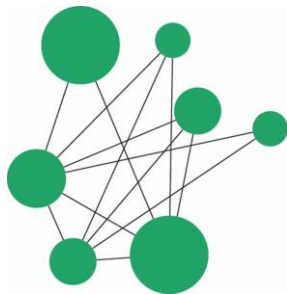
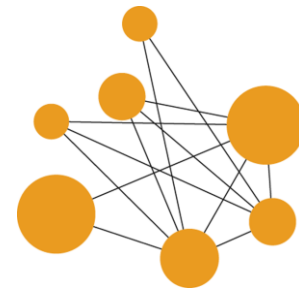




# Building a Mind for Life

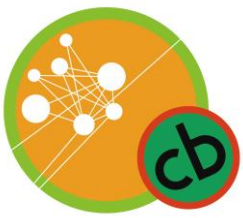
Lawrence Hunter, Ph.D.  
Director, Computational Bioscience Program  
University of Colorado School of Medicine



[Larry.Hunter@ucdenver.edu](mailto:Larry.Hunter@ucdenver.edu)

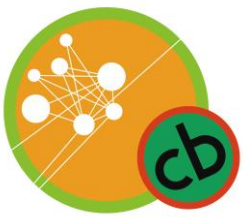
<http://compbio.ucdenver.edu/Hunter>

 @ProfLHunter



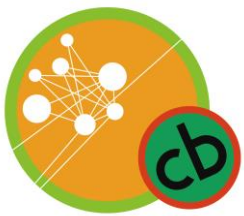
# Sometimes, two hard problems are easier to solve than one

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

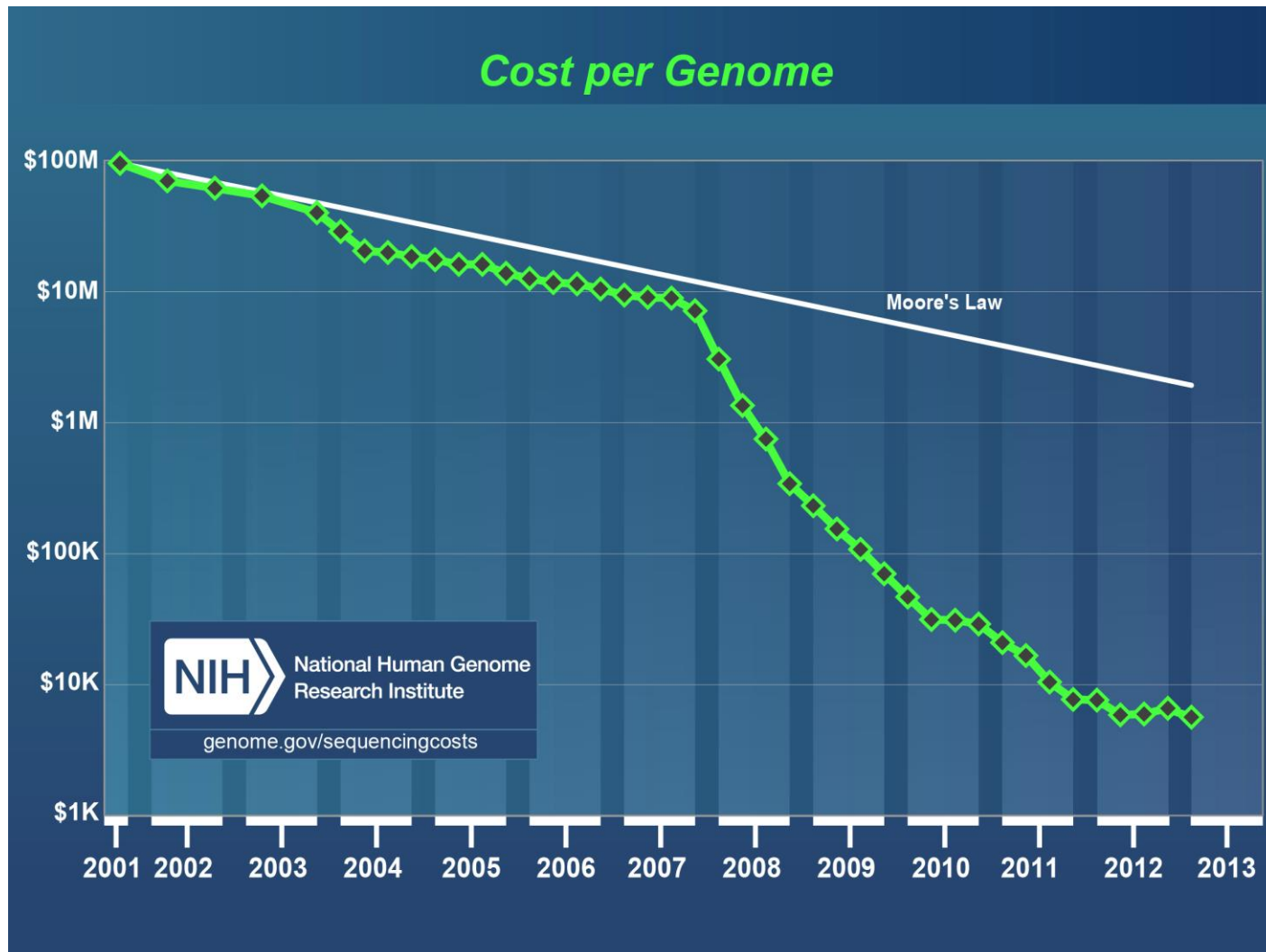


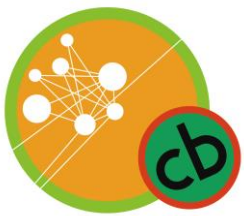
# The first artificial mind will think about molecular biology

