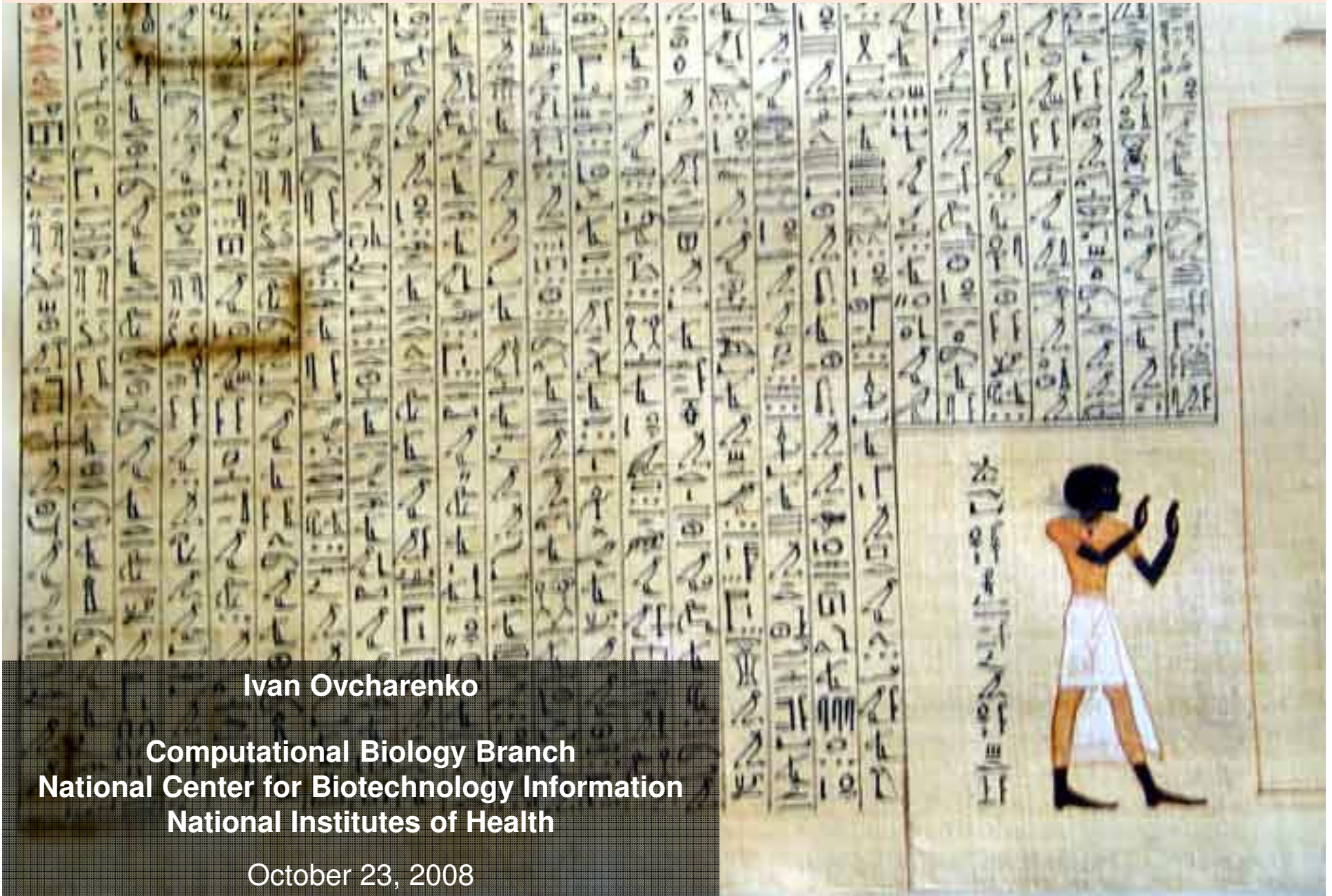


# The genetic code of gene regulatory elements



# Outline

- **Gene deserts and distant gene regulation**
- **Genetic encryption of gene regulation**
- **Heart regulatory code**
- **Regulation of regulators:**  
**how transcription factors regulate themselves**

# The Genome Sequence: The Ultimate Code of Life

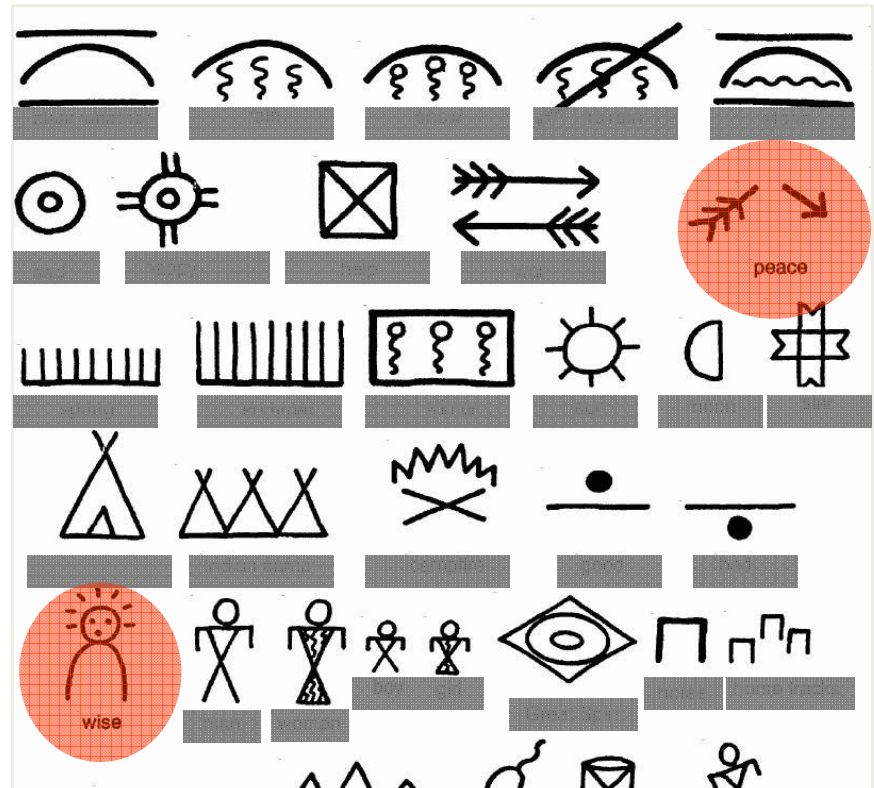


3 billion letters

~ **45% is "junk"** (repetitive elements)

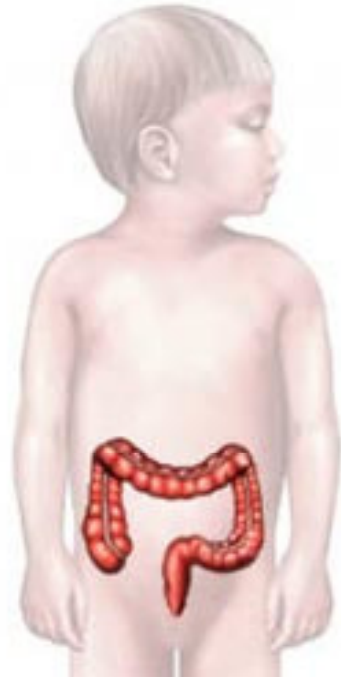
~ **3% is coding for proteins**

gene regulatory elements (REs) reside  
**SOMEWHERE** in the rest ~50%

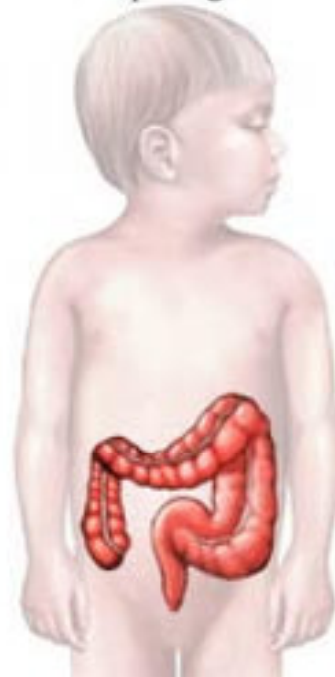


# Hirschsprung disease is associated with a noncoding SNP

Normal colon



Enlarged colon of Hirschsprung's Disease



articles

## A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk

Eileen Sproat Emison<sup>1</sup>\*, Andrew S. McCallion<sup>1</sup>\*, Carl S. Kashuk<sup>1</sup>, Richard T. Bush<sup>1</sup>, Elizabeth Grice<sup>1</sup>, Shin Lin<sup>1</sup>, Matthew E. Portnoy<sup>1</sup>, David J. Cutter<sup>1</sup>, Eric D. Green<sup>2,3</sup> & Aravinda Chakravarti<sup>1</sup>

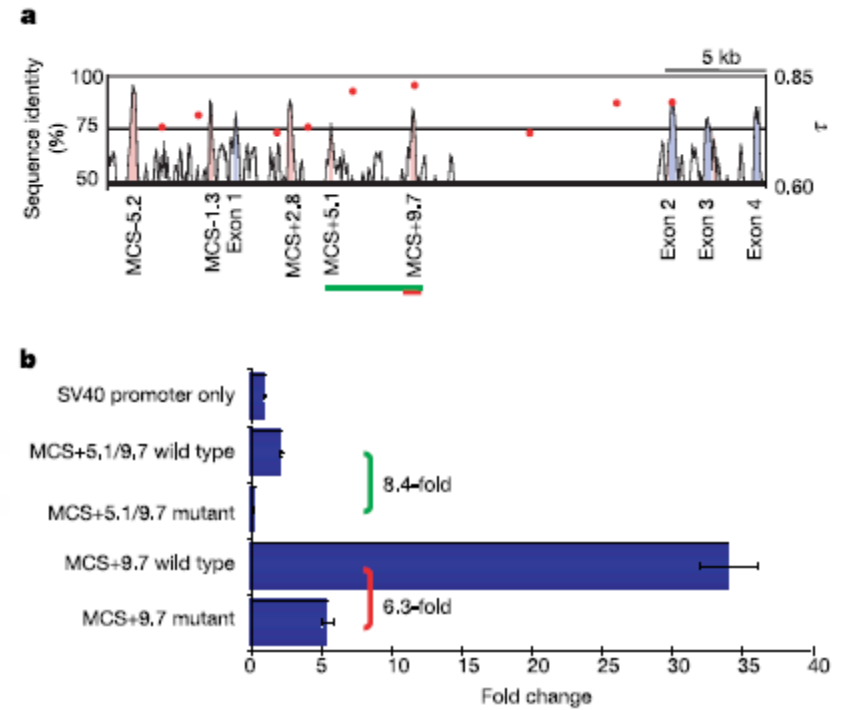
<sup>1</sup>McKusick – Nathan Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

