IMPRS workshop

Comparative Genomics

18th-21st of February 2013

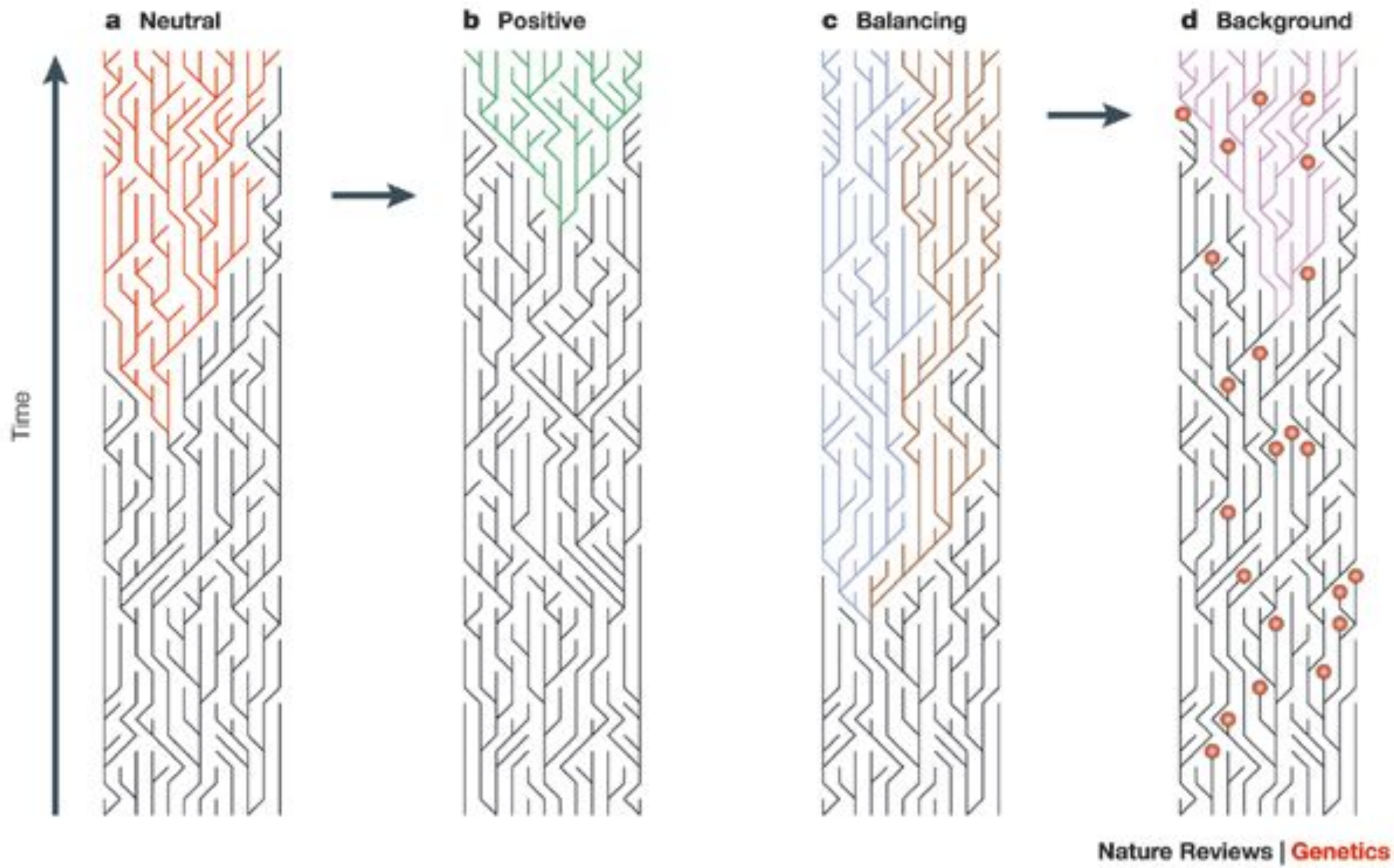
Lecture 4

Positive selection

What is positive selection?