- “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

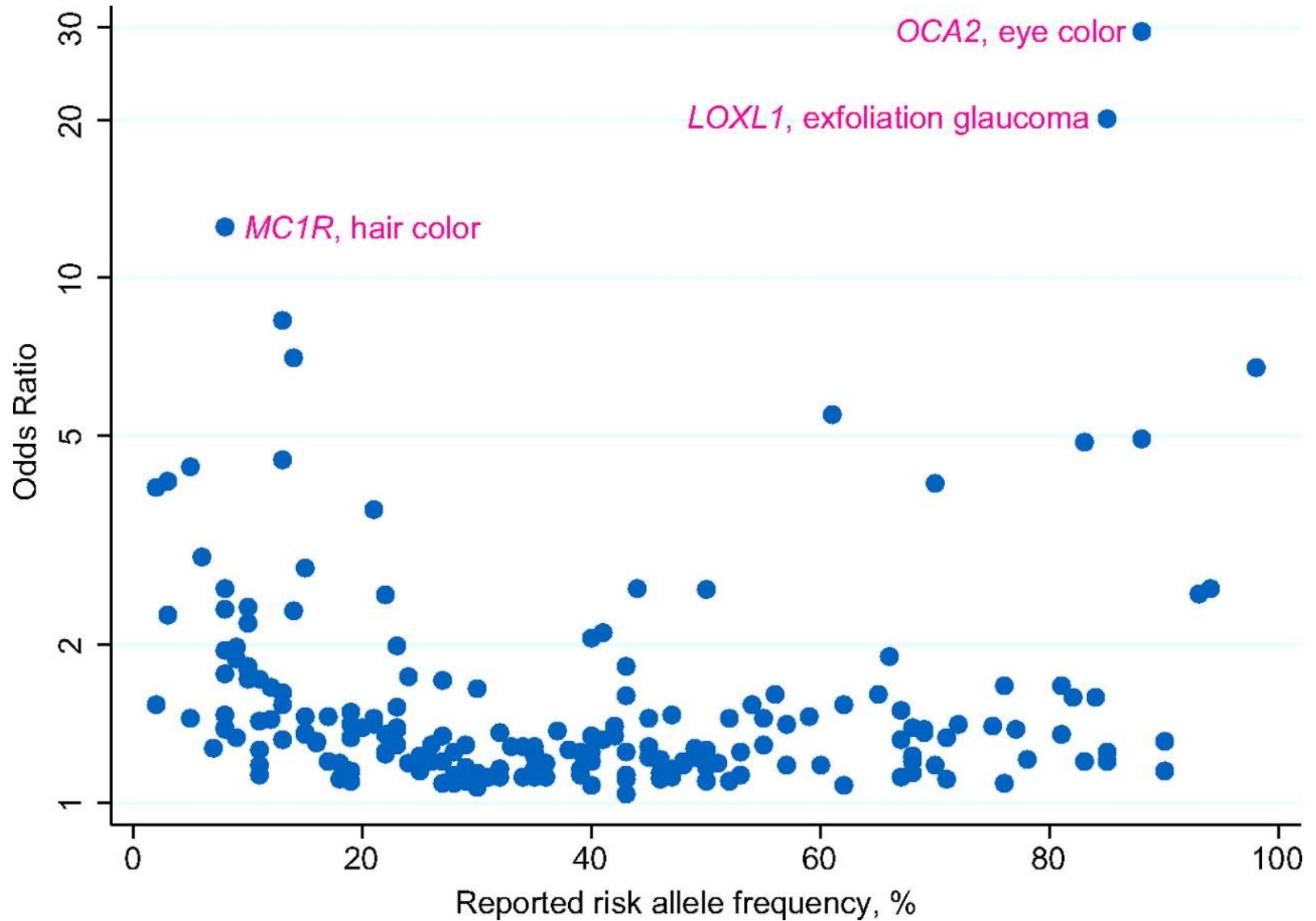


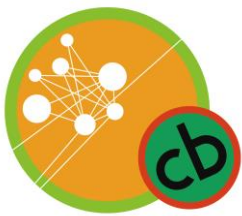
# The revolution in genomics





# *\$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  for

?

## Hypertension

- Related Diseases -

1537 genes have been reported with Hypertension

### Summary

Total Publications  
[3769](#)

Meta-Analyses  
[142](#)

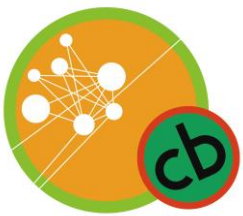
Genes  
1533

GWAS  
[76](#)

Group genes by [KEGG](#) | Display genes in [UCSC](#)

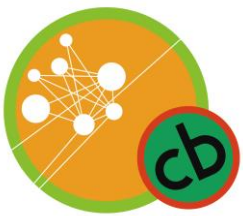
[Click  to re-sort the table]

