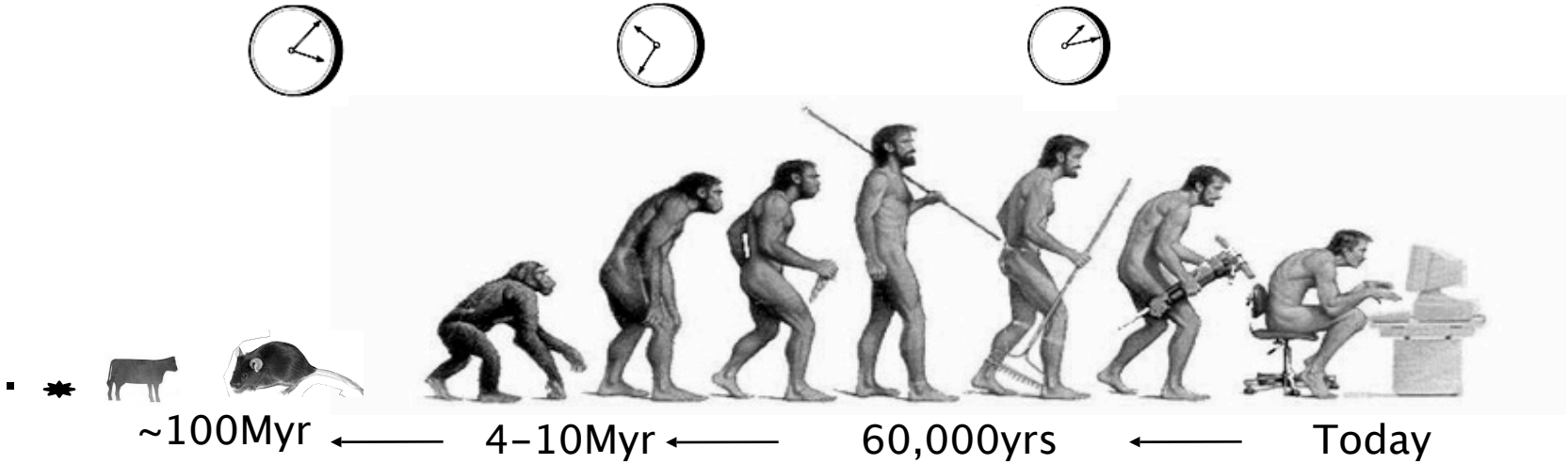
# Evolvability and protein divergence in the evolution of Homo sapiens

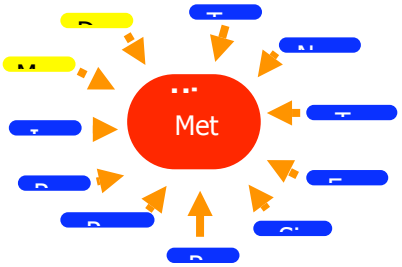



Sarah Teichmann

MRC Laboratory of Molecular Biology
Cambridge, UK

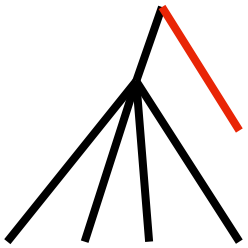
# Questions

- Are there certain groups of proteins with particular functions that evolve faster than others?




Nuria Lopez-Bigas (UPF, Barcelona)

- What is the selection pressure on different functional categories in the human lineage?



Subhajyoti De (MRC Laboratory of Molecular Biology)

# Protein Divergence Rates – Historical background

Dickerson 1971

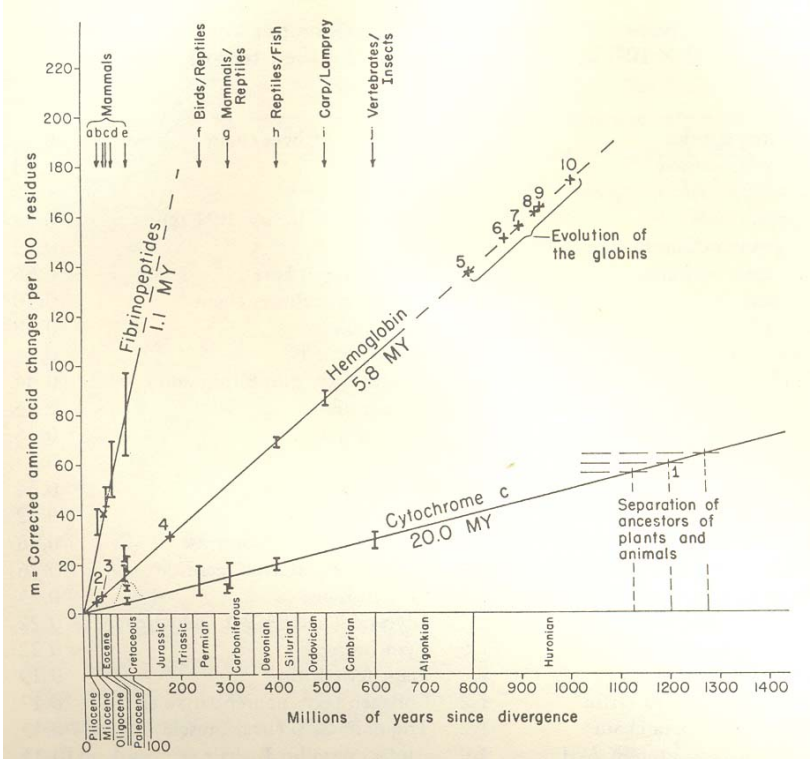



Figure 4.3. Rates of amino acid substitution in fibrinopeptides, hemoglobin, and cytochrome c. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of 1,200±75 MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome c curve. Points 2–10 refer to events in the evolution of the globin family. The $\delta/\beta$ separation is at point 3, $\gamma/\beta$ is at 4, and $\alpha/\beta$ is at 500 MY (carp/lamprey). From Dickerson (1971).

Dayhoff 1978

Table 4.4 Rates of amino acid substitutions per amino acid site per $10^9$ years ($\lambda \times 10^9$) in various proteins. Modified from Dayhoff (1978).

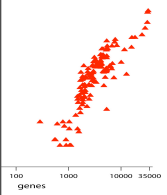
| *Protein* | *Rate* | *Protein* | *Rate* |
|---|---|---|---|
| Fibrinopeptides | 9.0 | Thyrotropin beta chain | 0.74 |
| Growth hormone | 3.7 | Parathyrin | 0.73 |
| Ig kappa chain C region | 3.7 | Parvalbumin | 0.70 |
| Kappa casein | 3.3 | Protease inhibitors, BP1 type | 0.62 |
| Ig gamma chain C region | 3.1 | Trypsin | 0.59 |
| Lutropin beta chain | 3.0 | Melanotropin beta | 0.56 |
| Ig lambda chain C region | 2.7 | Alpha crystallin A chain | 0.50 |
| Complement C3a anaphylatoxin | 2.7 | Endorphin | 0.48 |
| Lactalbumin | 2.7 | Cytochrome $b_5$ | 0.45 |
| Epidermal growth factor | 2.6 | Insulin (exc. guinea pig and coypu) | 0.44 |
| Somatotropin | 2.5 | Calcitonin | 0.43 |
| Pancreatic ribonuclease | 2.1 | Neurophysin 2 | 0.36 |
| Lipotropin beta | 2.1 | Plastocyanin | 0.35 |
| Haptoglobin alpha chain | 2.0 | Lactate dehydrogenase | 0.34 |
| Serum albumin | 1.9 | Adenylate kinase | 0.32 |
| Phospholipase $A_2$ | 1.9 | Triosephosphate isomerase | 0.28 |
| Protease inhibitor, PST1 type | 1.8 | Vasoactive intestinal peptide | 0.26 |
| Prolactin | 1.7 | Corticotropin | 0.25 |
| Pancreatic hormone | 1.7 | Glyceraldehyde 3-$PO_4$ dehydrogenase | 0.22 |
| Carbonic anhydrase C | 1.6 | Cytochrome c | 0.22 |
| Lutropin alpha chain | 1.6 | Plant ferredoxin | 0.19 |
| Hemoglobin alpha chain | 1.2 | Collagen (exc. nonrepetitive ends) | 0.17 |
| Hemoglobin beta chain | 1.2 | Troponin C, skeletal muscle | 0.15 |
| Lipid-binding protein A-II | 1.0 | Alpha crystallin B chain | 0.15 |
| Gastrin | 0.98 | Glucagon | 0.12 |
| Animal lysozyme | 0.98 | Glutamate dehydrogenase | 0.09 |
| Myoglobin | 0.89 | Histone H2B | 0.09 |
| Amyloid AA | 0.87 | Histone H2A | 0.05 |
| Nerve growth factor | 0.85 | Histone H3 | 0.014 |
| Acid proteases | 0.84 | Ubiquitin | 0.010 |
| Myelin basic protein | 0.74 | Histone H4 | 0.010 |

# Protein Divergence - Recent Studies

Complete genomes - eg chicken Hillier (2004)

Transcription factors vs target genes - Babu et al. 2006, Hershberg et al. 2006

Transcription factors are phylum-specific - Peregrin-Alvarez et al.




Expansion of functional categories with genome size - Ranea et al., van Nimwegen (2003), Vogel et al

# Methods::Genomes

vertebrates

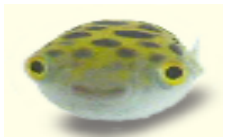
mammals

Mmus Rnor Cfam Btau Mdom

Ggal Xlae

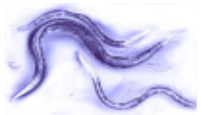
bony fish

Drer Trub Tnig



invertebrates

Agam Amel Dmel Cint Cele

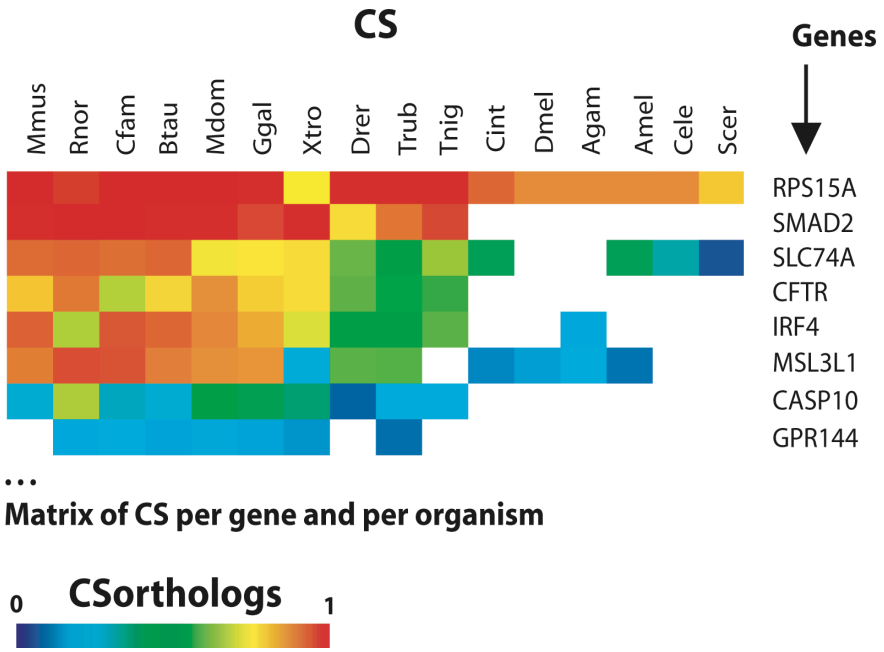


Scer

# Method::conservation score

Conservation Score (CS) orthologs $$CS = \frac{BLASTscore\_Ortholog}{BLASTscore\_itself}$$