<sup>2</sup>Genome Technology Branch and <sup>3</sup>NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

\*These authors contributed equally to this work

The identification of common variants that contribute to the genesis of human inherited disorders remains a significant challenge. Hirschsprung disease (HSCR) is a multifactorial, non-mendelian disorder in which rare high-penetrance coding sequence mutations in the receptor tyrosine kinase *RET* contribute to risk in combination with mutations at other genes. We have used family-based association studies to identify a disease interval, and integrated this with comparative and functional genomic analysis to prioritize conserved and functional elements within which mutations can be sought. We now show that a common non-coding *RET* variant within a conserved enhancer-like sequence in intron 1 is significantly associated with HSCR susceptibility and makes a 20-fold greater contribution to risk than rare alleles do. This mutation reduces *in vitro* enhancer activity markedly, has low penetrance, has different genetic effects in males and females, and explains several features of the complex inheritance pattern of HSCR. Thus, common low-penetrance variants, identified by association studies, can underlie both common and rare diseases.

*RET*



# Comparative Sequence Analysis

Biologically **functional** regions in the genome tend to stay **conserved** throughout the evolution.



Therefore, by **aligning** homologous sequences from different, but related species we can identify Evolutionary Conserved Regions (**ECRs**) with a putative functional importance

1880<sup>th</sup>



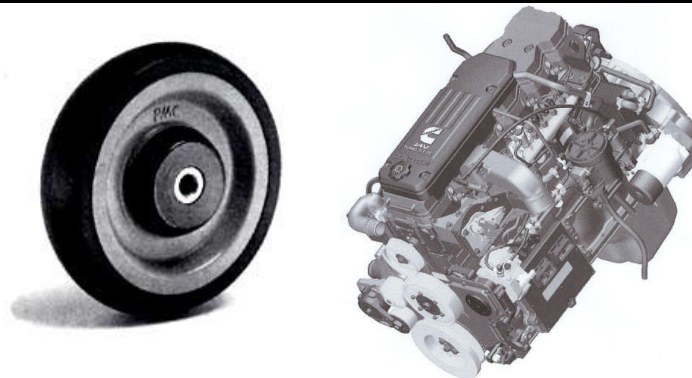
1920<sup>th</sup>



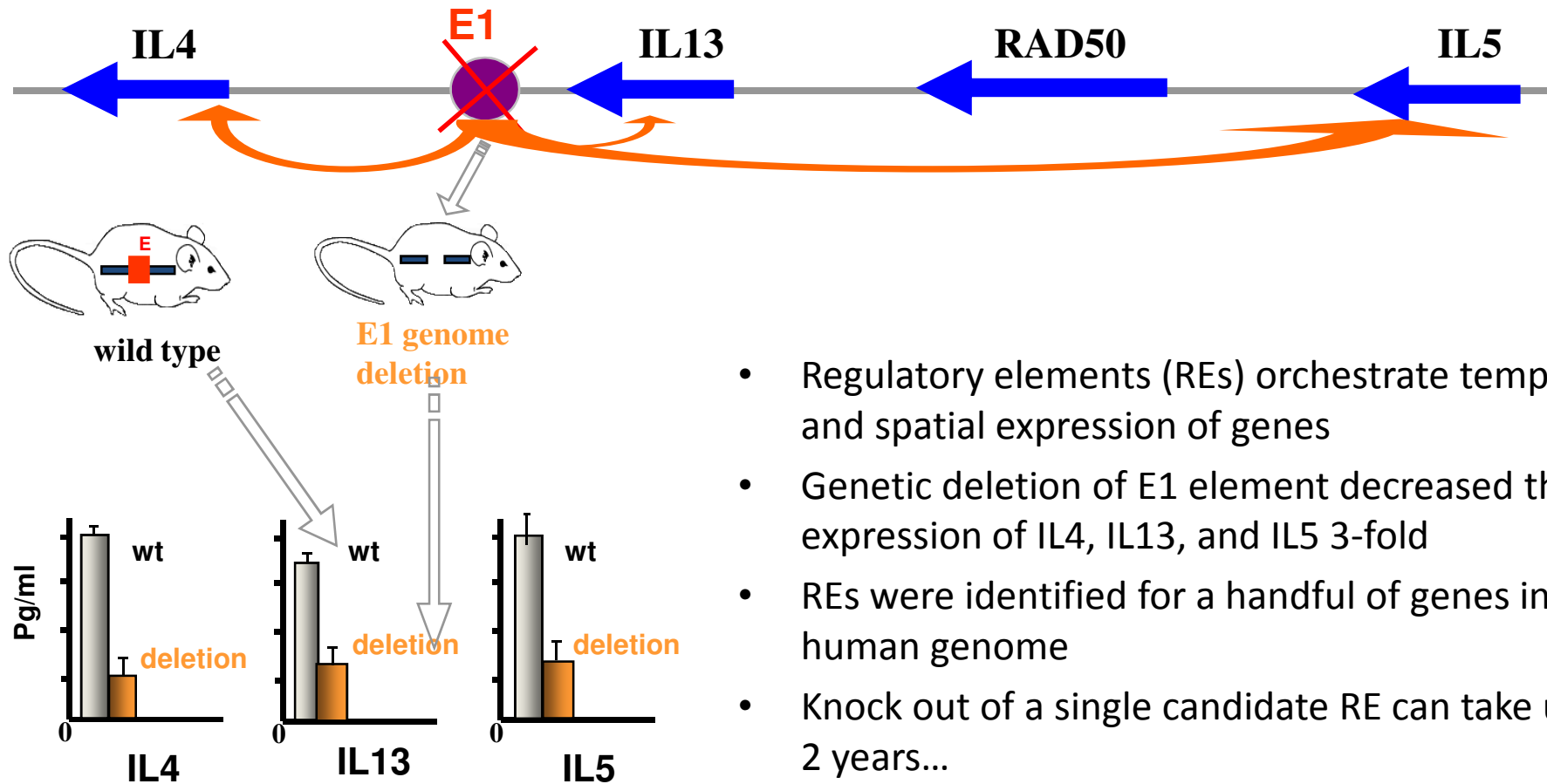
1950<sup>th</sup>



2000<sup>th</sup>

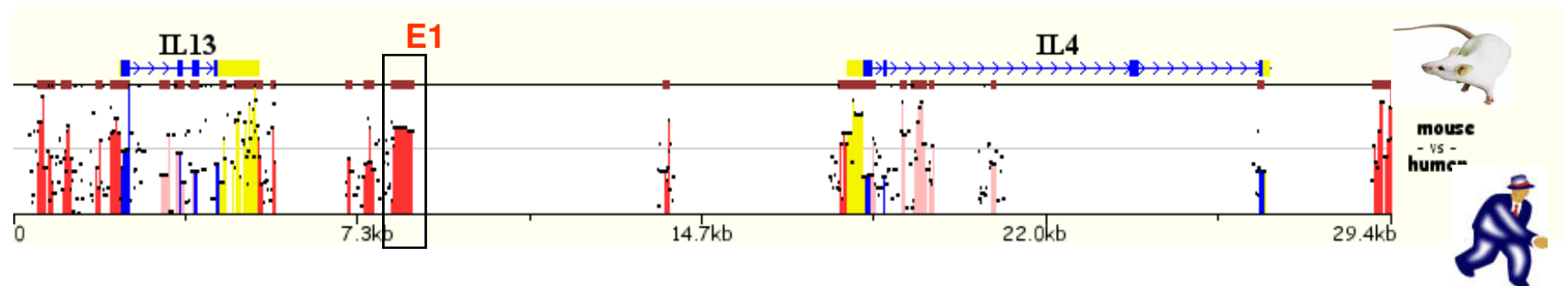


# In vivo validation of a regulatory element



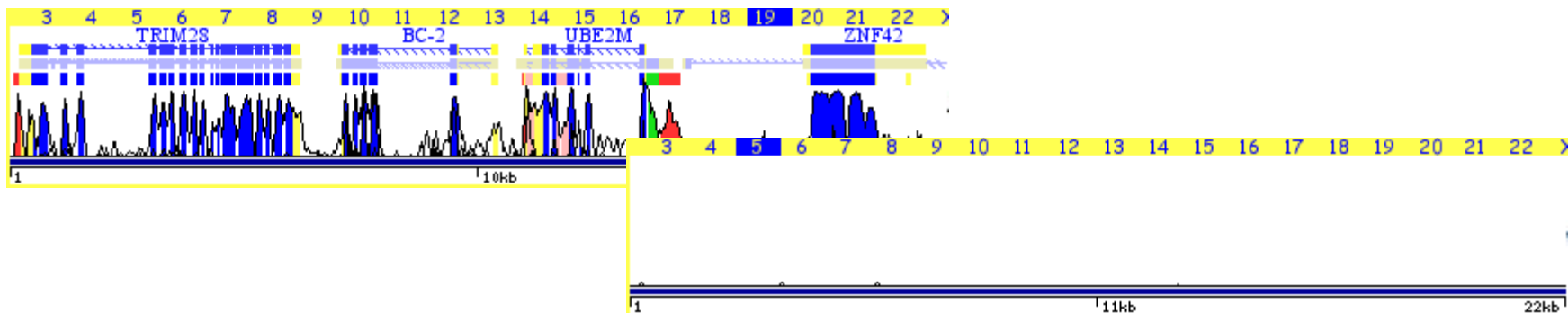
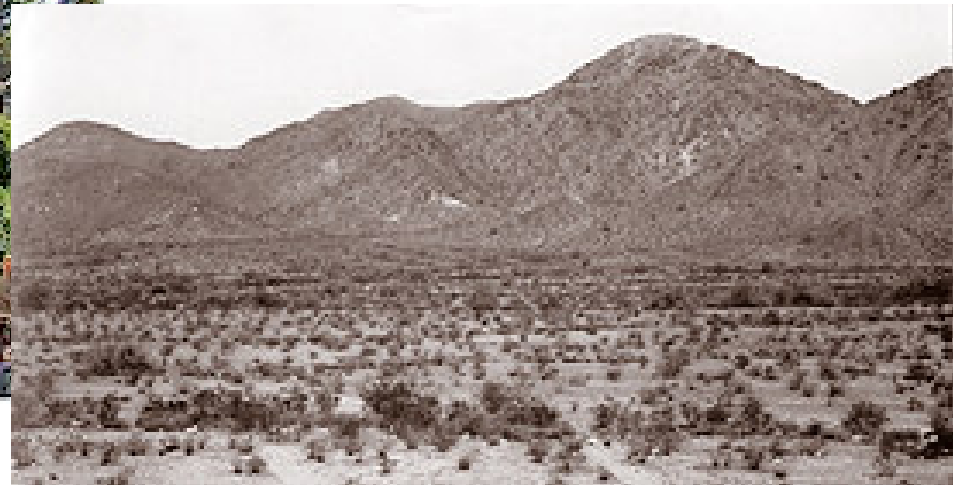
- Regulatory elements (REs) orchestrate temporal and spatial expression of genes
- Genetic deletion of E1 element decreased the expression of IL4, IL13, and IL5 3-fold
- REs were identified for a handful of genes in the human genome
- Knock out of a single candidate RE can take up to 2 years...