Gene	Total	Meta	GWAS	Gene-Env	Trend
<a href="#">ACE</a> <small>SNP</small>	<a href="#">567</a>	<a href="#">18</a>	<a href="#">1</a>	<a href="#">134</a>	<input type="checkbox"/>
<a href="#">AGT</a> <small>SNP</small>	<a href="#">347</a>	<a href="#">15</a>	0	<a href="#">73</a>	<input type="checkbox"/>
<a href="#">NOS3</a> <small>SNP</small>	<a href="#">255</a>	<a href="#">9</a>	0	<a href="#">52</a>	<input type="checkbox"/>
<a href="#">AGTR1</a> <small>SNP</small>	<a href="#">236</a>	<a href="#">5</a>	0	<a href="#">61</a>	<input type="checkbox"/>
<a href="#">MTHFR</a> <small>SNP</small>	<a href="#">169</a>	<a href="#">13</a>	<a href="#">1</a>	<a href="#">51</a>	<input type="checkbox"/>
<a href="#">CYP11B2</a> <small>SNP</small>	<a href="#">156</a>	<a href="#">7</a>	0	<a href="#">28</a>	<input type="checkbox"/>
<a href="#">APOE</a> <small>SNP</small>	<a href="#">152</a>	<a href="#">5</a>	<a href="#">1</a>	<a href="#">30</a>	<input type="checkbox"/>
<a href="#">GAB2</a> <small>SNP</small>	<a href="#">141</a>	<a href="#">6</a>	0	<a href="#">29</a>	<input type="checkbox"/>



# Analysis is the hard part

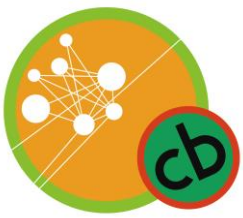
- “We are close to having a \$1,000 genome sequence, but this may be accompanied by a \$1,000,000 interpretation.”
  - Bruce Korf, president American College of Medical Genetics
- Not only is the cost of sequencing essentially free, but big computers and big storage are cheap, too. What will keep us busy for the next 50 years is understanding the data”
  - Russ Altman, chair of Biomedical Engineering at



# One Motivating Use Case

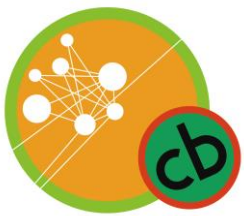
- Given a large set of genes (or the like) experimentally implicated in a phenomenon under study...
- Produce:
  - An explanation of the reasons that those genes are (or are not) relevant to the phenotype
  - Evidence to support the explanation(s)
  - Alternative explanations
  - Reasons to prefer one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



