Spring 2025

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 5

1

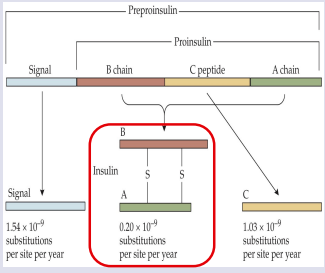
TODAY'S TOPICS

➢ Molecular Evolution - part 3

2

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$d_N(AB) = u_T(AB) * f_0(AB)$

$d_N(C) = u_T(C) * f_0(C)$

Both mutation rates (u_T) and selective constraints (f_0) affect nonsynonymous rates (d_N)

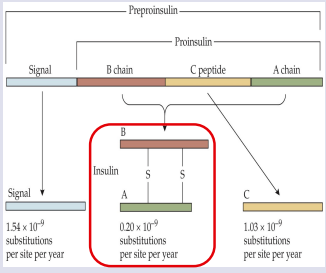
Note: In literatures, K_a and K_S are also used instead of d_N and d_S

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

3

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$d_N(AB) = u_T(AB) * f_0(AB)$

$d_N(C) = u_T(C) * f_0(C)$

Does it mean:

$u_T(AB) < u_T(C)$?

$f_0(AB) < f_0(C)$?

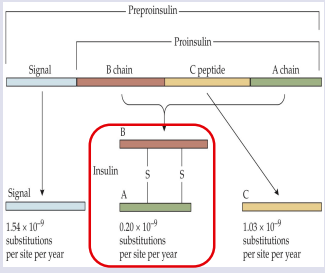
Both mutation rate (u_T) and selective constraints (f_0) can affect $d_N(AB) < d_N(C)$.

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

4

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$u_T(AB)*f_0(AB) < u_T(C)*f_0(C)$

If mutation rates are constant within a gene,

➜ $u_T(AB) = u_T(C)$

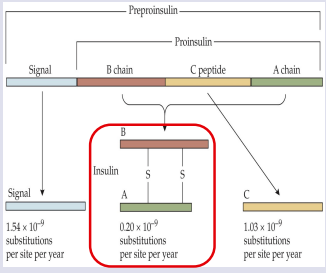
➜ $d_N(AB) < d_N(C)$ can be explained by $f_0(AB) < f_0(C)$

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

5

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$u_T(AB)*f_0(AB) < u_T(C)*f_0(C)$

If mutation rates are different

➜ $u_T(AB) \neq u_T(C)$

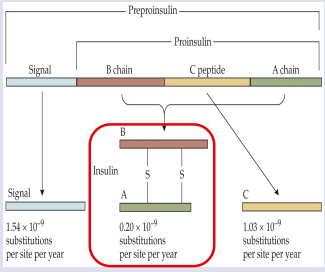
➜ $d_N(AB) < d_N(C)$ cannot be simply explained by $f_0(AB) < f_0(C)$

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

6

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$d_N(AB) = u_T(AB) * f_0(AB)$

$d_N(C) = u_T(C) * f_0(C)$

Synonymous substitutions are assumed to be neutral or near-neutral.

$d_S = u_T f_0 = u_T$

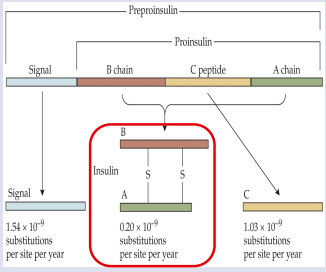
where $f_0 \approx 1$ (neutral)

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

7

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)



$d_N(AB) < d_N(C)$

$d_N(AB) = u_T(AB) * f_0(AB)$

$d_N(C) = u_T(C) * f_0(C)$

Synonymous substitutions are assumed to be neutral or near-neutral.