Definition of orthologues: ENSEMBL-Compara ML phylogeny



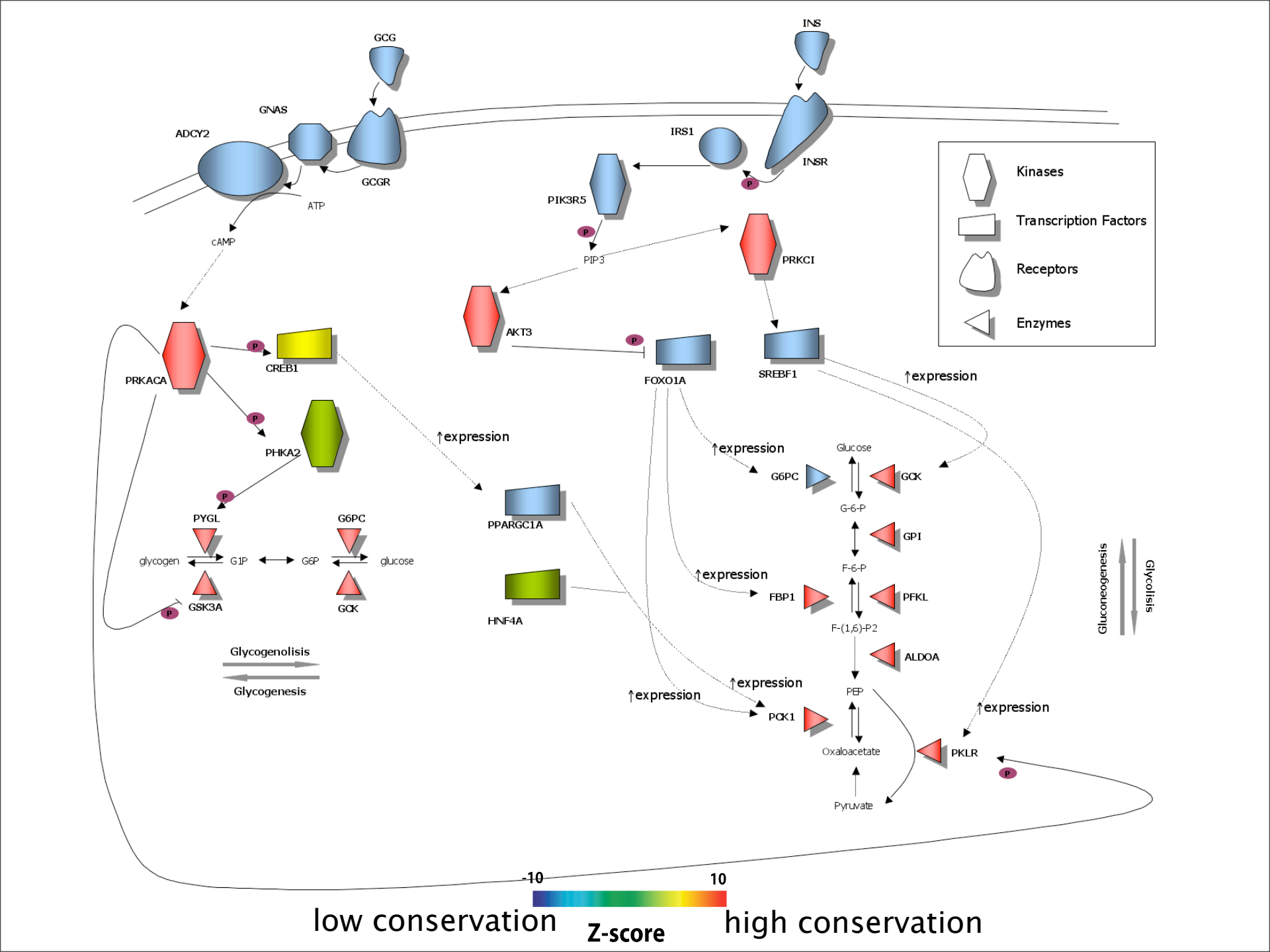
# Methods::four measures

| | CS homologs | CS orthologs | number of genes with homologs | number of genes with orthologs |
|---|---|---|---|---|
| CS homologs | - | 0.904 | 0.027 | 0.331 |
| CS orthologs | | - | 0.059 | 0.014 |
| number of genes with homologs | | | - | 0.007 |
| number of genes with orthologs | | | | - |

*Mus musculus*

| | CS homologs | CS orthologs | number of genes with homologs | number of genes with orthologs |
|---|---|---|---|---|
| CS homologs | - | 0.920 | 0.434 | 0.793 |
| CS orthologs | | - | 0.285 | 0.698 |
| number of genes with homologs | | | - | 0.760 |
| number of genes with orthologs | | | | - |

*Drosophila melanogaster*


GCG
INS
GNAS
ADCY2
GCGR
ATP
cAMP
IRS1
INSR
PIK3R5
PIP3
PRKCI
AKT3
FOXO1A
SREBF1
Kinases
Transcription Factors
Receptors
Enzymes
PRKACA
CREB1
PHKA2
PYGL
G6PC
glycogen
G1P
G6P
glucose
GSK3A
GCK
PPARGC1A
HNF4A
↑expression
Glucose
G6PC
GCK
G-6-P
GPI
F-6-P
FBP1
PFKL
F-(1,6)-P2
ALDOA
PEP
PCK1
Oxaloacetate
PKLR
Pyruvate
Glycogenolisis
Glycogenesis
Gluconeogenesis
Glycolysis
-10
10
low conservation
Z-score
high conservation

# What are the general patterns across different types of genes?

- Human genome – divide genes into
  >300 GO functional categories each with >100 genes


- Molecular function AND biological process

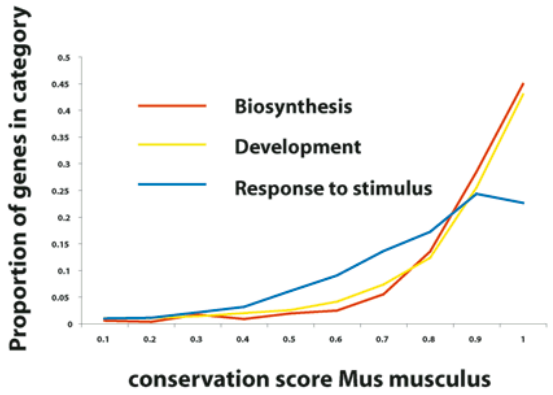
# Some functional classification databases for human

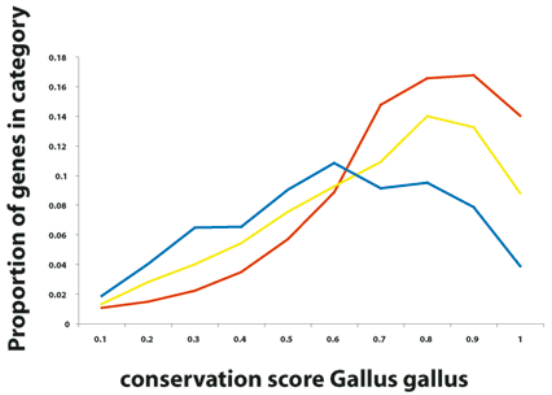
| Name | Description | URL |
|---|---|---|
| ***GO** | Gene Ontology | http://www.geneontology.org |
| ***Panther** | Protein ANalysis THrough Evolutionary Relationships | http://www.pantherdb.org |
| ***KOGs** | Functional categories and orthology groups | http://www.ncbi.nlm.nih.gov/COG |
| **FunCat** | Functional Catalog | http://mips.gsf.de/projects/funcat |
| **KEGG** | Kyoto Encyclopedia of Genes and Genomes | http://www.genome.jp/kegg |
| **BioCyc** | Collection of Pathway/Genome Databases | http://www.biocyc.org/ |
| **EC** | The Enzyme Commission Classification Scheme | http://www.expasy.org/enzyme |
| **REACTOME** | Curated resource of core pathways and reactions | http://www.reactome.org |

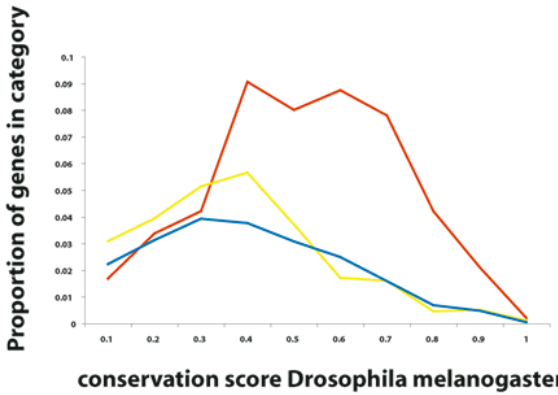
*Used in this work

# Distributions of Conservation Scores

Biological Process



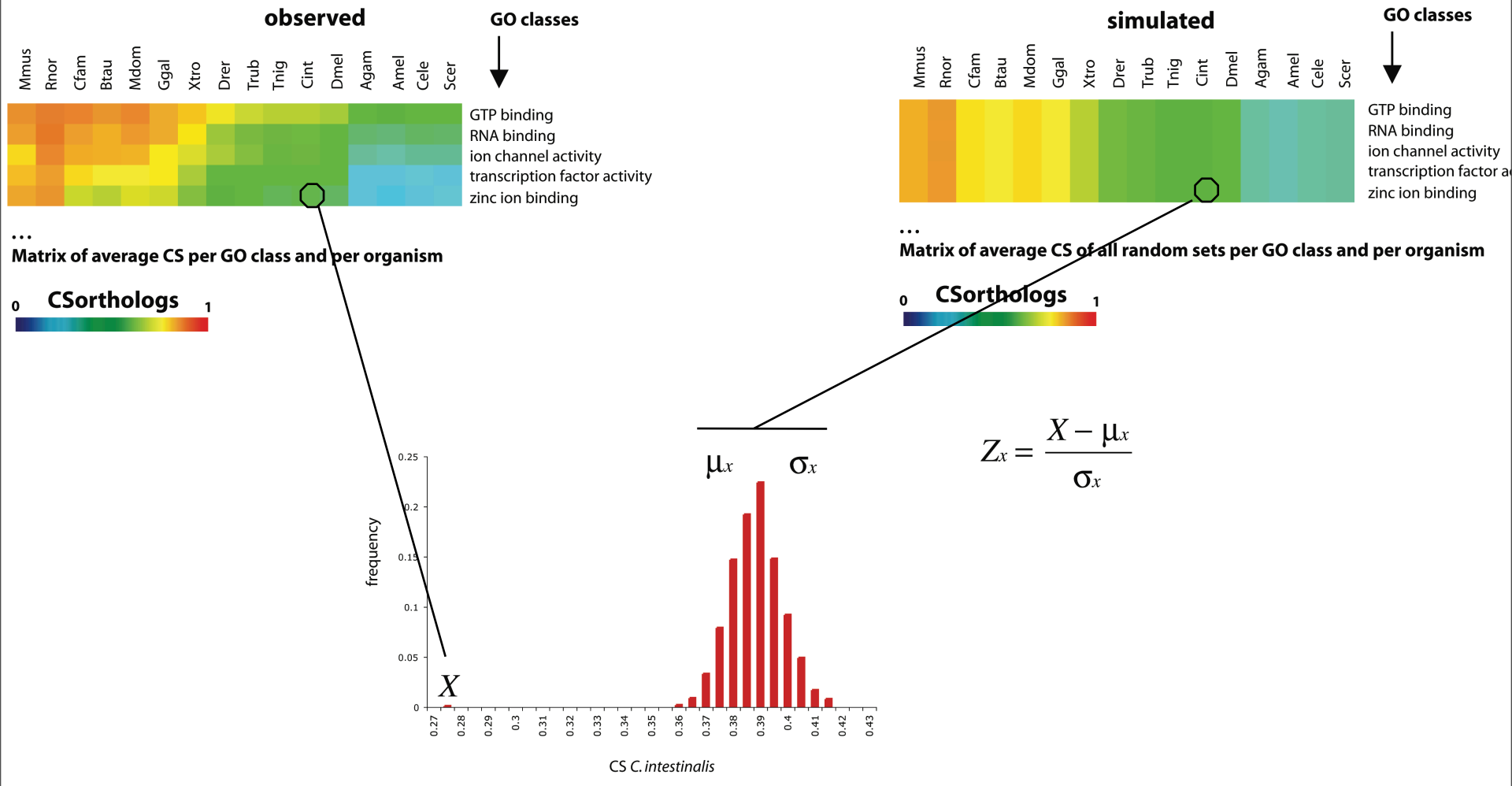




| | Average | Z-score | p-value |
|---|---|---|---|
| Biosynthesis | 0.840 | 8.06 | <1x10-04 |
| Development | 0.814 | 5.78 | <1x10-04 |
| Response to stimulus | 0.736 | -12.34 | <1x10-04 |