# Comparative genomics to predict regulatory elements



- Functionally important elements in genomes mutate at slower rates than the neutrally evolving background
  - 1% of sequence is conserved between humans and fish
  - 75% of genes are conserved between humans and fish
- RE E1 is highly conserved in human and mouse genomes.
- Comparative genomics can be utilized to prioritize functional elements

# Gene Deserts in the Human Genome



Gene deserts = **25%** of the human genome sequence

# Human chromosome 13

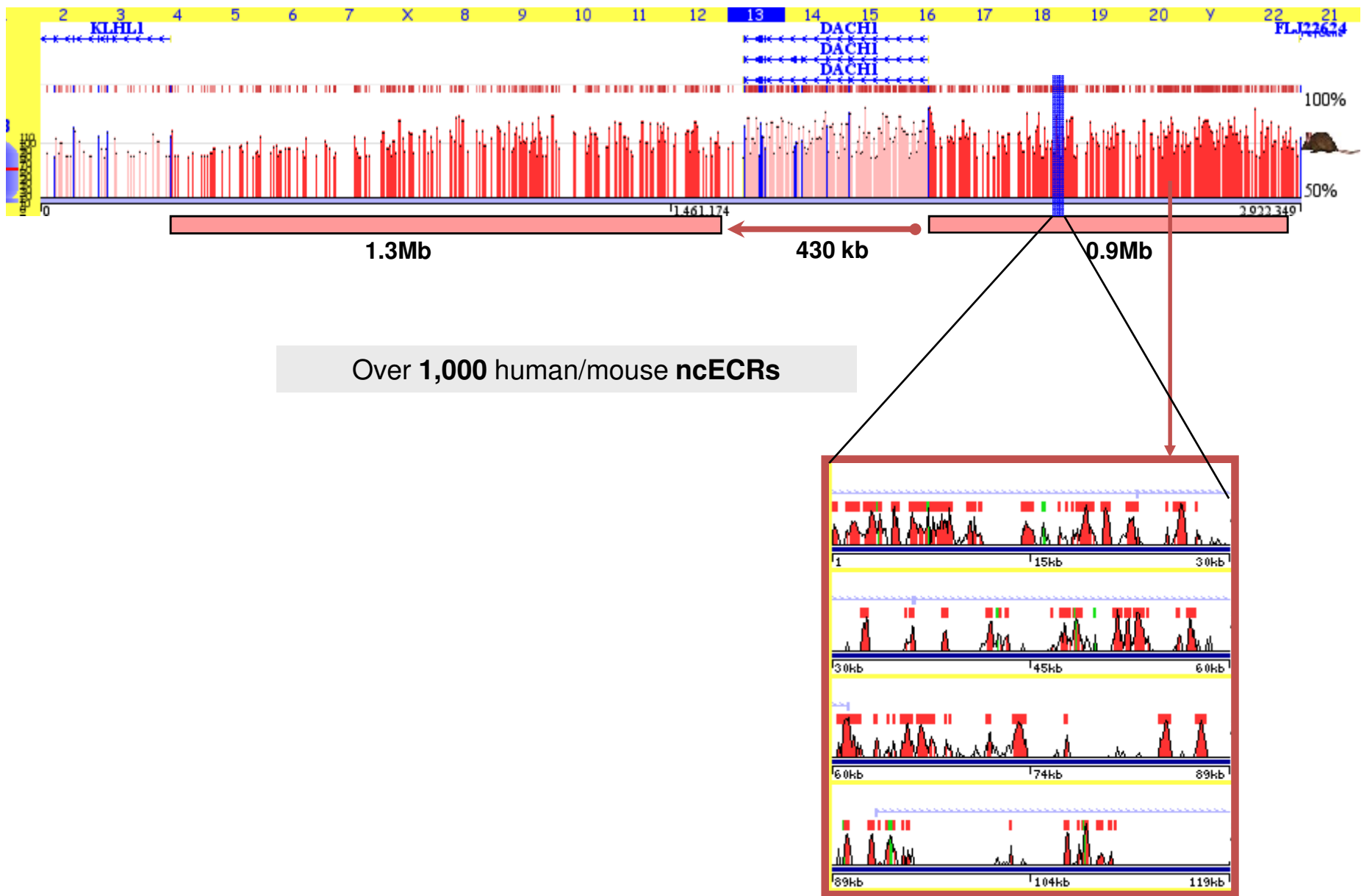


**~500 *gene deserts* in the human genome**

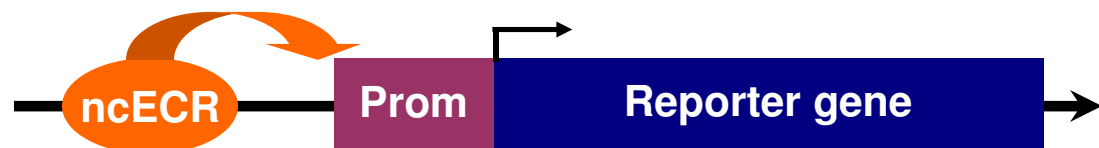
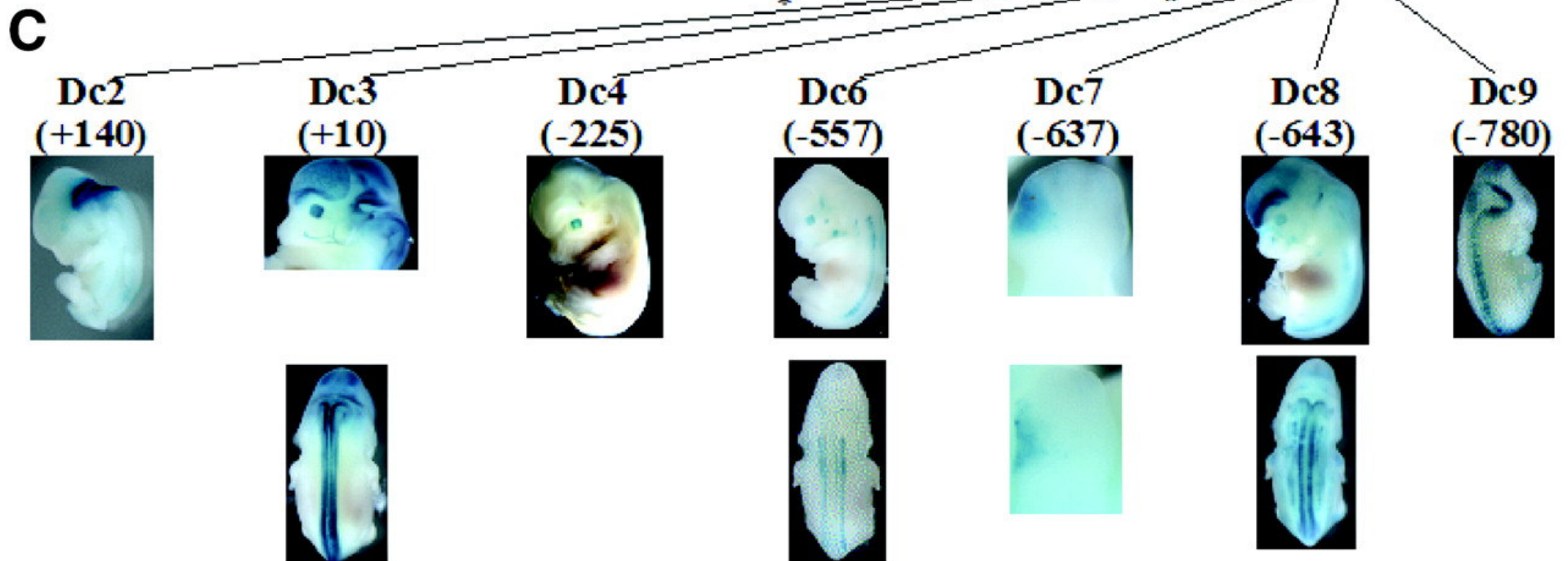
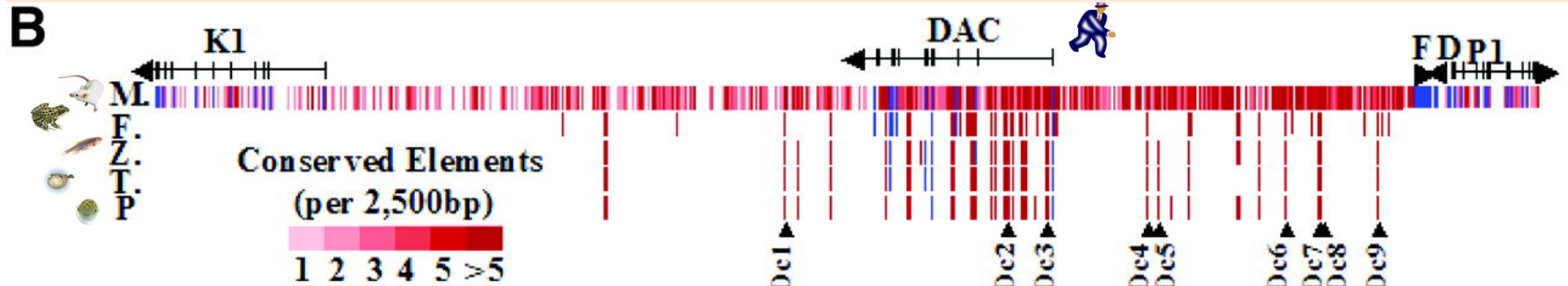
**~50% of the HSA13 consists of *gene deserts***

