BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 | Lecture 5

1

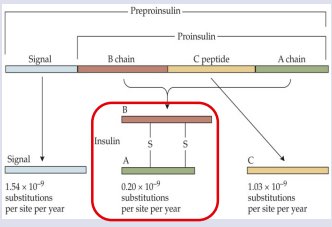
Today's topics

- Molecular Evolution - part 3

2

How to identify selection

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



Based on nonsynonymous substitution rates estimated from human/rat comparison (divergence time: 80 million years ago) from Graur (2016)

$d_N(AB) < d_N(C)$

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

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

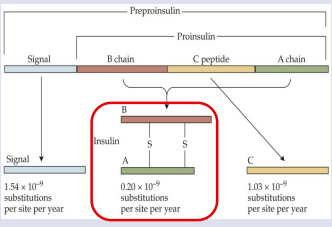
Both mutation rates (u_T) and selective constraints (f_0) affect nonsynonymous rates (d_N)

Note: In literatures, *Ka* and *Ks* are also used instead of d_N and d_S

3

How to isolate u_T vs. f_0

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



Based on nonsynonymous substitution rates estimated from human/rat comparison (divergence time: 80 million years ago) from Graur (2016)

$d_N(AB) < d_N(C)$

$\Downarrow$

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

What causes $d_N(AB) < d_N(C)$?

Both u_T and f_0 affect d_N:

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

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

or both?

4

How to isolate u_T vs. f_0 (continued)

$d_N(AB) < d_N(C)$

$\Downarrow$

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

If mutation rates are constant within a gene,

→ $u_T(AB) = u_T(C)$

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

If mutation rates are different,

→ $u_T(AB) \neq u_T(C)$

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

5

How to isolate f_0

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

$f_0 \approx 1$ (neutral)

→ $d_S = u_T f_0 \approx u_T$

$d_S(AB) = u_T(AB)$

$d_S(C) = u_T(C)$

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

$d_S(AB) = u_T(AB)$

$\Downarrow$

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

$\Downarrow$

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

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

$d_S(C) = u_T(C)$

$\Downarrow$

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

$\Downarrow$

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

6

How to isolate f_0 (continued)

If synonymous substitutions are assumed to be neutral or near-neutral: $\frac{d_N(AB)}{d_S(AB)} = f_0(AB)$ and $\frac{d_N(C)}{d_S(C)} = f_0(C)$

d_N/d_S can be used to identify selection

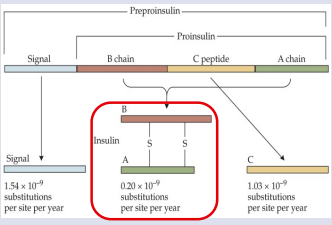
- Shows only the level of selective constrains (f_0)
- We don't need to worry about mutation rates (u_T)

→ even if $u_T(AB) \neq u_T(C)$, d_N/d_S can be compared

7

Identifying selection

d_N/d_S can be used to identify selection



Based on nonsynonymous substitution rates estimated from human/rat comparison (divergence time: 80 million years ago) from Graur (2016)

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

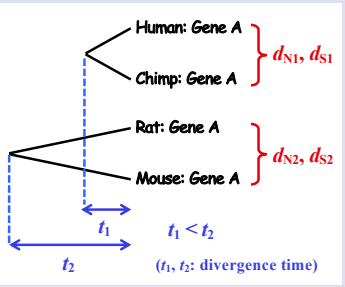
The better comparison is:

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

Even if mutation rates are different, we can compare: $f_0(AB)$ vs. $f_0(C)$

8

When divergence time is different



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

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.

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

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

9

Identifying selection

If nonsynonymous (replacement) substitutions are neutral:

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

10

Identifying selection (continued)

If nonsynonymous (replacement) substitutions are neutral:

$d_N = u_T * f_0$, where $f_0 \approx 1$

→ $d_N = u_T * f_0 \approx u_T$

Since $d_S = u_T$ → $d_N \approx d_S$

11

Identifying selection (continued)

If 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)

→ d_N is evaluated against d_S

12

Identifying selection (continued)

d_N/d_S can be used to identify selection

$d_N/d_S > 1.0$

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

$d_N/d_S < 1.0$

d_N/d_S 0

BIOS477/877 L5 - 13

13

Identifying selection (continued)

d_N/d_S can be used to identify selection

$d_N/d_S > 1.0$

Which indicates negative selection?