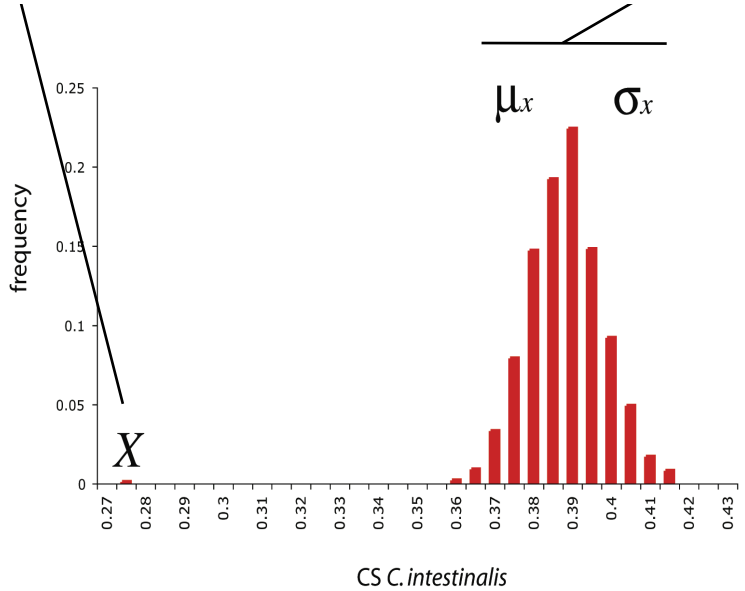
| Average | Z-score | p-value |
|---|---|---|
| 0.698 | 7.35 | <1x10-04 |
| 0.635 | -1.08 | 0.022 |
| 0.551 | -15.21 | <1x10-04 |

| Average | Z-score | p-value |
|---|---|---|
| 0.448 | 11.09 | <1x10-04 |
| 0.285 | -7.30 | <1x10-04 |
| 0.321 | -2.97 | 3x10-03 |

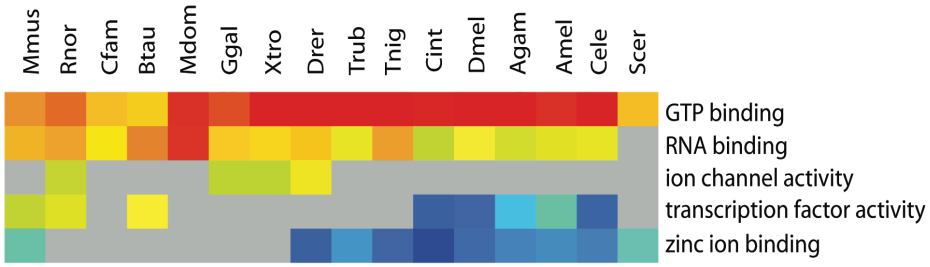
# **FRED**: Functional categories and their Relative Evolutionary Divergence



# **FRED**: Functional categories and their Relative Evolutionary Divergence



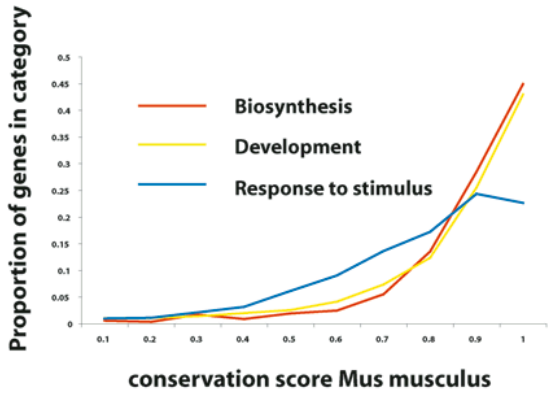

$$Z_x = \frac{X - \mu_x}{\sigma_x}$$

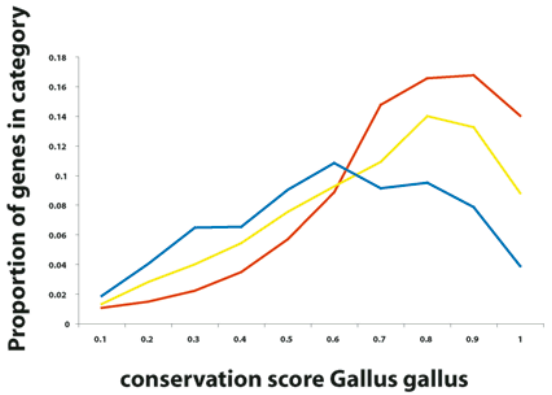


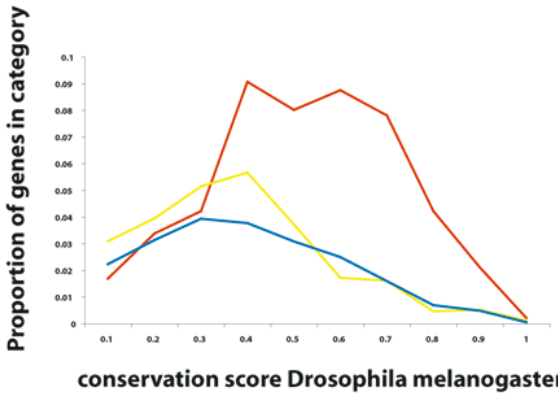

**Matrix of Z-score per GO class and per organism**

# Distributions of Conservation Scores

Biological Process



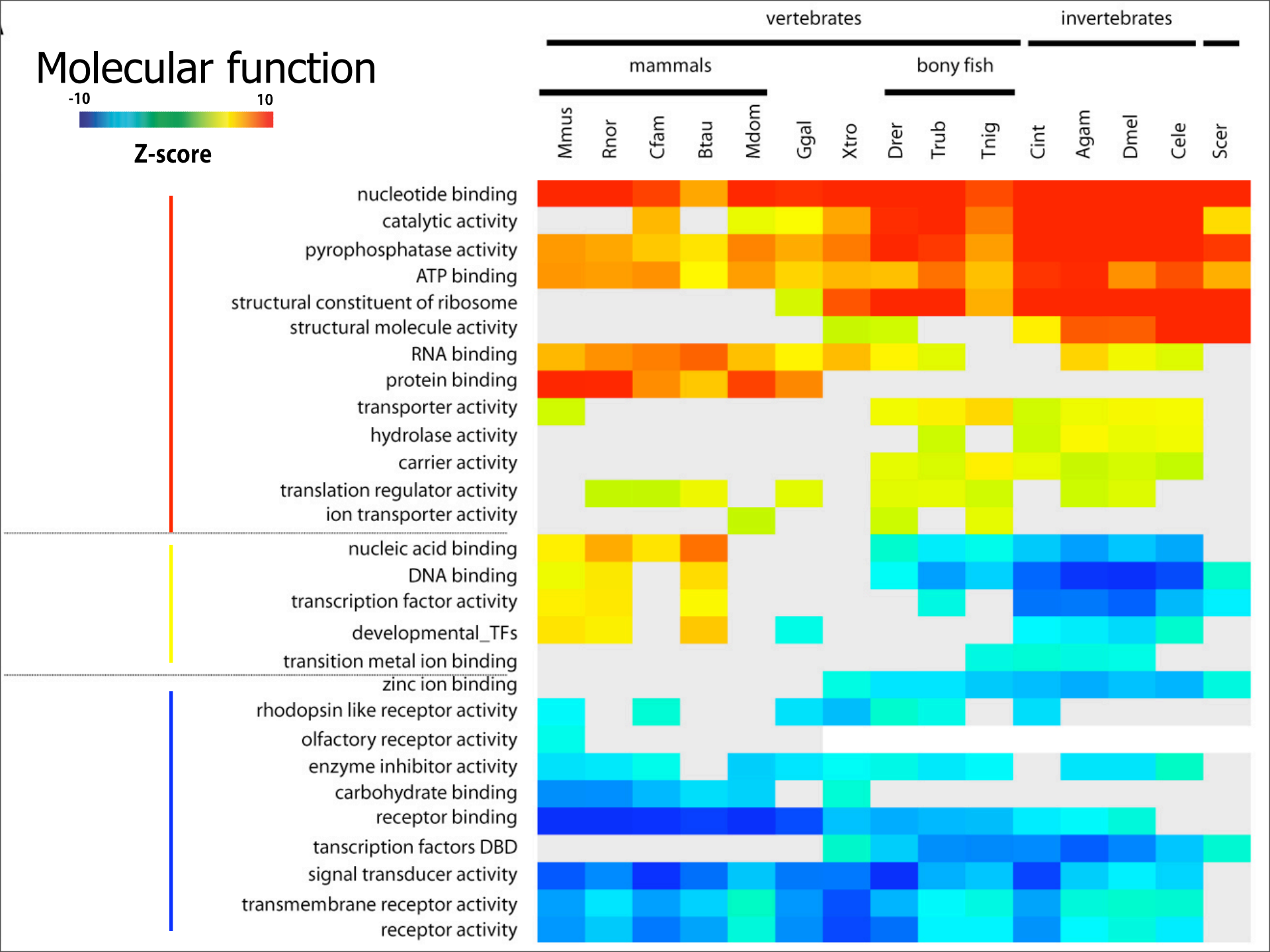




| | Average | Z-score | p-value |
|---|---|---|---|
| Biosynthesis | 0.840 | 8.06 | <1x10-04 |
| Development | 0.814 | 5.78 | <1x10-04 |
| Response to stimulus | 0.736 | -12.34 | <1x10-04 |

| Average | Z-score | p-value |
|---|---|---|
| 0.698 | 7.35 | <1x10-04 |
| 0.635 | -1.08 | 0.022 |
| 0.551 | -15.21 | <1x10-04 |

| Average | Z-score | p-value |
|---|---|---|
| 0.448 | 11.09 | <1x10-04 |
| 0.285 | -7.30 | <1x10-04 |
| 0.321 | -2.97 | 3x10-03 |


Molecular function
-10
10
Z-score
vertebrates
invertebrates
mammals
bony fish
Mmus
Rnor
Cfam
Btau
Mdom
Ggal
Xtro
Drer
Trub
Tnig
Cint
Agam
Dmel
Cele
Scer
nucleotide binding
catalytic activity
pyrophosphatase activity
ATP binding
structural constituent of ribosome
structural molecule activity
RNA binding
protein binding
transporter activity
hydrolase activity
carrier activity
translation regulator activity
ion transporter activity
nucleic acid binding
DNA binding
transcription factor activity
developmental_TFs
transition metal ion binding
zinc ion binding
rhodopsin like receptor activity
olfactory receptor activity
enzyme inhibitor activity
carbohydrate binding
receptor binding
tanscription factors DBD
signal transducer activity
transmembrane receptor activity
receptor activity

# Biological Process


Z-score

# Controls

- CS for homologs from ENSEMBL-Compara
- Presence/absence of orthologues and homologues
- CS for orthologues with same expression breadth across human tissue types
- KOGs functional classification
- CS for orthologues present across eukaryotes from KOGS
- Restrict to orthologues with one-to-one orthology


Development
Transcription Factors
Nucleic Acid Binding
Morphogenesis
Highly conserved core processes and functions
Metabolism
Cell structure
Transport
Physiological processes
Cellular processes
Binding
Transcription Regulation
Immune response
Response to Stimulus
Enzyme Regulation
Regulation of Biological Process
Signal Transducer Activity
Receptors

