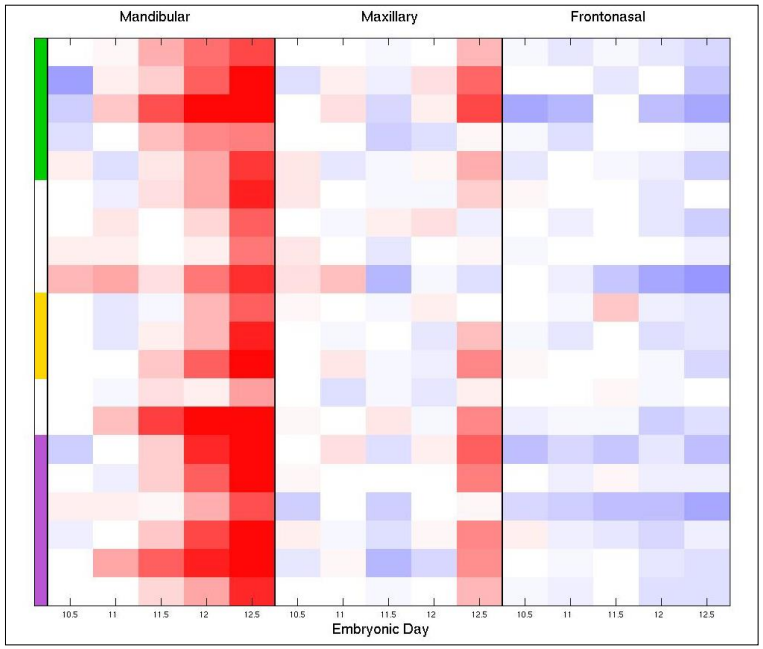
- Goal: Bring broad knowledge of molecular biology to bear on analyzing genome-scale datasets.
- Uses graphs to align genome-scale experimental results with knowledge about genes extracted from many databases (and, increasingly, directly from the biomedical literature).
- [Leach, et al., PLoS Comp Bio 2009]
  http://hanalyzer.sourceforge.org
  Search YouTube, for “Hanalyzer”