***Gene deserts* are NOT enriched in repetitive elements**

# DACH gene deserts on chromosome 13

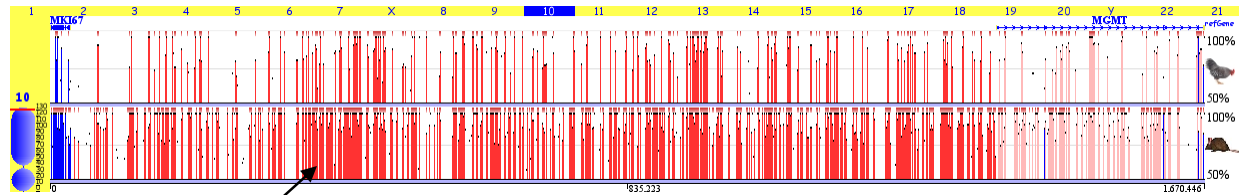


# Phylogenetic conservation of DACH gene deserts



# Dichotomy in the evolutionary conservation of gene deserts

**~200 stable**  
gene deserts



- DACH
- OTX2
- SOX2

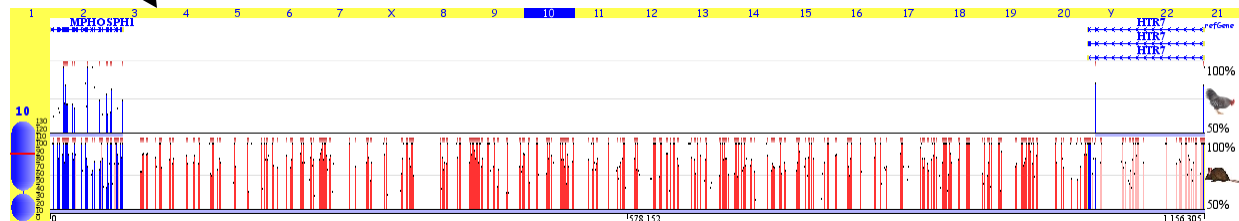


**-VS-**



**- Deletion does not lead to a phenotype**

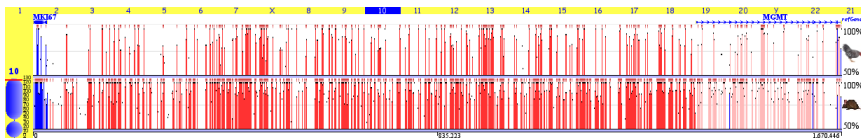
**~300 variable**  
gene deserts



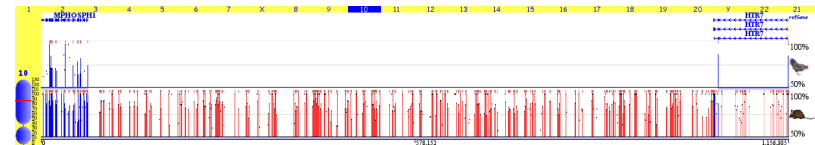
# Function of genes flanking gene deserts

## ***stable*** gene deserts

- transcription
- regulation of transcription
- regulation of metabolism
- development
- etc.



## ***variable*** gene deserts

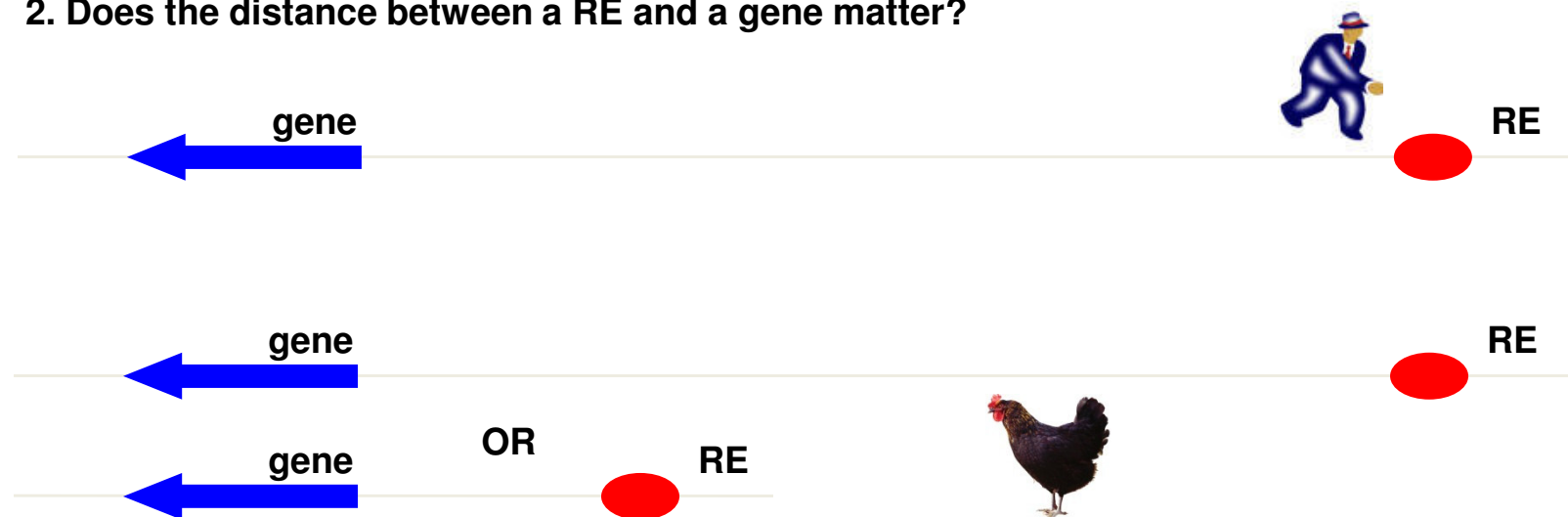


# Distant gene regulation

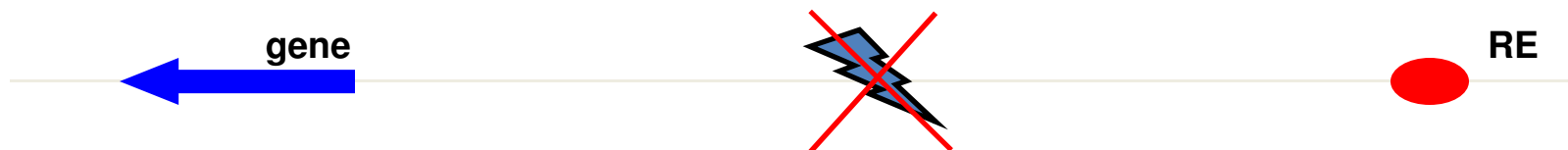
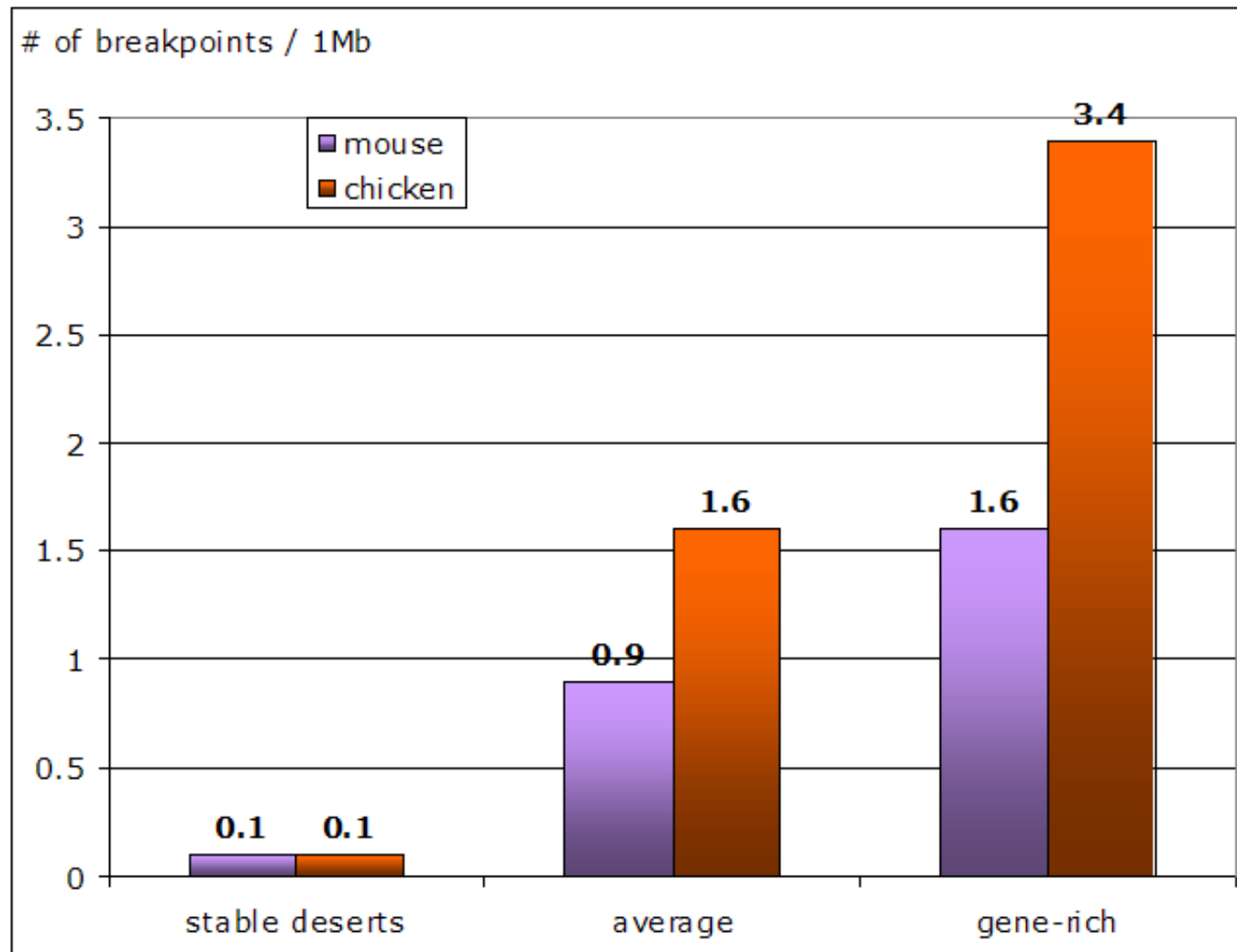
1. Do stable gene deserts harbor distant regulatory elements?