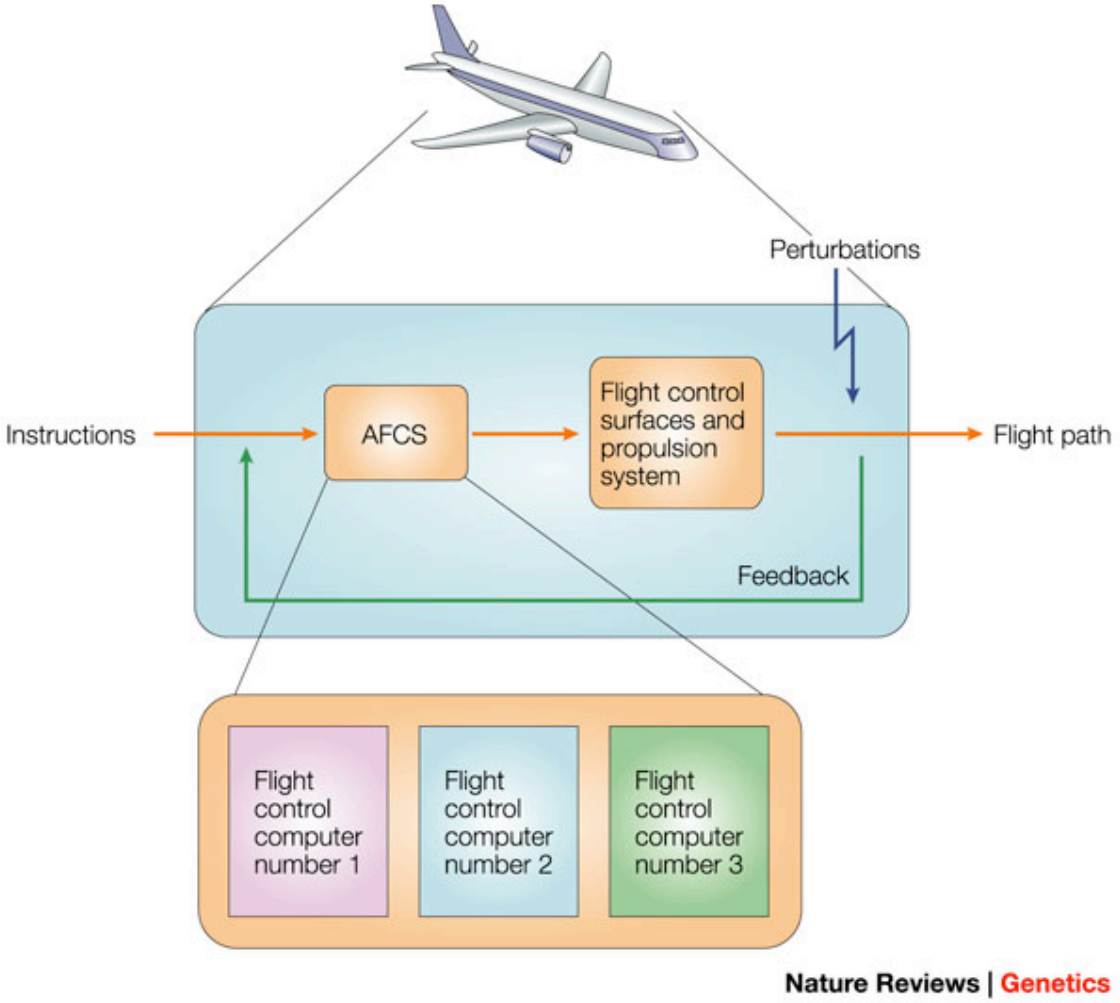
Kitano H. **Biological robustness.**
Nat Rev Genet. 2004 Nov;5(11):826-37.

**Modularity**

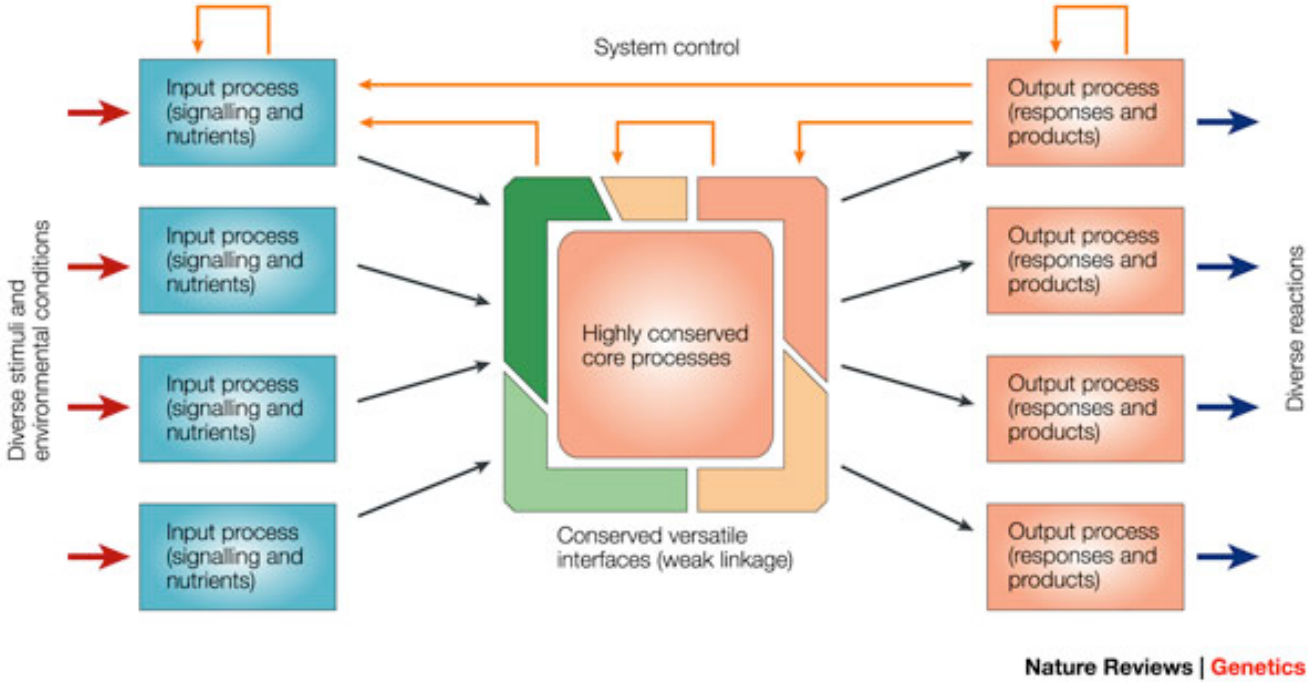
**Functional redundancy**

**Feedback-control systems.**

...



Kitano H. **Biological robustness.**
Nat Rev Genet. 2004 Nov;5(11):826-37.

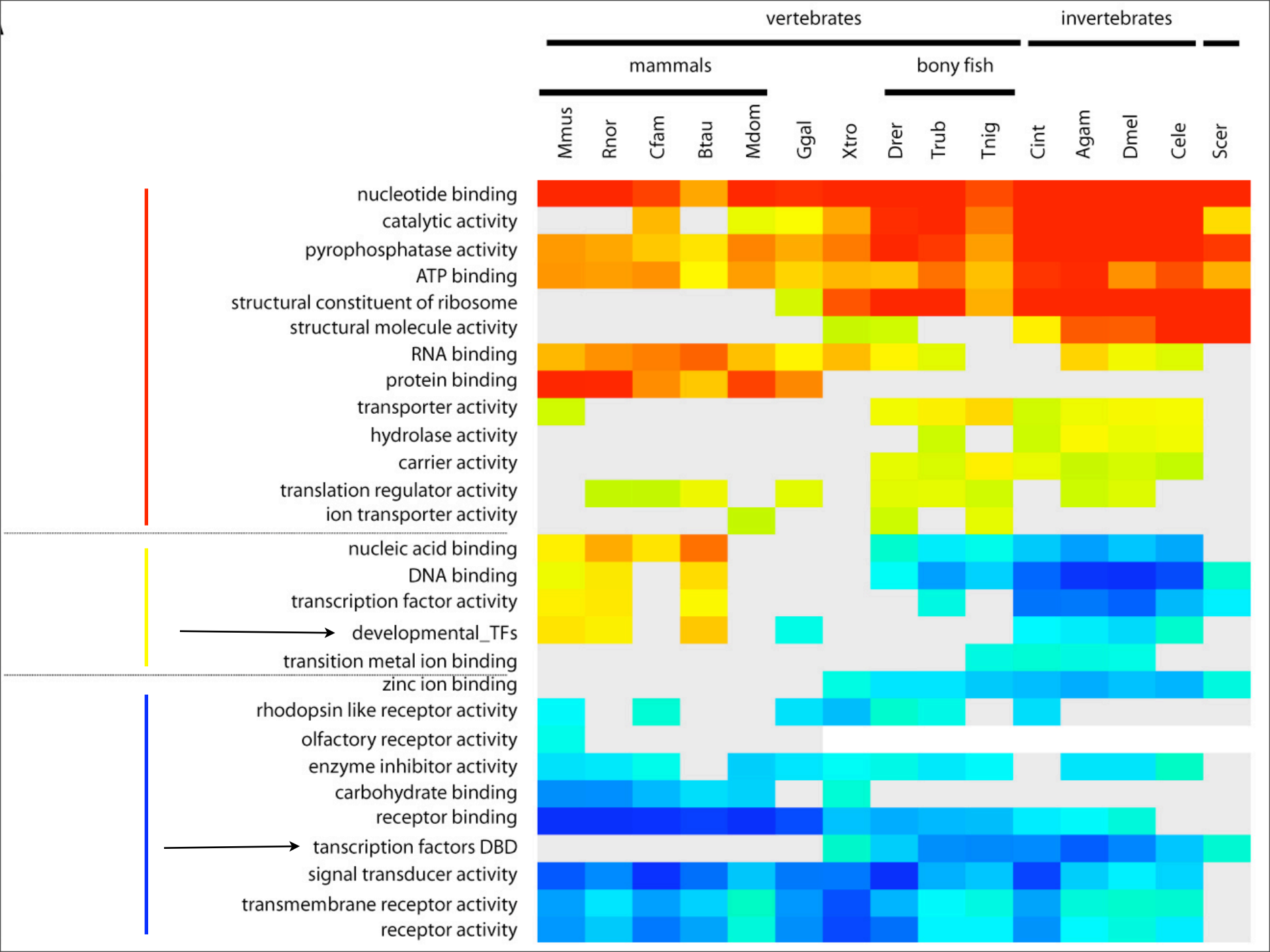



The architectural framework of evolvable systems

# Regulatory Protein Evolution - The Developmental Biology Viewpoint

Carroll SB (2005) Evolution at Two Levels: On Genes and Form. PLoS Biol 3(7): e245

"regulatory proteins are the most slowly evolving of all classes of proteins"


vertebrates
invertebrates
mammals
bony fish
Mmus
Rnor
Cfam
Btau
Mdom
Ggal
Xtro
Drer
Trub
Tnig
Cint
Agam
Dmel
Cele
Scer
nucleotide binding
catalytic activity
pyrophosphatase activity
ATP binding
structural constituent of ribosome
structural molecule activity
RNA binding
protein binding
transporter activity
hydrolase activity
carrier activity
translation regulator activity
ion transporter activity
nucleic acid binding
DNA binding
transcription factor activity
developmental_TFs
transition metal ion binding
zinc ion binding
rhodopsin like receptor activity
olfactory receptor activity
enzyme inhibitor activity
carbohydrate binding
receptor binding
tanscription factors DBD
signal transducer activity
transmembrane receptor activity
receptor activity

# DBD: Transcription factor prediction database

Version 1.2 α

Home | Browse Genomes | Browse Families | Search | About | Download | Links

***Saccharomyces cerevisiae***

Taxonomy: Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae;Saccharomyces

source (downloaded: 2003-03-19)

(Next 50) **Page:** 1 2 3 4 5

| Sequence ID | genome | DB | Domain architecture |
|---|---|---|---|
| YOR156C | sc | SF |   |
| | sc | PF |   |
| YDR409W | sc | SF |   |
| | sc | PF |   |
| YMR003W | sc | SF |   |
| | sc | PF |   |