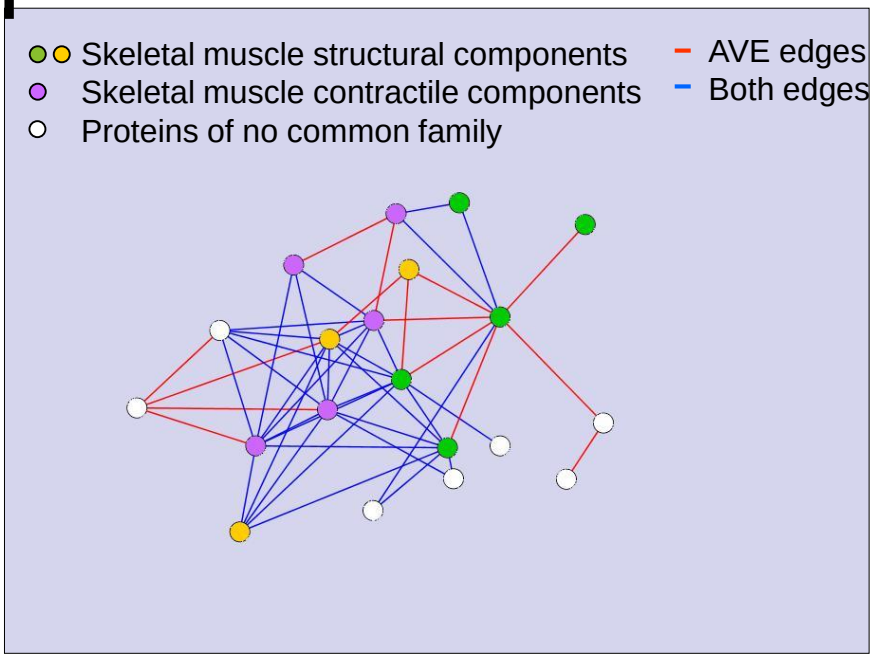






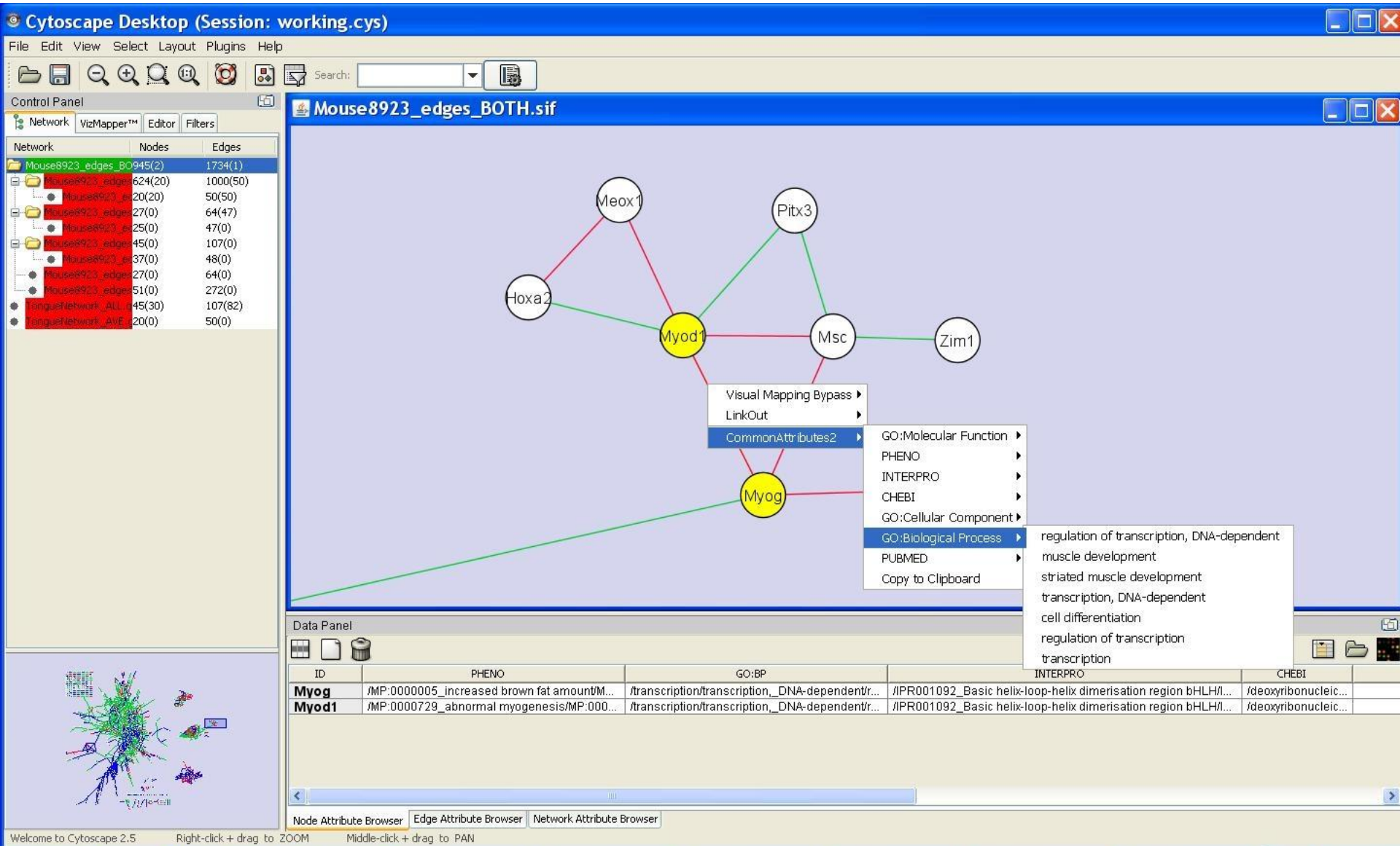
Strong data and background knowledge facilitate explanations