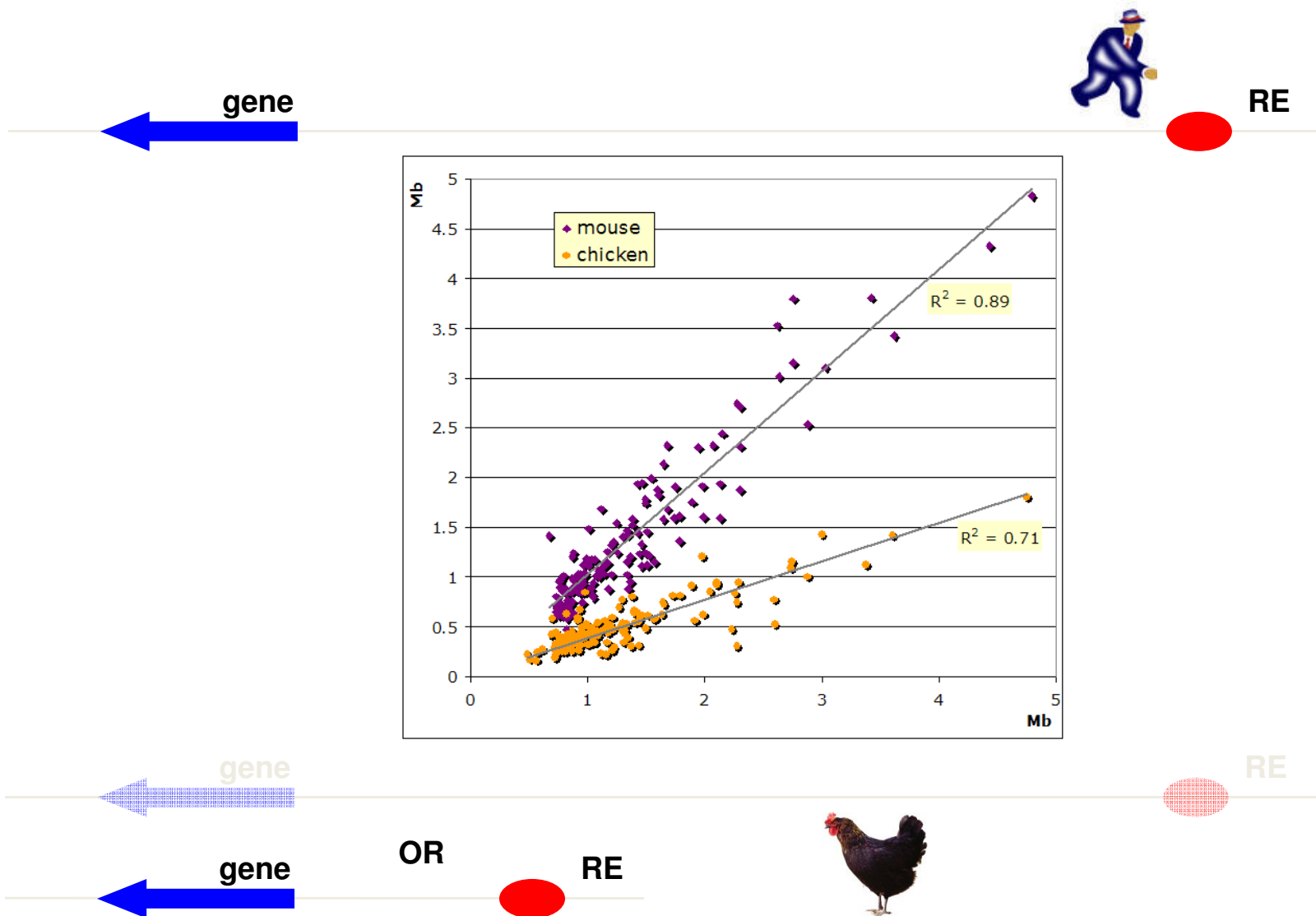
2. Does the distance between a RE and a gene matter?



# Chromosomal stability of gene deserts



# Distant REs are distance-independent



# Gene deserts :: Summary

**25% of the human genome consists of gene deserts**

**There are 2 classes of gene deserts – stable and variable gene deserts**

**Stable gene deserts are evolutionarily protected from chromosome rearrangements indicating presence of distant regulatory elements**

**Experimental validation of deeply conserved sequences in some stable gene deserts confirms their enhancer activity**

**Gene regulation based on distant regulatory elements does not probably depend on the distance between a regulatory element and the gene it regulates**