Positive selection is selection on a particular trait

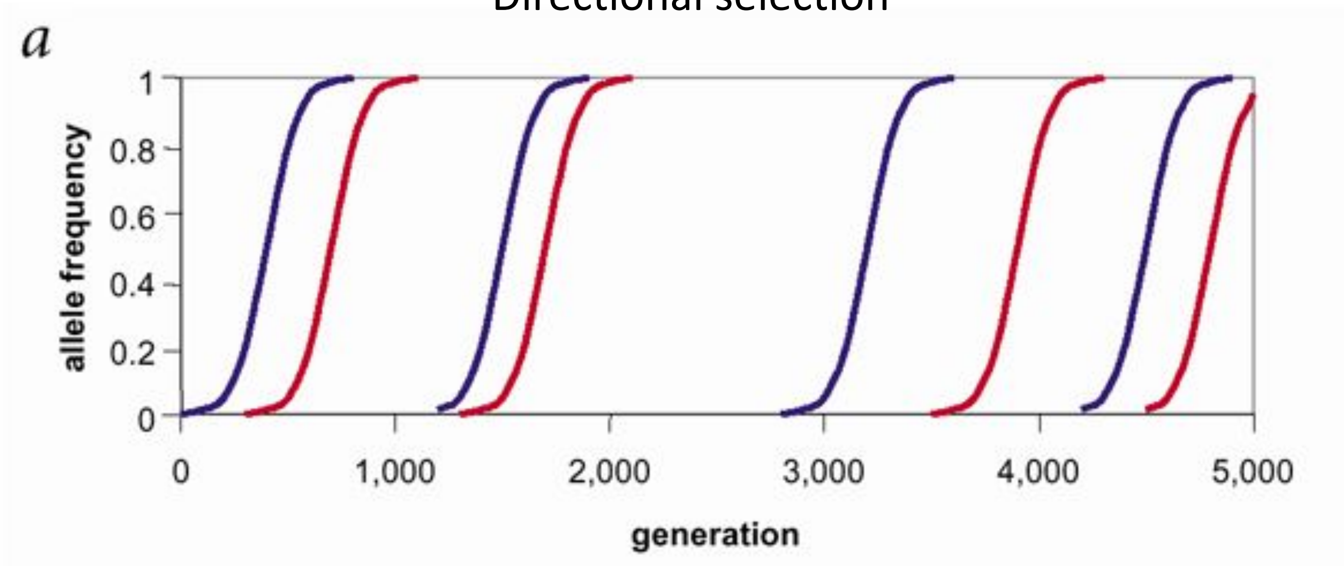
- and the increased frequency of an allele in a population