$d_S(AB) = u_T(AB)$

$d_S(C) = u_T(C)$

(Based on nonsynonymous substitution rates estimated from human-rat comparison)

8

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)

$d_N(AB)$ *vs.* $d_N(C)$

$d_N(AB) = u_T(AB) * f_0(AB)$ $\quad$ $d_N(C) = u_T(C) * f_0(C)$

$d_S(AB) = u_T(AB)$ $\quad$ $d_S(C) = u_T(C)$

$\frac{d_N(AB)}{d_S(AB)} = \frac{u_T(AB) * f_0(AB)}{u_T(AB)}$ $\quad$ $\frac{d_N(C)}{d_S(C)} = \frac{u_T(C) * f_0(C)}{u_T(C)}$

$\frac{d_N(AB)}{d_S(AB)} = f_0(AB)$ $\quad$ $\frac{d_N(C)}{d_S(C)} = f_0(C)$

9

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)

$\frac{d_N(AB)}{d_S(AB)} = f_0(AB)$ $\quad$ $\frac{d_N(C)}{d_S(C)} = f_0(C)$

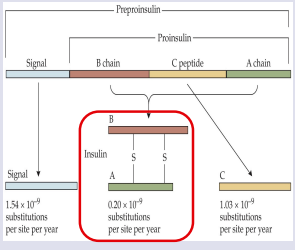
d_N/d_S can be used to identify selection

- Shows only the level of selective constrains (f_0)
- We don't need to worry about mutation rates (u_T)
  - → Even if $u_T(AB) \neq u_T(C)$, d_N/d_S can be compared

10

Identifying Selection

- Number of synonymous substitutions per site: d_S (or K_S)
- Number of nonsynonymous substitutions per site: d_N (or K_N)
  - → d_N/d_S can be used to identify selection



~~$d_N(AB) < d_N(C)$~~

$d_N(AB)/d_S(AB)$? $d_N(C)/d_S(C)$

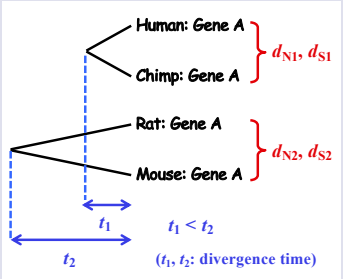
This is the better comparison

Regardless of the mutation rates, we can compare:

$f_0(AB)$ vs. $f_0(C)$

11

Identifying Selection



$d_{N1} < d_{N2}$, $d_{S1} < d_{S2}$

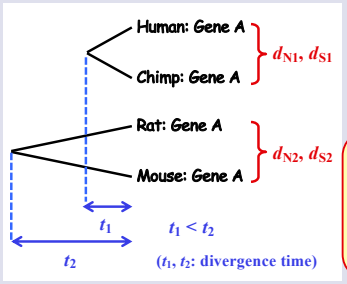
This may be caused simply by the difference in divergence time ($t_1 < t_2$)

$d_{N1} = u_{T1} * f_1 * 2t_1$, $d_{S1} = u_{T1} * 2t_1$

$d_{N2} = u_{T2} * f_2 * 2t_2$, $d_{S2} = u_{T2} * 2t_2$

12

Identifying Selection



$d_{N1} < d_{N2}$, $d_{S1} < d_{S2}$

This may be caused simply by the difference in divergence time ($t_1 < t_2$)

d_{N1}/d_{S1} *vs.* d_{N2}/d_{S2}

By comparing d_N/d_S, the time effect can be cancelled out.

We can compare the selective constraints between genes and between lineages.

$d_{N1} = u_{T1} * f_1 * 2t_1$, $d_{S1} = u_{T1} * 2t_1$ ➜ $d_{N1}/d_{S1} = u_{T1} * f_{01} * 2t_1 / (u_{T1} * 2t_1) = f_{01}$

$d_{N2} = u_{T2} * f_2 * 2t_2$, $d_{S2} = u_{T2} * 2t_2$ ➜ $d_{N2}/d_{S2} = u_{T2} * f_{02} * 2t_2 / (u_{T2} * 2t_2) = f_{02}$

13

Identifying Selection

- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N

➜ d_N/d_S can be used to identify selection

If nonsynonymous (replacement) substitutions are neutral:

$d_N > d_S$ or $d_N \approx d_S$ or $d_N < d_S$?

