Phenotype Data as an invaluable source for the exploration of biology in the post-genomic era

Ben-Yang Liao (廖本揚)
Division of Biostatistics & Bioinformatics
Institute of Population Health Sciences
National Health Research Institutes

Outline

Section 1:

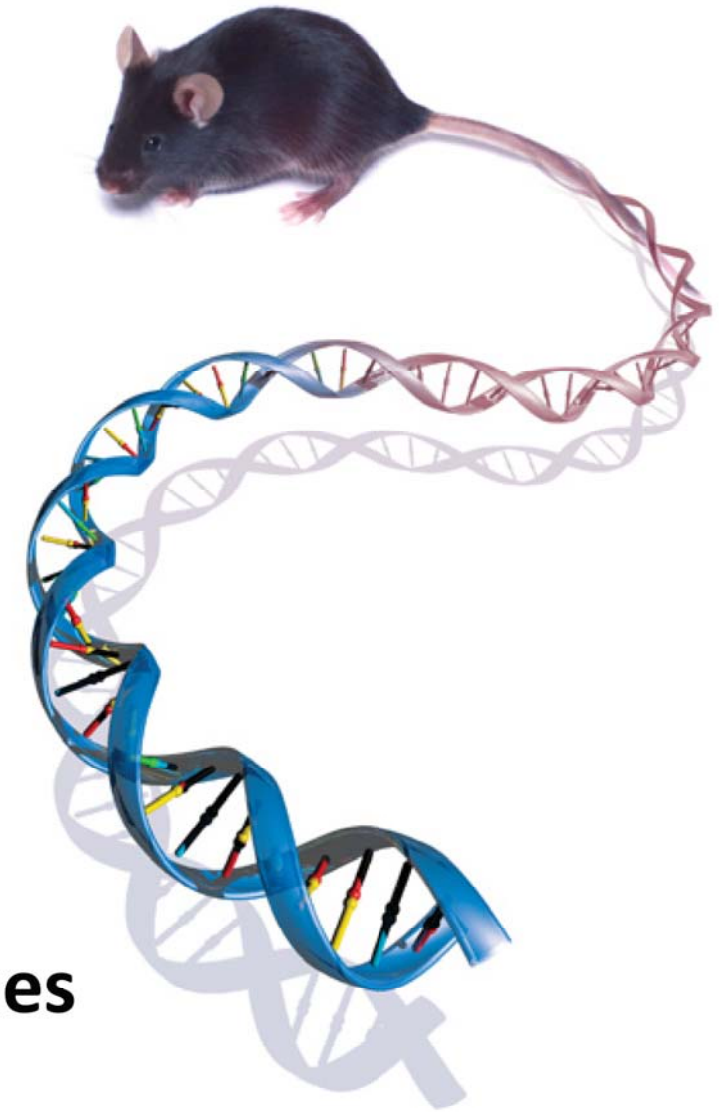
- Phenotype records in major biological data repositories

Section 2:

- Example studies that have utilized the phenotype data

Section 3:

- Developed tools for gene set enrichment analysis on phenotypes



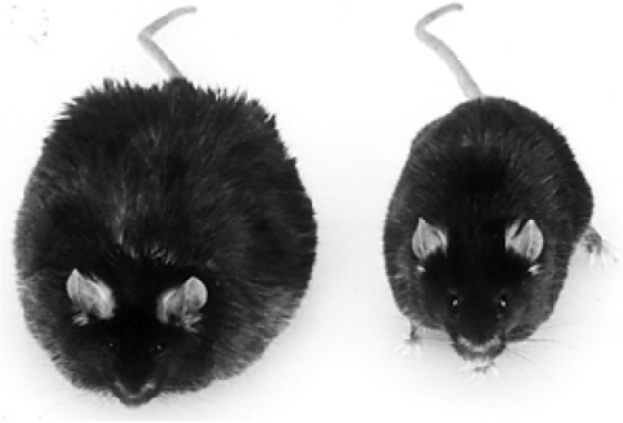
Section 1

Phenotype records in major biological data repositories

- mouse as an example (MGI)

Phenotype:

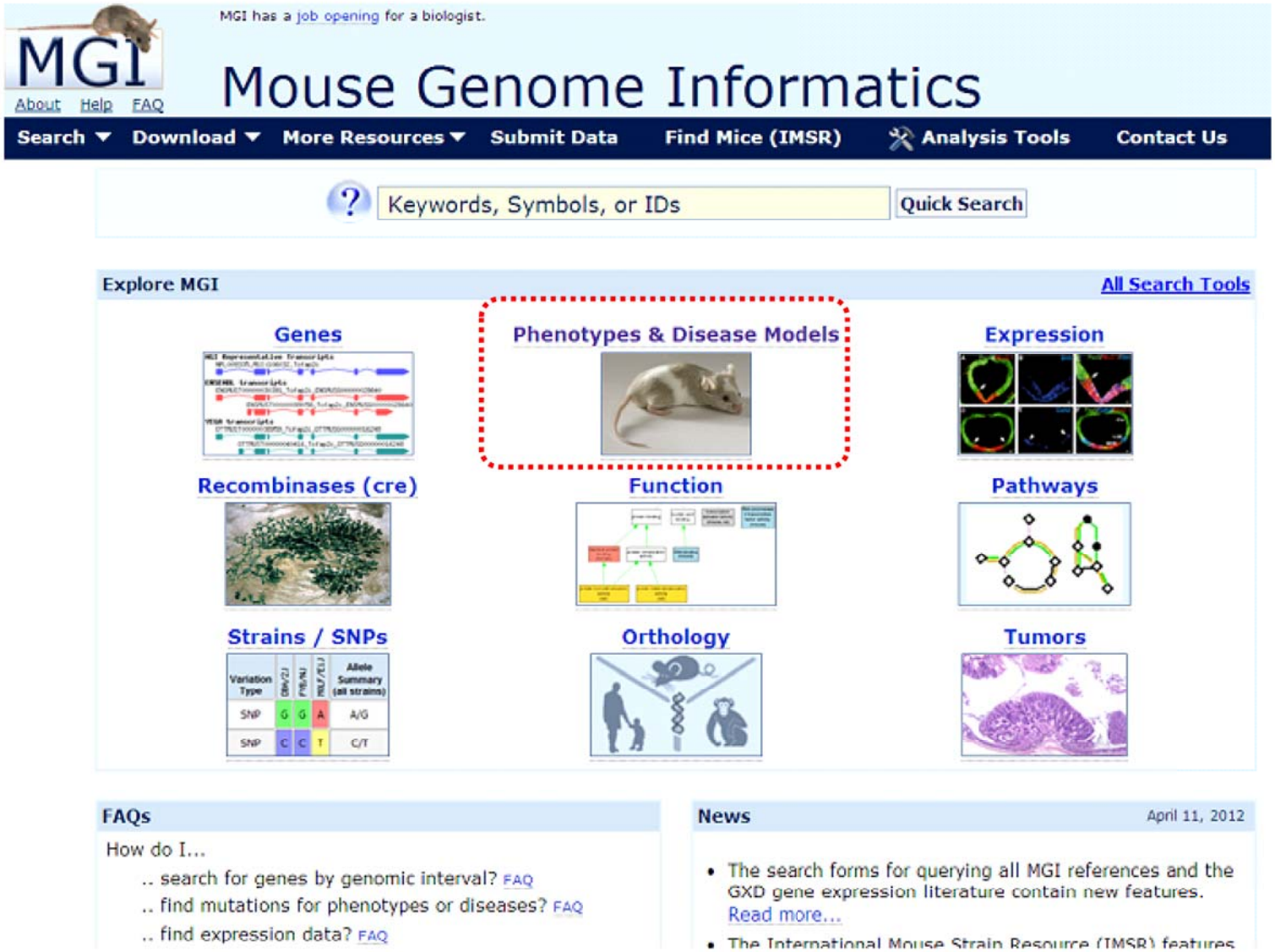
The composite of an organism's observable characteristics or traits, such as its morphology, development, physiological properties, behavior etc.



Obese Leptin Knockout Mouse (left)
Normal wildtype mouse (right)

Mouse Genome Informatics:

http://www.informatics.jax.org/



Hierarchical Structure of Phenotypic “Codes”

Section 2

Example studies that have utilized the mouse phenotype data

Example #1

PNAS PNAS PNAS

Null mutations in human and mouse orthologs frequently result in different phenotypes

Ben-Yang Liao and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

Edited by David J. Lipman, National Institutes of Health, Bethesda, MD, and approved March 10, 2008 (received for review January 14, 2008)

One-to-one orthologous genes of relatively closely related species are widely assumed to have similar functions and cause similar phenotypes when deleted from the genome. Although this assumption is the foundation of comparative genomics and the basis for the use of model organisms to study human biology and disease, its validity is known only from anecdotes rather than from systematic examination. Comparing documented phenotypes of null mutations in humans and mice, we find that >20% of human essential genes have nonessential mouse orthologs. These changes of gene essentiality appear to be associated with adaptive evolution at the protein-sequence, but not gene-expression, level. Proteins localized to the vacuole, a cellular compartment for waste management, are highly enriched among essentiality-changing genes. It is probable that the evolution of the prolonged life history in humans required enhanced waste management for proper cellular function until the time of reproduction, which rendered these vacuole proteins essential and generated selective pressures for their improvement. If our gene sample represents the entire genome, our results would mean frequent changes of phenotypic effects of one-to-one orthologous genes even between relatively closely related species, a possibility that should be considered in comparative genomic studies and in making cross-species inferences of gene function and phenotypic effect.

biomedical research, the international genetics community recently initiated the Knockout Mouse Project (KOMP) to individually knock out every gene in the mouse genome and acquire phenotypic data (16). Our analysis will be valuable in guiding the proper use of the KOMP data.

In the present study, we focus on one of the most dramatic types of change in a gene's phenotypic effect, namely, a change in gene essentiality. A gene is said to be essential to an organism if the loss of its function renders the fitness of the organism zero; otherwise, the gene is said to be nonessential. We show that >20% of human essential genes have nonessential mouse orthologs and elucidate the mechanisms underlying the changes of gene essentiality in evolution.

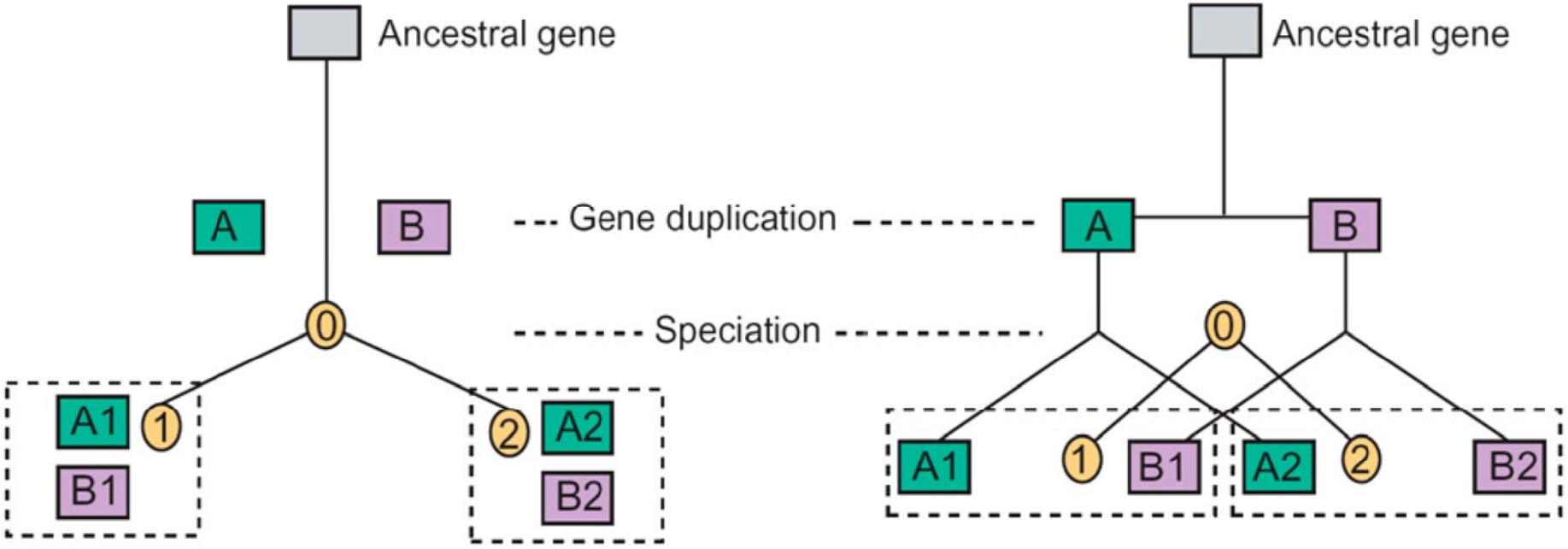
Results and Discussion

Many Human Essential Genes Have Nonessential Mouse Orthologs. From Online Mendelian Inheritance in Man (OMIM) (19), we identified 1,716 human genes with clear gene–disease associations, in which 1,450 genes have unambiguous one-to-one orthologs in the mouse genome (see *Methods*). This set contains 756 human genes whose mouse orthologs have been experimentally deleted with the resulting phenotypes cataloged in the database of Mouse Genome Informatics (MGI). For the 594 human genes associated with mild

EVOLUTION

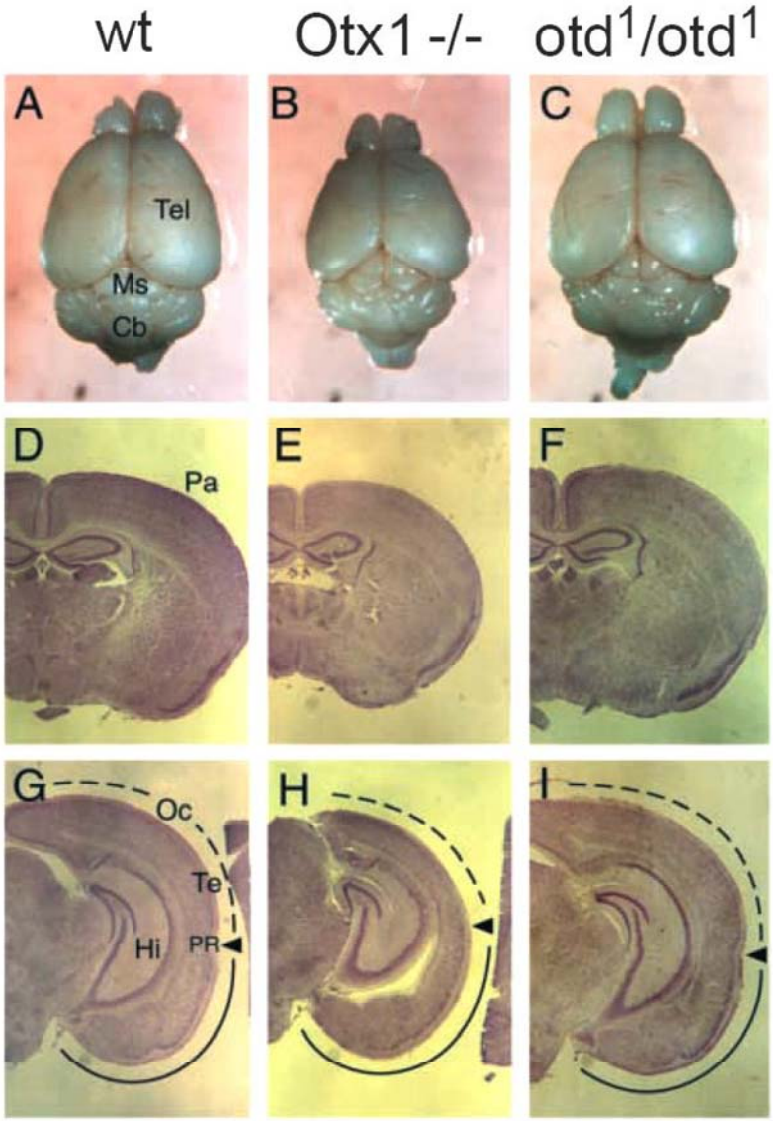
Orthologs are originated from speciation …

and are thought to be functionally equivalent



from Jensen RA (2001)

