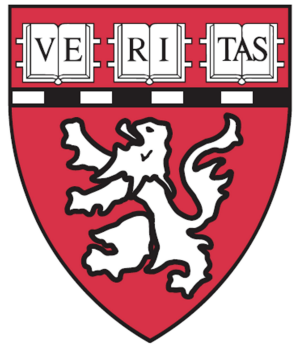
Conservation of tissue-specific gene expression predicts transcription factor binding sites in human and mouse

Martin Hemberg

Children's Hospital Boston, Harvard Medical School
Department of Ophthalmology
Swartz Center for Theoretical Neuroscience

Kreiman Lab

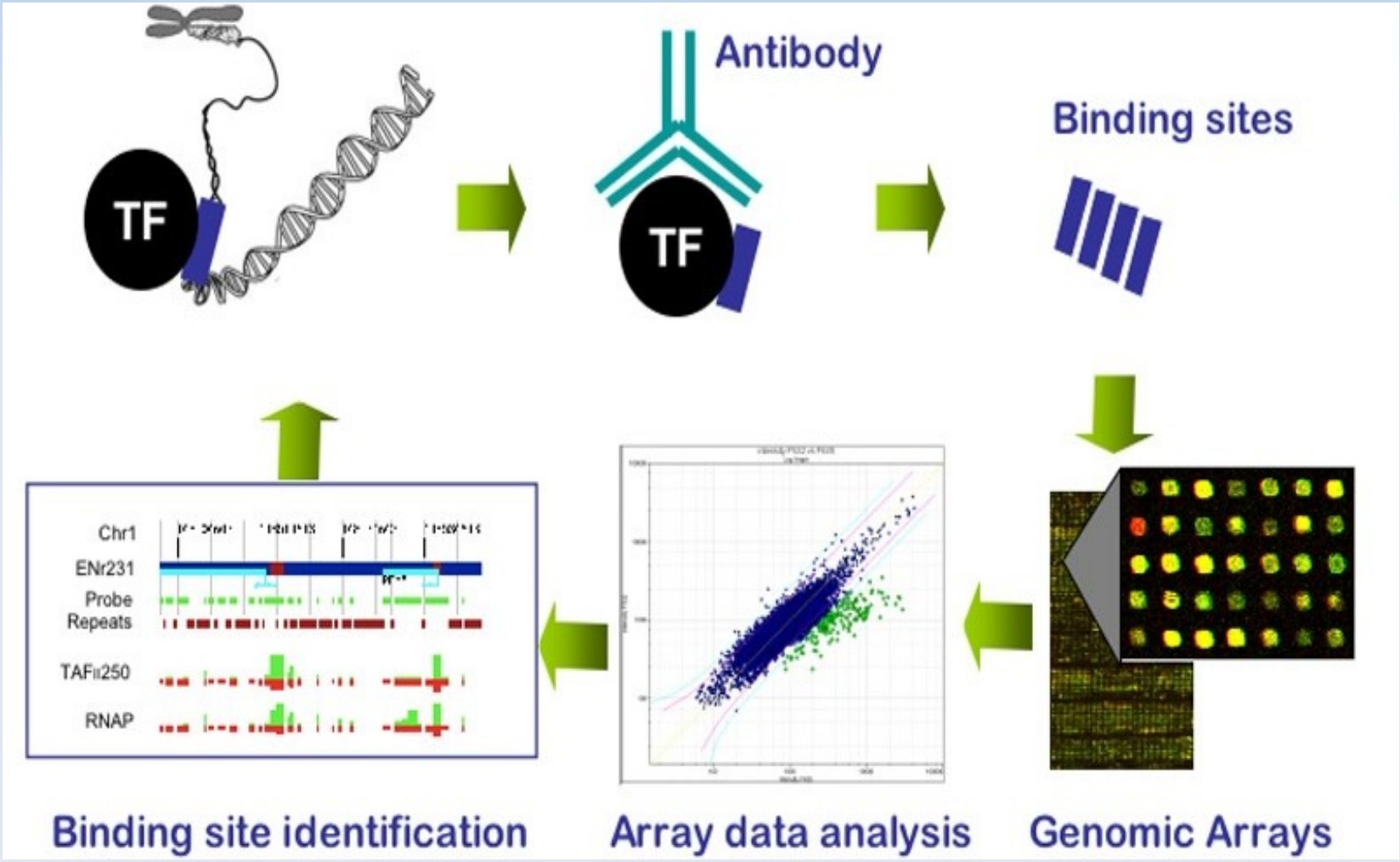


A talk of two parts

- Using gene expression to predict conservation of transcription factor binding sites (TFBSs)
- Using conservation of TFBSs to predict gene expression

Genome wide TFBSs

- ChIP-chip, ChIP-Seq experiments



(www.chiponchip.org)