14

Identifying Selection

- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N

➜ d_N/d_S can be used to identify selection

If nonsynonymous (replacement) substitutions are neutral:

$d_N > d_S$ or $d_N \approx d_S$ or $d_N < d_S$?

$d_N = u_T * f_0$

$d_S = u_T$

If nonsynonymous substitutions are neutral,

➜ $f_0 = 1$,

➜ $d_N = u_T * f_0 = u_T = d_S$

15

Identifying Selection

- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N

➜ d_N/d_S can be used to identify selection

When nonsynonymous (replacement) substitutions are neutral,

$d_N \approx d_S$ ➜ $d_N/d_S \approx 1$

➜ d_S is used as a control
(substitution rate at neutral or near-neutral)

16

Identifying Selection

- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N

➜ d_N/d_S can be used to identify selection

d_N/d_S

$d_N/d_S > 1.0$

$d_N/d_S \approx 1.0$ Neutral (*e.g.*, non-coding regions, pseudogenes)

$d_N/d_S < 1.0$

0

➜ d_S is used as a control
(substitution rate at neutral or near-neutral)

17

Identifying Selection

- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N

➜ d_N/d_S can be used to identify selection

d_N/d_S

$d_N/d_S > 1.0$

1.0

$d_N/d_S < 1.0$

0

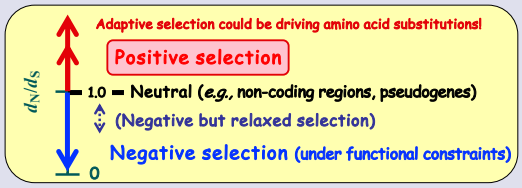
Which indicates negative selection?

➜ d_S is used as a control
(substitution rate at neutral or near-neutral)

18

Identifying Selection

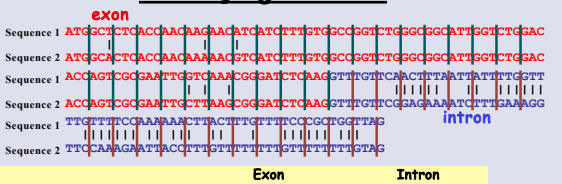
- Number of synonymous substitutions per site: d_S
- Number of nonsynonymous substitutions per site: d_N
  - ➜ d_N/d_S can be used to identify selection



- ➜ d_S is used as a control
  (substitution rate at neutral or near-neutral)

20

Identifying Selection



	Exon			Intron		
[Codon position]	1st	2nd	3rd	1st	2nd	3rd
# nucleotide sites	31	31	31	23	22	22
# nucleotide substitutions	2	0	4	12	12	14
Nucleotide substitutions/site	0.06	0	0.13	0.52	0.55	0.64

Exon(2nd) < Exon(3rd) | Intron(2nd) ≲ or ≈ Intron(3rd)

Exon(2nd) / Exon(3rd) 0 << Intron(2nd) / Intron(3rd) 0.86 (≈1)

(Similar to d_N/d_S analysis)

21

Identifying Selection

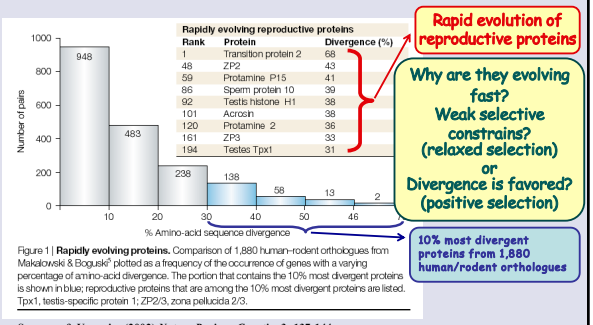


Figure 1 | **Rapidly evolving proteins.** Comparison of 1,880 human-rodent orthologues from Makalowski & Boguski[5] plotted as a frequency of the occurrence of genes with a varying percentage of amino-acid divergence. The portion that contains the 10% most divergent proteins is shown in blue; reproductive proteins that are among the 10% most divergent proteins are listed. Tpx1, testis-specific protein 1; ZP2/3, zona pellucida 2/3.