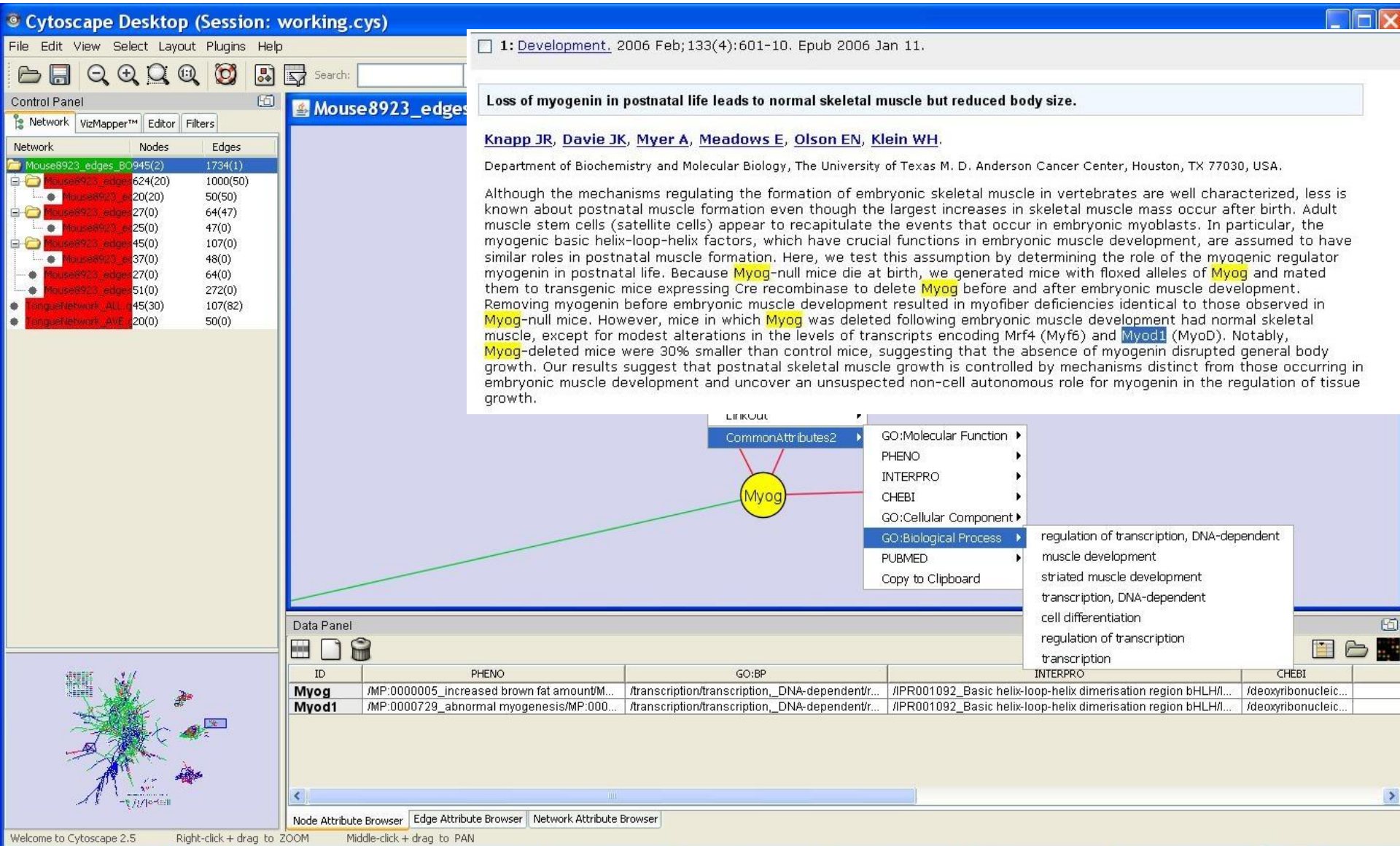


- Goal is abductive inference: why are these genes doing this?
  - Specifically, why the increase in mandible before the increase in maxilla, and not at all in the frontonasal prominence?