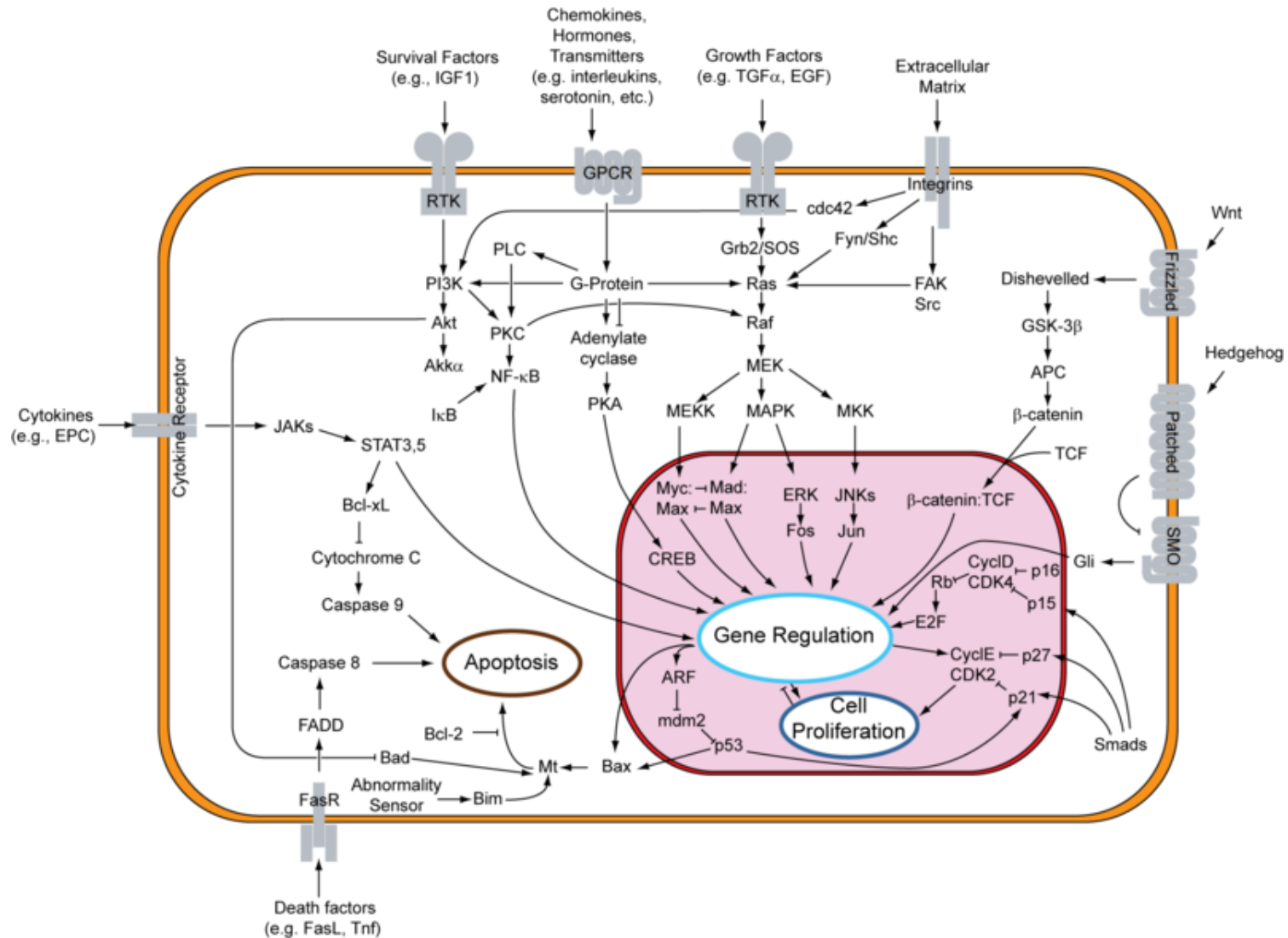


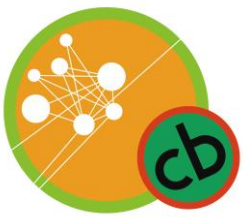
# Biomedical Explanation

- 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

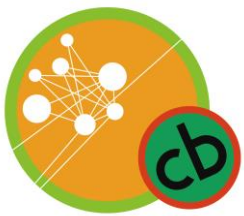




# Abduction, and AI

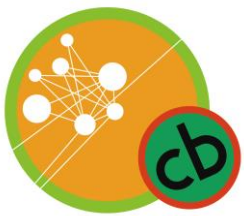
- Charles Sanders Peirce, 1931:  
“However man may have acquired his faculty of divining the ways of Nature, it has certainly not been by a self-controlled and critical logic. Even now he cannot give any exact reason for his best guesses.... For though it goes wrong oftener than right, yet the relative 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.”





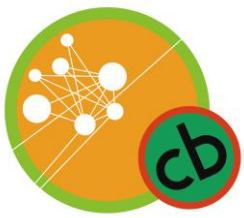
# Why abductive AI has failed (so far)

- 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

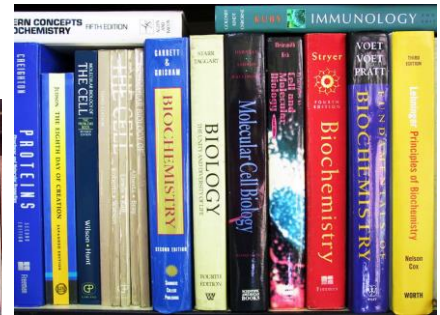
- Interestingness is a key judgment about an explanation (likelihood is the other)
- Virtuous cycle:
  - Judgments about explanations get better explanations
  - Explanations of judgments lead to better judgments
- The features of interestingness in our use case:
  - Open questions, state of the field, 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 everything known about molecular biology



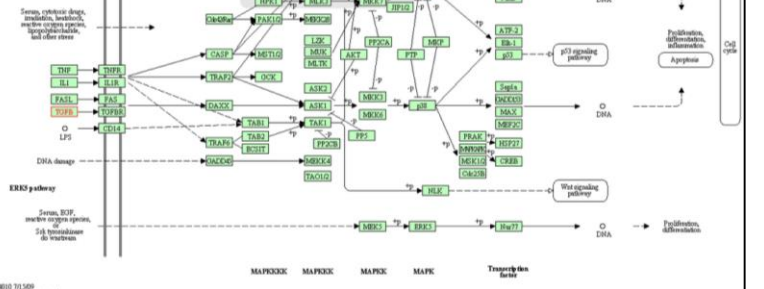
- Formal representation of a biology textbook:  
<http://www.ai.sri.com/halo/halobook2010/exported-kb/biokb.html>

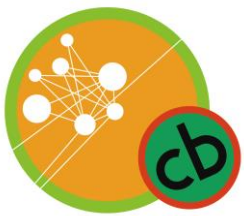


- 1,560 peer-reviewed gene-related databases in 2013 *Nucleic Acids Research* database issue.
- >1M peer-reviewed biomedical journal articles published in 2012.
- About 10,000 biomedical textbooks

Protein Interactors	Name of Interactor	Experiment Type
Decorin	In Vivo	In Vivo
Elastase, leukocyte	In Vivo	In Vivo
Fibrillin 2	In Vivo	In Vivo
Lysyl oxidase like 1	In Vivo - Yeast 2 Hybrid	In Vivo
Fibrillin 1	In Vivo	In Vivo
Fibrulin 1	In Vivo	In Vivo
Fibrulin 2	In Vivo	In Vivo
Lysyl oxidase	In Vivo	In Vivo
Glaucosyl 3	In Vivo - In Vivo	In Vivo
Microfibrillar associated protein 2	In Vivo	In Vivo
SPIN1	In Vivo	In Vivo
Proteinase 3	In Vivo	In Vivo
Biotinylase	In Vivo	In Vivo

At about the time of the second molt *Adh-2* [7] expression commences and continues to express *Adh-1*. Coexpression of *C-EBP alpha* greatly stimulates expression of the *AD* only weakly stimulated expression in H4IIE-C3 cells. In *D. melanogaster* and its sibling species, *D. obscura*, the *Adh* gene has a *Ddh-1* (-) and an adult form (*Adh-2* (-)). The partitioning of these ESTs into paralogous genes revealed there are two *Adh1*. In these *Drosophila* species there are two functional *Adh* loci, an adult (*Adh-1* (-)). Here we show that phylogenetic trees produced from either the nucleotide genes consisted of two main clusters, with *Adh* sequences of the same form a cluster, and *Adh2* (-) sequences form a second one), as expected speciation within the family Tephritidae. Furthermore, it was shown for the first time that carboxycolexib forms *dehydrogenase* actually *ADH1* [7] and/or *ADH2* [7]. We conclude that the activity measured with 6-methoxy-2-naphthaldehyde (*ADH1* (-) isozyme, and the activity detected with 4-methoxy-1-naphthyl isozyme. *Adh1* and *Adh2* [7] (-) intron sequences cannot be aligned, and we therefore carried out separate analyses of *ADH1A* and *ADH2* [7] genes using exon and intron sequences together.