Rapid evolution of reproductive proteins

Why are they evolving fast?
Weak selective constrains? (relaxed selection)
or
Divergence is favored? (positive selection)

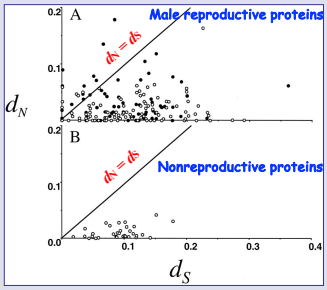
10% most divergent proteins from 1,880 human/rodent orthologues

Swanson & Vacquier (2002) *Nature Reviews Genetics* **3: 137-144**

22

Selection and functions

- Elevated d_N/d_S ratios are often found in reproductive genes



(*e.g.*, genes involved in mating behavior, fertilization, spermatogenesis, ejaculation, or sex determination)

Swanson *et al.* (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *PNAS* 98: 7375-7379.

See also

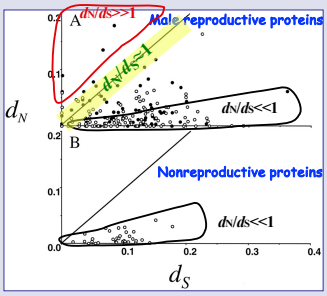
Findlay *et al.* (2010) Proteomics enhances evolutionary and functional analysis of reproductive proteins. *BioEssays* 32: 26-36.

Wilburn and Swanson (2016) From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *J. Proteomics* 135: 12-25.

23

Selection and functions

- Elevated d_N/d_S ratios are often found in reproductive genes



(*e.g.*, genes involved in mating behavior, fertilization, spermatogenesis, ejaculation, or sex determination)

Swanson *et al.* (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *PNAS* 98: 7375-7379.

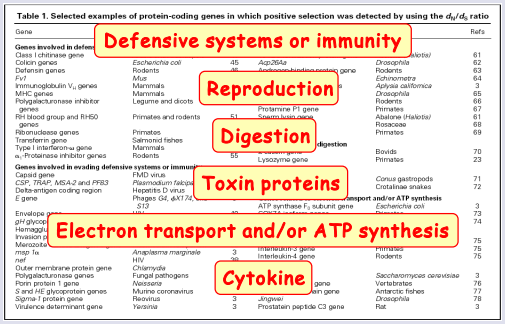
See also

Findlay *et al.* (2010) Proteomics enhances evolutionary and functional analysis of reproductive proteins. *BioEssays* 32: 26-36.

Wilburn and Swanson (2016) From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *J. Proteomics* 135: 12-25.

24

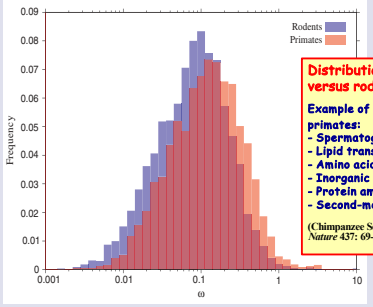
Selection and functions



Yang & Bielawski (2000) Statistical methods for detecting molecular adaptation. *TREE* 15: 496-503.

25

Identifying Selection



Distribution of ω (d_N/d_S) in primates versus rodents.

Example of genes whose ω is larger in primates:
- Spermatogenesis
- Lipid transport
- Amino acid transport
- Inorganic anion transport
- Protein amino acid glycosylation
- Second-messenger-mediated signaling

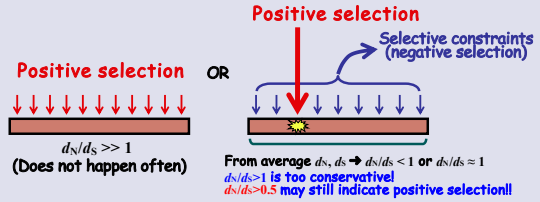
(Chimpanzee Sequencing and Analysis Consortium, 2005 *Nature* 437: 69-87)

Gibbs *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316: 222-34.