Orthologs are originated from speciation …
and are thought to be functionally equivalent



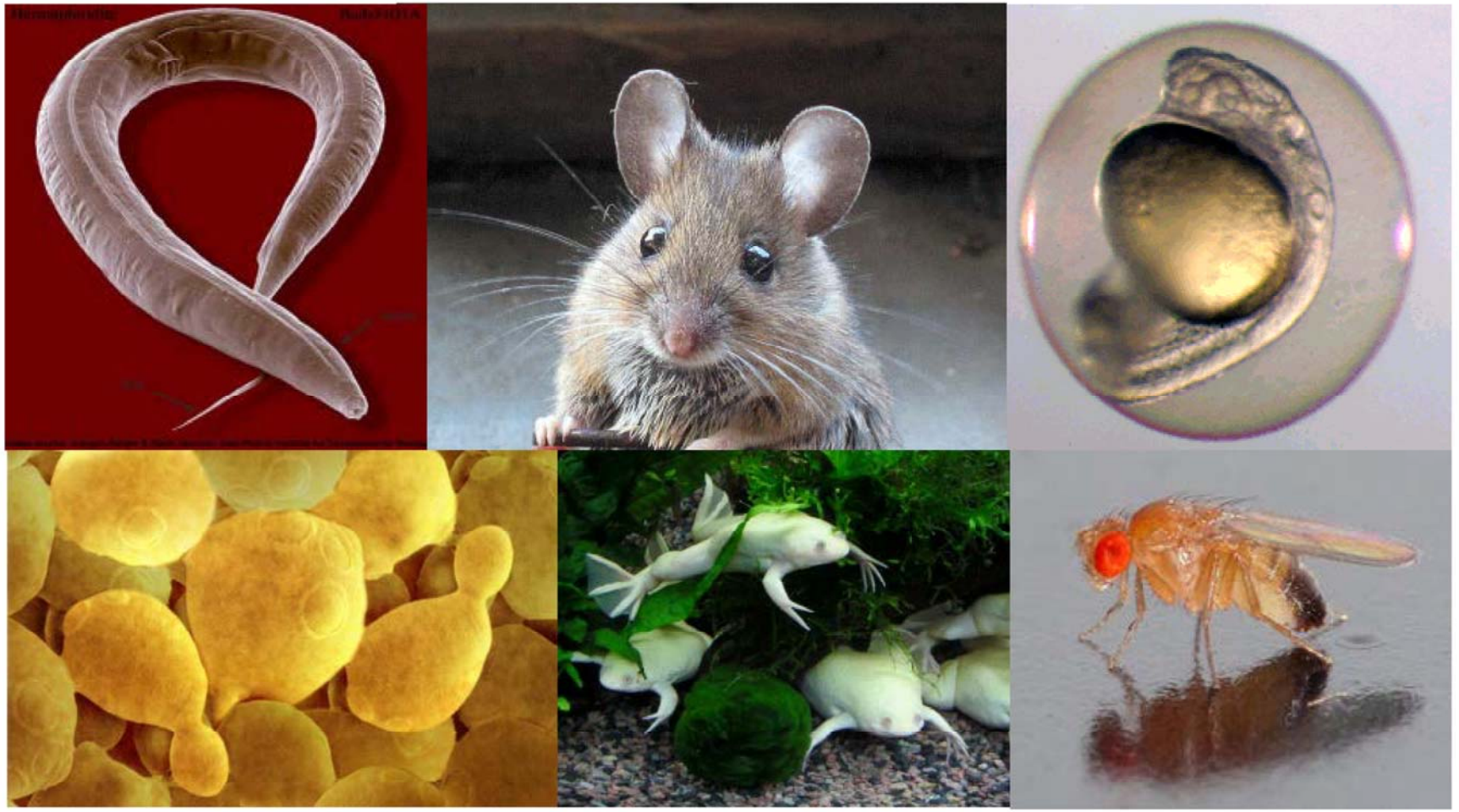
***Otx1*: mouse allele**

***otd*: fruitfly allele**

Epilepsy and corticogenesis defects due to the absence of *Otx1* were fully rescued in homozygous *otd* mice

Acampora et al. (1998) PNAS

Functional equivalence of orthologs is the fundamental assumption of using model organisms to study human biology and diseases, but this assumption has never been systematically verified.



Two main questions to be addressed in comparing humans and mice:

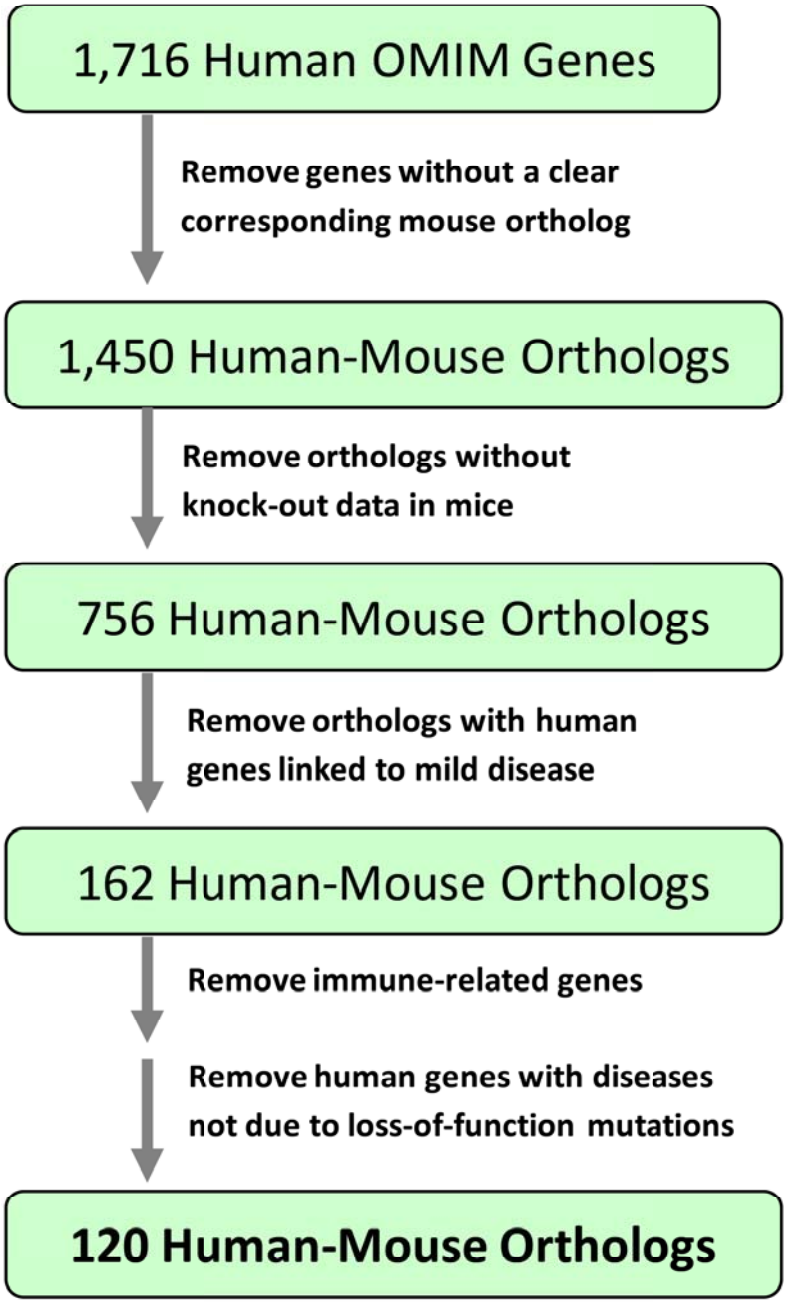
How different in terms of knockout effect the human-mouse orthologous genes are?

Criteria: change of gene essentiality

What factors associate with such differences?

Examined factors: gene duplication, coding-sequence change, gene expression change

Human Essential Genes and the Essentiality of Mouse Orthologs



Among these 120 human-mouse orthologs:

93 Human-Essential-Mouse-Essential (H_1M_1) orthologs

27 Human-Essential-Mouse-Nonessential (H_1M_0) orthologs

22.5%

19.2%

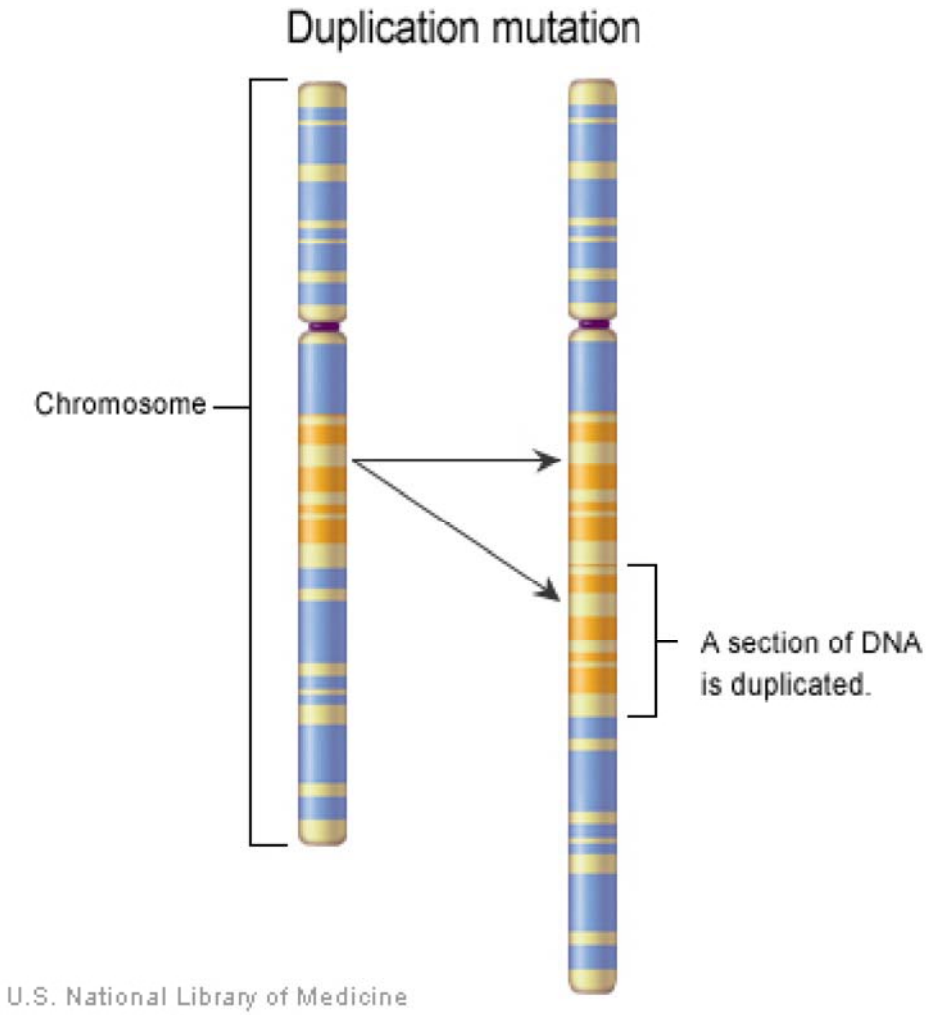
24 Mouse knockout strains can breed as successful as the wild types at least before 6 months

Examples of the essentiality change between human-mouse orthologs

Gene Symbol	Human Disease	Knockout Mouse Phenotypes
DNAH11	Primary ciliary dyskinesia; Kartagener syndrome	normal fertility; abnormal left-right axis patterning
ST3GAL5	Amish infantile epilepsy syndrome	normal viability & fertility; hypoglycemia; increased insulin sensitivity; abnormal lipid level

What factors associate with the change of essentiality?

Lineage (mouse)-specific gene duplication?



(backup because of redundancy)

What factors associate with the change of essentiality?

Lineage (mouse)-specific gene duplication? **NO**

a. Higher proportion of genes with paralogs in H_1M_0 group?

H_1M_0: 18/27=66.7%; H_1M_1: 55/93=59.1% (P = 0.48, χ^2 test)

b. More paralogs found in H_1M_0 duplicates?

H_1M_0: avg. = 4.33; H_1M_1: avg. = 3.78 (P = 0.415, Mann-Whitney U test)

c. More similar the closest paralog for H_1M_0 duplicates?

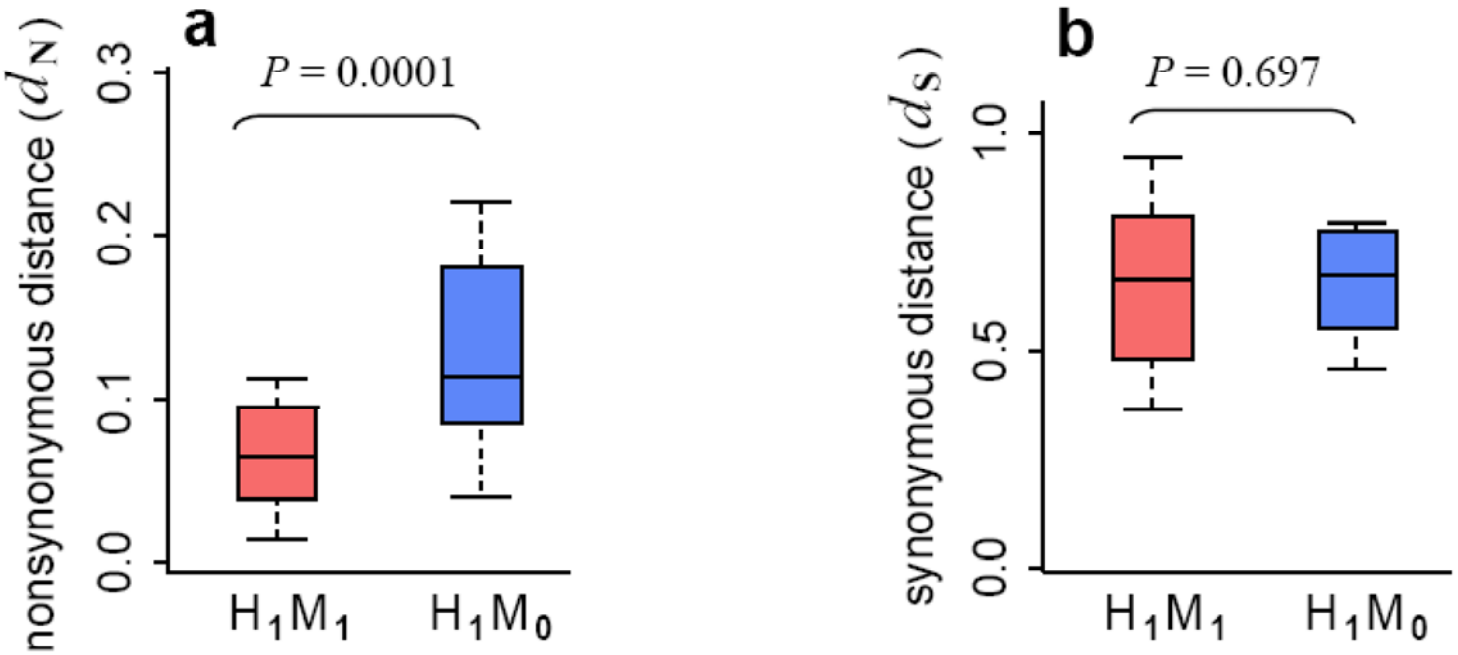
H_1M_0 : pep%=56.2%; H_1M_1: pep%=58.3% (P = 0.568, Mann-Whitney U test)

H_1M_0 : Human-Essential-Mouse-Nonessential orthologs

H_1M_1: Human-Essential-Mouse-Essential orthologs

What factors associate with the change of essentiality?

Protein sequence divergence?