Genome wide TFBSs

- ChIP-chip, ChIP-Seq experiments
  - Widespread binding of TFs

The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells

Yuin-Han Loh[1,2,7], Qiang Wu[1,7], Joon-Lin Chew[1,2,7], Vinsensius B Vega[3], Weiwei Zhang[1,2], Xi Chen[1,2], Guillaume Bourque[3], Joshy George[3], Bernard Leong[3], Jun Liu[4], Kee-Yew Wong[5], Ken W Sung[3], Charlie W H Lee[3], Xiao-Dong Zhao[4], Kuo-Ping Chiu[3], Leonard Lipovich[3], Vladimir A Kuznetsov[3], Paul Robson[2,5], Lawrence W Stanton[5], Chia-Lin Wei[4], Yijun Ruan[4], Bing Lim[5,6] & Huck-Hui Ng[1,2]

Cell, Vol. 122, 947–956, September 23, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.cell.2005.08.020

Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells

Genome wide TFBSs

- ChIP-chip, ChIP-Seq experiments
  - Widespread binding of TFs
    - Functional significance of TFBSs?

The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells

Yuin-Han Loh[1,2,7], Qiang Wu[1,7], Joon-Lin Chew[1,2,7], Vinsensius B Vega[3], Weiwei Zhang[1,2], Xi Chen[1,2], Guillaume Bourque[3], Joshy George[3], Bernard Leong[3], Jun Liu[4], Kee-Yew Wong[5], Ken W Sung[3], Charlie W H Lee[3], Xiao-Dong Zhao[4], Kuo-Ping Chiu[3], Leonard Lipovich[3], Vladimir A Kuznetsov[3], Paul Robson[2,5], Lawrence W Stanton[5], Chia-Lin Wei[4], Yijun Ruan[4], Bing Lim[5,6] & Huck-Hui Ng[1,2]

Extensive low-affinity transcriptional interactions in the yeast genome

Amos Tanay

Cell, Vol. 122, 947–956, September 23, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.cell.2005.08.020

Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells

Genome wide TFBSs

- ChIP-chip, ChIP-Seq experiments
  - Widespread binding of TFs
    - Functional significance of TFBSs?
  - Low degree of conservation between species
    - Regulation and function of a gene decoupled?

Divergence of Transcription Factor Binding Sites Across Related Yeast Species

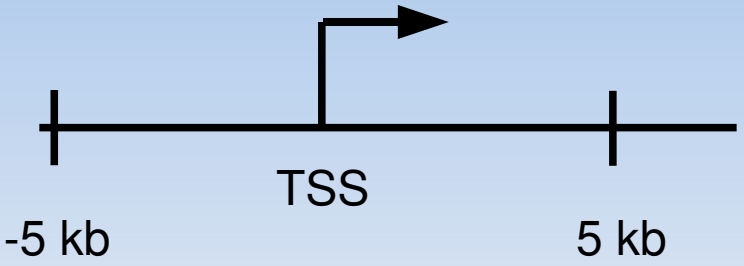
Anthony R. Borneman,[1*] Tara A. Gianoulis,[2] Zhengdong D. Zhang,[3] Haiyuan Yu,[3] Joel Rozowsky,[3] Michael R. Seringhaus,[3] Lu Yong Wang,[4] Mark Gerstein,[2,3,5] Michael Snyder[1,2,3†]

The Evolution of Combinatorial Gene Regulation in Fungi

Brian B. Tuch[1,2☯], David J. Galgoczy[1,2☯], Aaron D. Hernday[1,2], Hao Li[1*], Alexander D. Johnson[1,2*]