26

Identifying Selection

- Positive selection usually works within a limited region (*e.g.*, a few particular amino acid sites: binding sites)



$d_N/d_S >> 1$ (Does not happen often)

From average d_S, $d_S \rightarrow d_N/d_S < 1$ or $d_N/d_S \approx 1$

$d_N/d_S>1$ is too conservative!

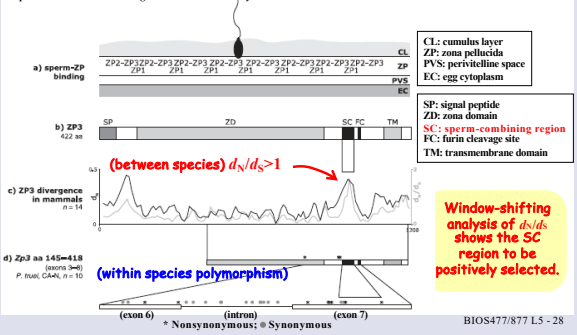
$d_N/d_S>0.5$ may still indicate positive selection!!

- If we can identify sites under positive selection, we can predict amino acid positions that are important in adaptive evolution

27

Identifying Selection

Turner and Hoekstra (2008) Reproductive protein evolution within and between species: maintenance of divergent ZP3 alleles in *Peromyscus*. *Mol. Ecol.* 17: 2616-2628.



CL: cumulus layer
ZP: zona pellucida
PVS: perivitelline space
EC: egg cytoplasm

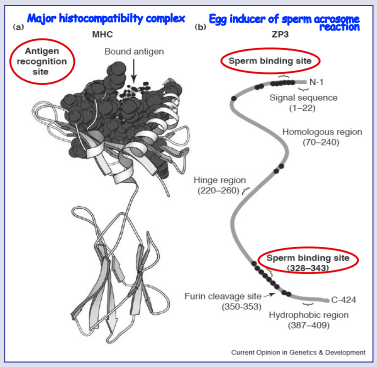
SP: signal peptide
ZD: zona domain
SC: sperm-combining region
FC: furin cleavage site
TM: transmembrane domain

Window-shifting analysis of d_N/d_S shows the SC region to be positively selected.

* Nonsynonymous; ⊕ Synonymous

28

Identifying Selection



Swanson (2003) Adaptive evolution of genes and gene families. *Curr Opin Genet Dev* 13:617–622.

[Originally from: Swanson *et al.* (2001) Darwinian selection drives the evolution of several female reproductive proteins in mammals. *PNAS* 98: 2509-2514.]

● sites:
Predicted to be subjected to positive selection

by using
a maximum likelihood model that allows d_N/d_S to vary among amino acid sites

PAML:
http://abacus.gene.ucl.ac.uk/software/paml.html
(reviewed in Yang, 2007)

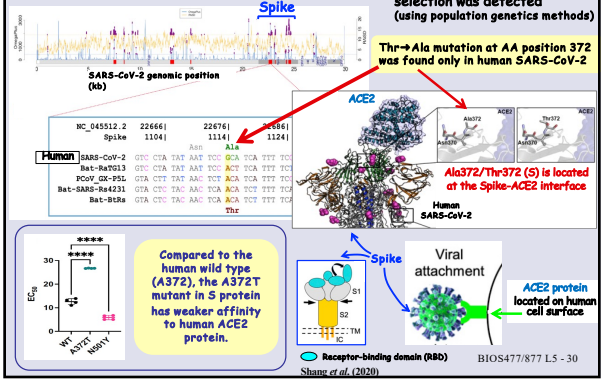
See also HyPhy and Datamonkey:
http://hyphy.org
http://www.datamonkey.org
(characterize various selective pressures)

29

A selective sweep in the *Spike* gene has driven SARS-CoV-2 human adaptation

Kang *et al.*, (2021) *Cell* 184, 4392–4400

Lin Kang,[1,2] Guijuan He,[1] Amanda K. Sharp,[1] Xiaofeng Wang,[1] Anne M. Brown,[4,5,6] Pawel Michalak,[1,7,8,*] and James Weger-Lucarelli[1,9,*]



: Regions where positive selection was detected (using population genetics methods)

Thr→Ala mutation at AA position 372 was found only in human SARS-CoV-2

Shang *et al.* (2020)

30

Identifying Selection - Data Quality

The quality of the sequence, the degree of misannotation, and ambiguities in the multiple sequence alignment, all affect identification of positive selection!