# Practical advantages of a Life Mind AI research

## agenda

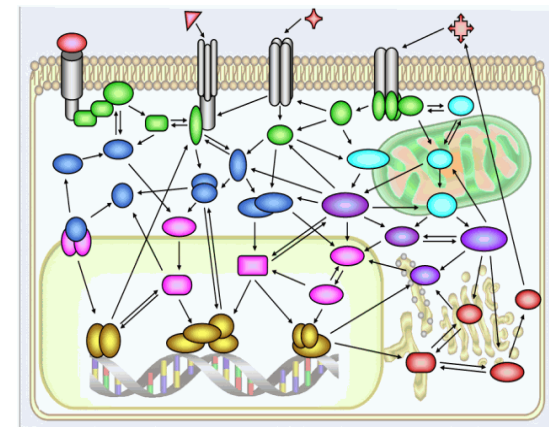
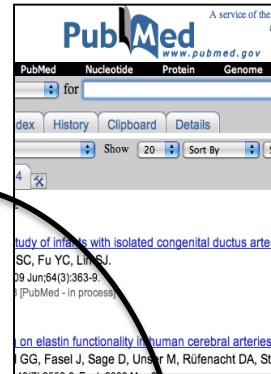
- User community desperate for help
  - Clear criteria for success (publication)
- Community-curated ontologies developed and used by molecular biologists (GO, BFO, etc.)
  - Fiducial, in that arguments among experts about meaning have been resolved.
  - OWL, but not yet much inference
- Biomedical language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

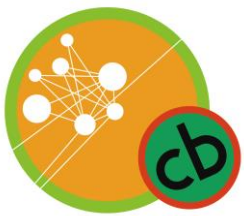




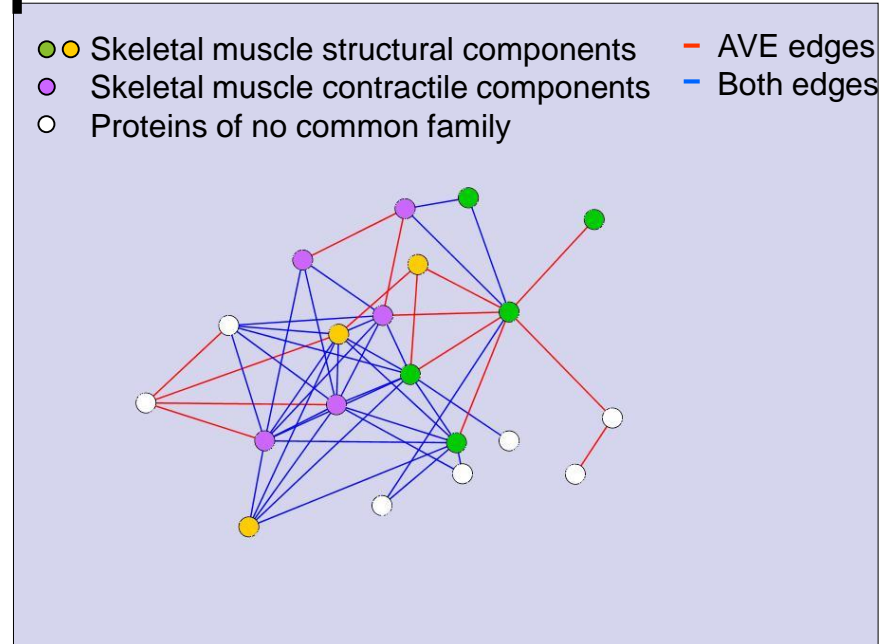
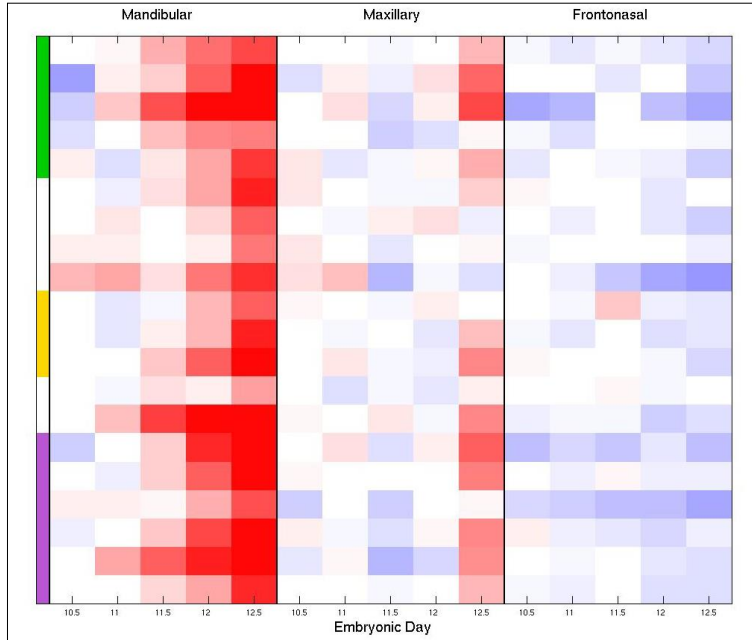
# Hanalyzer PoC

- **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”

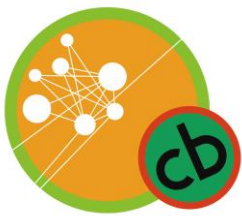




# Strong data and background knowledge facilitate explanations



- 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

**Cytoscape Desktop (Session: working.cys)**

File Edit View Select Layout Plugins Help

Search:

Control Panel

Network VizMapper™ Editor Filters

Network	Nodes	Edges
Mouse8923_edges_BO945(2)	1734(1)	
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	624(20)	1000(50)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	20(20)	50(50)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	27(0)	64(47)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	25(0)	47(0)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	45(0)	107(0)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	37(0)	48(0)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	27(0)	64(0)
Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2) > Mouse8923_edges_BO945(2)	51(0)	272(0)
SingleNetwork_ALL	45(30)	107(82)
SingleNetwork_AVE	20(0)	50(0)