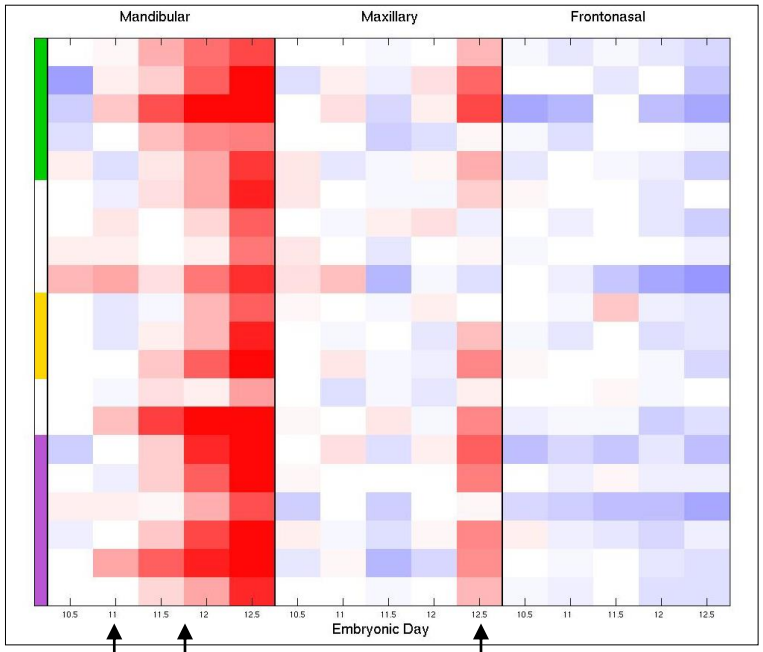
Exploring the knowledge network

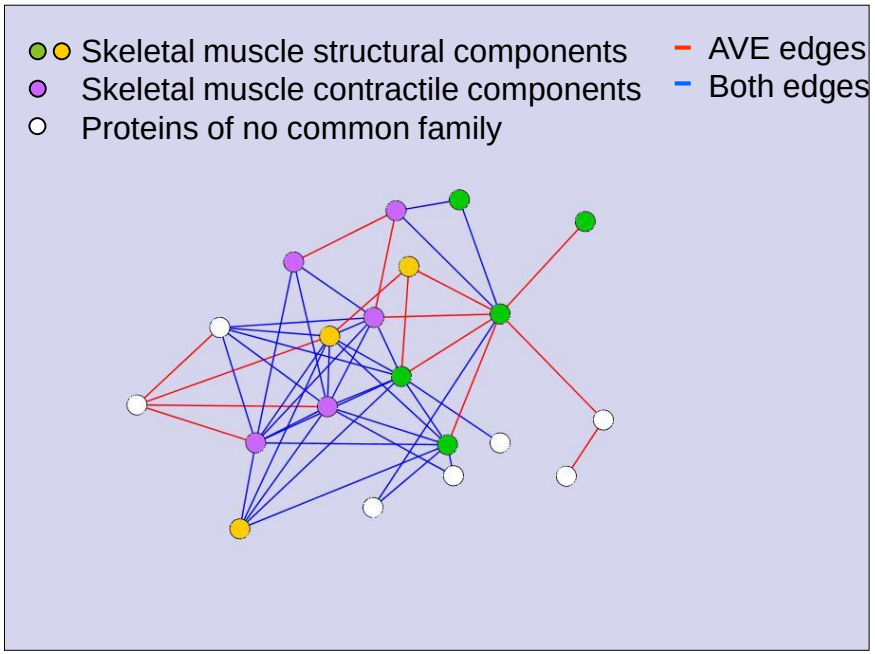


Exploring the knowledge network



Scientist + aide + literature → explanation: tongue development



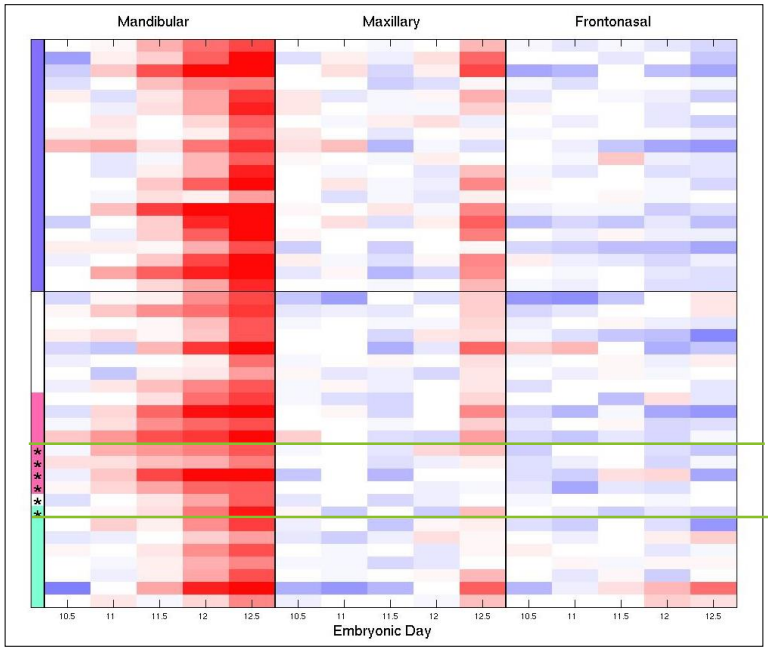