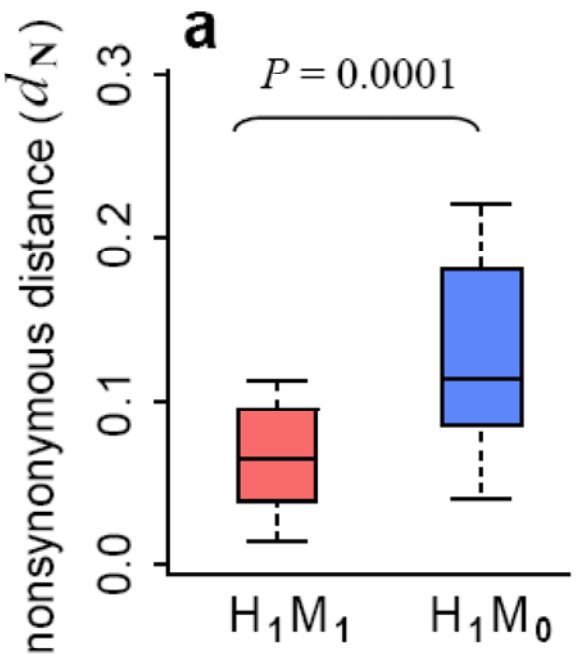


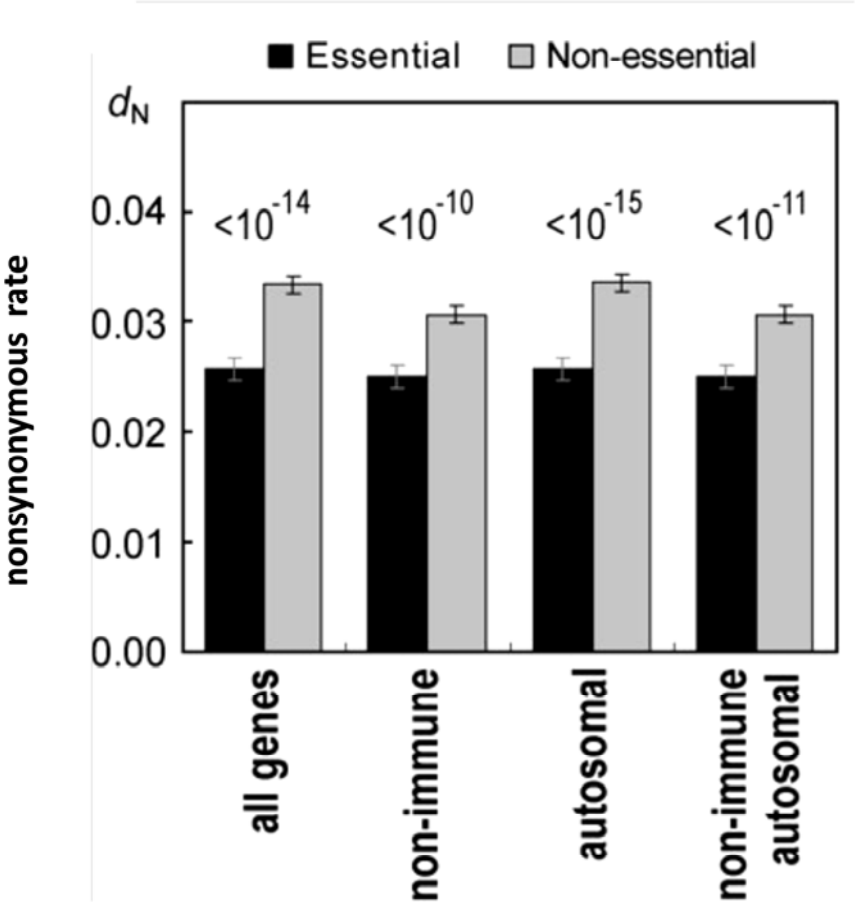
H_1M_0: Human-Essential-Mouse-Nonessential orthologs

H_1M_1: Human-Essential-Mouse-Essential orthologs

What factors associate with the change of essentiality?

Protein sequence divergence?





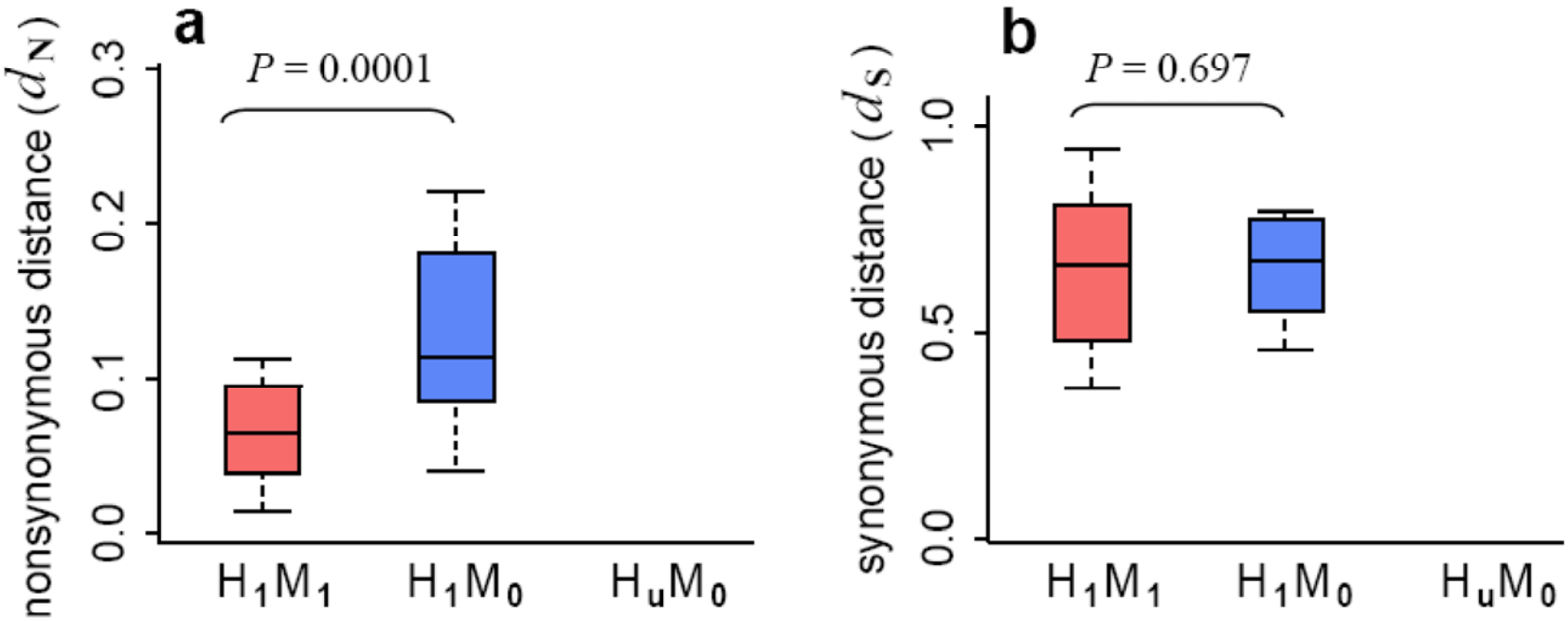
Nonessential genes evolve rapidly!

Liao and Zhang (2006) *Mol Biol Evol* **23**: 2072-2080.

What factors associate with the change of essentiality?

Protein sequence divergence?

YES



H_1M_0: Human-Essential-Mouse-Nonessential orthologs

H_1M_1: Human-Essential-Mouse-Essential orthologs

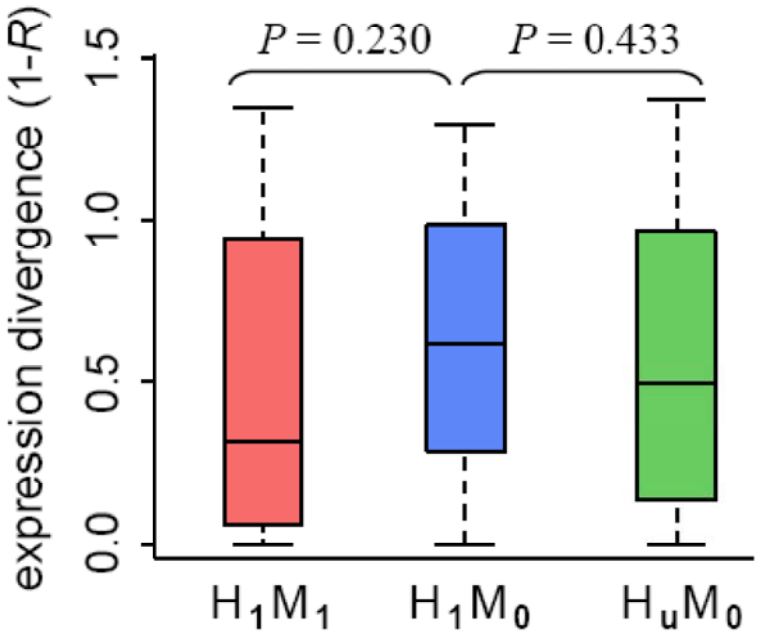
H_uM_0: Human-unknown-Mouse-Nonessential orthologs

What factors associate with the change of essentiality?

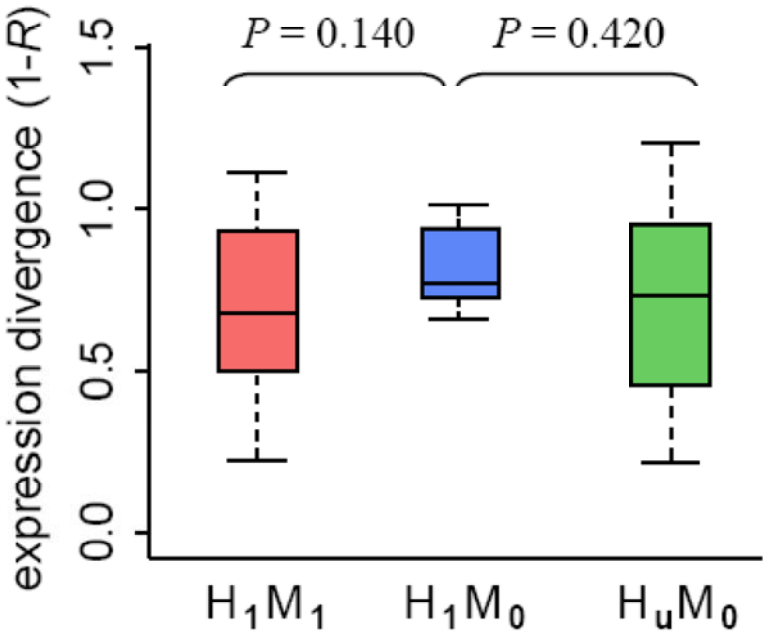
Gene expression divergence?

NO

Dataset: Gene Atlas v2



Dataset: ExonArray



H_1M_0: Human-Essential-Mouse-Nonessential orthologs

H_1M_1: Human-Essential-Mouse-Essential orthologs

H_uM_0: Human-unknown-Mouse-Nonessential orthologs

Functional difference between H_1M_0 and H_1M_1 genes

GENE ONTOLOGY (GO)

Biological Process:

No difference between H_1M_0 genes and H_1M_1 genes

Molecular Function:

No difference between H_1M_0 genes and H_1M_1 genes

Cellular Component:

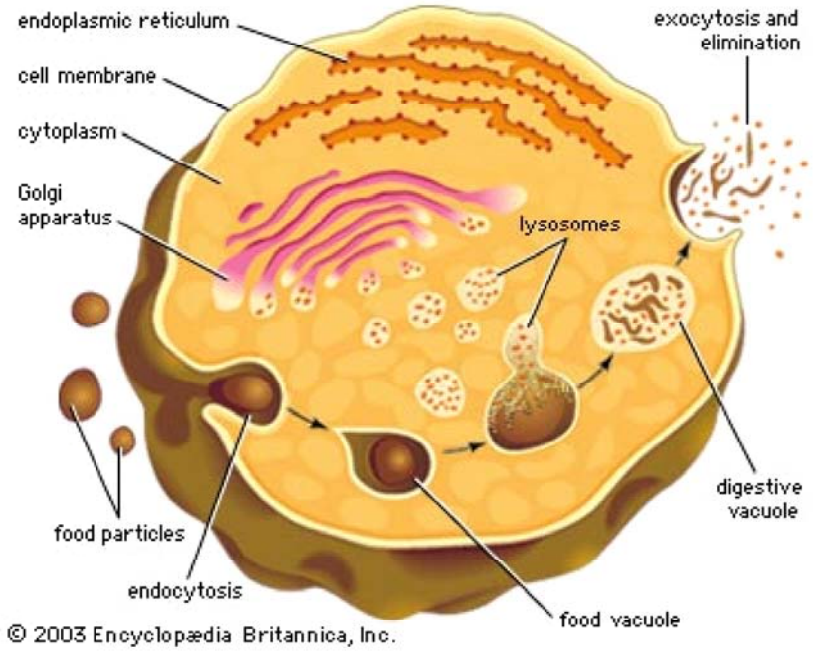
A significantly greater proportion of H_1M_0 genes (11/27 = 40.7%)

localized in vacuole in comparison to H_1M_1 genes (5/93 = 5.4%)

(FDR *P* = 2.11×10-3) !!

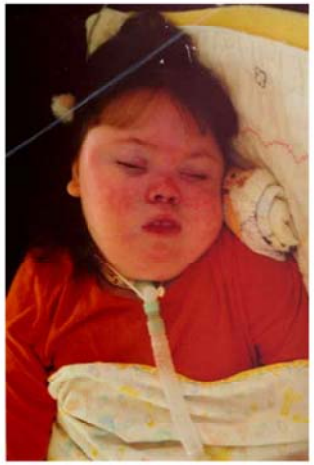
Functions of vacuoles in an animal cell

- Maintaining a balance between biogenesis and degradation
- Removing and exporting unwanted structural debris
- Isolating materials that might be harmful or a threat to the cell
- Containing waste products
- Maintaining an acidic internal pH
- Containing small molecules
- Enabling the cell to change shape



Diseases caused by Dysfunction of Vacuole (lysosomal) proteins

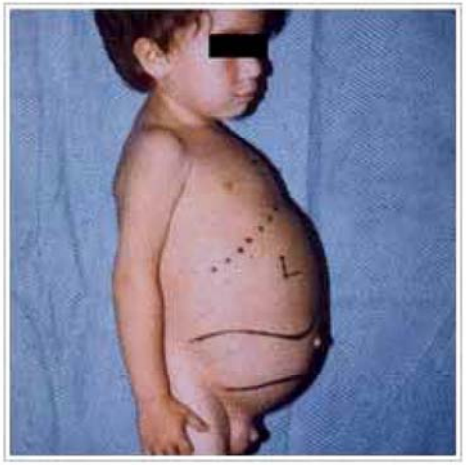
Sandhoff Disease



HEXB (beta subunit of hexosaminidase)

- onset: 6 moths of age; death < 3 yrs
- accumulation of lipids in the brain and organs
- early blindness
- progressive mental and motor deterioration
- macrocephaly
- seizures
- enlarged liver and spleen.

Niemann-Pick Disease, type A/B



SMPD1 (acid sphingomyelinase)

- onset: <6 moths of age; death < 3 yrs
- persistent early jaundice
- accumulation sphingomyelin causes the death of ganglion cells
- retarded physical and mental growth
- severe neurologic disturbances
- enlarged liver and spleen.

Diseases caused by Dysfunction of Vacuole (lysosomal) proteins

Knockout Mouse Phenotypes

- normal growth and fertility
- spasticity, muscle weakness, rigidity, tremors, and ataxia begin around 4 months of age
- death occurs around 5.5 months of age

- males could breed until 20 weeks of age and females until 10 weeks of age with normal litter size;
- lifespan of 4-8 months
- impaired coordination
- mild tremor and ataxia after 8 weeks of age
- abnormal lipid homeostasis
- decreased body weight

HEXB (beta subunit of hexosaminidase)

- onset: 6 moths of age; death < 3 yrs
- accumulation of lipids in the brain and organs
- early blindness
- progressive mental and motor deterioration
- macrocephaly
- seizures
- enlarged liver and spleen.

SMPD1 (acid sphingomyelinase)

- onset: <6 moths of age; death < 3 yrs
- persistent early jaundice
- accumulation sphingomyelin causes the death of ganglion cells
- retarded physical and mental growth
- severe neurologic disturbances
- enlarged liver and spleen.