$d_N/d_S < 1.0$

d_N/d_S 0

BIOS477/877 L5 - 14

14

Type of selection (continued)

d_N/d_S can be used to identify selection

Adaptive selection could be driving amino acid substitutions!

Positive selection

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

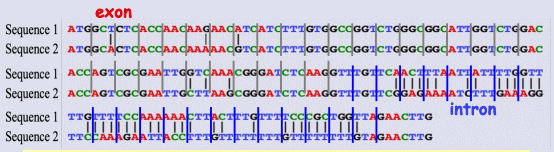
(Negative but relaxed selection)

0 Negative selection
(under strong functional constraints)

BIOS477/877 L5 - 16

16

Identifying selection (exon vs. intron)



	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

<<

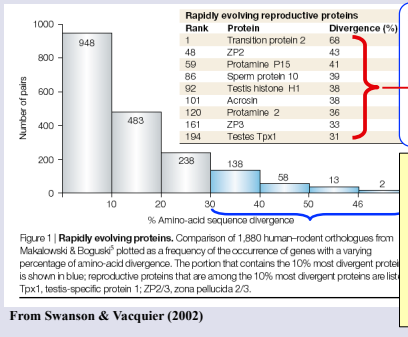
Intron(2nd) / Intron(3rd)
0.86 (≈1)

(Similar to d_N/d_S analysis)

BIOS477/877 L4 - 17

17

Positive selection (reproductive proteins)



Rapidly evolving reproductive proteins

Rank	Protein	Divergence (%)
1	Transition protein 2	68
48	ZP2	43
59	Protamine P15	41
86	Sperm protein 10	39
92	Testis histone H1	38
101	Acrosin	38
120	Protamine 2	36
161	ZP3	33
194	Testes Tpx1	31

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 protei is shown in blue; reproductive proteins that are among the 10% most divergent proteins are list Tpx1, testis-specific protein 1; ZP2/3, zona pellucida 2/3.

From Swanson & Vacquier (2002)

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

Many rapidly evolving proteins are associated to reproduction system

Why are they evolving fast?

Weak selective constrains?
(relaxed selection)

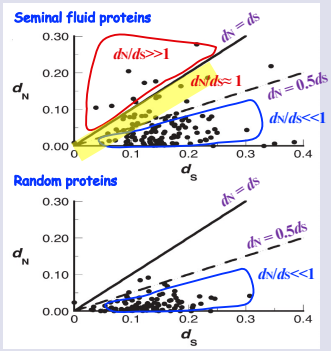
or

Divergence is favored?
(positive selection)

BIOS477/877 L5 - 18

18

Positive selection (reproductive proteins)



Elevated d_N/d_S ratios are often found in reproductive genes

e.g., genes involved in:
mating behavior, fertilization, spermatogenesis, ejaculation, sex determination

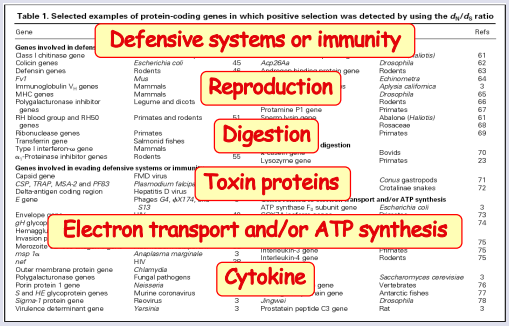
Possibly due to sexual selection, preventing hybridization, etc.