Table 1
Inferred Percentage of PSGs as a Function of Sequencing Coverage (PSG: positively selected genes)

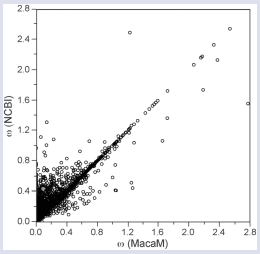
Many are based on errors

	Higher quality→ Coverage ≥3×			Coverage <3× ←Lower quality			
	Total	PSG	% PSG	Total	PSG	% PSG	$P(\chi^2)$
Human	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Chimp	1,144	9	0.8	1,836	74	4.0	9.6×10^{-8}
Macaque	896	32	3.6	2,084	488	23.4	8.1×10^{-46}
Mouse	2,493	77	3.1	487	37	7.6	1.3×10^{-6}
Rat	1,841	93	5.1	1,139	217	19.1	3.2×10^{-38}
Dog	1,568	212	13.5	1,412	481	34.1	1.7×10^{-49}
Cow	1,086	54	5.0	1,894	249	13.1	4.8×10^{-14}
Total	9,028	477	5.3	8,852	1,546	17.5	2.3×10^{-166}

Schneider *et al.* (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* 1: 114-8.

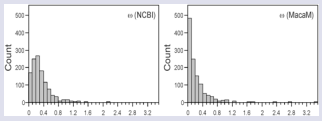
31

Identifying Selection - Data Quality



Gradnigo *et al.* (2016) Advantages of an improved rhesus macaque genome for evolutionary analyses. *PLOS ONE* 11: e0167376.

Low quality annotation of the rhesus macaque genome inflated estimated ω (*dN*/*dS*).

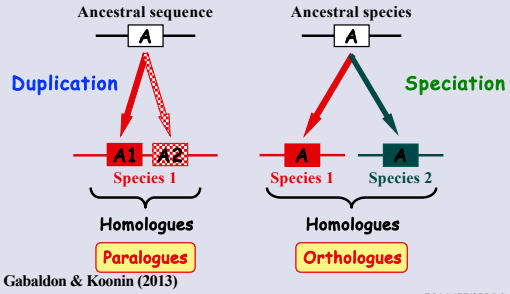


- Human vs. rhesus genes
- NCBI annotation: draft quality rhesus genome
- MacaM annotation: new high quality rhesus genome (same assembly but different gene annotation)

32

Homologue, Orthologue, Paralogue

➢ Homologues: sequences that share a common ancestor

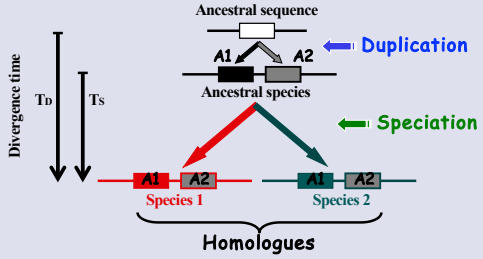


Read Gabaldon & Koonin (2013)
Mini-reviews: Koonin (2001)/Jensen (2001)/Theißen (2002)

33

Homologue, Orthologue, Paralogue

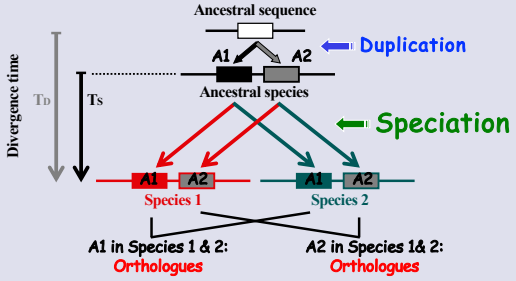
➢ Homologues: sequences that share a common ancestor