Conservation of TFBSs in liver

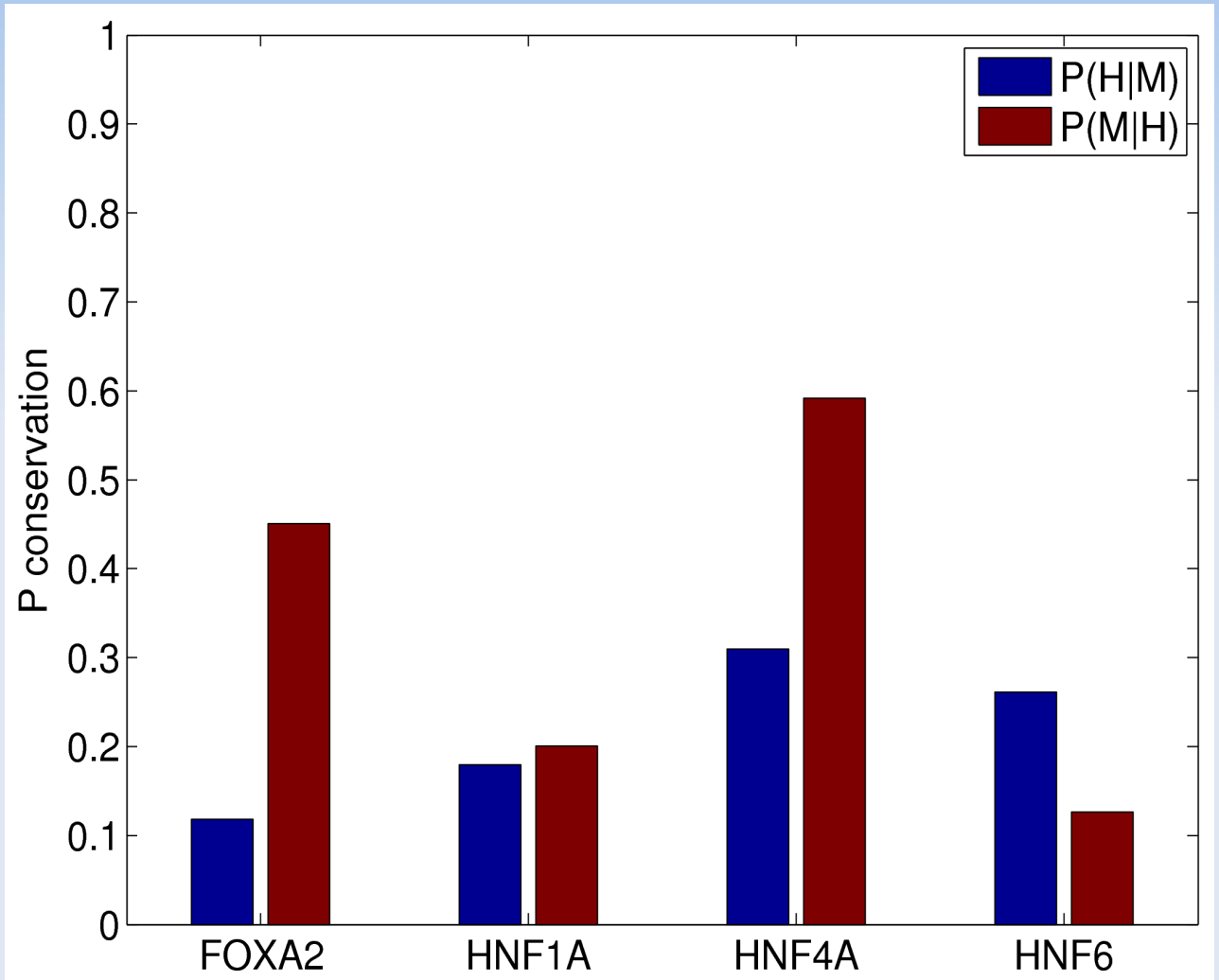
- Liver
  - Expression profiles similar
  - Homogeneuos cell types
- Compare TF binding sites
  - ChIP-chip
  - **4022** homologous genes

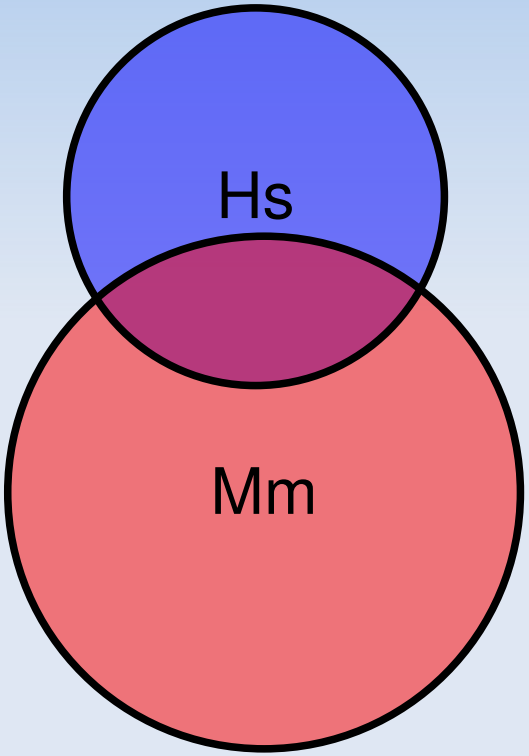


Tissue-specific transcriptional regulation has diverged significantly between human and mouse

Duncan T Odom[1,5,6], Robin D Dowell[2,6], Elizabeth S Jacobsen[1], William Gordon[3], Timothy W Danford[2], Kenzie D MacIsaac[4], P Alexander Rolfe[2], Caitlin M Conboy[1,5], David K Gifford[1,2] & Ernest Fraenkel[2,3]

Poor Conservation of TFBSs





P(H|M) = *probability of observing TFBS in Hs if TFBS in Mm*

= *#conserved TFBS / #Mm TFBS*

Higher Conservation Expected for Genes Expressed in Both Species

- Functional aspects not considered
- Conservation of TFBS for genes not expressed in tissue?
  - Binding **not** expected for genes **not** expressed
  - These TFs are known to be activators in liver