Mouse8923\_edges\_BOTH.sif

```

graph TD
    Meox1 --- Myod1
    Pitx3 --- Myod1
    Pitx3 --- Msc
    Hoxa2 --- Myod1
    Myod1 --- Myog
    Myod1 --- Msc
    Msc --- Zim1
  
```

Visual Mapping Bypass ▶  
 LinkOut ▶  
 CommonAttributes2 ▶  
 GO:Molecular Function ▶  
 PHENO ▶  
 INTERPRO ▶  
 CHEBI ▶  
 GO:Cellular Component ▶  
 GO:Biological Process ▶  
 PUBMED ▶  
 Copy to Clipboard

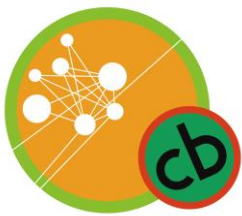
regulation of transcription, DNA-dependent  
 muscle development  
 striated muscle development  
 transcription, DNA-dependent  
 cell differentiation  
 regulation of transcription  
 transcription

Data Panel

ID	PHENO	GO:BP	INTERPRO	CHEBI
Myog	/MP:0000005_increased brown fat amount/M...	/transcription/transcription_DNA-dependent/r...	/IPR001092_Basic helix-loop-helix dimerisation region bHLH/...	/deoxyribonucleic...
Myod1	/MP:0000729_abnormal myogenesis/MP:000...	/transcription/transcription_DNA-dependent/r...	/IPR001092_Basic helix-loop-helix dimerisation region bHLH/...	/deoxyribonucleic...

Node Attribute Browser Edge Attribute Browser Network Attribute Browser

Welcome to Cytoscape 2.5 Right-click + drag to ZOOM Middle-click + drag to PAN



# Exploring the knowledge network

**Cytoscape Desktop (Session: working.cys)**

File Edit View Select Layout Plugins Help

Control Panel

Network VizMapper™ Editor Filters

Network Nodes Edges

Network	Nodes	Edges
Mouse8923_edges_BO945(2)	1734(1)	
Mouse8923_edges_624(20)	1000(50)	
Mouse8923_edges_20(20)	50(50)	
Mouse8923_edges_27(0)	64(47)	
Mouse8923_edges_25(0)	47(0)	
Mouse8923_edges_45(0)	107(0)	
Mouse8923_edges_37(0)	48(0)	
Mouse8923_edges_27(0)	64(0)	
Mouse8923_edges_51(0)	272(0)	
SingleNetwork_ALL_45(30)	107(82)	
SingleNetwork_AVE_20(0)	50(0)	

1: [Development](#), 2006 Feb;133(4):601-10. Epub 2006 Jan 11.

**Loss of myogenin in postnatal life leads to normal skeletal muscle but reduced body size.**

**Knapp JR, Davie JK, Myer A, Meadows E, Olson EN, Klein WH.**

Department of Biochemistry and Molecular Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

Although the mechanisms regulating the formation of embryonic skeletal muscle in vertebrates are well characterized, less is known about postnatal muscle formation even though the largest increases in skeletal muscle mass occur after birth. Adult muscle stem cells (satellite cells) appear to recapitulate the events that occur in embryonic myoblasts. In particular, the myogenic basic helix-loop-helix factors, which have crucial functions in embryonic muscle development, are assumed to have similar roles in postnatal muscle formation. Here, we test this assumption by determining the role of the myogenic regulator myogenin in postnatal life. Because **Myog**-null mice die at birth, we generated mice with floxed alleles of **Myog** and mated them to transgenic mice expressing Cre recombinase to delete **Myog** before and after embryonic muscle development. Removing myogenin before embryonic muscle development resulted in myofiber deficiencies identical to those observed in **Myog**-null mice. However, mice in which **Myog** was deleted following embryonic muscle development had normal skeletal muscle, except for modest alterations in the levels of transcripts encoding Mrf4 (Myf6) and **Myod1** (MyoD). Notably, **Myog**-deleted mice were 30% smaller than control mice, suggesting that the absence of myogenin disrupted general body growth. Our results suggest that postnatal skeletal muscle growth is controlled by mechanisms distinct from those occurring in embryonic muscle development and uncover an unsuspected non-cell autonomous role for myogenin in the regulation of tissue growth.

LINKOUT

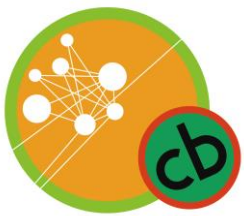
- CommonAttributes2
- GO:Molecular Function
- PHENO
- INTERPRO
- CHEBI
- GO:Cellular Component
- GO:Biological Process
  - regulation of transcription, DNA-dependent
  - muscle development
  - striated muscle development
  - transcription, DNA-dependent
  - cell differentiation
  - regulation of transcription
  - transcription
- PUBMED
- Copy to Clipboard