Based on comparisons between *D. melanogaster* and *D. simulans*; Swanson *et al.* (2001), Findlay and Swanson (2010)

BIOS477/877 L5 - 19

19

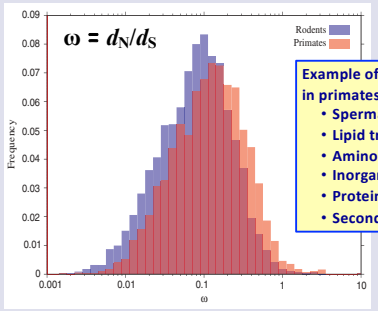
Proteins under positive selection



Yang & Bielawski (2000)

20

Distribution of d_N/d_S: primates vs. rodents



Example of genes whose ω is larger in primates:

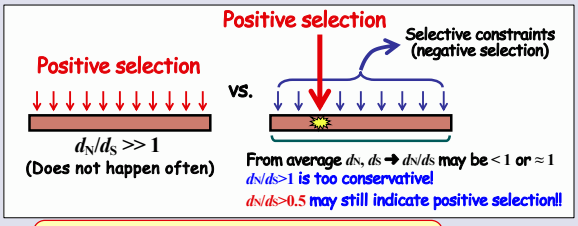
- Spermatogenesis
- Lipid transport
- Amino acid transport
- Inorganic anion transport
- Protein amino acid glycosylation
- Second-messenger-mediated

Gibbs *et al.* (2007); also see Chimpanzee Sequencing and Analysis Consortium (2005)
Rodents: mouse vs. rat genomes;
Primates: human, chimpanzee, vs. rhesus macaque genomes

21

Identifying positive selection

Positive selection usually works within a limited region, (e.g., a few particular amino acid), not for the entire coding region

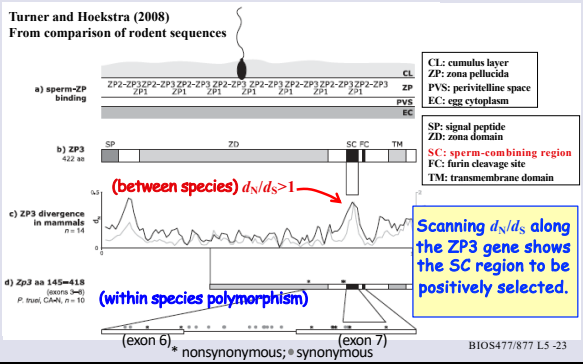


If we can identify specific sites under positive selection, we can predict amino acid positions that are important in adaptive evolution

22

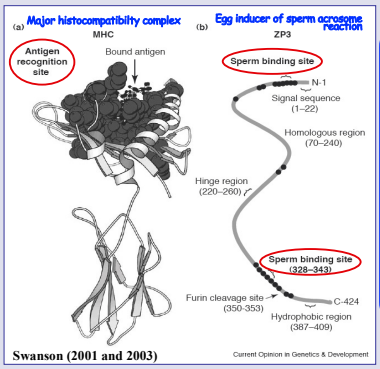
Scanning d_N/d_S

Turner and Hoekstra (2008)
From comparison of rodent sequences



23

Identifying sites under positive selection



● sites: Predicted to be under positive selection

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

- PAML (see Alvarez-Carretero et al. 2023 for a guide)

also

- HyPhy
- Datamonkey

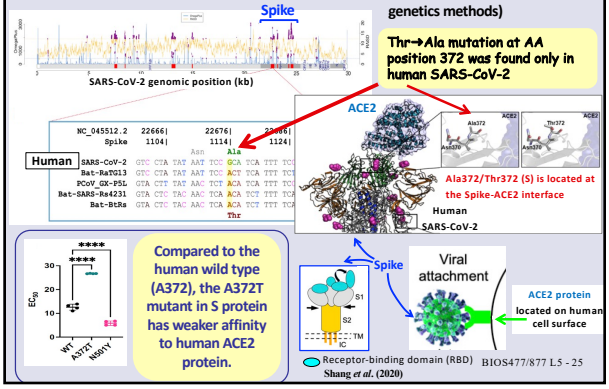
(various selective pressures can be characterized)

24

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

Kang *et al.*, (2021)

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



25

Detection of selection and 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

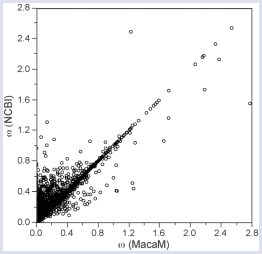
Higher quality → Coverage ≥3× | Coverage <3× ← Lower quality