Hypothesis for the association between evolutionary changes of proteins sequence and gene essentiality

Facts:

(i) vacuole is the cellular compartment primarily responsible for containing and degrading wastes and toxins.

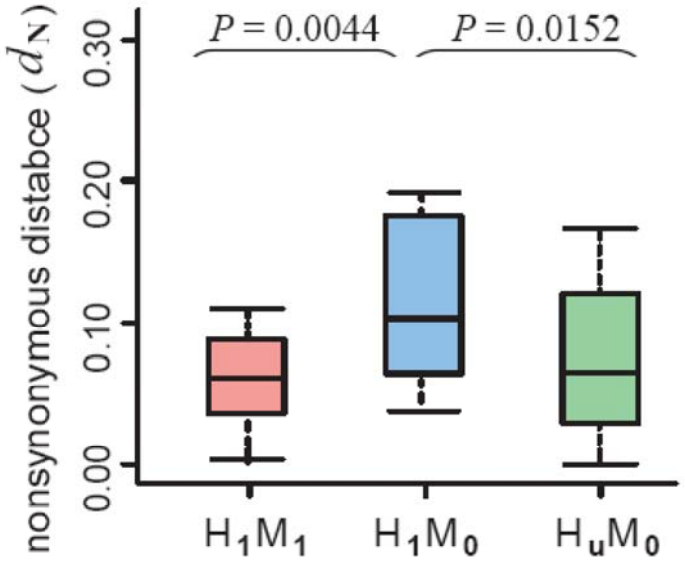
(ii) the loss of vacuole proteins in humans tends to cause the accumulation of cellular wastes and toxins that often leads to fatal neurological diseases

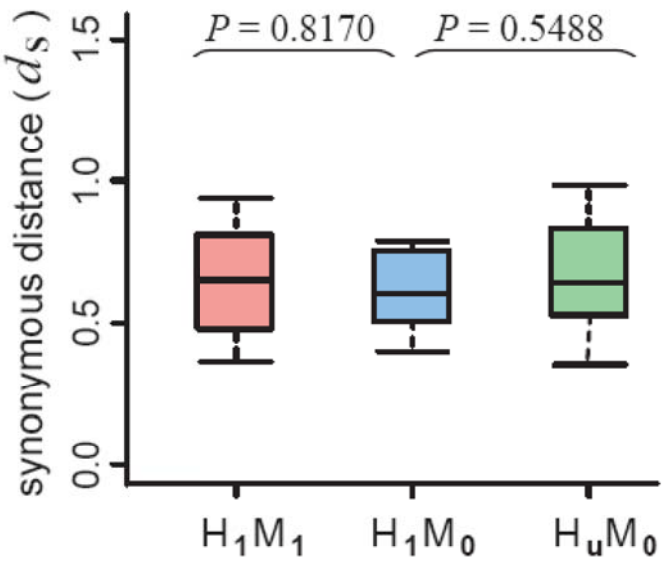
(iii) human reproductive age is >100 times that of the mouse.

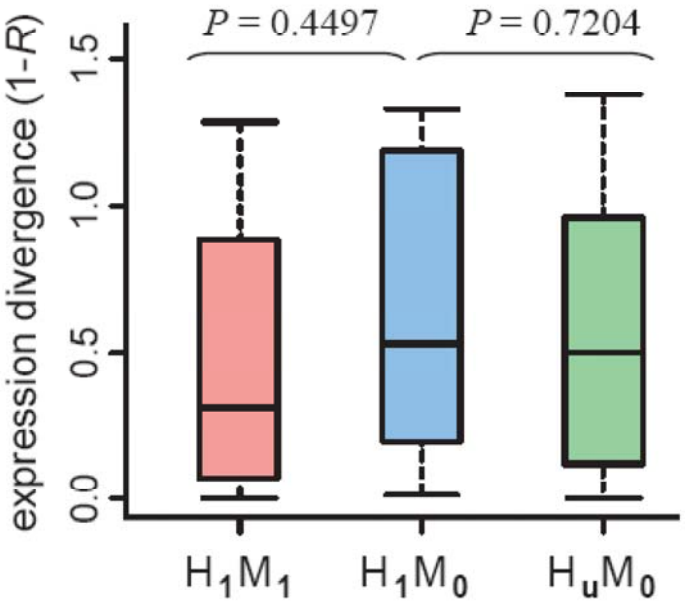
Hypothesis:

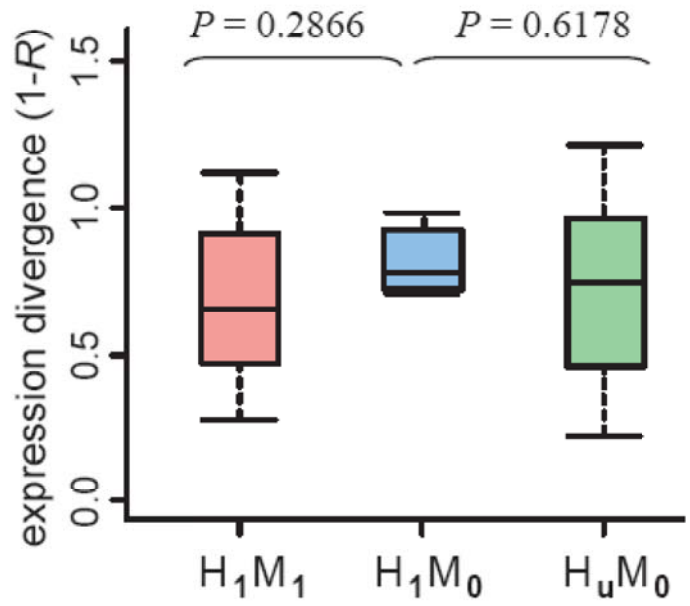
The evolution of the prolonged life history of humans generated selective pressures for better vacuole proteins

Reanalysis after removing Vacuole Proteins from the dataset









Conclusions.

1. There are distinct functional difference between human-mouse orthologous genes in terms of gene essentiality
2. Gene duplication plays negligible role in the evolutionary change of gene essentiality of mammalian genes
3. The evolutionary change of gene essentiality is associated with protein sequence changes rather than expression changes

Example #2

Contrasting genetic paths to morphological and physiological evolution

Ben-Yang Liao[a,1], Meng-Pin Weng[a], and Jianzhi Zhang[b,1]

[a]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350, Taiwan, Republic of China; and [b]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

Edited by Sean B. Carroll, University of Wisconsin, Madison, WI, and approved January 6, 2010 (received for review September 9, 2009)

The relative importance of protein function change and gene expression change in phenotypic evolution is a contentious, yet central topic in evolutionary biology. Analyzing 5,199 mouse genes with recorded mutant phenotypes, we find that genes exclusively affecting morphological traits when mutated (dubbed "morphogenes") are grossly enriched with transcriptional regulators, whereas those exclusively affecting physiological traits (dubbed "physiogenes") are enriched with channels, transporters, receptors, and enzymes. Compared to physiogenes, morphogenes are more likely to be essential and pleiotropic and less likely to be tissue specific. Morphogenes evolve faster in expression profile, but slower in protein sequence and gene gain/loss than physiogenes. Thus, morphological and physiological changes have a differential molecular basis; separating them helps discern the genetic mechanisms of phenotypic evolution.

evolutionary rate | gene expression | molecular evolution | phenotypic evolution

Nearly 35 years ago, King and Wilson remarked that, despite the large phenotypic difference, human and chimpanzee have virtually identical protein sequences, which prompted their proposal that gene expression change plays a more important role than protein function change in phenotypic evolution, including human origins (1). We now know that, between these two species, there are on average ~2 amino acid differences per protein and >70% of their

Results

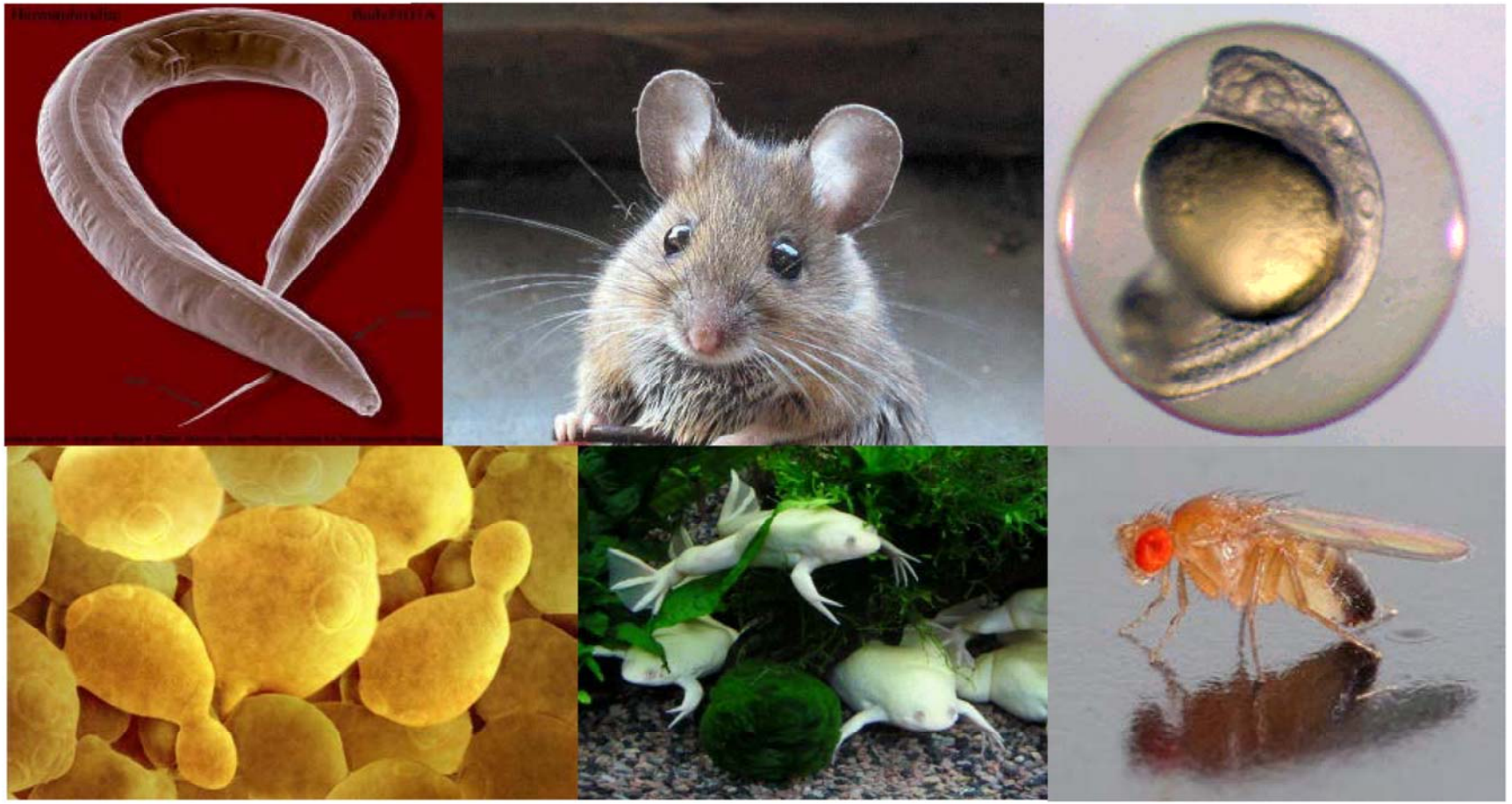
Morphogenes and Physiogenes Have Distinct Molecular Functions. We use the mouse *Mus musculus* as our focal organism because of the availability of its genome and transcriptome data as well as those of related species, presence of numerous well-characterized morphological and physiological traits (Table S1), and, most importantly, extensive documentation of its mutant phenotypes. At the time of our study, there were 5,199 mouse genes with recorded mutant phenotypes in the Mouse Genome Informatics (MGI) database, of which 821 affected only morphological traits and 912 affected only physiological traits (*Materials and Methods*). These genes are referred to as "morphogenes" and "physiogenes," respectively (Table S2).

By definition, morphogenes and physiogenes differ in certain biological processes they participate in, such as "anatomical structure development" and "immune response" (Table S3). However, it is interesting to note that morphogenes are much more frequently associated with the biological process of "transcription" than physiogenes ($P < E\text{-}29$ after correction for multiple testing), although this is not expected a priori. In addition, the molecular function of "transcriptional regulator activity" is grossly overrepresented among morphogenes, whereas those of "ion transporter activity," "channel or pore class transporter activity," "receptor activity," and "catalytic activity" are enriched among physiogenes (Fig. 1). Not unexpectedly, "structural molecule

EVOLUTION

The Use of Model Organisms in Understanding Humans

Model organisms are species that are extensively studied to understand biological phenomena, with the expectation that discoveries made in the organism model will provide insight into the workings of other organisms. (Fields and Johnston 2005 *Science*)



Mice are most commonly used model organisms to study human genetics

- mice are mammals
- 99% mouse genes have human orthologs
- the nucleotide sequence identity in coding regions between human and mouse is 85%
- small body size
- rapid breeding cycle

Phenotypic Differences between Humans and Mice

- in morphology

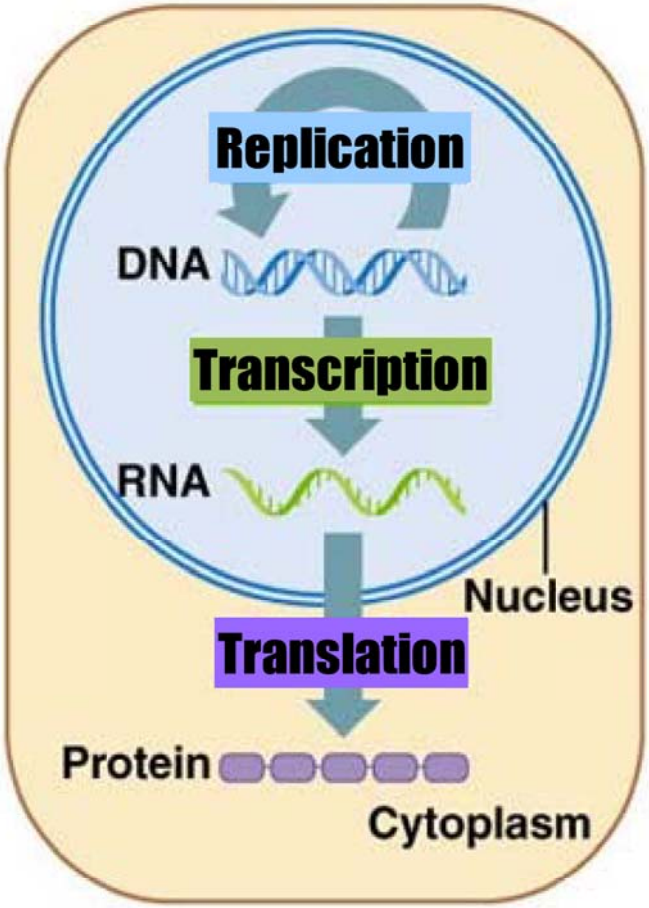
- in physiology

  e.g. Mice have the vomeronasal organ which detects pheromones, but humans do not.

- in behavior

  e.g. Human use tools, but mice do not.

Genetic Factors are Primarily Responsible for Inter-species Differences in Phenotypes



Two components determining functions of a gene

protein coding sequence

regulatory elements

Organismal Evolution by:

1) coding-sequence evolution

2) gene expression evolution

RNase (defensive → digestive) in douc langur (Asian leaf-eating monkey)

Zhang, J. *Nat. Genet.* **38**:819-823 (2006).

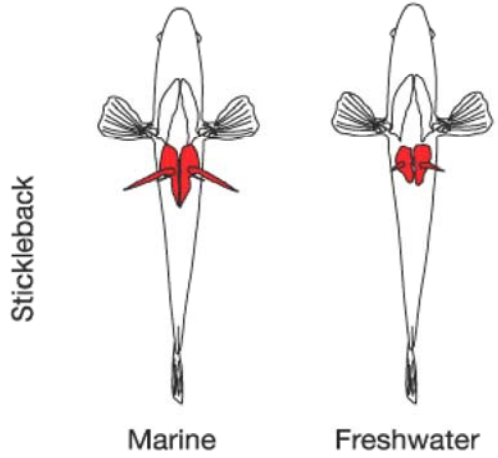
Organismal Evolution by:

1) coding-sequence evolution

2) gene expression evolution

Pitx1 in Stickleback

Shapiro, MD. *Nature*. **428:**717-723 (2004).