Data Panel

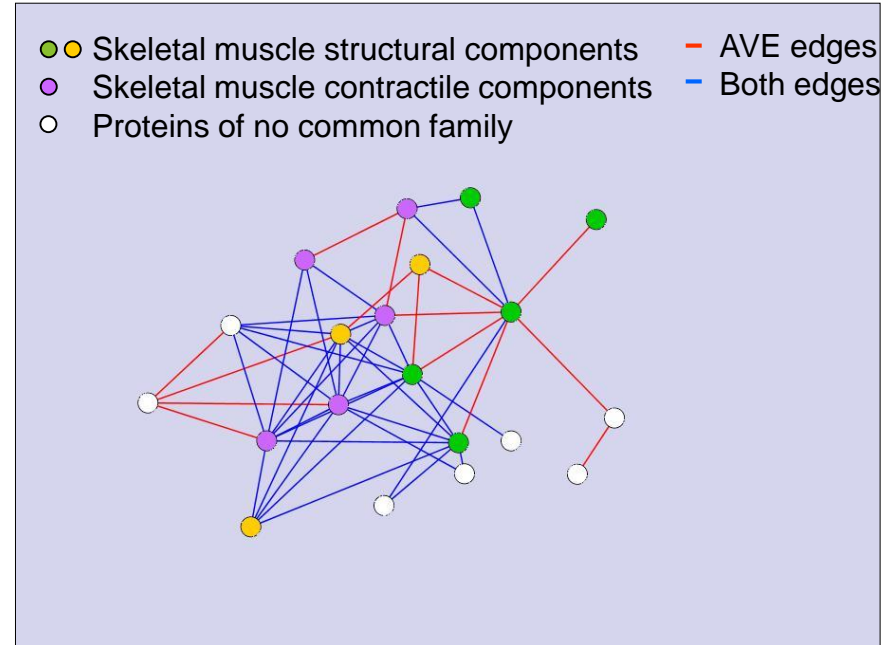
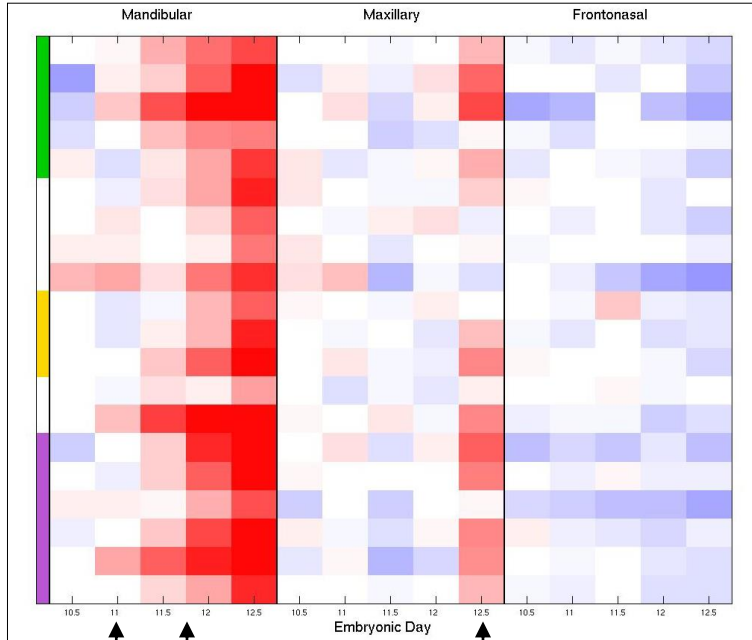
ID	PHENO	GO:BP	INTERPRO	CHEBI
<b>Myog</b>	/MP:0000005_increased brown fat amount/M...	/transcription/transcription,_DNA-dependent/r...	/IPR001092_Basic helix-loop-helix dimerisation region bHLH/...	/deoxyribonucleic...
<b>Myod1</b>	/MP:0000729_abnormal myogenesis/MP:000...	/transcription/transcription,_DNA-dependent/r...	/IPR001092_Basic helix-loop-helix dimerisation region bHLH/...	/deoxyribonucleic...

Node Attribute Browser Edge Attribute Browser Network Attribute Browser

Welcome to Cytoscape 2.5 Right-click + drag to ZOOM Middle-click + drag to PAN



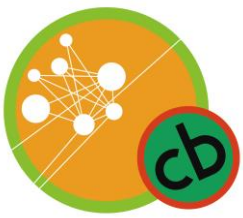
# Scientist + aide + literature → explanation: tongue development



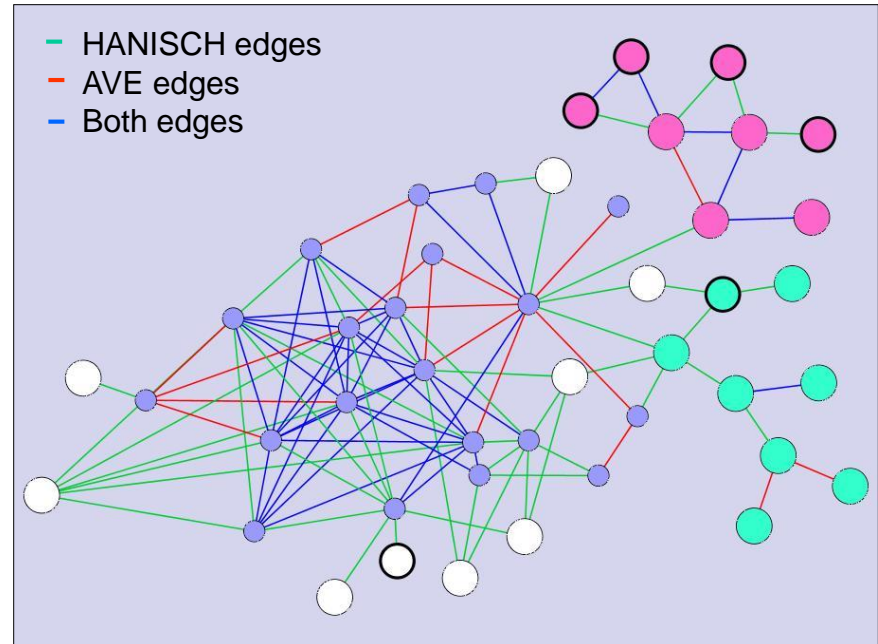
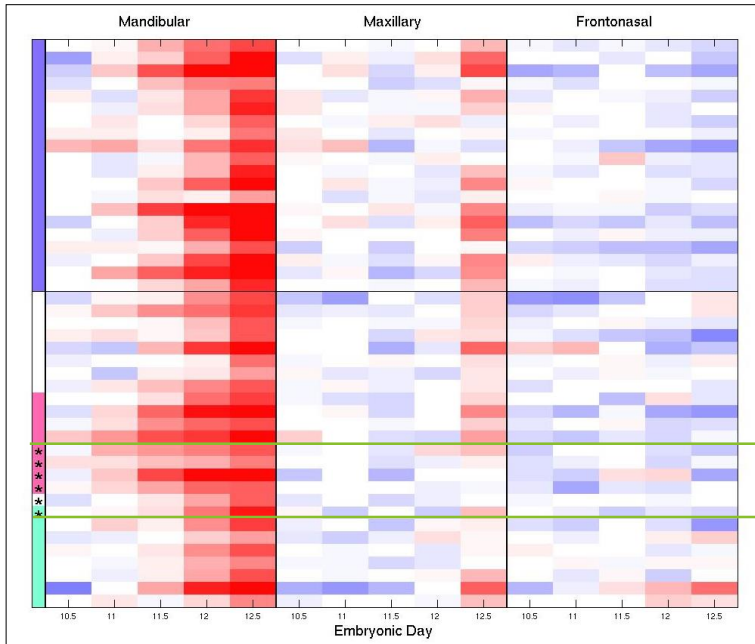
The delayed onset, at E12.5, of the same group of proteins during mastication muscle development.