	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}

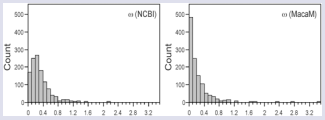
Schneider *et al.* (2009)

26

Detection of selection and data quality
(continued)



Low quality annotation of the rhesus macaque genome inflated estimated ω (d_N/d_S)



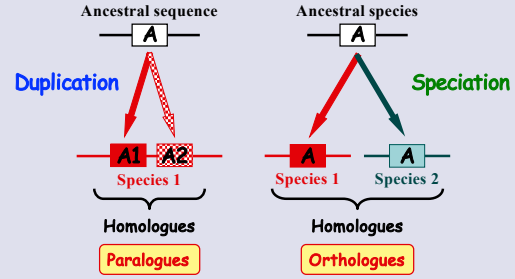
- Comparisons: human vs. rhesus genes
- Genome annotations:
  - NCBI: draft quality rhesus genome
  - MacaM: new high quality rhesus genome (same assembly but different gene annotation)

Gradnigo *et al.* (2016)

27

Homologue, orthologue, paralogue

➢ Homologues: sequences that share a common ancestor

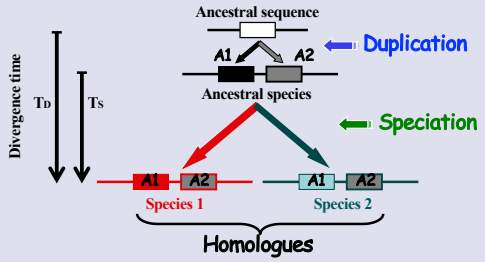


Read Theißen (2002) [very short review]
Gabaldon & Koonin (2013)/

28

Homologue, orthologue, paralogue
(a more complex gene history)

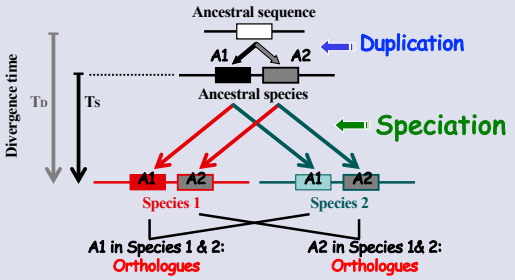
➢ Homologues: sequences that share a common ancestor



29

Homologue, orthologue, paralogue
(identifying orthologues)

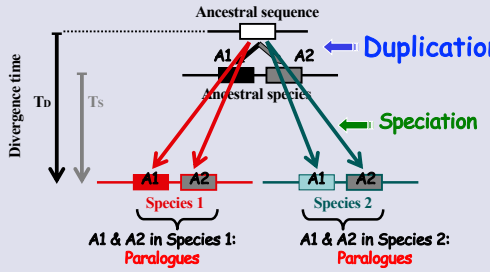
➢ Orthologues: derived from a speciation event



30

Homologue, orthologue, paralogue
(identifying paralogues)

➢ Paralogues: derived from a duplication event

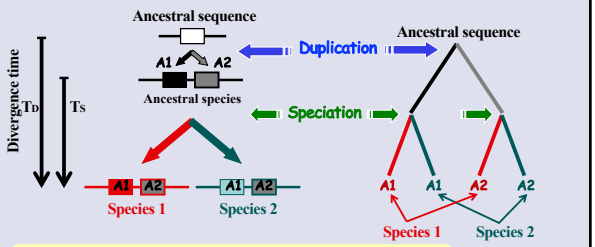


A1:Species 1 & A2:Species 2, A2:Speices 1 & A1:Species 2 are also paralogues (paralogues can be in different genomes)

31

Gene duplication and gene family

Duplication and speciation events can be identified in a phylogenetic tree of a gene family



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

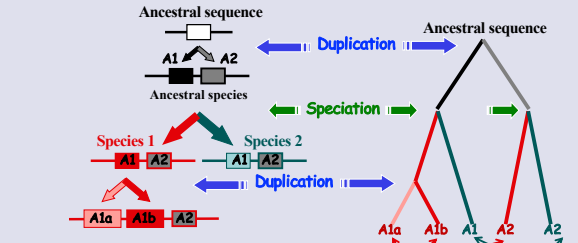
Duplication: joined by a horizontal line
Speciation: joined by an inverted Y