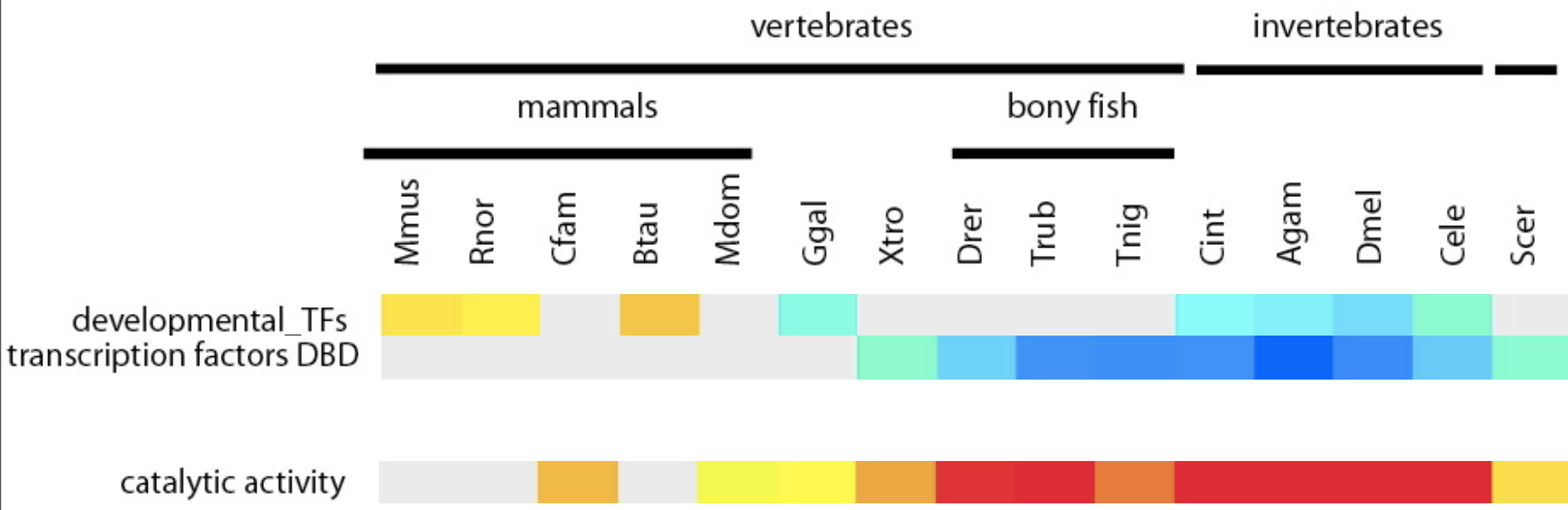
(Next 3) **Page:** 1 2 3 4 5

Authors: Sarah K. Kummerfeld and Sarah A. Teichmann

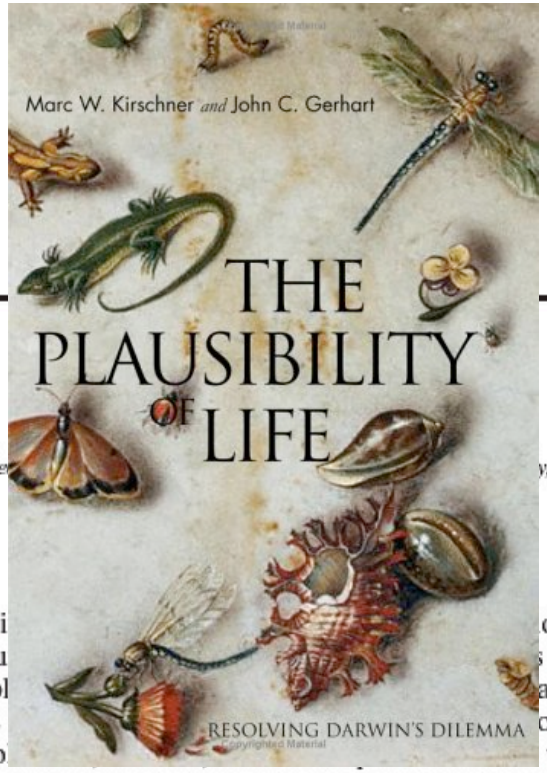
# www.transcriptionfactor.org

Kummerfeld & Teichmann (2006) Nucleic Acids Res 34, D74–81.

# Transcription factors - developmental vs other



*Proc. Natl. Acad. Sci. USA*
Vol. 95, pp. 8420–8427, July 1998




**Perspective**

# Evolvability

*Marc Kirschner*† and John Gerhart‡*

**Department of Cell Biology, Harvard Medical School, Boston, MA 02115; and ‡Department of Mole… CA 94720*

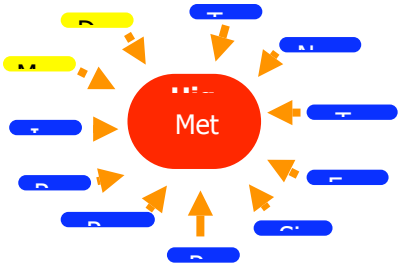
*Contributed by John C. Gerhart, April 7, 1998*

**ABSTRACT** **Evolvability is an organism's capacity to generate heritable phenotypic variation. Metazoan evolution is marked by great morphological and physiological diversification, although the core genetic, cell biological, and developmental processes are largely conserved. Metazoan diversification has entailed the evolution of various regulatory processes controlling the time, place, and conditions of use of the conserved core processes. These regulatory processes, and certain of the core processes, have special properties relevant to evolutionary change. The properties of versatile protein elements, weak linkage, compartmentation, redundancy, and exploratory behavior reduce the interdependence of components and confer robustness and flexibility on processes during embryonic development and in adult physiology. They also confer evolvability on the organism by reducing constraints on change and allowing the accumulation of nonlethal variation. Evolvability may have been generally selected in the course of selection for robust, flexible processes suitable for complex development and physiology and specifically selected in lineages undergoing repeated radiations.**

Evolvabili…odels by several au…s confirm in princip…ation can affect the …olvability at the mo… with the conviction that it is more clearly demonstrable at these levels than at the level of morphology.

It is difficult to evaluate how the particular characteristics of cellular, developmental, and physiological mechanisms affect the quantity and quality of phenotypic variation after genetic change and hence affect evolvability. To understand the consequence of mutation for a protein's activity, one needs to understand the interactions of that protein with many other cell components. A current view is that conserved core processes constrain phenotypic variation, acting as a barrier to evolution (4, 6). Many core processes are conserved throughout metazoa (e.g., many signaling pathways and genetic regulatory circuits), others throughout eukaryotes (e.g., the cytoskeleton and cdk/cyclin-based cell cycle, and yet others throughout all life forms (e.g., metabolism and replication). It is natural to assume that highly conserved mechanisms are

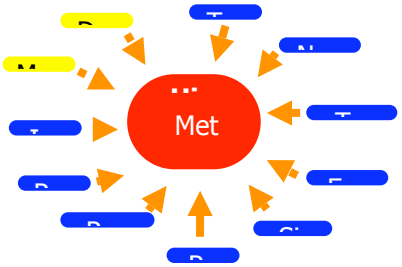
# Summary I – FRED




- Conserved core of structural and metabolic molecules at cell level from human to yeast
- Transcription factors involved in development conserved between human and other mammals
- Rapid sequence divergence of other TFs, receptors, signal transducers, stress and immune response

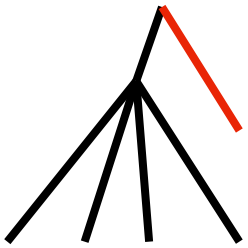
# Questions

- Are there certain groups of proteins with particular functions that evolve faster than others?



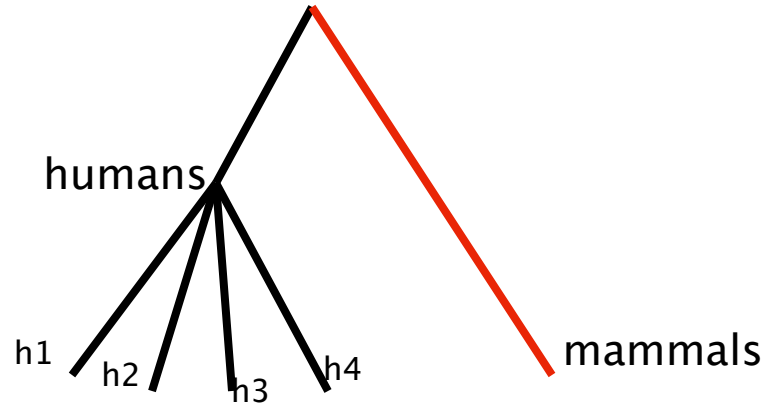

Nuria Lopez-Bigas (UPF, Barcelona)

- What is the selection pressure on different functional categories in the human lineage?



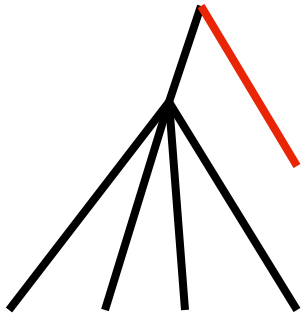
Subhajyoti De (MRC Laboratory of Molecular Biology)

# Possible scenarios of divergence/diversity at a base:



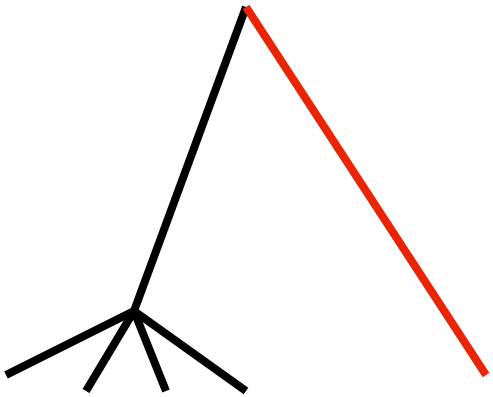

High diversity and high divergence

Weak negative selection



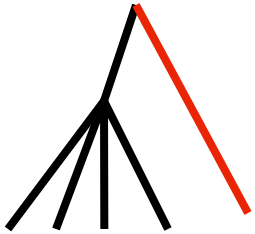
High diversity and low divergence

Human-specific positive selection



Low diversity and high divergence

Human specific fixation

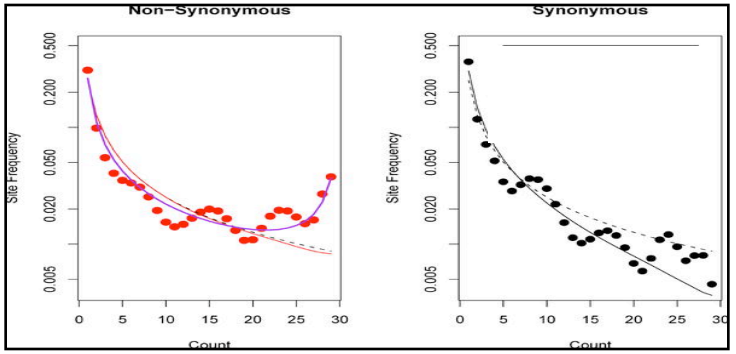


Low diversity and Low divergence

Strong negative selection

# Methods for identifying selection I – dN/dS

- Ratio of nonsynonymous to synonymous substitution rates

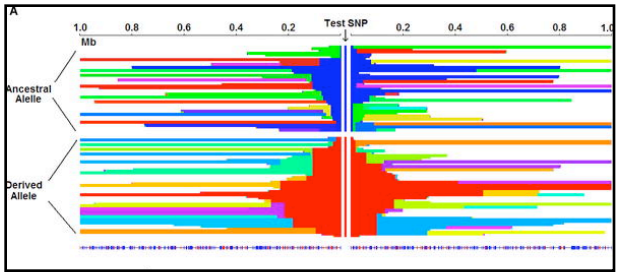



- e.g. Nielsen et al (PLoS Biology, 2005) comparing human and chimpanzee (4-10 Myrs)

- Immunity, reproduction, olfactory genes

# Methods for identifying selection II

- McDonald - Kreitman test: compares nonsynonymous to synonymous mutation rate distributions
- e.g. Bustamante et al (Nature, 2005) resequencing genes in human individuals
- Linkage disequilibrium: recombination rate for a base
- e.g. Voight et al (PLoS Biology, 2006)- very recent evolution

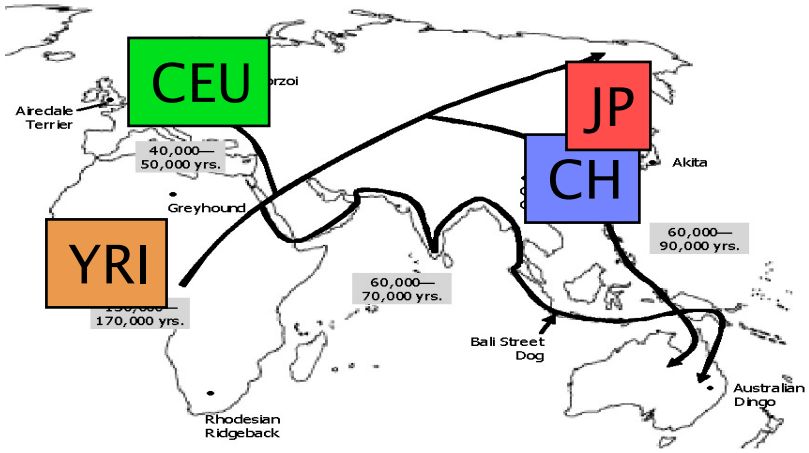


# Single Nucleotide Polymorphisms (SNPs) in humans

HapMap Project:

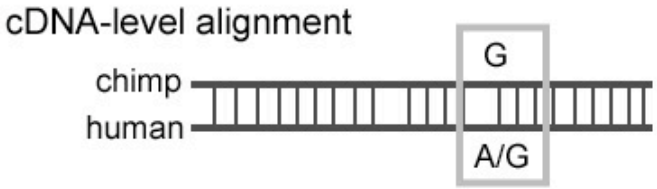
Sequencing genomic positions every 5kb for 270 individuals:

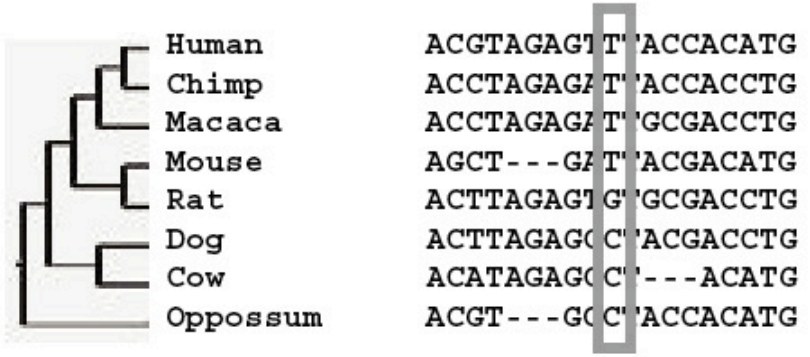
- Han Chinese population from Beijing (CH).
- Japanese population from Tokyo (JP)
- Yoruba population from Ibadan (YRI)
- Caucasian European population from Utah (CEU)



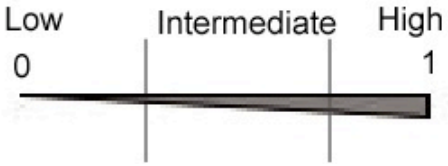

⇒Allele frequencies for each SNP in all four

# BaseDiver: Comparing Divergence and Diversity at each Base

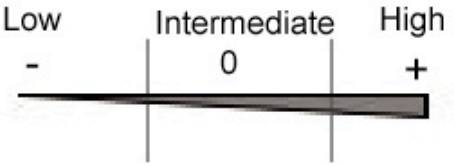


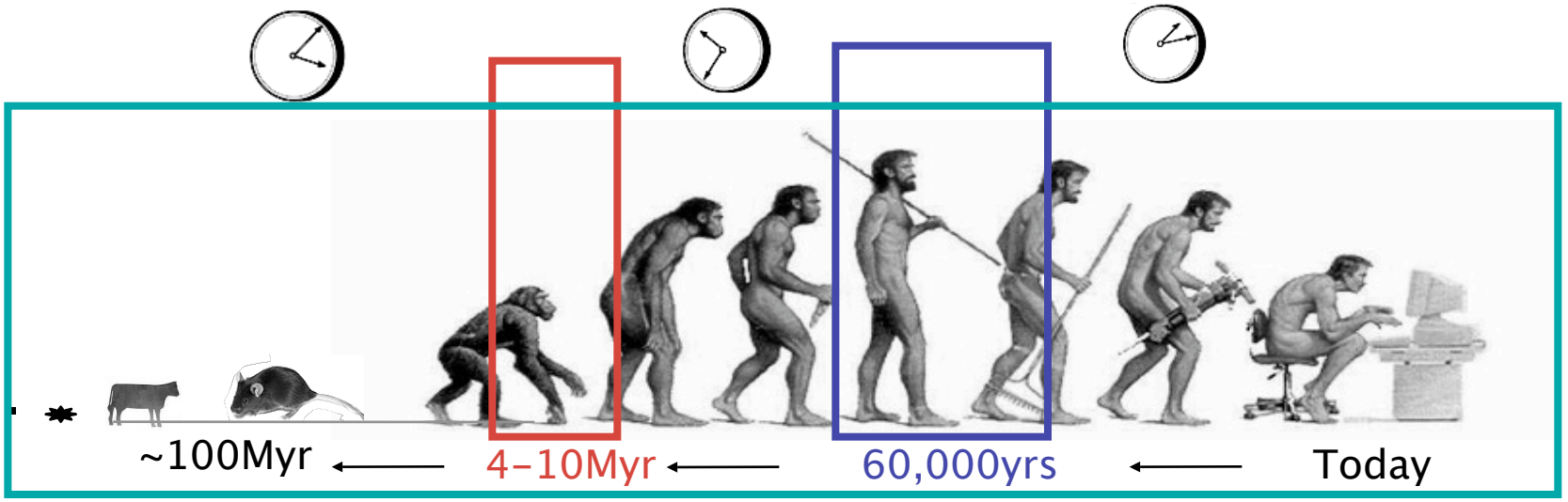

Diversity: extent of deviation in human population from the ancestral base (Derived allele frequency: DAF)




Divergence: mutation rate in mammals at a base position (GERP score; Cooper et al. 2005)



# Methods for identifying selection – BaseDiver




Nielsen et al (PLoS Biology, 2005) and Bustamante et al (Nature, 2005).

Voight et al (PLoS Biology, 2006)

BaseDiver: variation within humans and across mammals

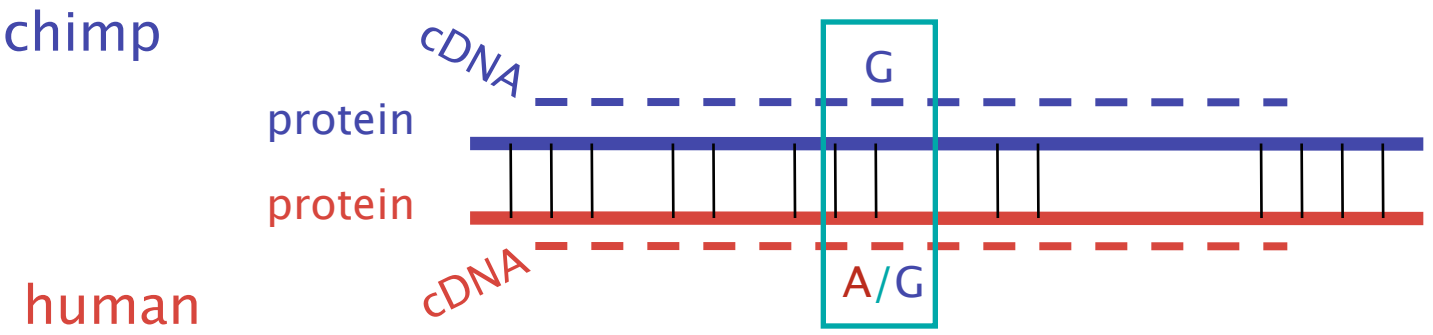
# Measure of divergence at each base position

GERP score: Cooper et al (Genome Res., 2005)


Calculated using 8 mammalian genomes for all coding positions.

Conserved position: GERP score <0
Neutral position: GERP score ~0
Divergent position: GERP score >0

# Diversity – Derived Allele Frequency (DAF)

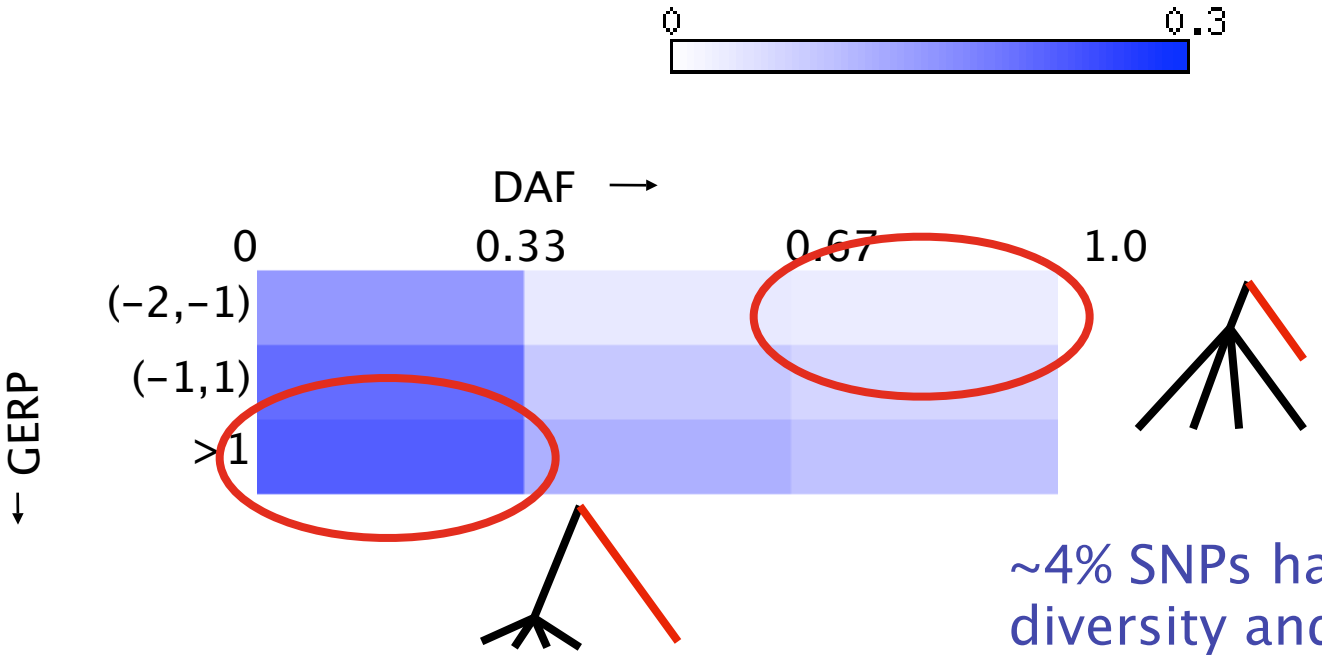



The allele present in the chimp genome was considered to be the ancestral allele.

The derived allele frequency – DAF – is a measure of diversity and varies from 0 to 1.

About 8000 SNPs in our human protein data set from HapMap.

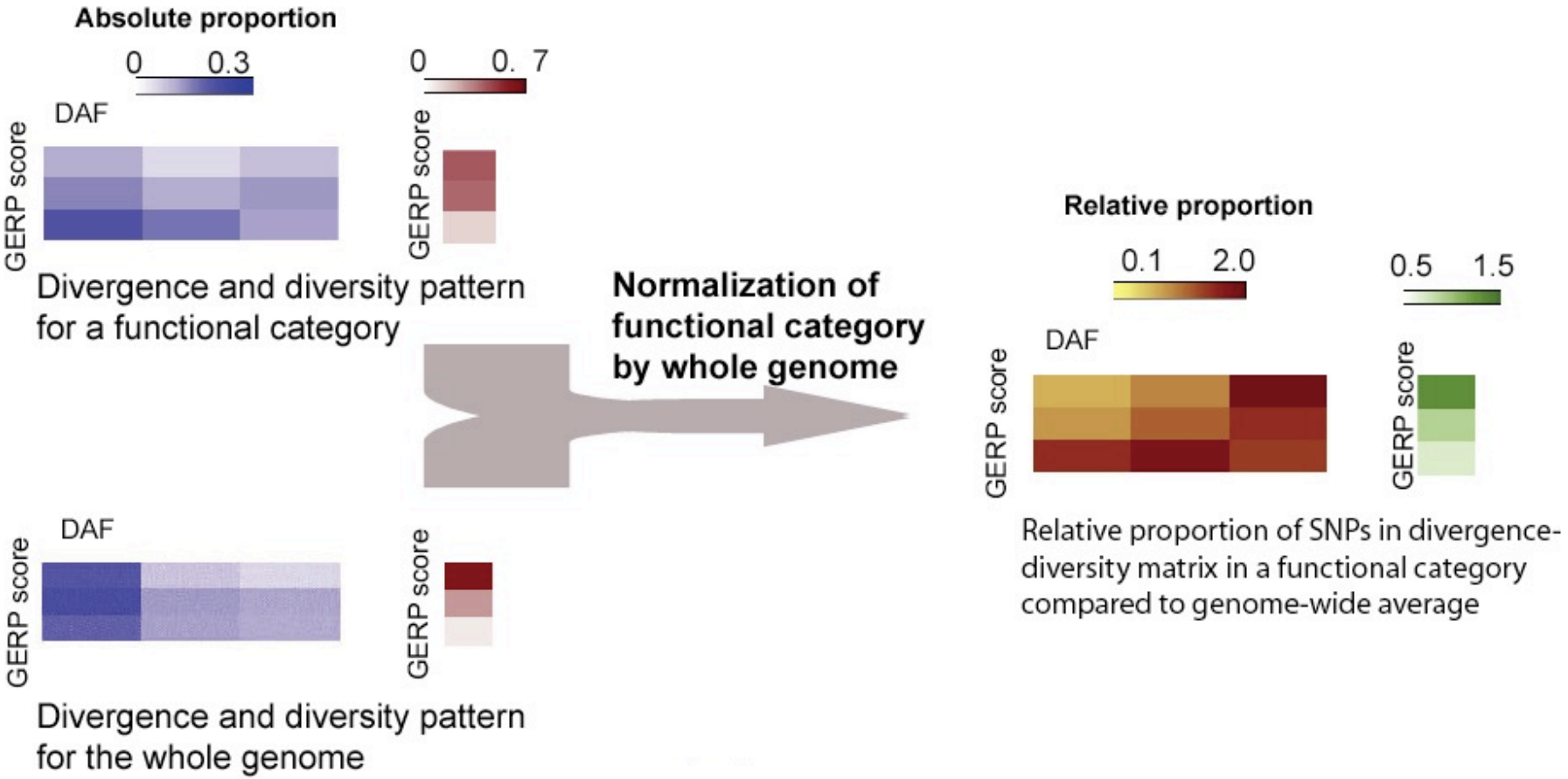
# Diversity and Divergence: genome-wide frequency distribution of SNPs