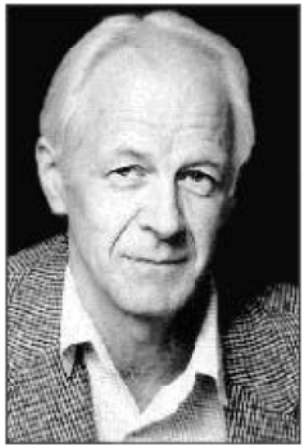


Evolution at Two Levels in Humans and Chimpanzees

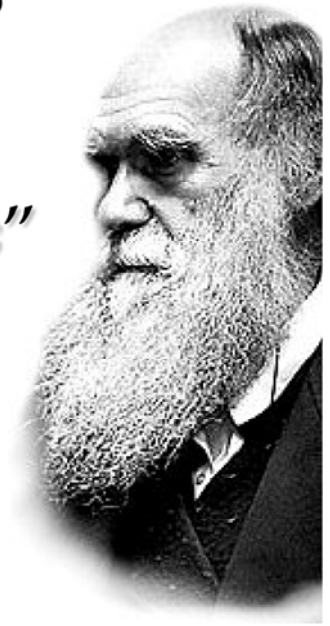
***Science* 1975, 188: 107-116**

"Their macromolecules are so alike that regulatory mutations may account for their biological differences"

Mary-Claire King



Allan Wilson



We now know that human-chimp orthologs differ by on average ~2 residues per protein, and >70% of their proteins are non-identical in sequence.

Debates on the Molecular Basis underlying Phenotypic Evolution

Two groups recently compiled all cases of phenotypic evolution with known genetic mechanisms, but reached different conclusions about the relative importance of protein function change versus gene expression change (esp. *cis*-regulatory change) in phenotypic evolution.

Protein sequence evolution

"Neither the theoretical arguments nor the data from nature, then, support the claim for a predominance of *cis*-regulatory mutations in evolution"

Hoekstra, HE and Coyne JA. *Evolution* **61**:995-1016 (2007)

Gene regulation evolution

A long-standing belief of Evo-Devo biologists

"*cis*-regulatory changes contribute more on the evolution of morphological traits."

Stern, DL and Orgogozo V. *Evolution* **62**:2155-2177 (2008)

Deciphering the Genetics of Evolution

Powerful personalities lock horns over how the genome changes to set the stage for evolution

The zeal with which some biologists have embraced this so-called cis-regulatory hypothesis rubbed Hoekstra and Coyne the wrong way. In a 2007 commentary in *Evolution*, they urged caution, arguing that the idea was far from proven. The article sparked a sharp debate, with accusations from both sides that the other was misrepresenting and misinterpreting the literature. "What really got people upset is the tone of

8 AUGUST 2008 VOL 321 **SCIENCE** www.sciencemag.org
Published by AAAS

"I am not trying to say that regulatory sequence is the most important thing in evolution." But for morphological changes, "it's a shutout" in favor of cis elements.
—SEAN CARROLL

"I'm distressed that Sean Carroll is preaching … that we know how evolution works based on such thin evidence."
— JERRY COYNE

Debates on the Molecular Basis underlying Phenotypic Evolution

Two groups recently compiled all cases of phenotypic evolution with known genetic mechanisms, but reached different conclusions about the relative importance of protein function change versus gene expression change (esp. *cis*-regulatory change) in phenotypic evolution.

Although these meta-analyses offer summaries of case studies, they may provide distorted pictures, because the case studies are potentially biased by preferences for certain methods, phenotypes, genes, and types of mutations in research.

A Genomic Approach

(1) Identify genes that affect certain phenotypes when mutated.

(2) Analyze properties and evolutionary patterns of these genes.

We are particularly interested in testing whether a distinction exists in the genetic basis of morphological and physiological evolution, which was previously proposed based on case studies and some theoretical considerations (Carroll 2005).

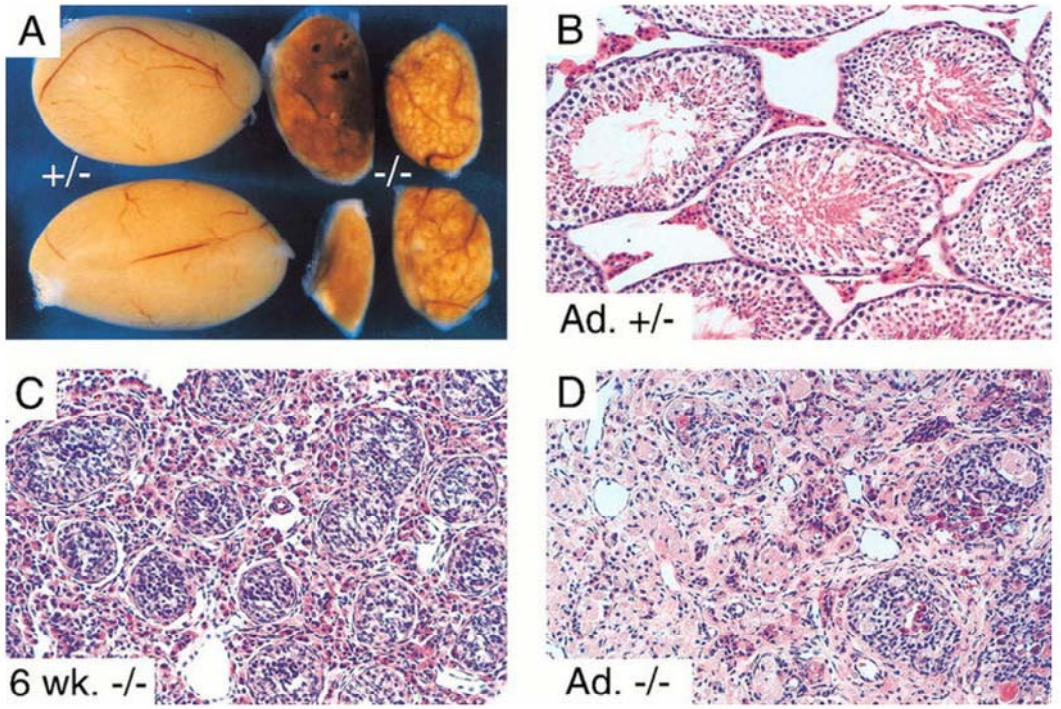
Defining two groups of genes based on the phenotypes of mouse mutants

Morphogenes: genes exclusively affect morphological traits when mutated.

e.g.

Dmrt1 (**doublesex and mab-3 related transcription factor 1**)

Mutant phenotypes: **Males homozygous for null mutations are sterile and exhibit a complete loss of germ, disorganized seminiferous tubules, and degeneration of Leydig cells.**



Source: MGI (Mouse Genome Informatics, Jackson Lab)

Defining two groups of genes based on the phenotypes of mouse mutants

Physiogenes: genes exclusively affect physiological traits when mutated.

e.g.
Oprk1 **(opioid receptor, kappa 1)**

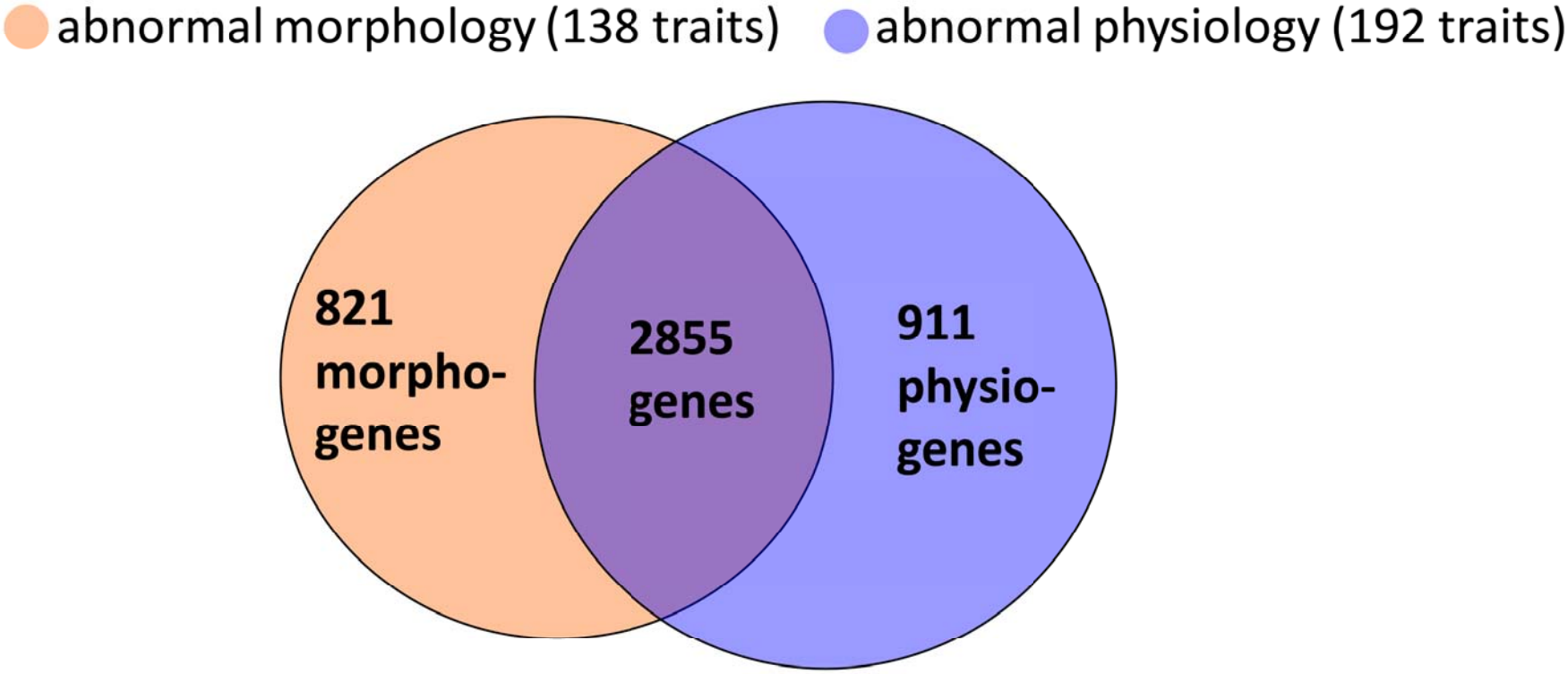
Mutant phenotypes: **Mice homo-zygous for a knock-out allele exhibit impaired response to morphine and an opioid agonist, abnormal pain threshold, and increased litter size.**

Source: MGI (Mouse Genome Informatics, Jackson Lab)

Defining two groups of genes based on the phenotypes of mouse mutants

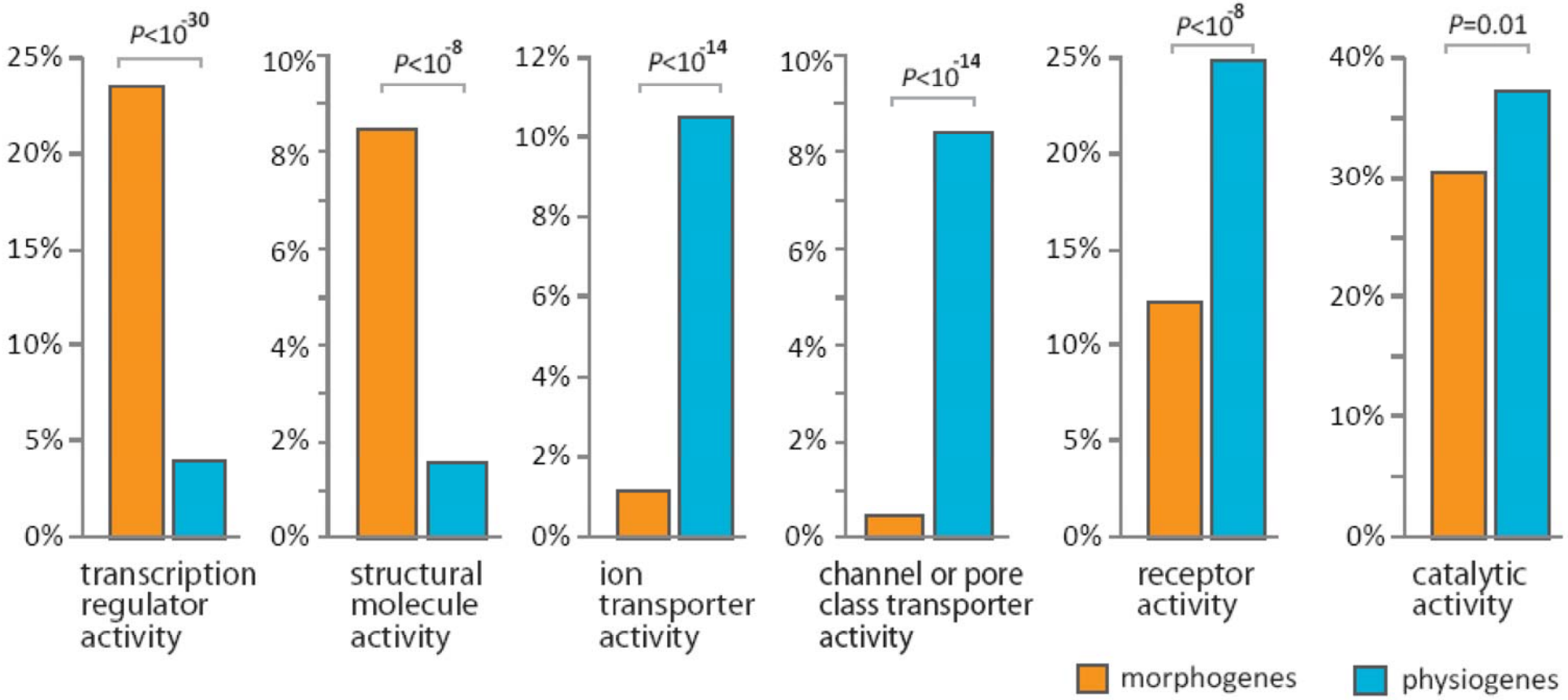
Morphogenes: genes exclusively affect morphological traits when mutated.

Physiogenes: genes exclusively affect physiological traits when mutated.