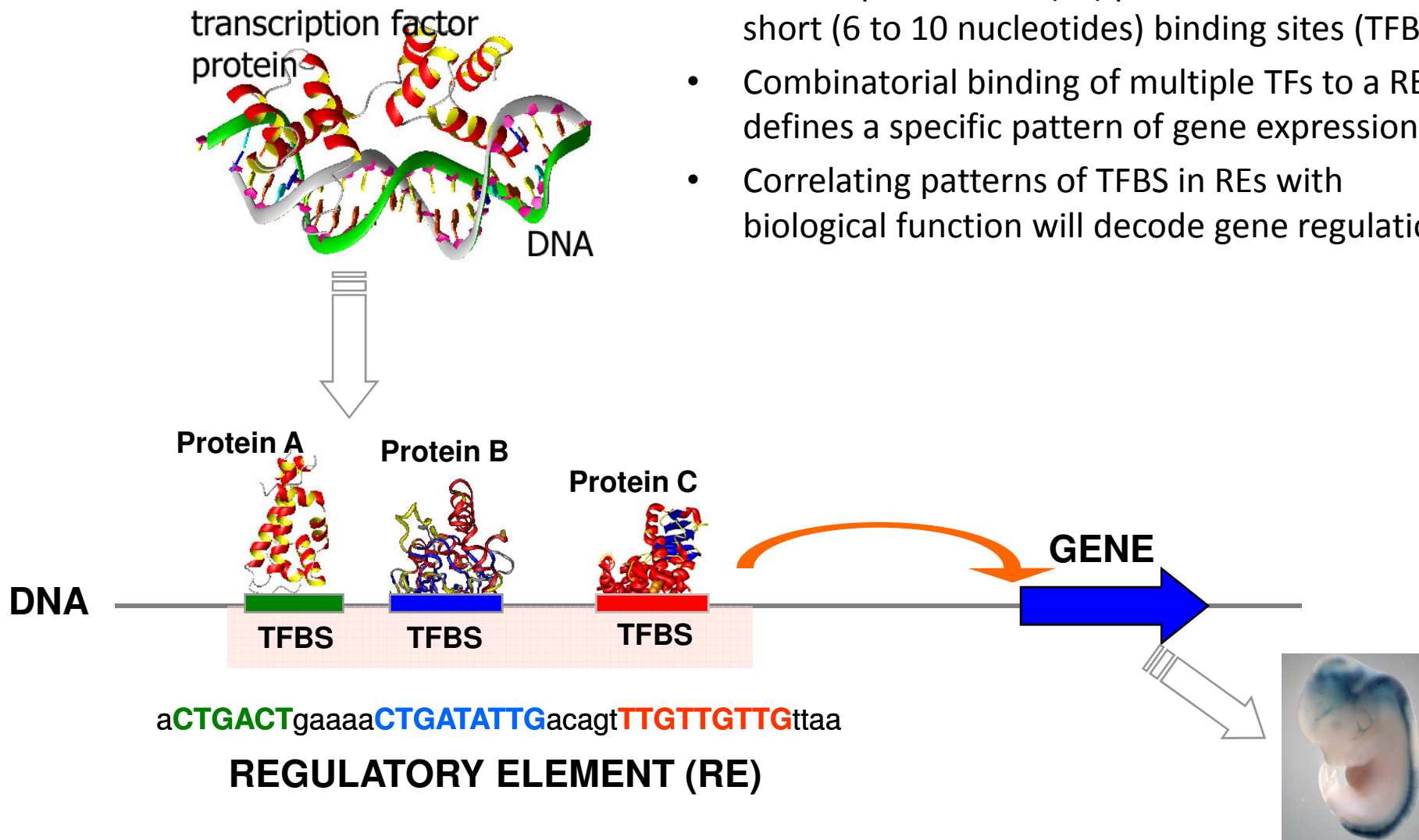
# Outline

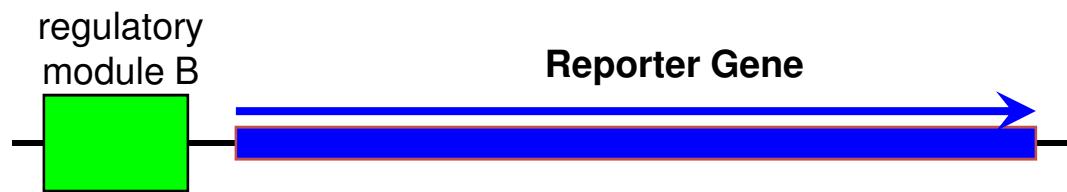
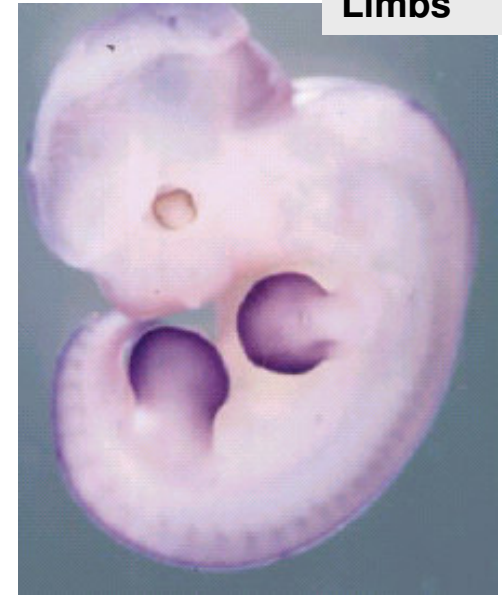
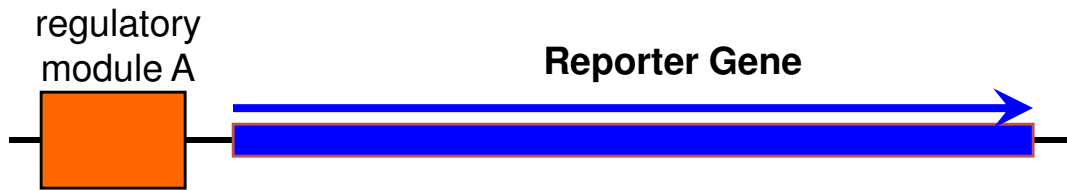
- Gene deserts and distant gene regulation
- **Genetic encryption of gene regulation**
- Heart regulatory code
- Regulation of regulators:  
how transcription factors regulate themselves



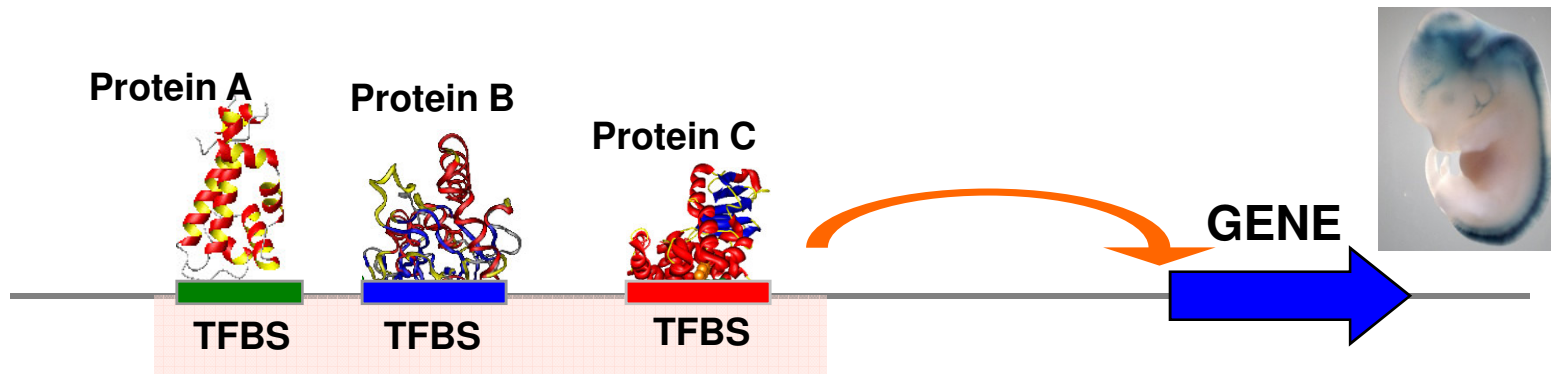
# Patterns of transcription factor binding sites define biological functions of REs

- Transcription factor (TF) proteins bind to very short (6 to 10 nucleotides) binding sites (TFBS)
- Combinatorial binding of multiple TFs to a RE defines a specific pattern of gene expression
- Correlating patterns of TFBS in REs with biological function will decode gene regulation





# Computational identification of TFBS



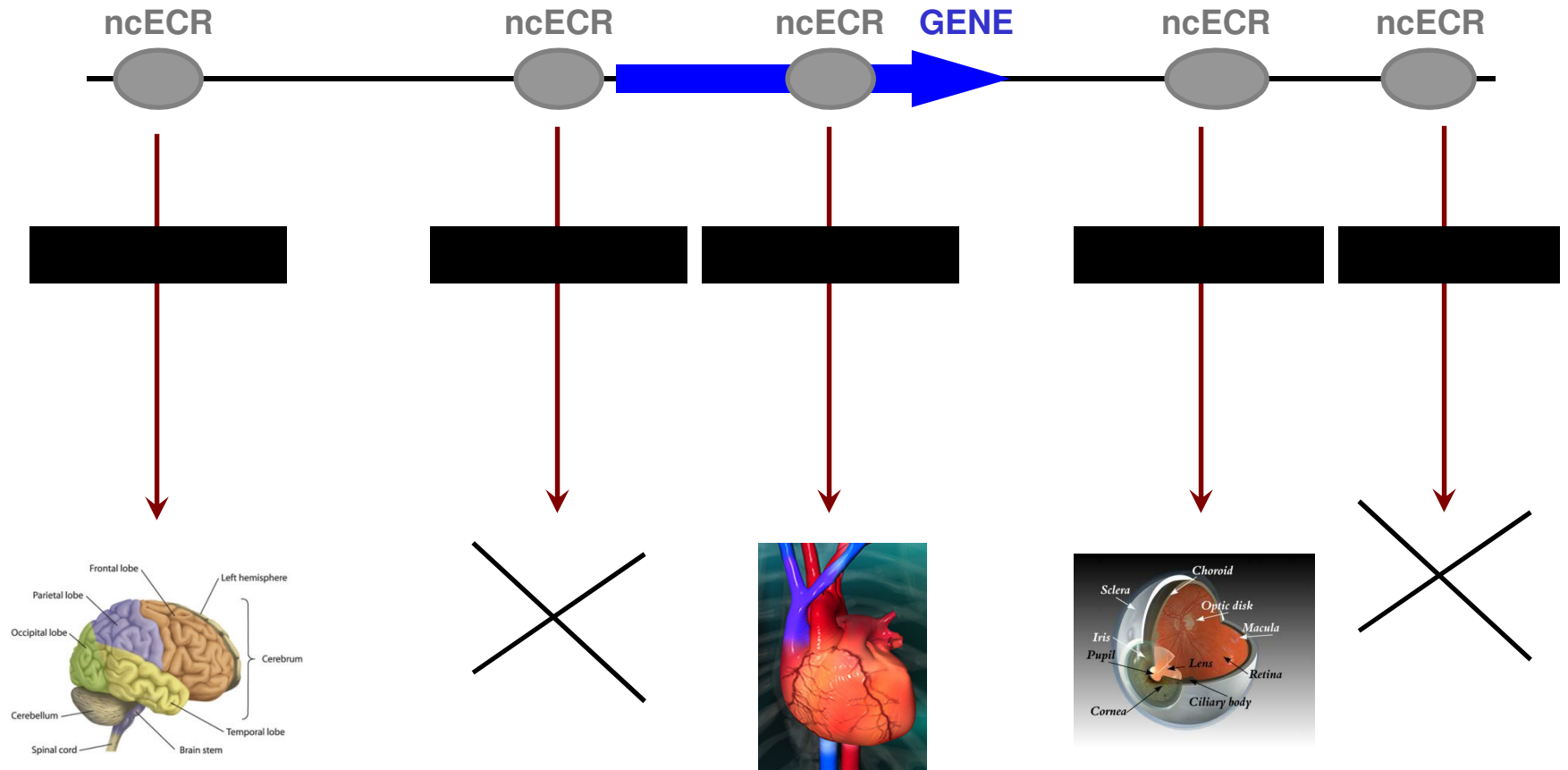
a**CTGACT**gaaaa**CTGATATTG**acagt**TTGTTGTTG**ttaa

**REGULATORY ELEMENT (RE)**

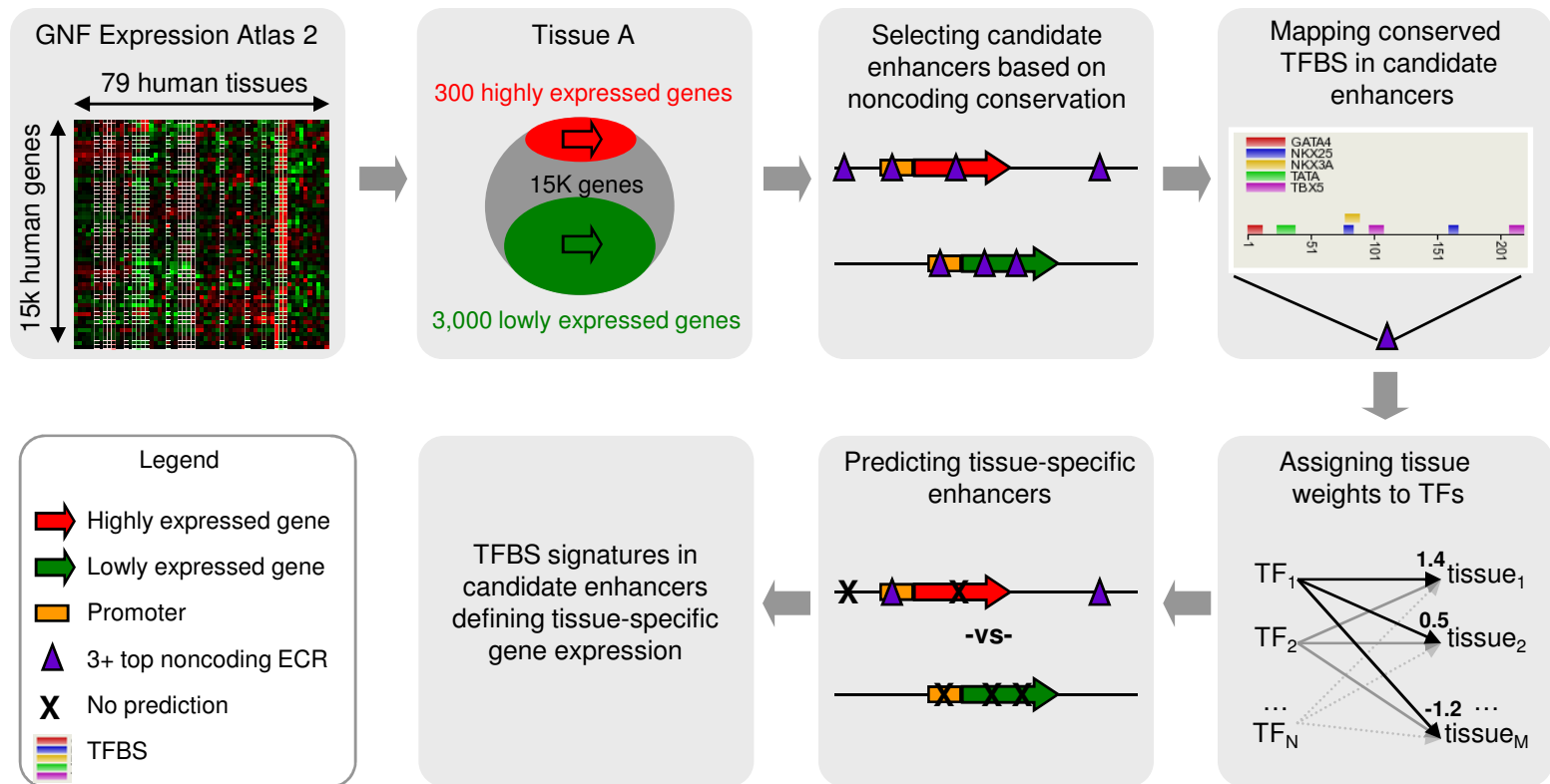
-- Transcription factor binding sites are very short (~ 6-10 bp)