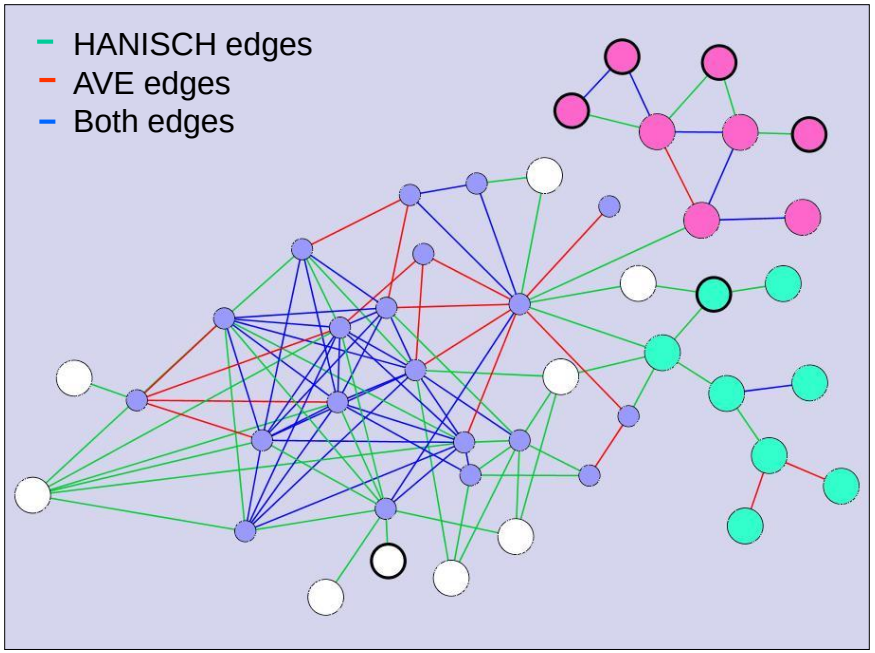
The delayed onset, at E12.5, of the same group of proteins during mastication muscle development.

Myoblast differentiation and proliferation continues until E15 at which point the tongue muscle is completely formed.