~4% SNPs have high diversity and low divergence

~20% SNPs have low diversity and high divergence

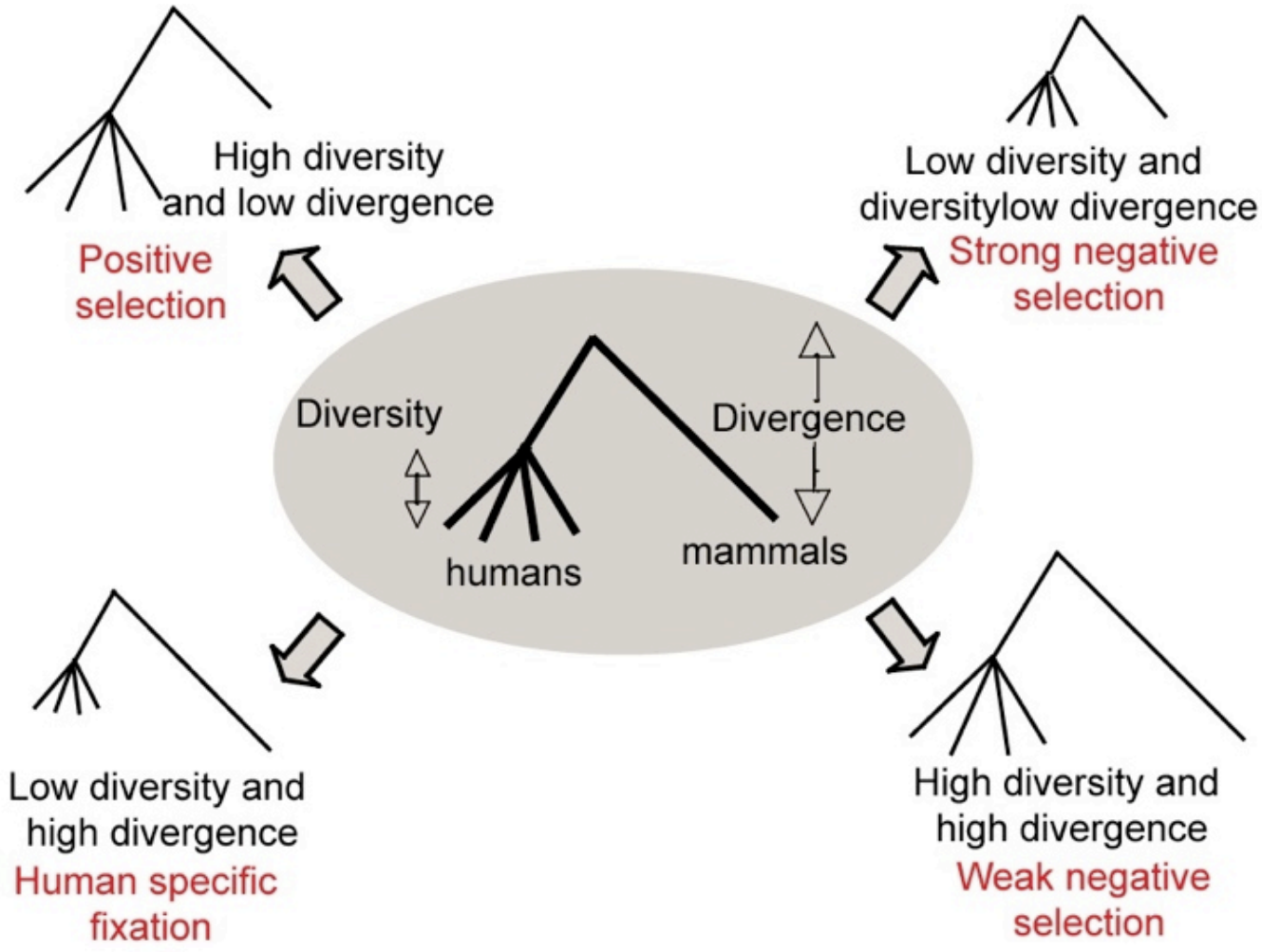
# Assessing modes of selection



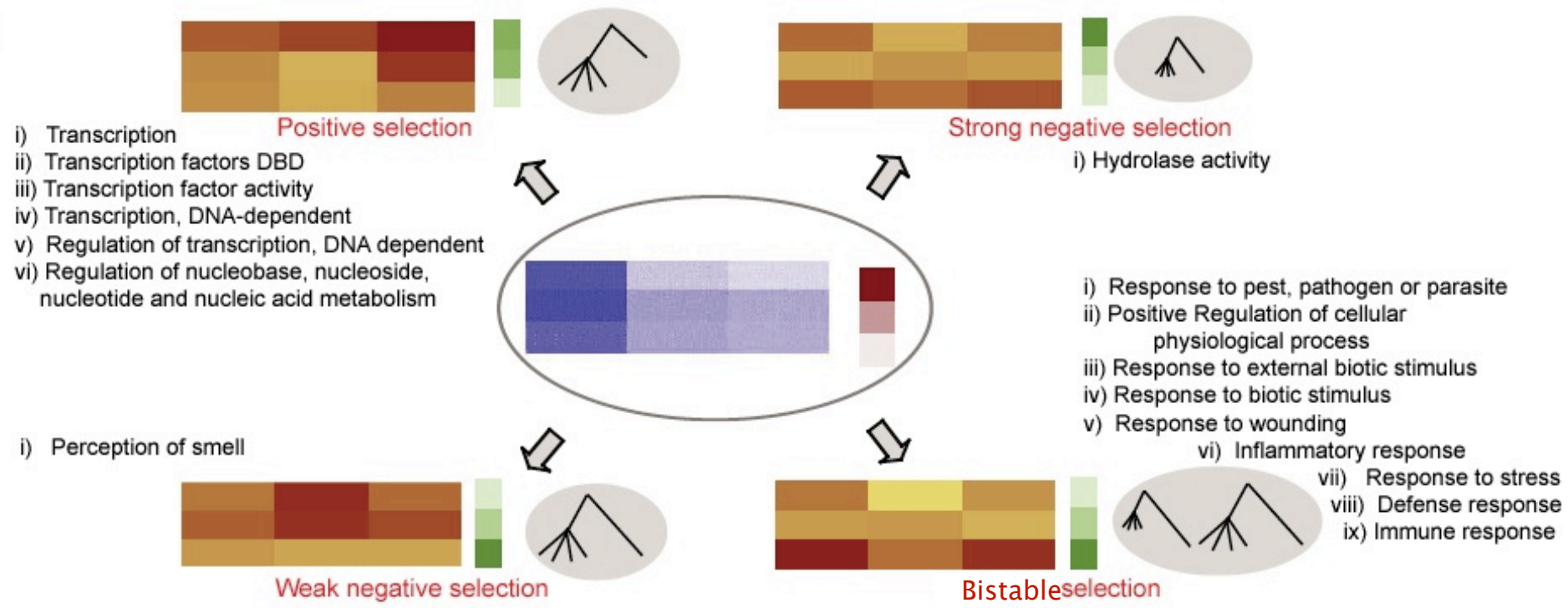

1. Significant difference between functional category and genome-wide distributions?

2. Inferring mode of selection from DAF-GERP a GERP distributions

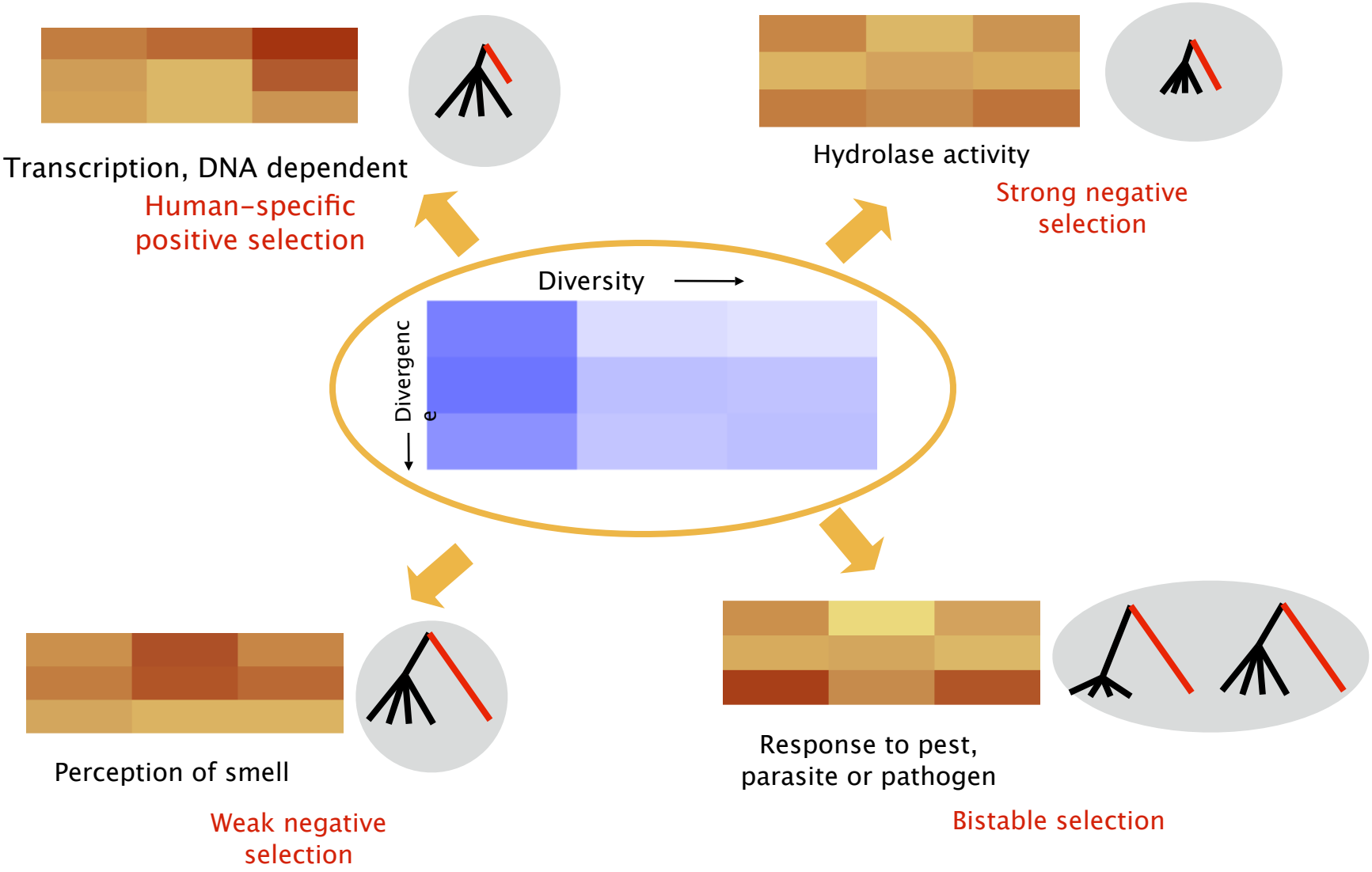
# Four possible modes of selection



# Signatures of selection on different functional categories



# Signatures of selection on different functional categories

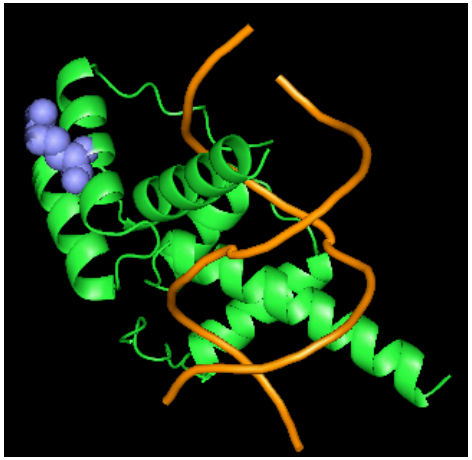


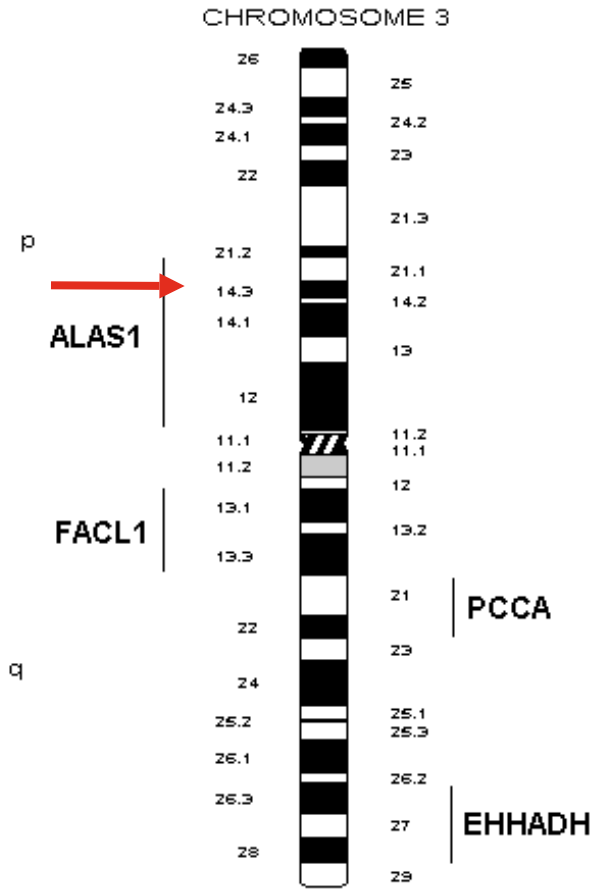
# HESX1

Development of forebrain, eye, pituitary glands etc.
Related to de Morsier Syndrome.