34

Homologue, Orthologue, Paralogue

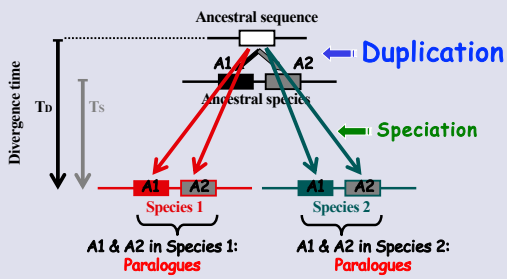
➢ Orthologues: derived from a speciation event



35

Homologue, Orthologue, Paralogue

➢ Paralogues: derived from a duplication event

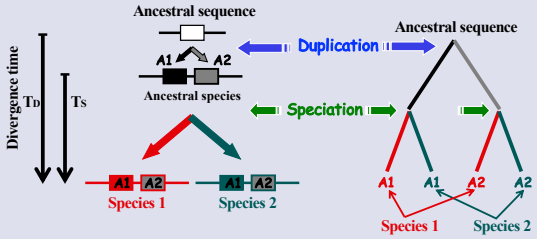


A1:Species 1 & A2:Species 2, A2:Speices 1 & A1:Species 2 are also paralogues (paralogues can be in different genomes)

36

Homologue, Orthologue, Paralogue

➢ Phylogenetic relationship of gene families



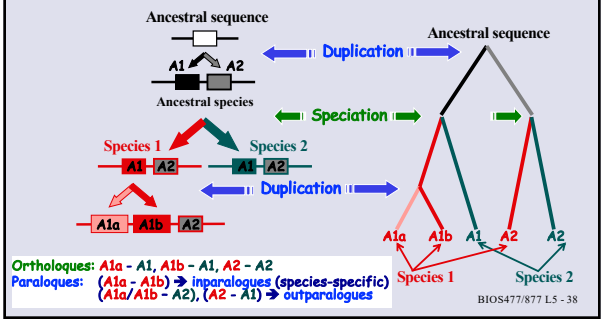
A diagram to clarify duplication vs. speciation (a visual help, informal)

Duplication: joined by a horizontal line
Speciation: joined by an inverted Y
See Jensen (2001)

37

Homologue, Orthologue, Paralogue

➢ Duplications can happen anytime



Orthologues: A1a - A1, A1b - A1, A2 - A2
Paraloques: (A1a - A1b) ➔ inparalogues (species-specific)
(A1a/A1b - A2), (A2 - A1) ➔ outparalogues

38

Gene Duplication

➢ Gene duplication creates large gene families

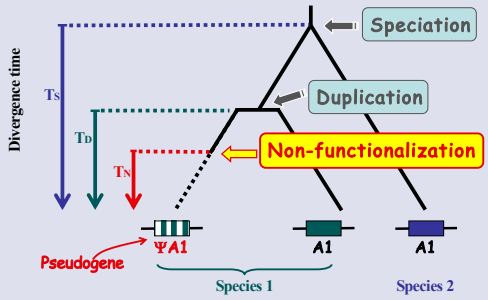
➢ Important means of providing a substrate on which evolution can work

- 15% or more of human genes are duplicates!

1) Duplicated copies: Redundancy
   - ➔ Weaker functional constraint (neutral or near neutral)
   - ➔ One copy can be degenerated to become a pseudogene
   - ➔ Non-functionalization … likely to be lost
2) One copy can acquire a novel gene function
   - ➔ Neo-functionalization, rare but important
3) Sub-functionalization?
   - ➔ Functional complementation between duplicated genes