Higher Conservation Expected for Genes Expressed in Both Species

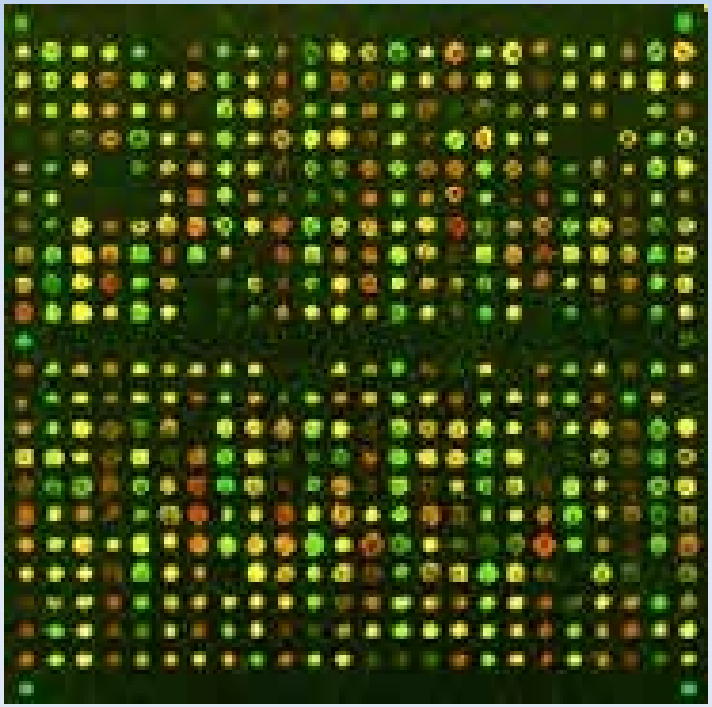
- Functional aspects not considered
- Conservation of TFBS for genes not expressed in tissue?
  - Binding **not** expected for genes **not** expressed
  - These TFs are known to be activators in liver

Hypothesis: Greater conservation for subset of genes which are expressed in both species

$$P(T_{Hs}|T_{Mm},E_{Hs},E_{Mm})>P(T_{Hs}|T_{Mm})$$

Microarray Expression Data

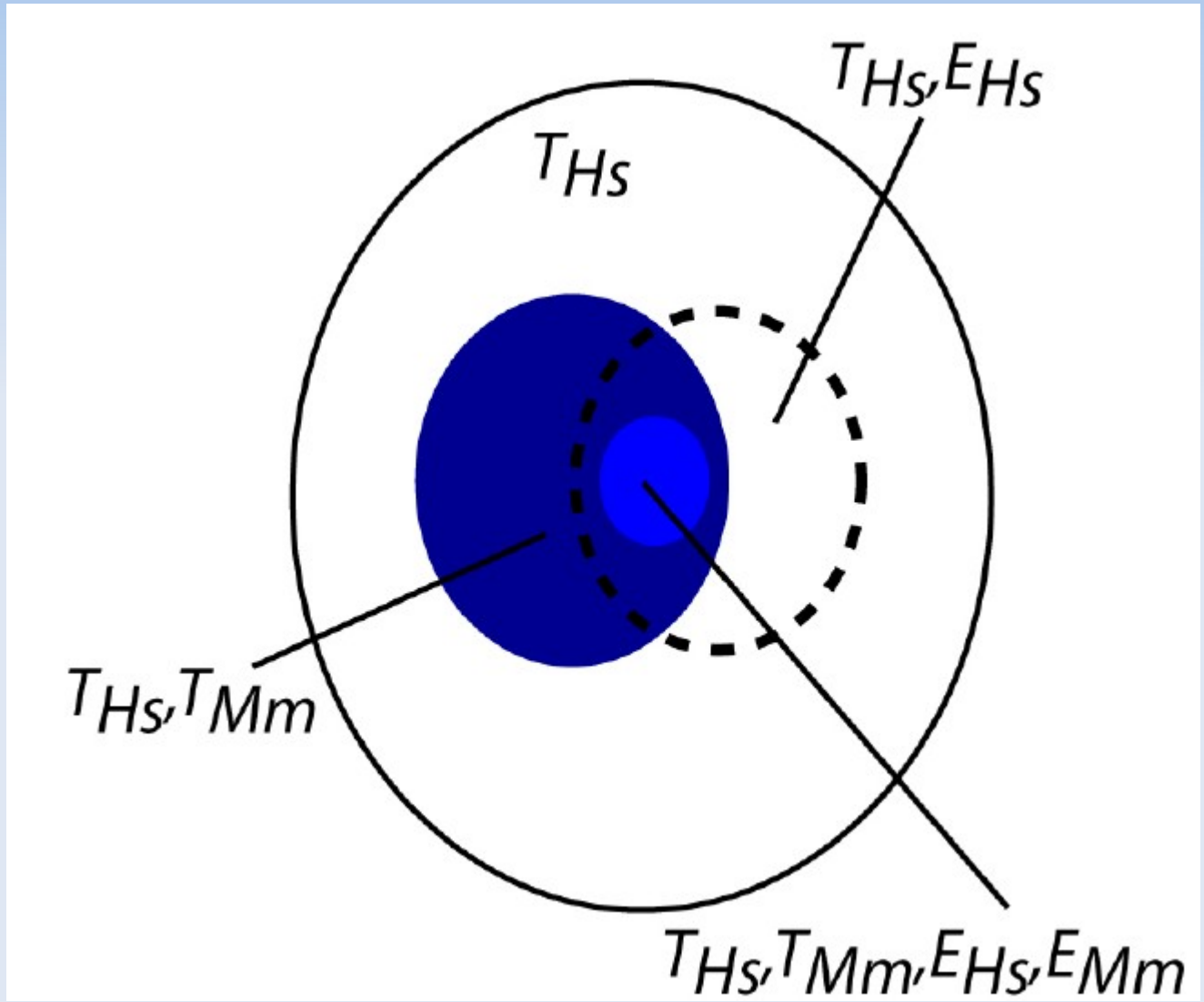
- GNF SymAtlas for liver
  - Microarray data from human and mouse
    - **3051**/4022 pairs with expression for both