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

Myogenic cells invade the tongue primordia ~E11



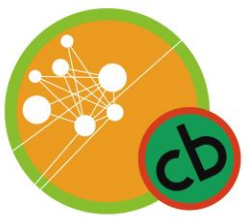
# On to Discovery



- 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

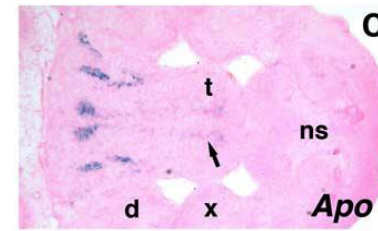
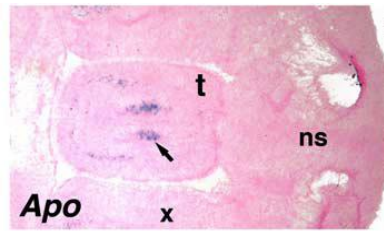
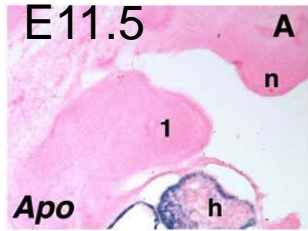
Transverse, E12.5

Sagittal,  
E11.5

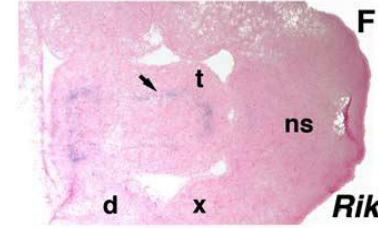
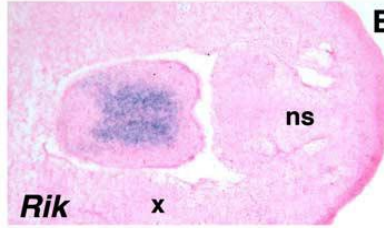
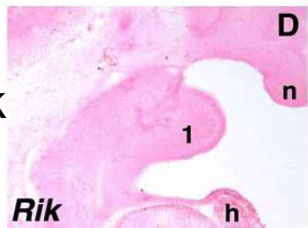
More rostral

More caudal

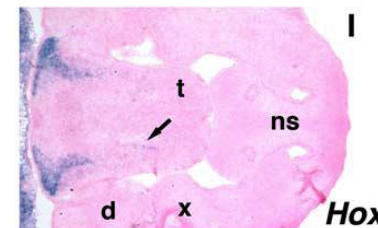
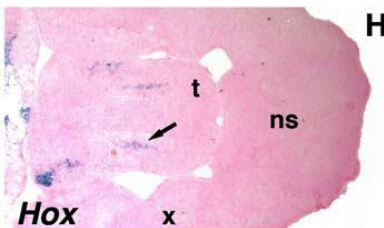
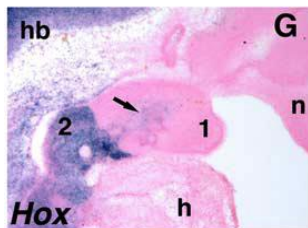
*Apobec2*



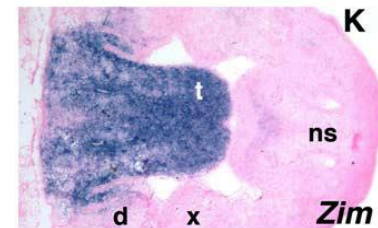
*E430002G05Rik*

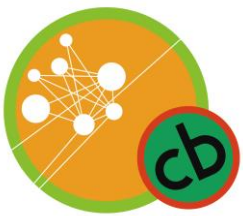


*Hoxa2*



*Zim1*





# 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, Tversky

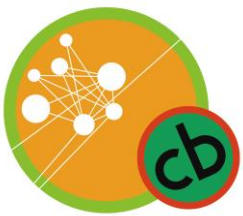




# Social test for mind

- 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](mailto:Larry.Hunter@ucdenver.edu)  
[@ProfLHunter](#)