Myogenic cells invade the tongue primodia ~E11

On to Discovery



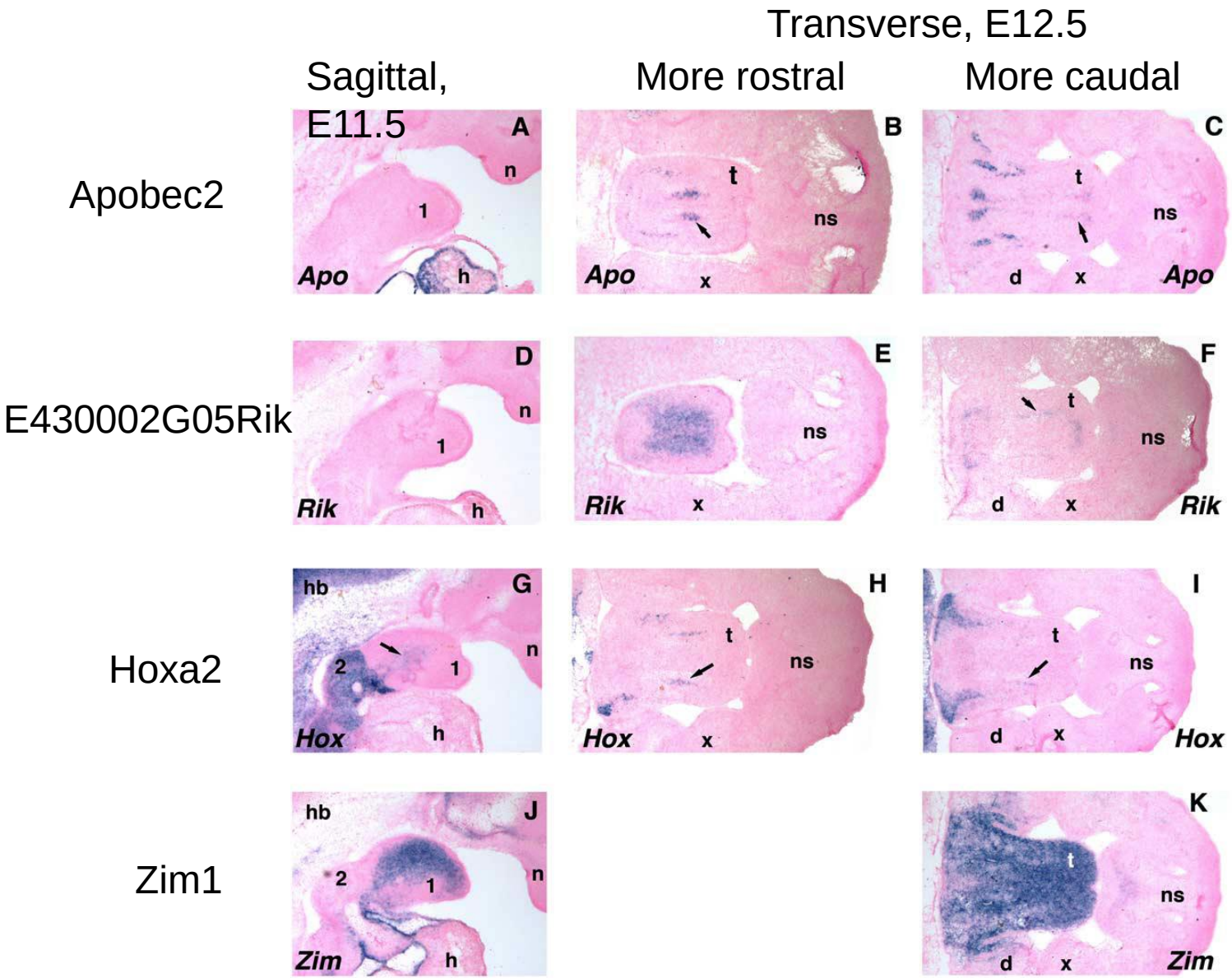


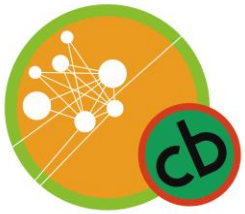
- inferred synapse signaling proteins
- Inferred myogenic proteins

- Proteins of no common family
- Proteins in the previous AVE based sub-network

- Add the strong data, weak background knowledge (Hanisch) edges to the previous network, bringing in new genes.
- Four of these genes not previously implicated in facial muscle development (1 almost completely unannotated)

Biological validation



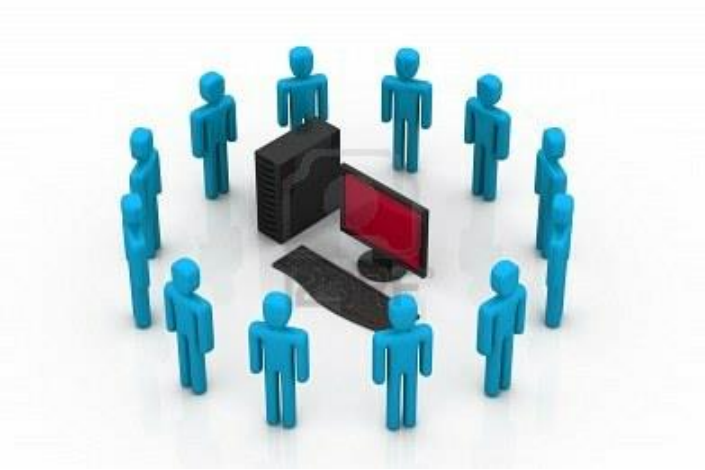


Central hypothesis

- Main challenges for building an artificial mind:
  - *Explanation*: Developing an all-encompassing (or nearly so) characterization of causation
    - Prospective (use for selecting actions)
    - Internally consistent, defines “surprise”, in terms of causes and intentions
  - *Judgment*: Comparing any two states of the world, determining a goal- (or value-) based preference
    - Watson’s most significant contribution
    - Analogous “state mapping” from Kahneman, Lakoff

Social test for mind

- Extended, collaborative relationships between people and a program provide evidence regarding its ability to think.



- Evaluation criteria:
  - Judgments of people interacting with the program long term
  - Do ongoing interactions with a program generate significant new insights, explanations, hypotheses?
  - Are the program's contributions original, interesting or surprising?

Want to take this on?

- Lots of opportunity:
  - NIH “Big Data to Knowledge”
  - NSF “Discovery Informatics”
- Learn some biology
- Contact me:

Larry.Hunter@ucdenver.edu
@ProfLHunter