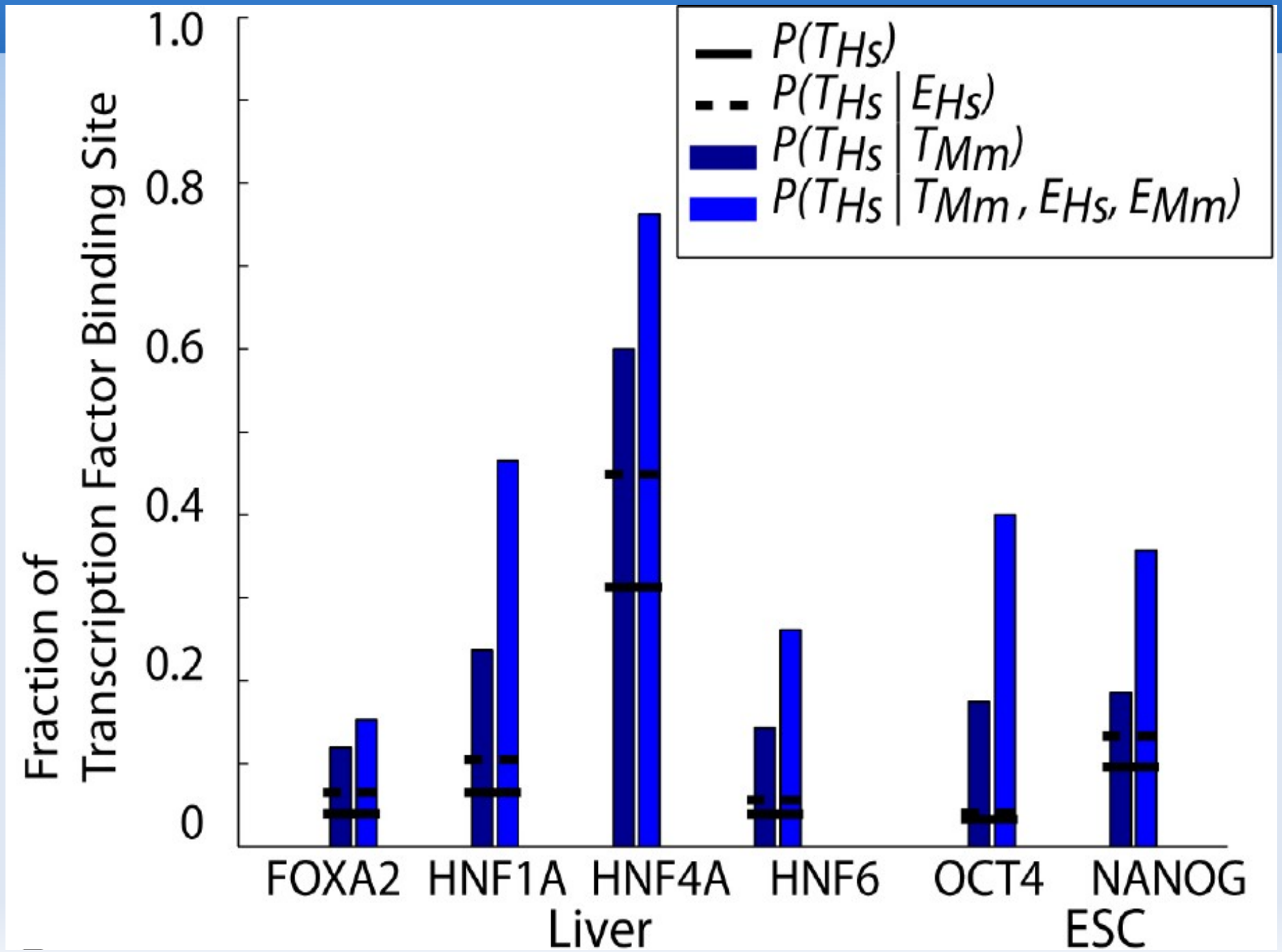


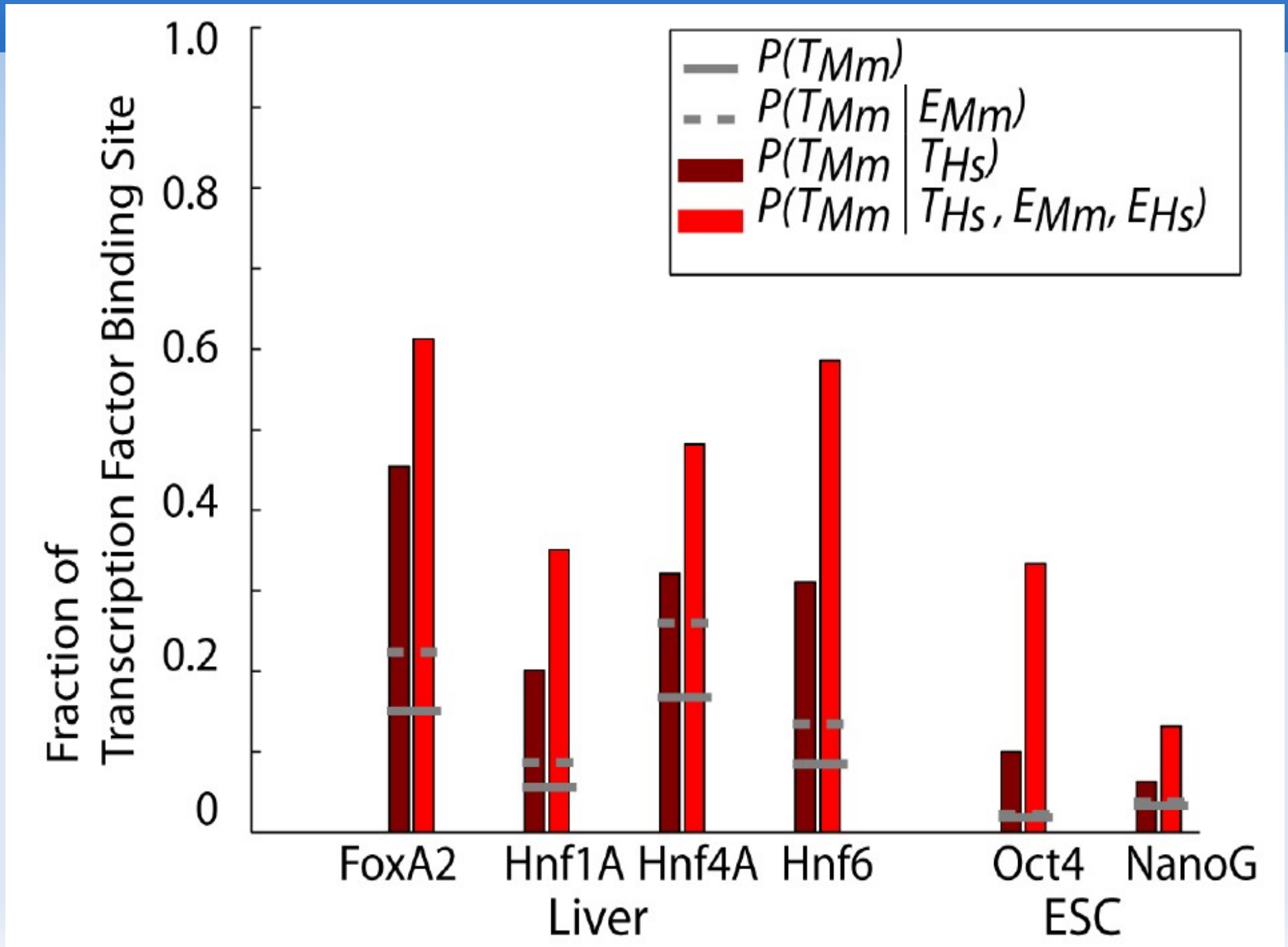
A gene atlas of the mouse and human protein-encoding transcriptomes

Andrew I. Su*†, Tim Wiltshire*†, Serge Batalov*†, Hilmar Lapp*, Keith A. Ching*, David Block*, Jie Zhang*, Richard Soden*, Mimi Hayakawa*, Gabriel Kreiman*‡, Michael P. Cooke*, John R. Walker*, and John B. Hogenesch*§¶