DAF = 0.35 in YRI
No contact with DNA

GERP score: -1.92






dN/dS = 3/3 =1
Situated near HAR-10 (Chr 3:p26.3)
(Pollard et al, Nature, 2006)

Repressor expressed early in development. Targets are unknown.

# Controls

- Bonferroni correction
- Consistent difference of a functional category in 3 out of 4 HapMap populations
- Use Panther functional classification – consistent with GO
- Use only one-to-one orthologues
- Compare to findings of other studies

# Summary – Positive or Negative Selection?

Of categories conserved in protein sequence:

Strong negative selection:

Catalytic activity, catabolism, hydrolase act

Of divergent categories:

Odorant receptors

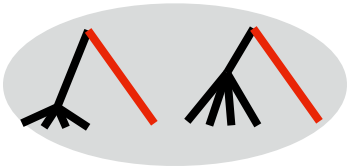
Weak negative selection:

Transcriptional regulation

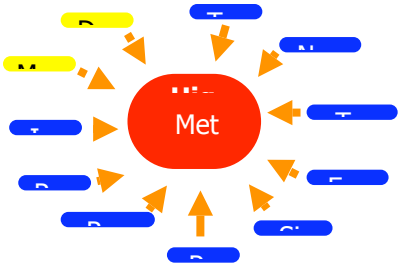
Positive selection:

Stress response, immune response

Bistable selection:



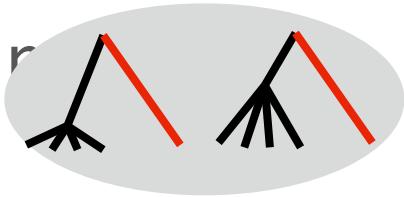
# Summary I – FRED




- Conserved core of structural and metabolic molecules at cell level from human to yeast
- Transcription factors involved in development conserved between human and other mammals
- Rapid sequence divergence of other TFs, receptors, signal transducers, stress and immune response

# Summary II – BaseDiver

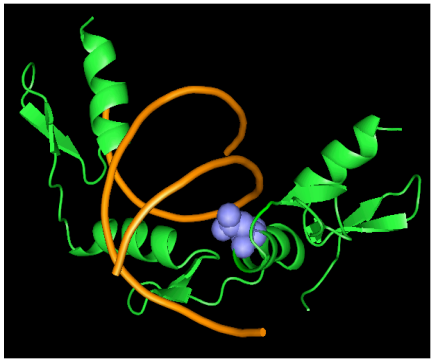
- General method applicable to any (SNP) sequence data – independent of assumptions of neutral model
- Transcription factors under positive selection in human lineage
- Stress response and immune response u bistable selection



# Acknowledgements

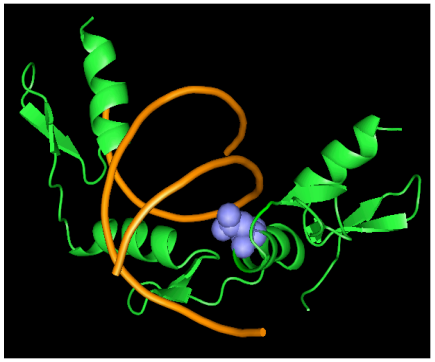

- Nuria Lopez-Bigas, UPF, Barcelona
- Subhajyoti De, MRC-LMB, Cambridge

# ZNF228



High expression in smooth muscles and hematopoietic cells.

# ZNF228

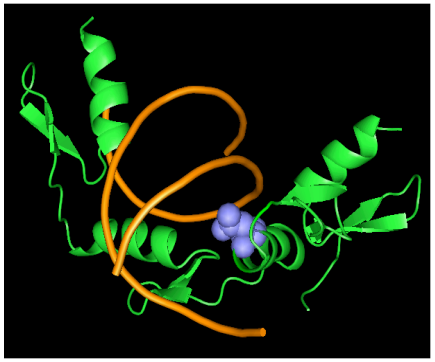


High expression in smooth muscles and hematopoietic cells.

Diversity ~ (0.5 - 0.75)  Divergence: -0.81

One N-Syn SNP in DNA-binding residue.

# ZNF228

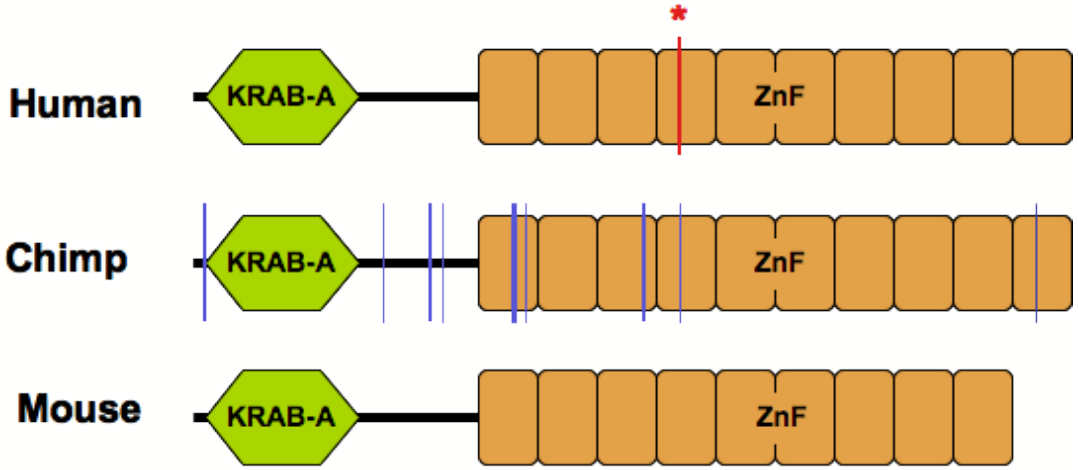


High expression in smooth muscles and hematopoietic cells.

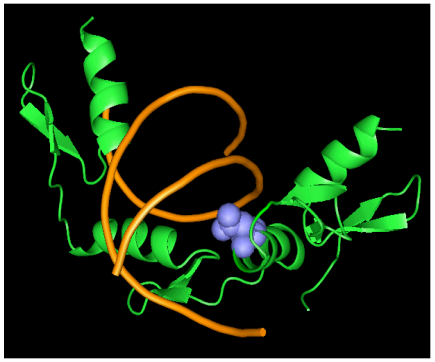
Diversity ~ (0.5 – 0.75)  Divergence: -0.81

One N-Syn SNP in DNA-binding residue.

Additional zinc finger domain in primates.



# ZNF228



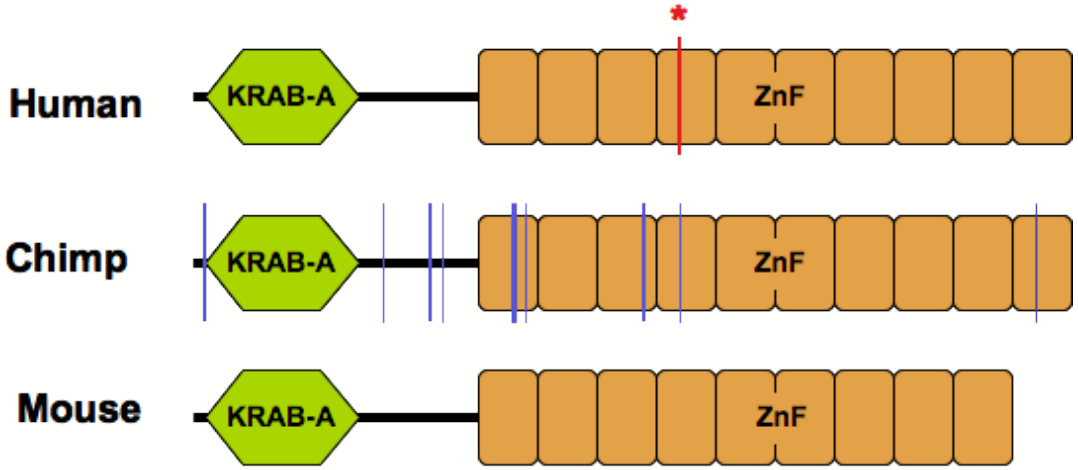
High expression in smooth muscles and hematopoietic cells.

Diversity ~ (0.5 – 0.75)  Divergence: -0.81

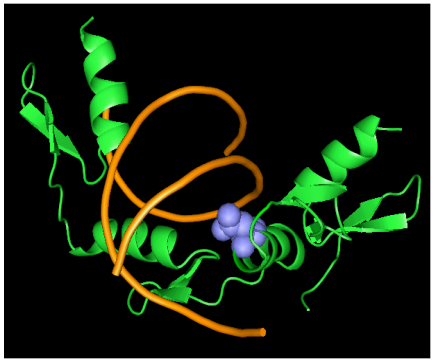
One N-Syn SNP in DNA-binding residue.

Additional zinc finger domain in primates.

High LD value in Asian and European population showing very recent selection (Voight et al, PLoS Biol. 2006)



# ZNF228



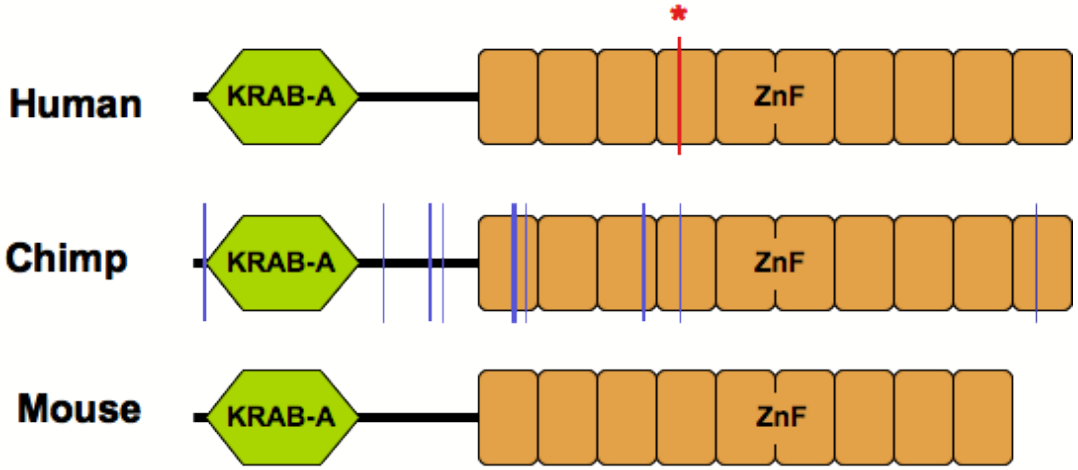
High expression in smooth muscles and hematopoietic cells.

Diversity ~ (0.5 – 0.75)  Divergence: -0.81

One N-Syn SNP in DNA-binding residue.

Additional zinc finger domain in primates.

High LD value in Asian and European population showing very recent selection (Voight et al, PLoS Biol. 2006)




Putative TF expressed in CD4+ cells but downstream targets are unknown (Gomes et al, Blood, 2002)