-- Computational predictions of transcription factor binding sites are overwhelmed with **false positives**

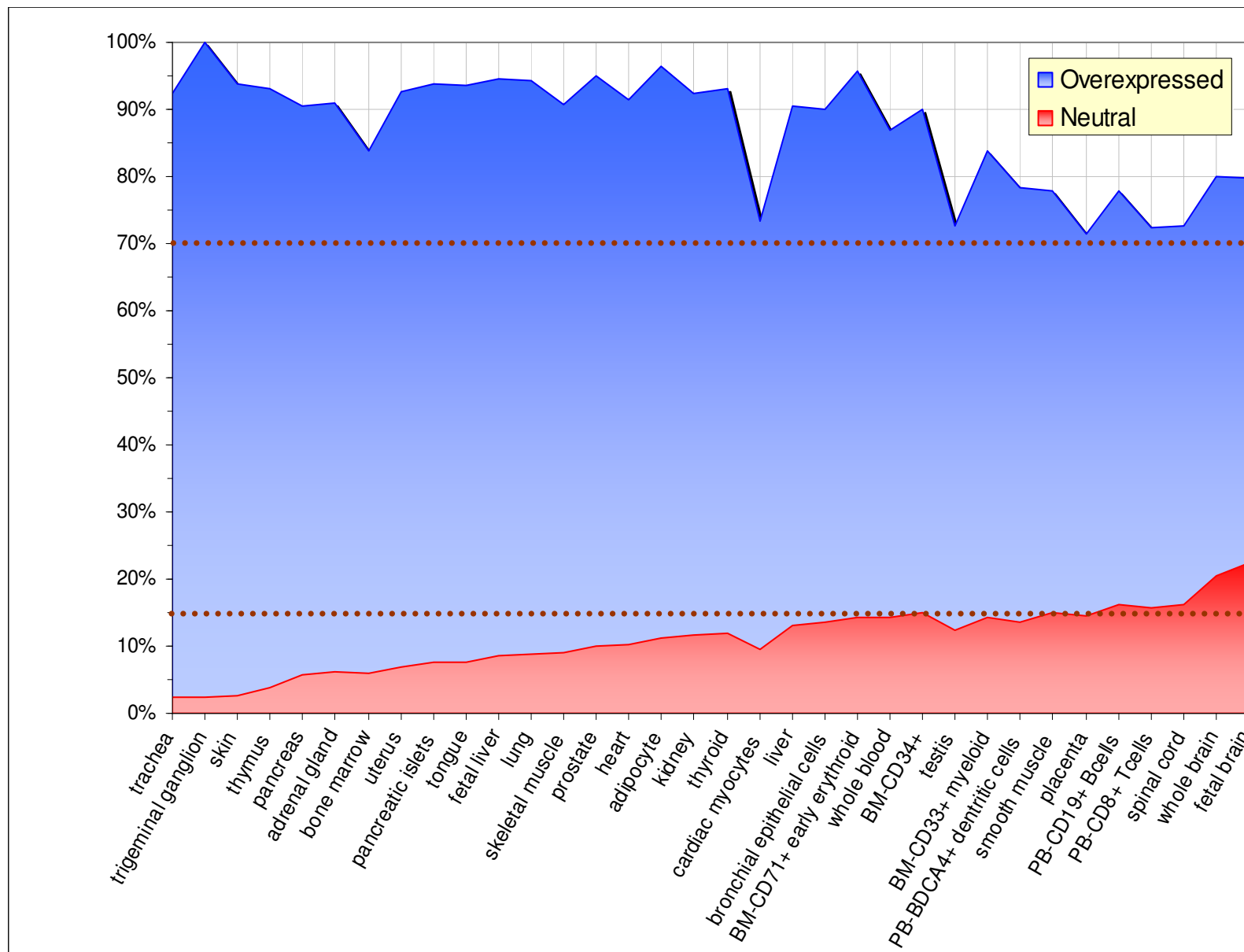
# Deciphering the genetic code of enhancers



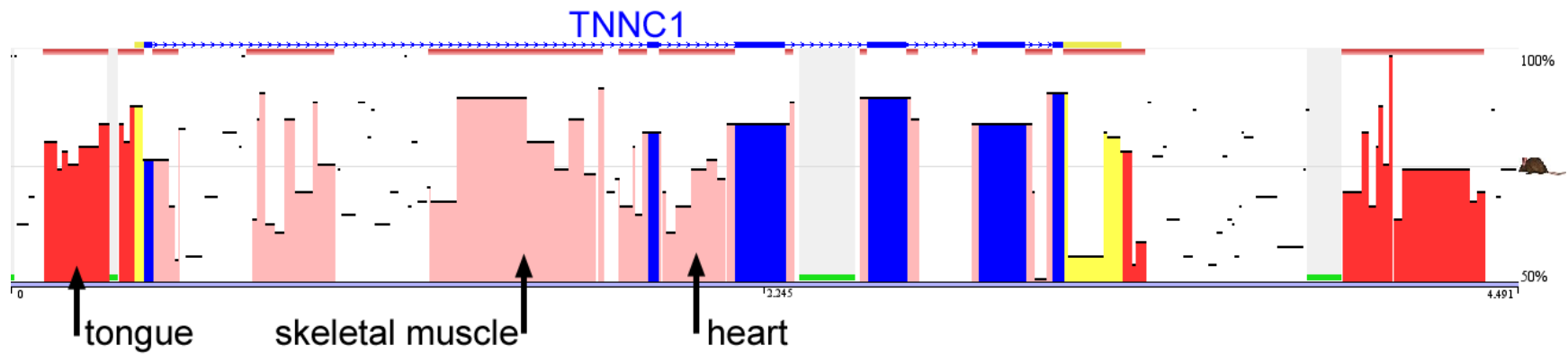
# Enhancer Identification (EI) method



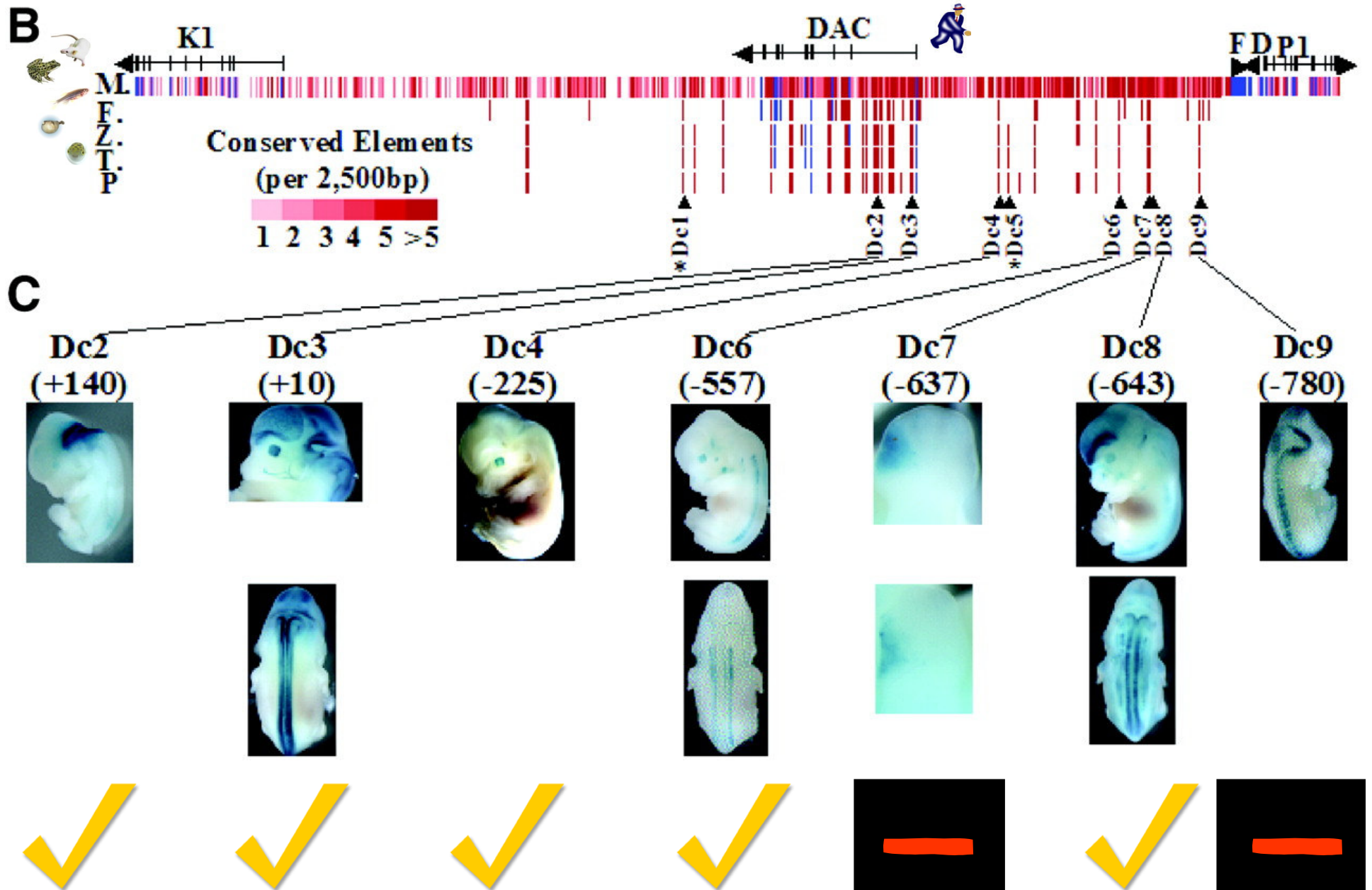
# Tissue-specificity noncoding signatures