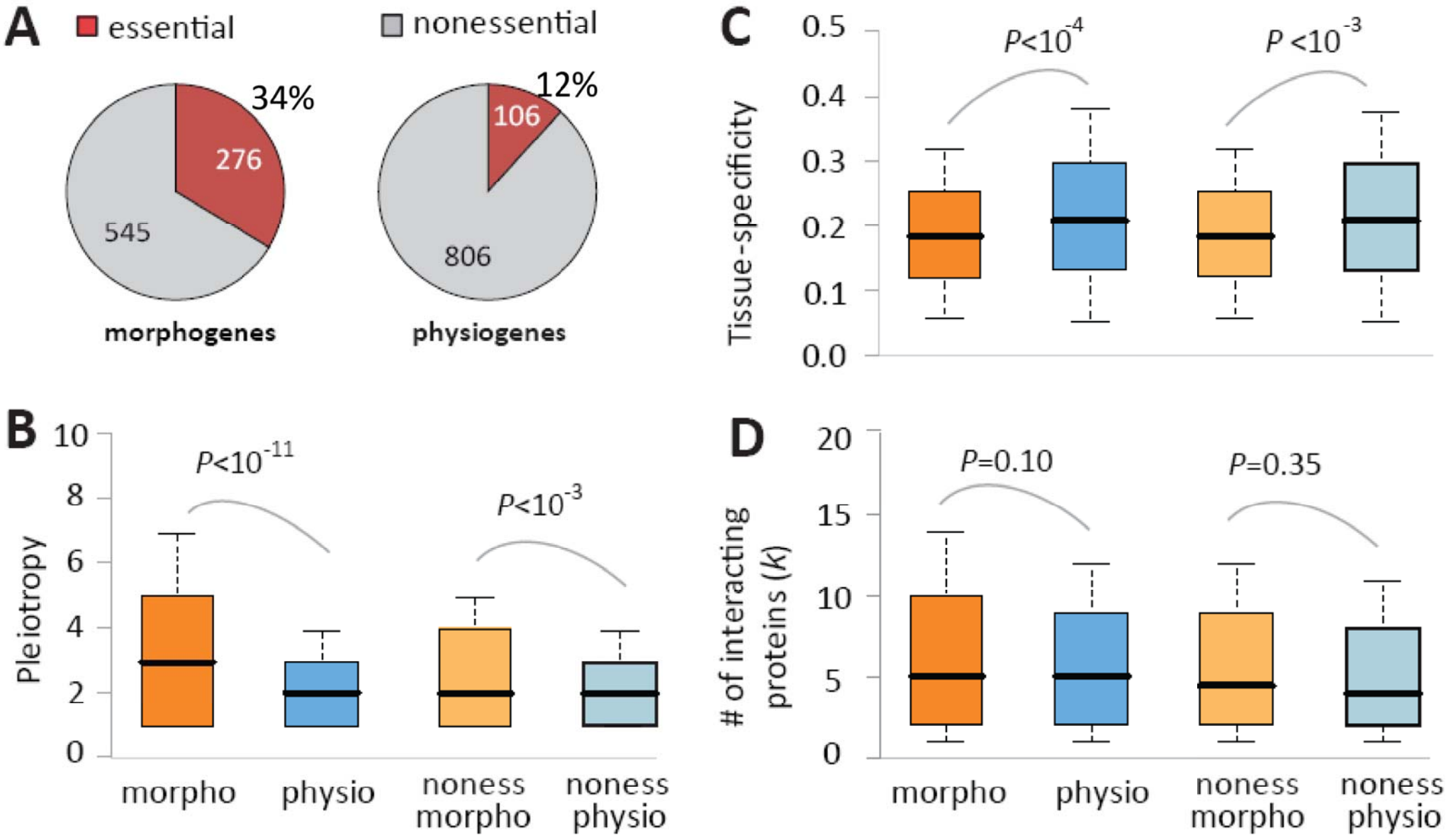
Source: MGI (Mouse Genome Informatics, Jackson Lab)

Morphogenes and physiogenes differ greatly in molecular function



(*P*-values adjusted for multiple testing)

They also differ in gene essentiality, tissue-specificity, and pleiotropy



Index for the features of gene expression

Expression Level:

Averaged $S_M(i,j)$

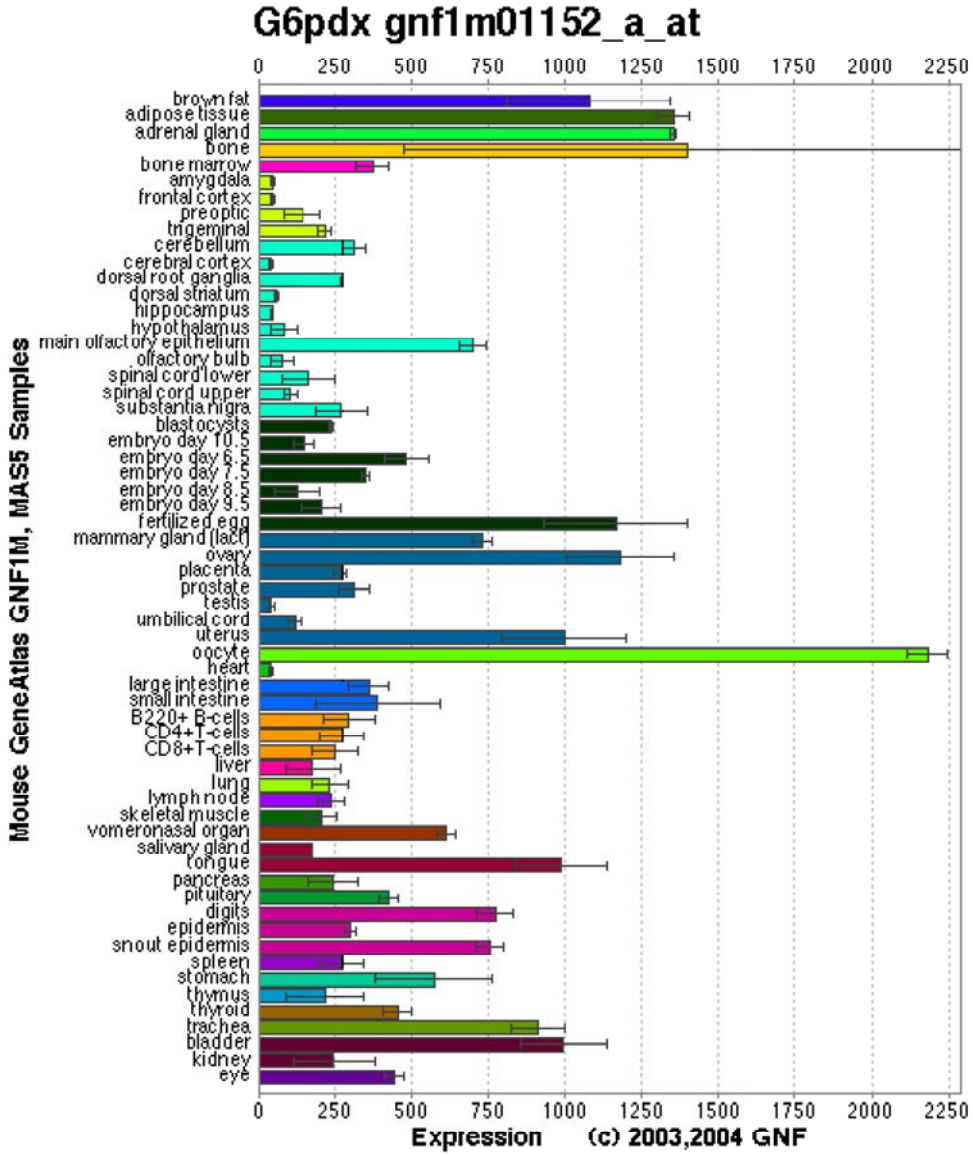
or Maximum $S_M(i,j)$

Tissue-specificity:

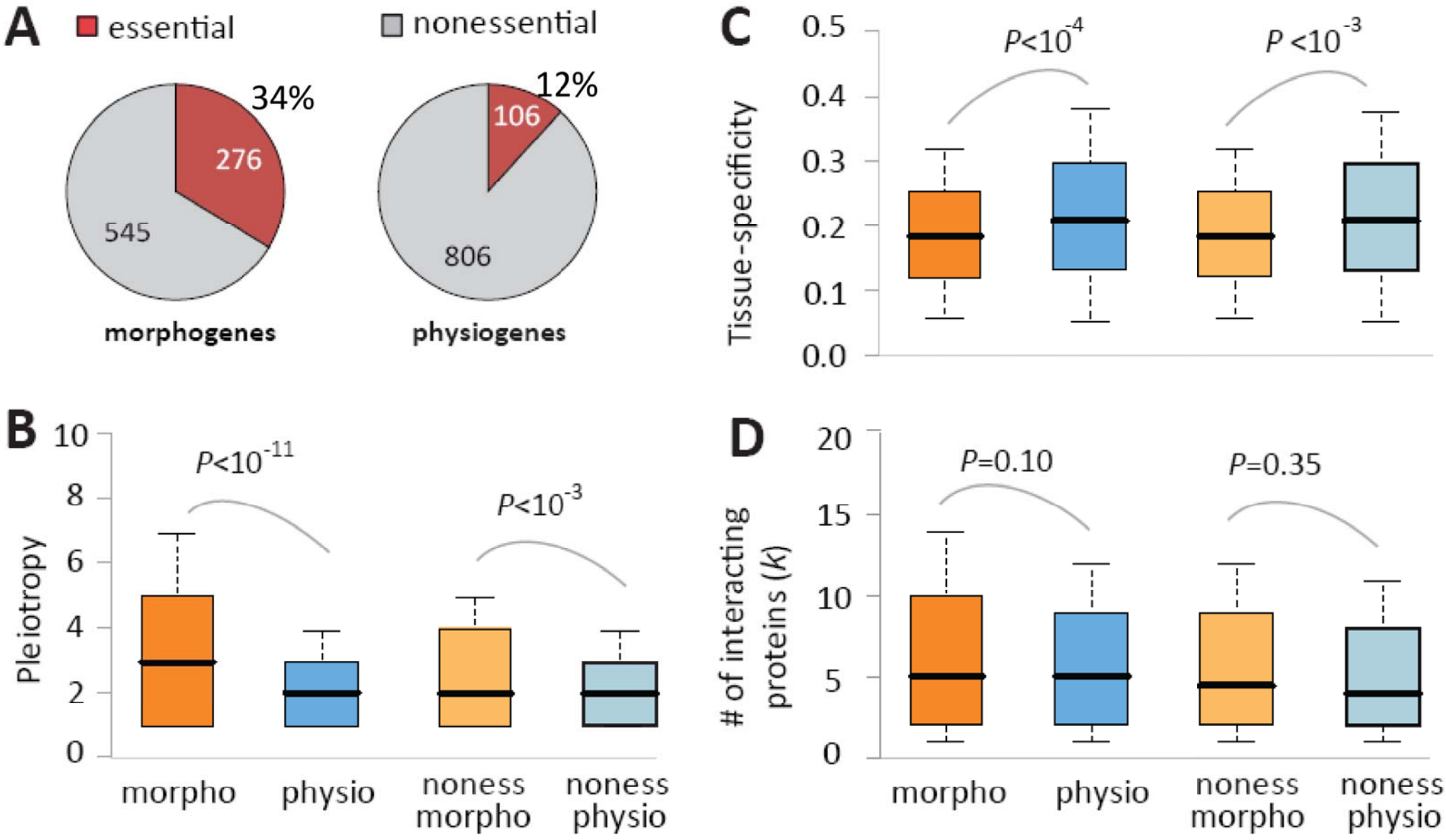
$$\tau_M = \frac{\sum_{j=1}^{n_M}\left(1-\left[\frac{\log_2 S_M(i,j)}{\log_2 S_M(i,\max)}\right]\right)}{n_M - 1}$$

$$let\ S_M(i,j) = 100,\ if\ \ S_M(i,j) < 100$$

Yanai et al. (2005) *Bioinformatics* **21**: 650-659.



They also differ in gene essentiality, tissue-specificity, and pleiotropy



Do they differ in the rate of evolution?