Subsets of TFBSs




1.0
0.8
0.6
0.4
0.2
0
Fraction of Transcription Factor Binding Site
P(THs)
P(THs | EHs)
P(THs | TMm)
P(THs | TMm , EHs, EMm)
FOXA2
HNF1A
HNF4A
HNF6
OCT4
NANOG
Liver
ESC


1.0
0.8
0.6
0.4
0.2
0
Fraction of Transcription Factor Binding Site
P(TMm)
P(TMm | EMm)
P(TMm | THs)
P(TMm | THs , EMm, EHs)
FoxA2
Hnf1A
Hnf4A
Hnf6
Oct4
NanoG
Liver
ESC

Predicting Expression from TFBS

- Use linear regression to predict expression
  - Real value for expression *E* of gene *i*
  - Binary variable *I* for binding of TF *j* at gene *i*

$$\log E_i = \sum_{j=1}^{4} a_j I_{ij}$$

(4 parameters)

$$\log E_i = \sum_{j=1}^{4} \sum_{c=0}^{1} a_j^c I_{ij}^c$$

(8 parameters)

Predicting Expression from TFBS

- Use linear regression to predict expression
  - Real value for expression E of gene i
  - Binary variable I for binding of TF j at gene i

$$\log E_i = \sum_{j=1}^{4} a_j I_{ij}$$

(4 parameters)

$$\log E_i = \sum_{j=1}^{4} \sum_{c=0}^{1} a_j^c I_{ij}^c$$

(8 parameters)

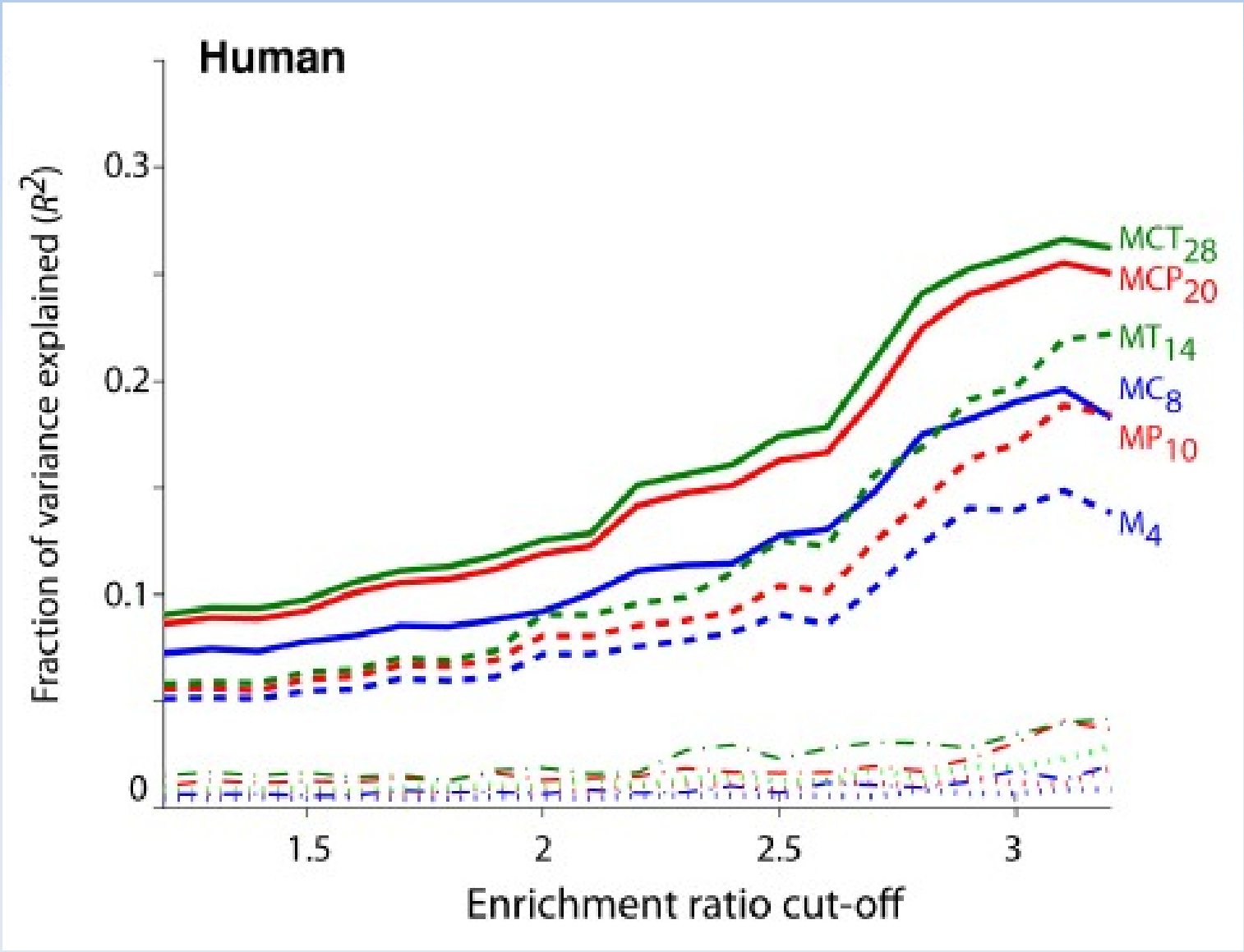
$$\log E_i = \sum_{j=1}^{4} \left[a_j I_{ij} + \sum_{k=j+1}^{4} a_{jk} I_{ijk} \right]$$