See Jensen (2001)

32

A complex history of gene family

Duplications can happen anytime



Orthologues: A1a - A1, A1b – A1, A2 – A2

Paralogues: (A1a - A1b) ➔ inparalogues (species-specific)
(A1a/A1b – A2), (A2 - A1) ➔ outparalogues

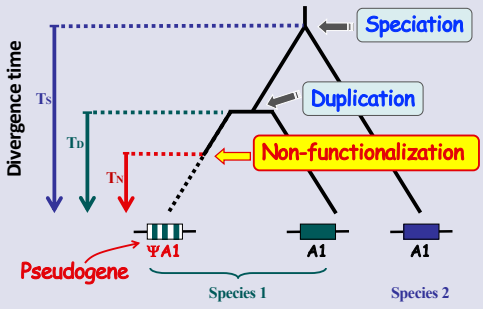
33

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 (not very well understood)
     - Functional complementation between duplicated genes

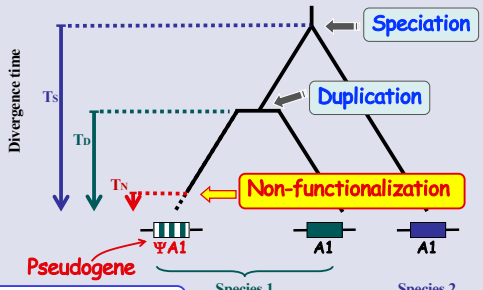
34

Pseudogene



35

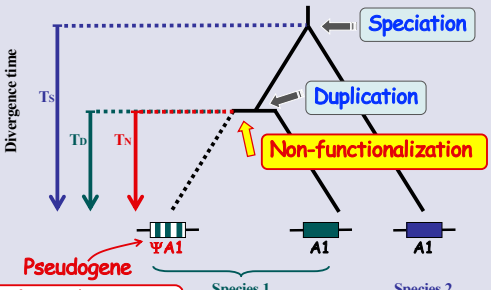
Pseudogene
(recent non-functionalization)



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

36

Pseudogene
(non-functionalization long-time ago)

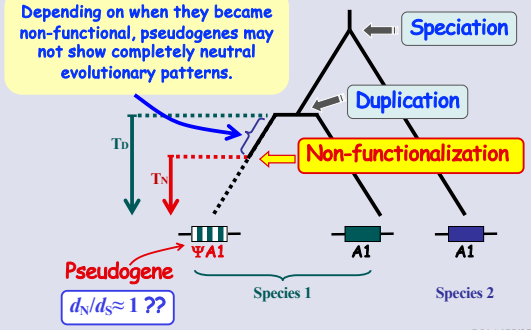


Lost function long time ago ($d_N/d_S \approx 1$)

37

Pseudogene and d_N/d_S

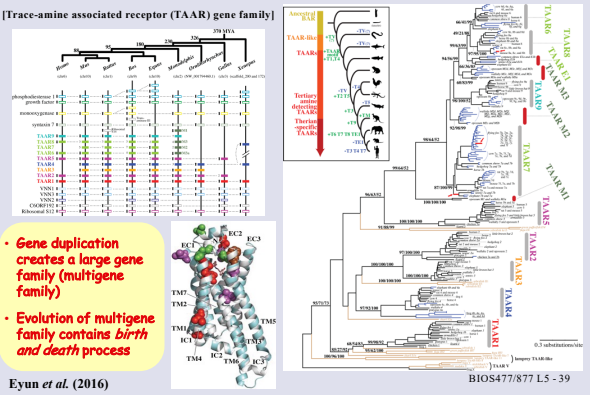
Depending on when they became non-functional, pseudogenes may not show completely neutral evolutionary patterns.



38

Multigene family

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



- Gene duplication creates a large gene family (multigene family)
- Evolution of multigene family contains *birth and death* process

Eyun *et al.* (2016)

39

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

40