# Known skeletal muscle enhancer



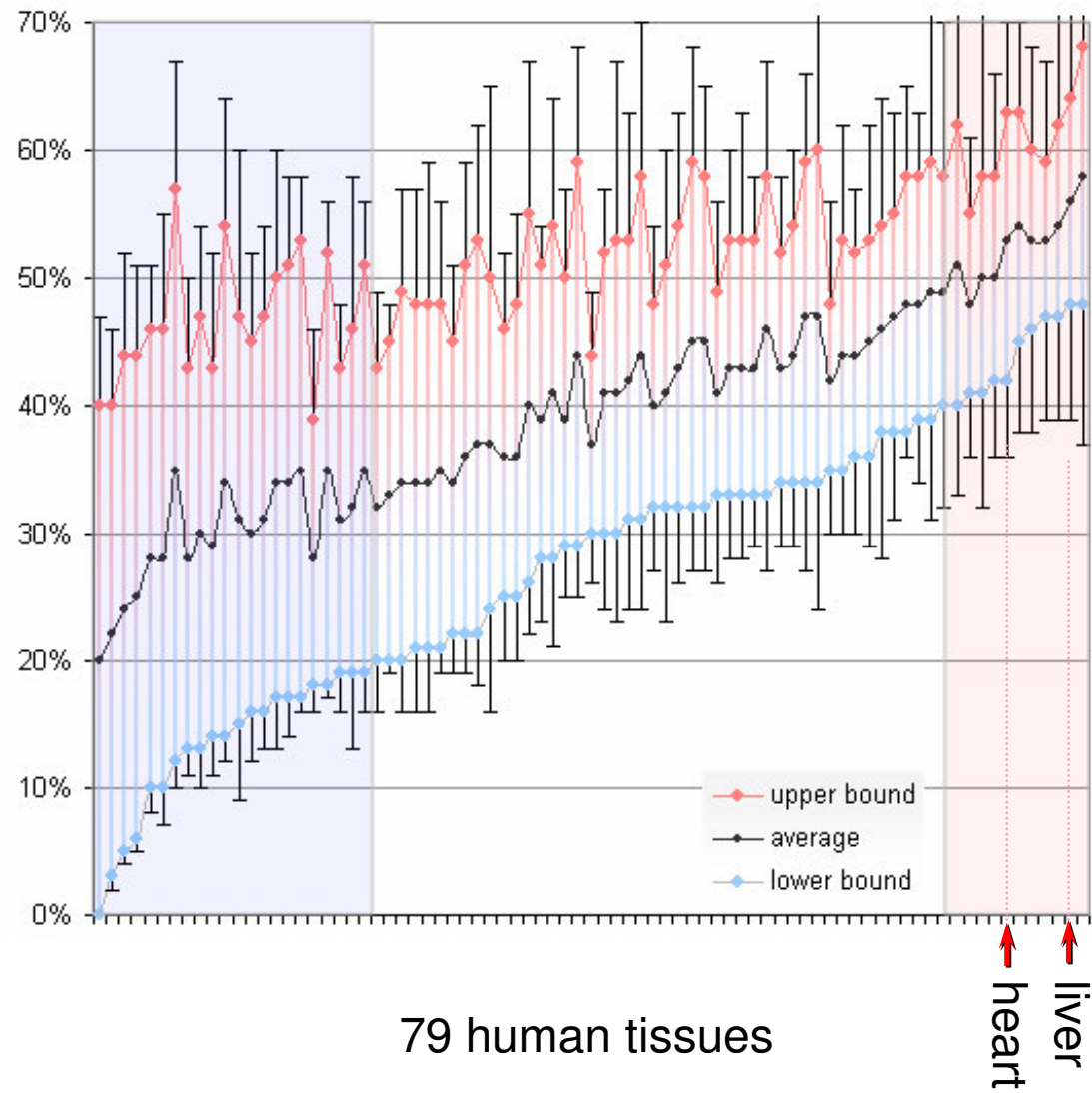
# DACH brain/CNS enhancers



# EI Performance

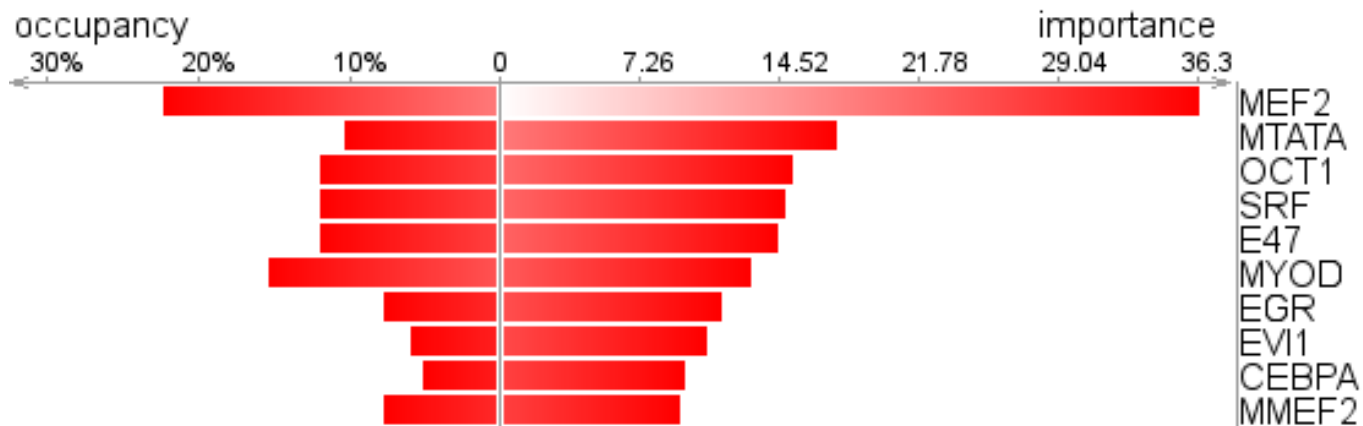
## Precision

(the fraction of predicted tissue-specific enhancers that are correct):

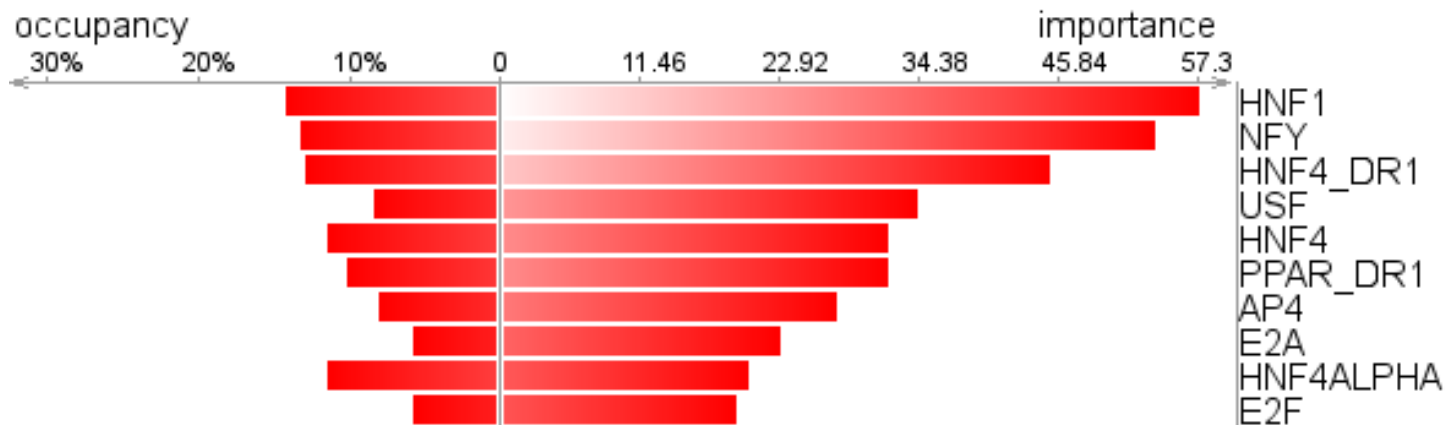


# Associating TFs with tissue specificities

## Skeletal muscle



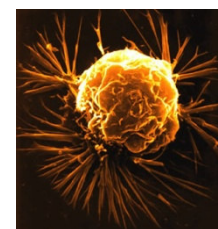
## Liver



# Database of enhancers in the human genome



~4k *heart* candidate enhancers



~8k *cancer* candidate enhancers

# Gene regulatory code :: Summary

**Combination of comparative genomics, gene expression data, and TFBS clustering provides a tool to define the sequence code of tissue-specific enhancers**

**Over 7,000 candidate tissue-specific enhancers had been predicted in the human genome**

**Enhancers for several tissues (including heart and liver) were predicted with high precision**

**Prediction of TFs associated with the regulation of tissue-specific expression provides the means to move from microarray expression data directly to the identification of active TFs**