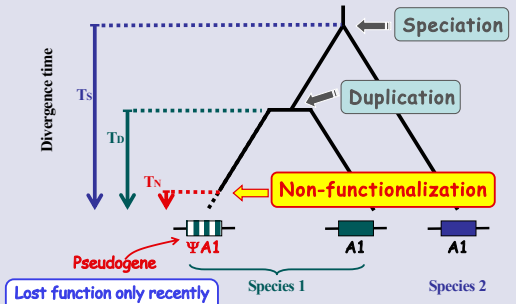
39

Pseudogene



40

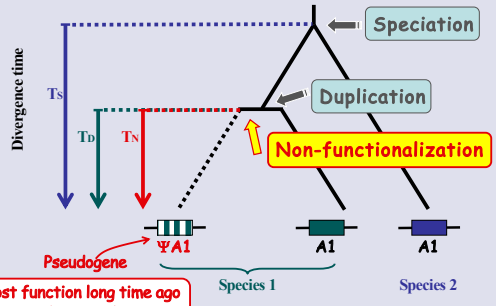
Pseudogene



Lost function only recently ($d_N/d_S \ll 1$)

41

Pseudogene

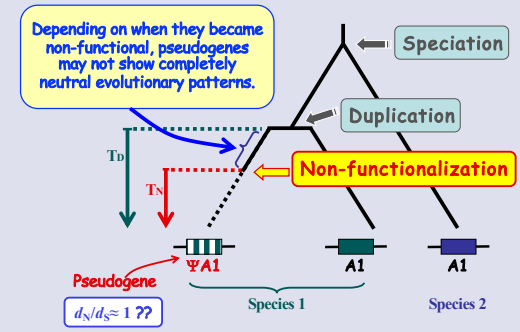


Lost function long time ago ($d_N/d_S \approx 1$)

42

Pseudogene

Depending on when they became non-functional, pseudogenes may not show completely neutral evolutionary patterns.

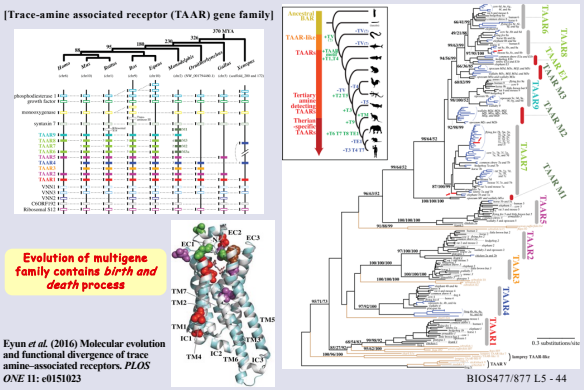


$d_N/d_S \approx 1$??

43

Gene duplication creates gene family: multigene family

[Trace-amine associated receptor (TAAR) gene family]



Evolution of multigene family contains *birth and death* process

Eyun *et al.* (2016) Molecular evolution and functional divergence of trace amine–associated receptors. *PLOS ONE* 11: e0151023

44

Homology vs. Similarity

- **Similarity**: the extent to which sequences are related.
  → makes no statement about descent from a common ancestor
- **Homology**: sequence similarity that can be attributed to descent from a common ancestor

Homology ≠ Similarity !!

- Sequences can be either homologous or non-homologous, but not in between (*e.g.*, you cannot say two genes are 10% homologous!)
- Homology is not directly measurable or observable.
- Similarity is a direct measurement.

Higgs and Attwood (2005) Chapter 1 page 8
Read also Petsko (2001), Koonin (2001), Jensen (2001)

45

Next time ...

- Pairwise sequence comparison by dotplot
  - Dotlet JS https://dotlet.vital-it.ch
  - Dotlet (old Java version) https://myhits.isb-sib.ch/cgi-bin/dotlet
    [For security reasons, old Java programs are not available within UNL network]
  - DotMatcher (A program in EMBOSS)
    http://emboss.bioinformatics.nl/cgi-bin/emboss/dotmatcher
    (see course Web for other EMBOSS links)
  - JDotter: Java Dot Plot Alignments
    http://pgrc.ipk-gatersleben.de/jdotter/
  - YASS: Visualization of local pairwise alignments
    https://bioinfo.univ-lille.fr/yass/index.php

See the Course website for more programs

46