1. Protein sequence

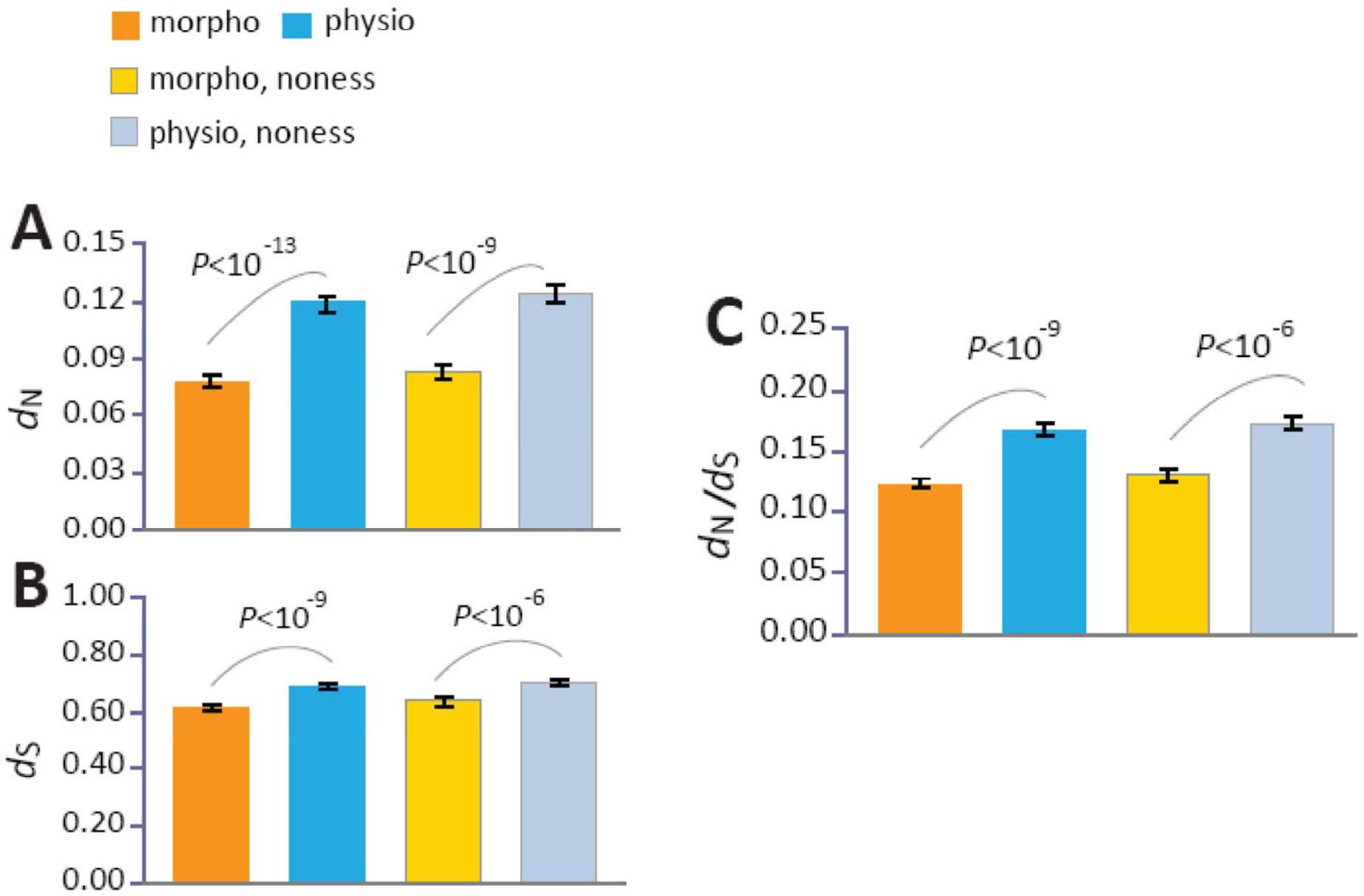
2. Expression-profile

3. cis-regulatory sequence

4. Gene relocation

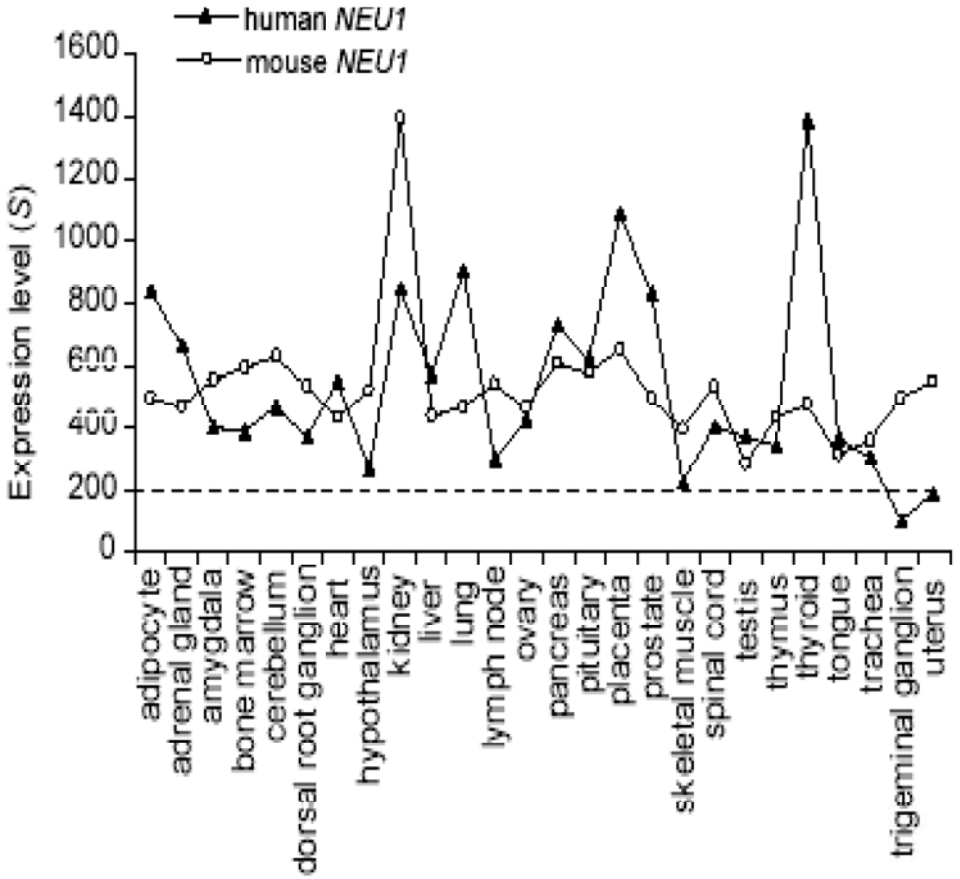
5. Gene family expansion/contraction

Physiogenes evolve faster than morphogenes in protein sequence

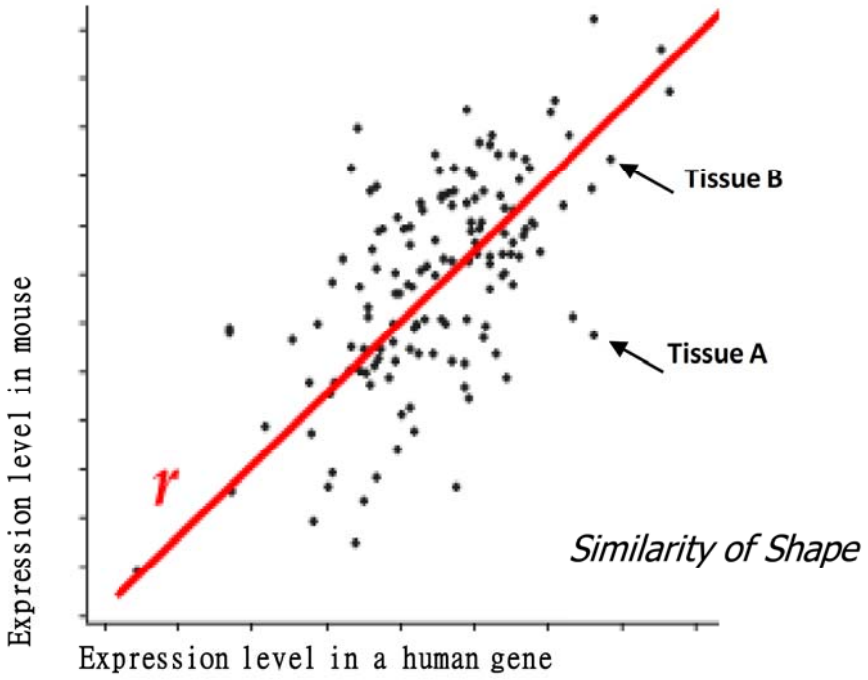


Measuring the rate of gene expression evolution

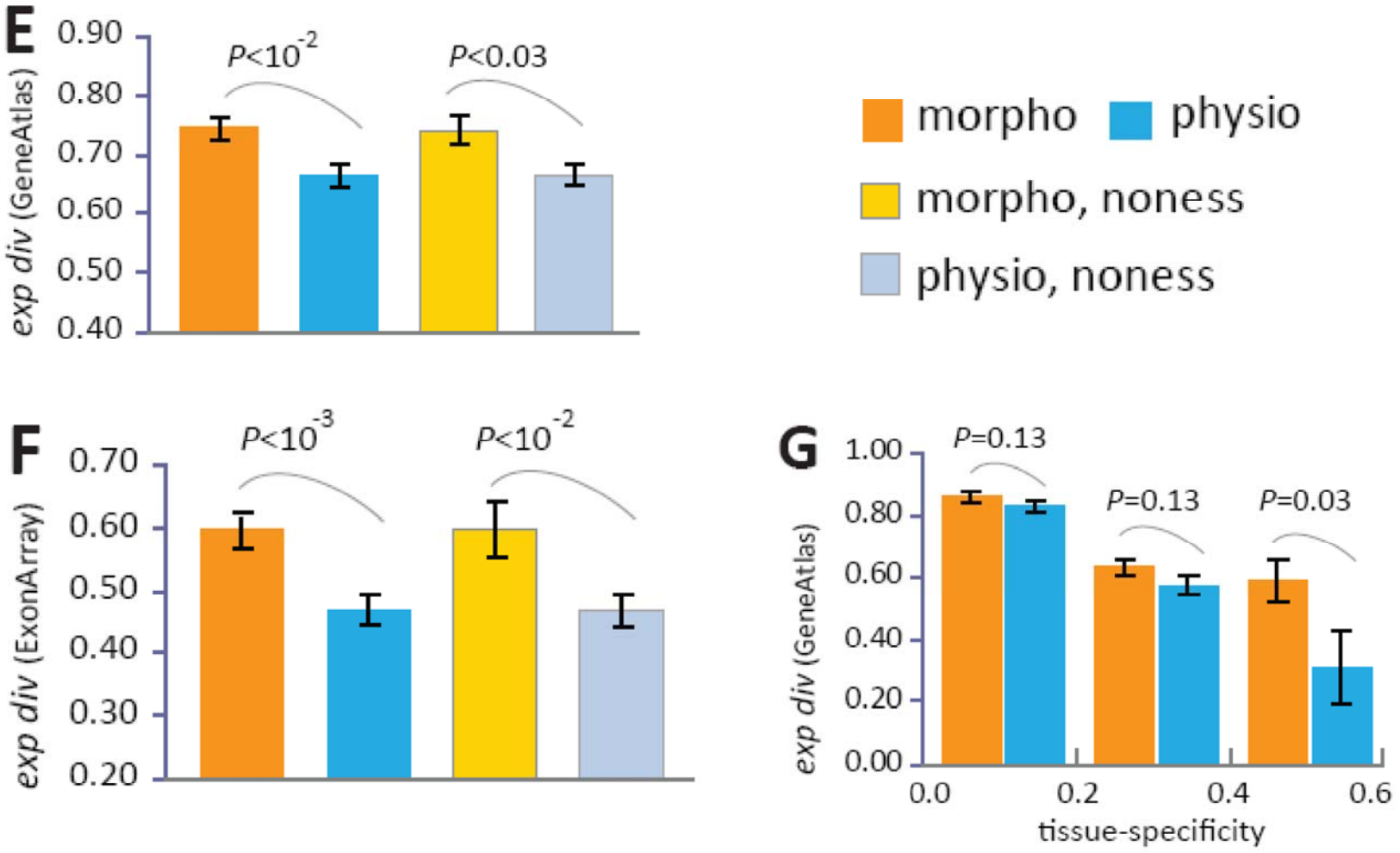
Divergence of gene expression-profiles between mouse and human



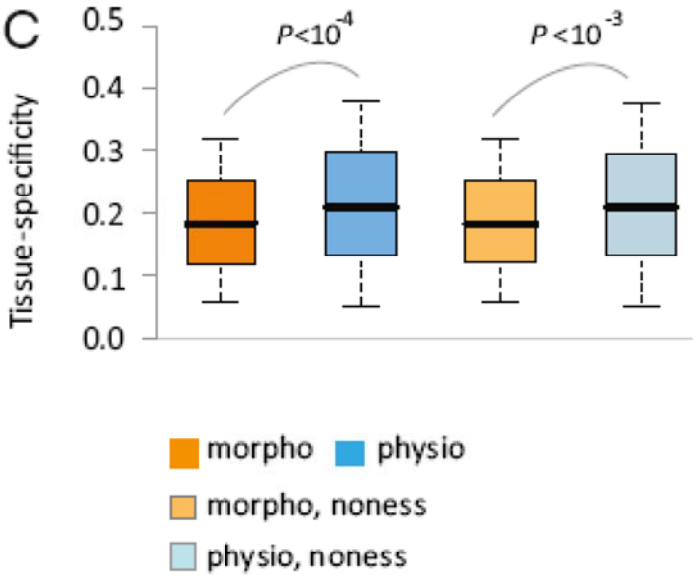
1 - Pearson's *r*

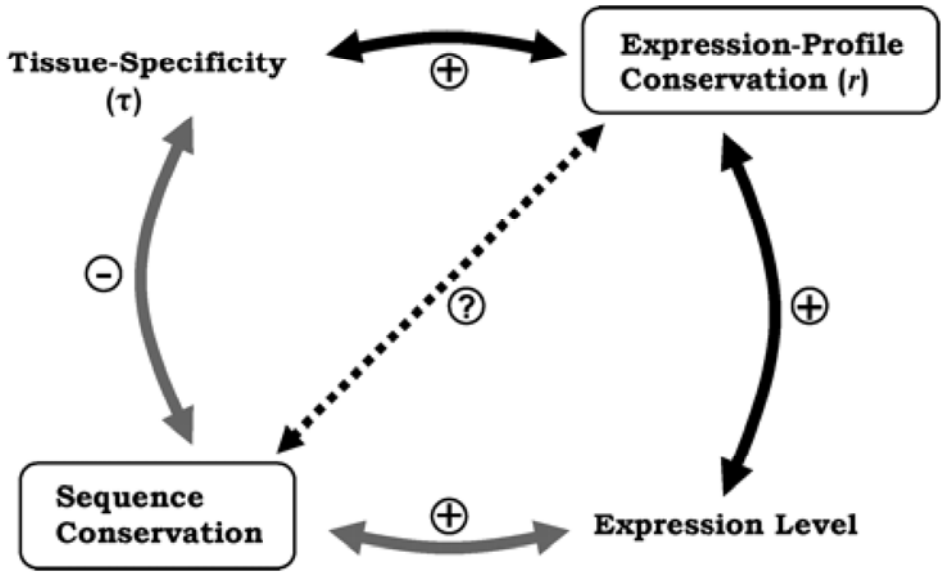


Morphogenes evolve faster than physiogenes in expression-profile



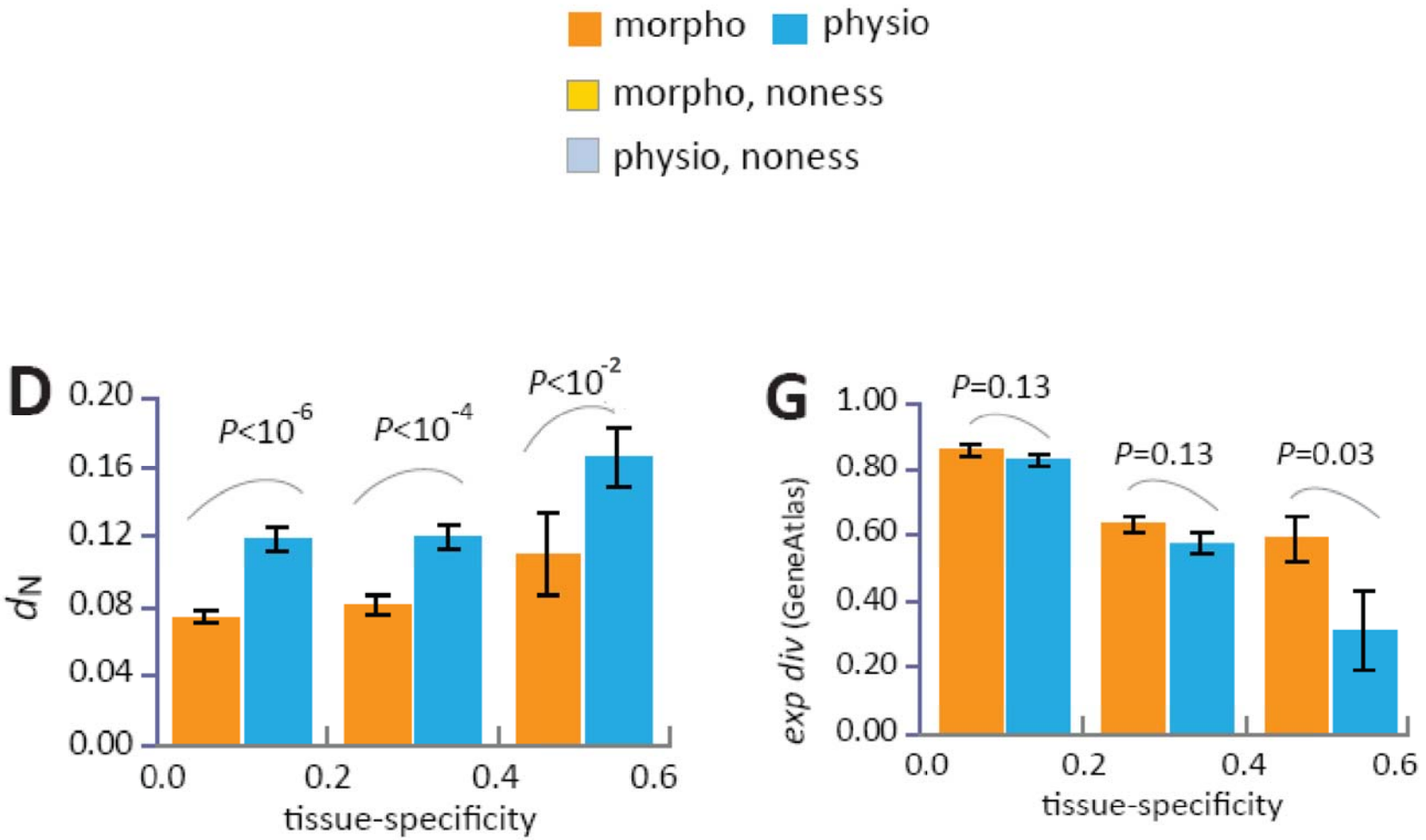
Tissue-specificity as the Potential Contributor to the Observations





Liao and Zhang, *Mol. Biol. Evol.* **23**:1119-28 (2006).

Consistent Results after Controlling for Tissue-specificity

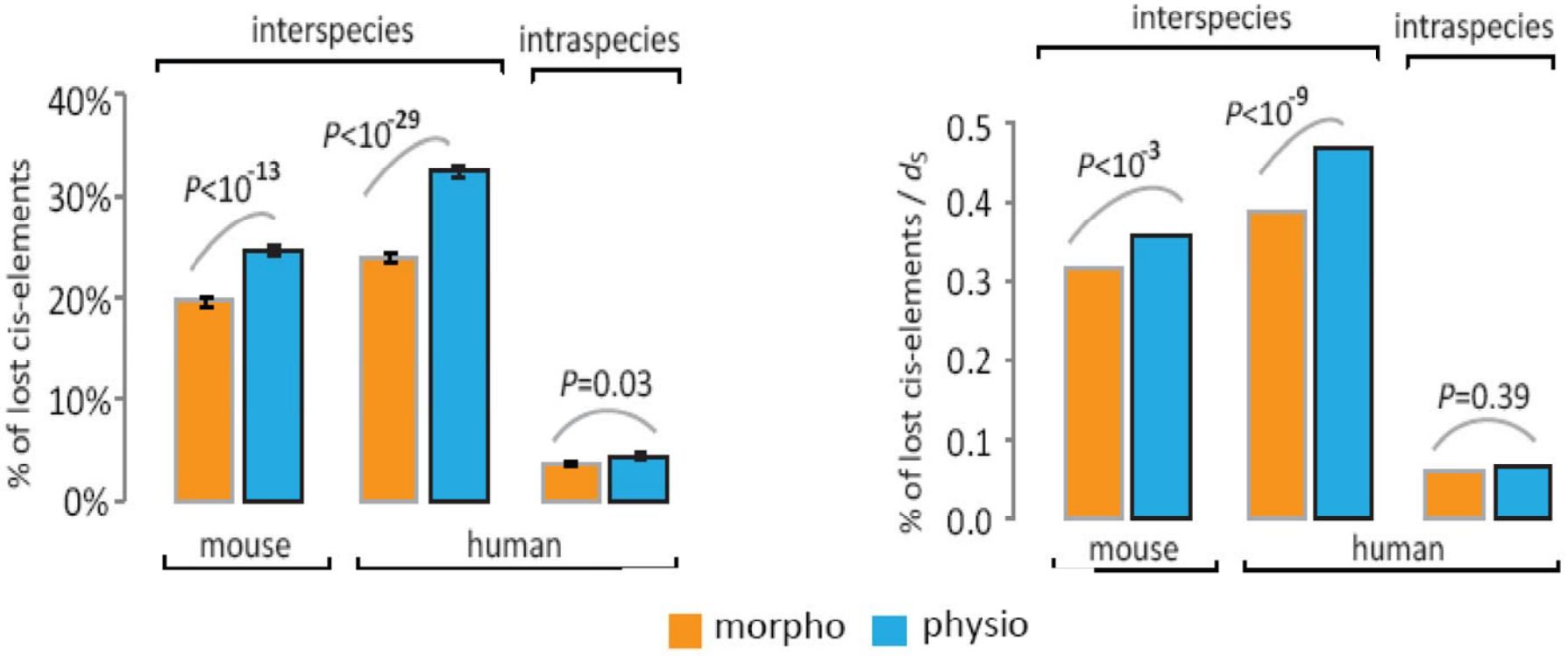


Is the higher rate of expression evolution of morphogenes due to faster *cis*-regulatory changes ??