(10 parameters)

$$\log E_i = \sum_{j=1}^{4} \sum_{c=0}^{1} \left[a_j^c I_{ij}^c + \sum_{k=j+1}^{4} a_{jk}^c I_{ijk}^c \right]$$

(20 parameters)

Predicting Expression from TFBSs



Model complexity

- Adding parameters improves prediction
- Use AIC to penalize model complexity

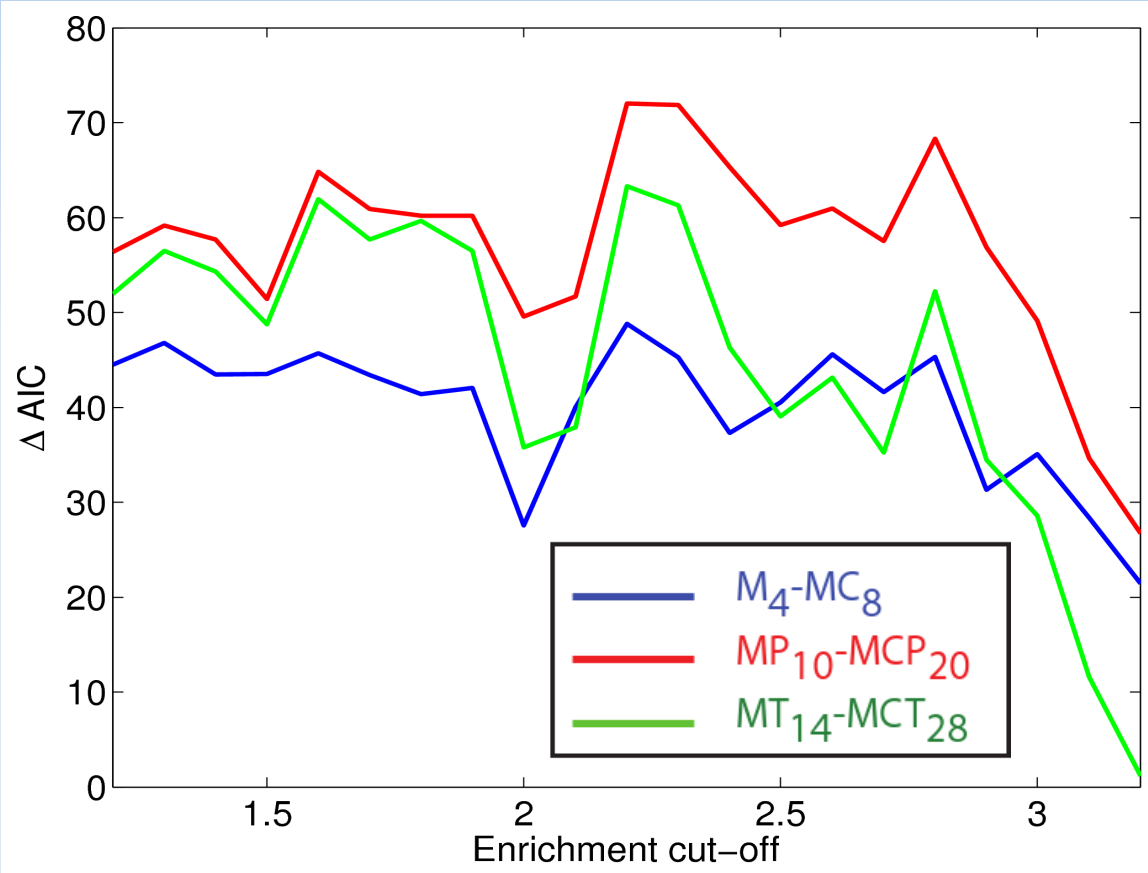
$$AIC = 2k + n\log(RSS/n)$$

$$L \propto \exp(-\Delta AIC/2)$$

k - #parameters

n - #genes

- Coeffs for conserved TFBSs 8 times higher



Summary

- Combined several studies containing TF binding and gene expression data
- Genes expressed in both human and mouse twice as likely to have conserved TFBSs
  - Expression proxy for function
- Conserved TFBSs have a significantly greater impact on expression
  - Prediction of expression levels increased by ~30%