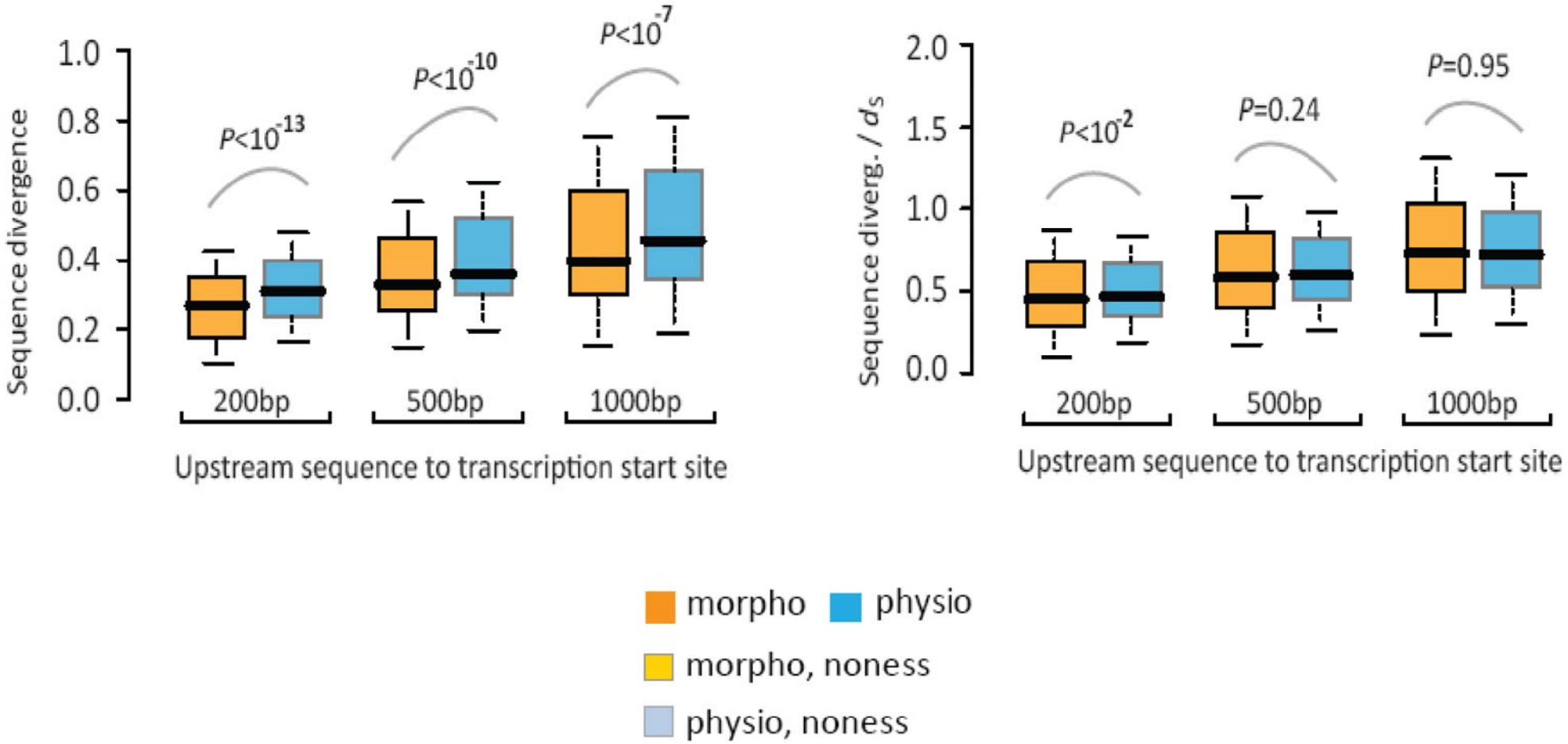
"I am not trying to say that regulatory sequence is the most important thing in evolution." But for morphological changes, "it's a shutout" in favor of cis elements.
—SEAN CARROLL

Analysis of annotated *cis*-regulatory motifs

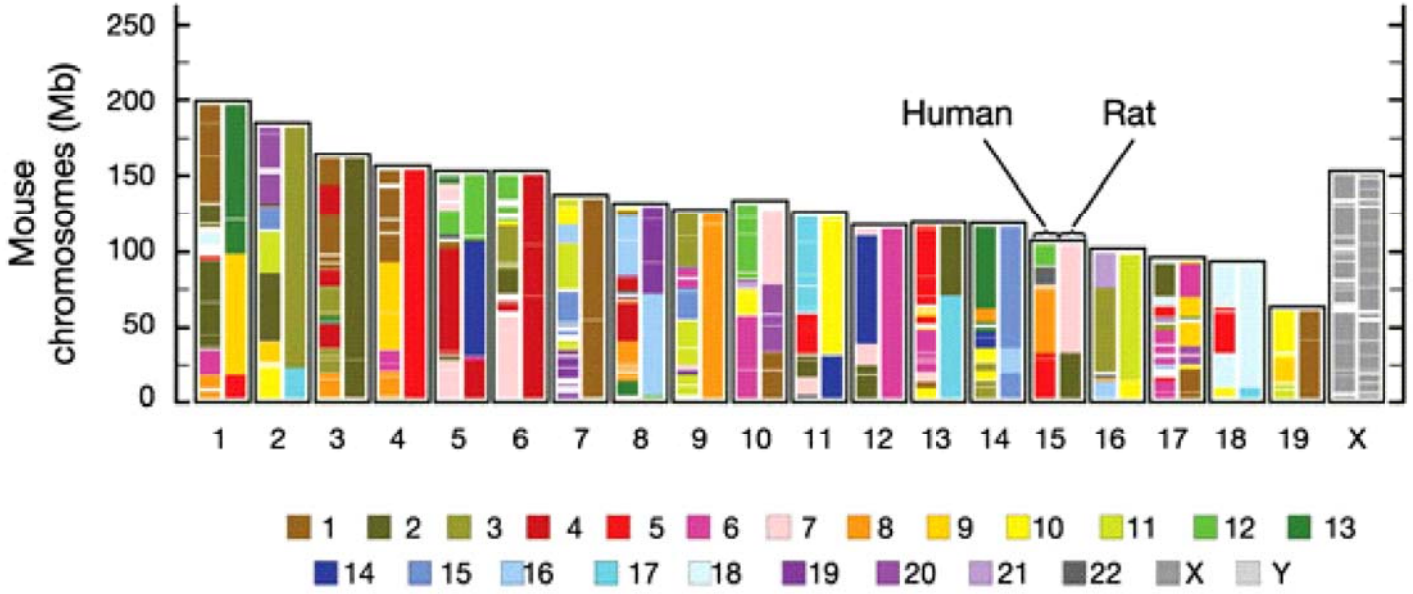
		Mouse	Human
cis-regulatory motifs of	morphogenes	8,440	7,688
	physiogenes	7,082	7,215



Analysis of noncoding sequences upstream of transcription start site



Analysis of the rate of gene relocation



Frequency of relocation resulting in the change of neighboring genes between human and mouse:

Physiogenes: 11.9%
Morphogenes: 9.7% **(*P* =0.16, χ2 test).**

Physiogenes evolve faster than morphogenes in gene family expansion/contraction

$$D_{fam} = | N_M - N_H | / \sqrt{(N_M + 1)(N_H + 1)}$$

N_M: number of paralogs of the focal gene in the mouse genome
N_H: number of paralogs of the focal gene in the human genome

Physiogenes: 0.072
Morphogenes: 0.046 (P <0.01)

$$D'_{fam} = | N_M - N_H | / [(N_M + 1)(N_H + 1)]$$

Physiogenes: 0.032
Morphogenes 0.022 (P <0.01)

Robustness of the results

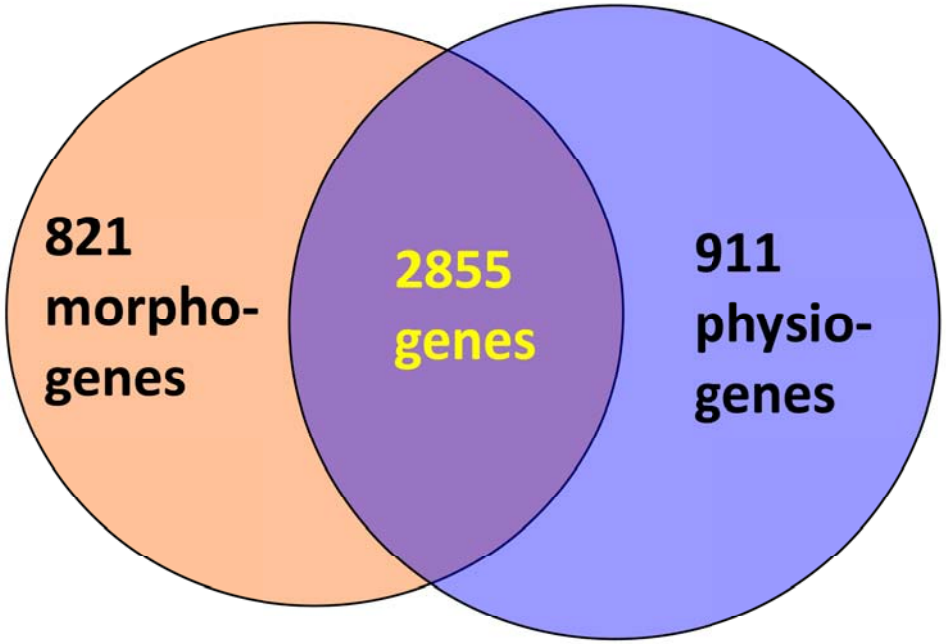
Because the number of traits examined in each mutant mouse is limited and somewhat arbitrary, one wonders whether the distinctions between morphogenes and physiogenes that we found are robust.

We randomly removed 30% of phenotypes of each gene, re-identified morphogenes and physiogenes, and repeated all the analyses. All the results are still valid, albeit with slightly reduced statistical support. This finding suggests that our results would be statistically more significant with additional phenotyping. In other words, our results are conservative.

Robustness of the results

Analysis on overlapping genes….

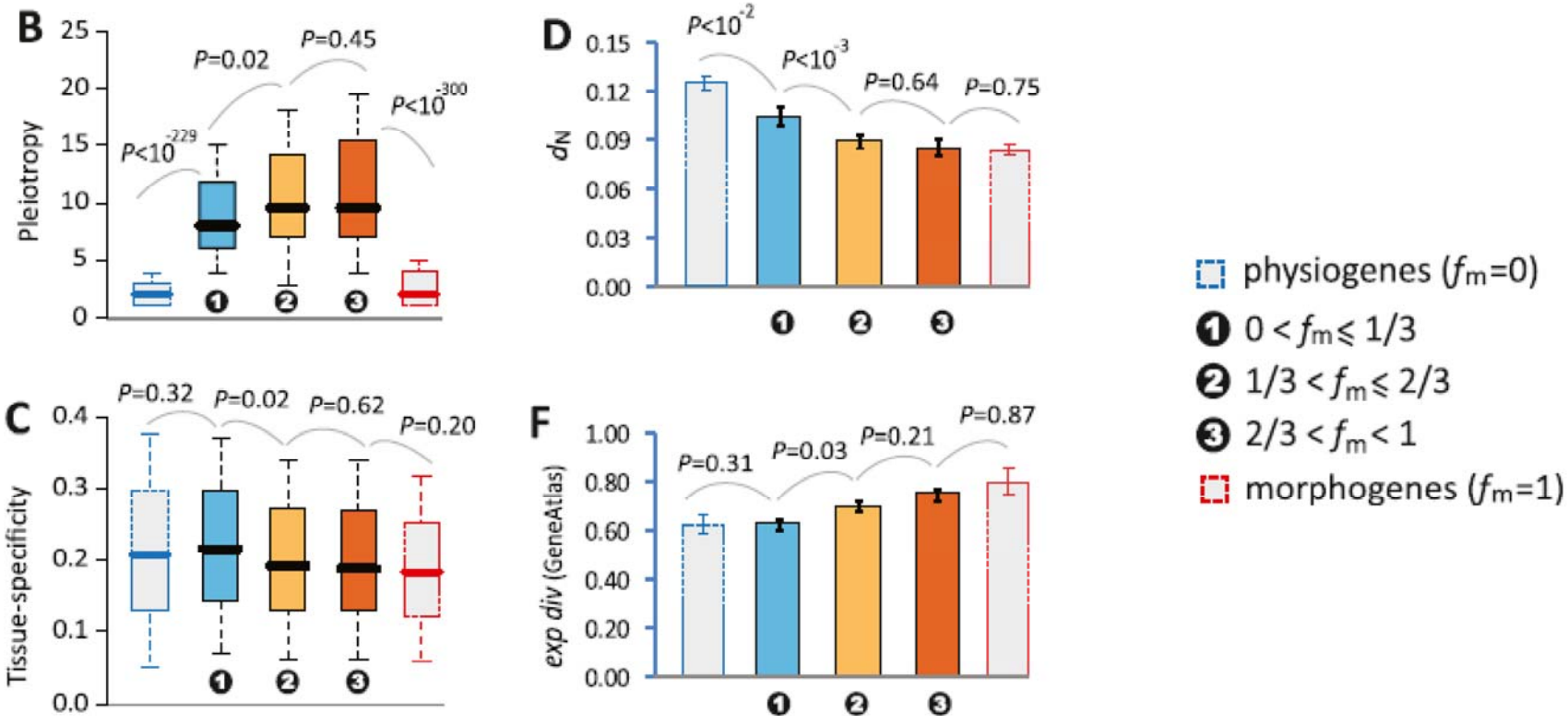
f_m: the fraction of phenotypes being morphological



Robustness of the results

Analysis on overlapping genes….

f_m: the fraction of phenotypes being morphological



Take-home messages (II)

1. The separation between morphogenes and physiogenes is both possible and biologically meaningful.

2. Although our morphogenes and physiogenes are classified largely by the phenotypes of their strongly deleterious mutants, it is reasonable to assume that most beneficial or neutral mutations in morphogenes affect morphological traits rather than physiological traits and vice versa.

   Thus, we predict that morphological evolution more often involves transcriptional regulators and gene expression changes while physiological evolution more likely involves transporters, channels, receptors, and enzymes and protein sequence changes or gene gains/losses.

Take-home messages (II) (contd.)

3. Although we showed that gene expression evolves faster for morphogenes than physiogenes, we did not find *cis*-regulatory elements or regions to evolve faster in morphogenes

4. Our analysis of the knockout mouse data suggests that morphological defects are more likely due to problems with gene expression. This knowledge could help identify the disease-causing mutations more quickly, because it narrows the set of candidate genes and mutations that one needs to search from.