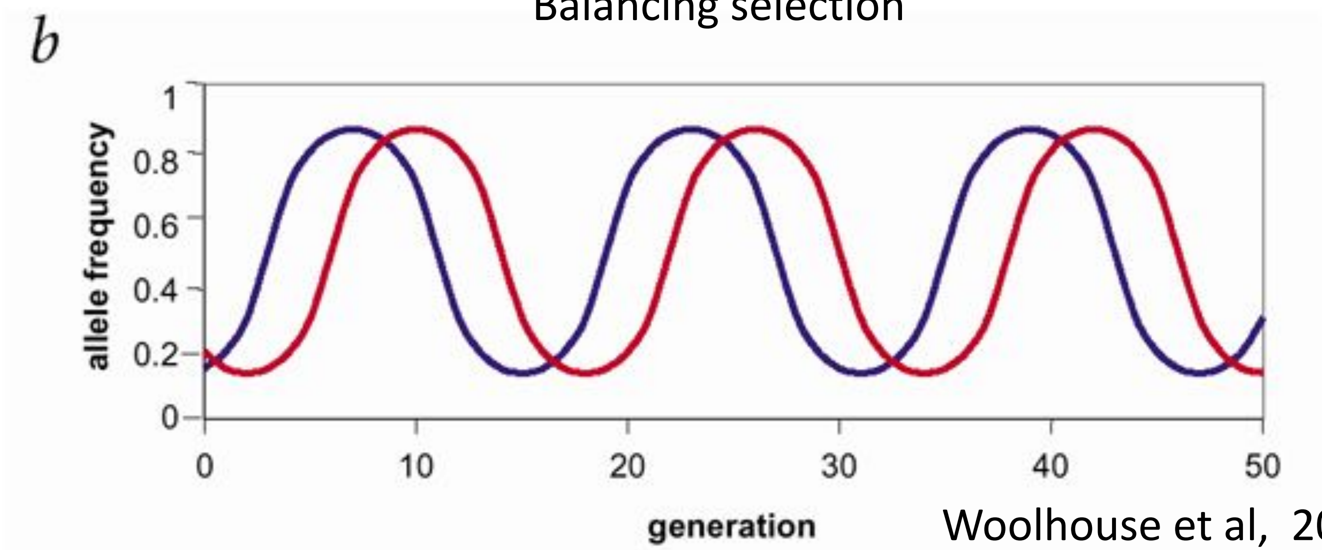
Population level

Positive selection can drive the changes in frequencies of two alleles

Directional selection



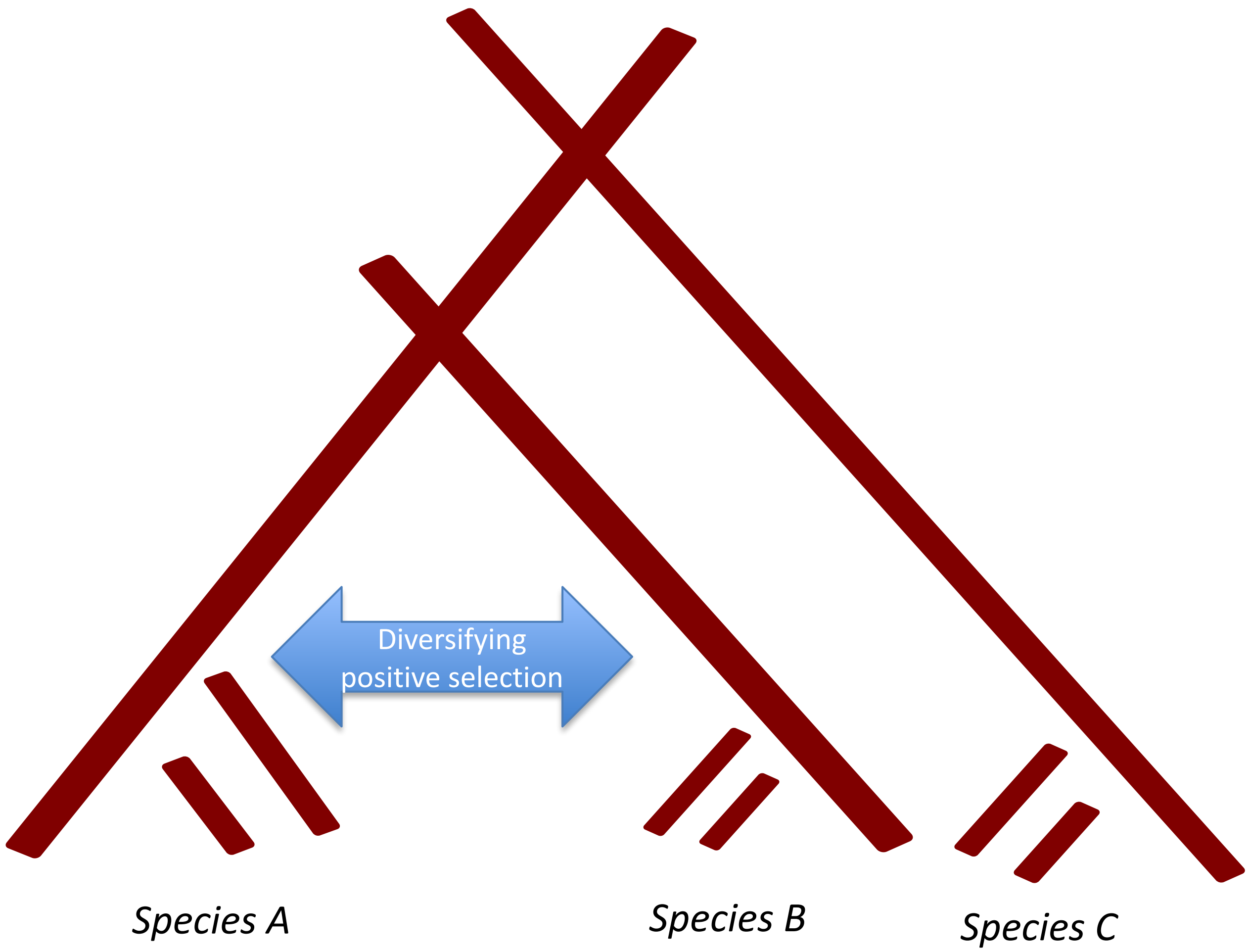
Balancing selection



Woolhouse et al, 2002. Nat.

Interspecific level

Positive selection driving divergence



Why is it interesting to identify traits which have undergone or are under positive selection?

Function

Evolution

Environment

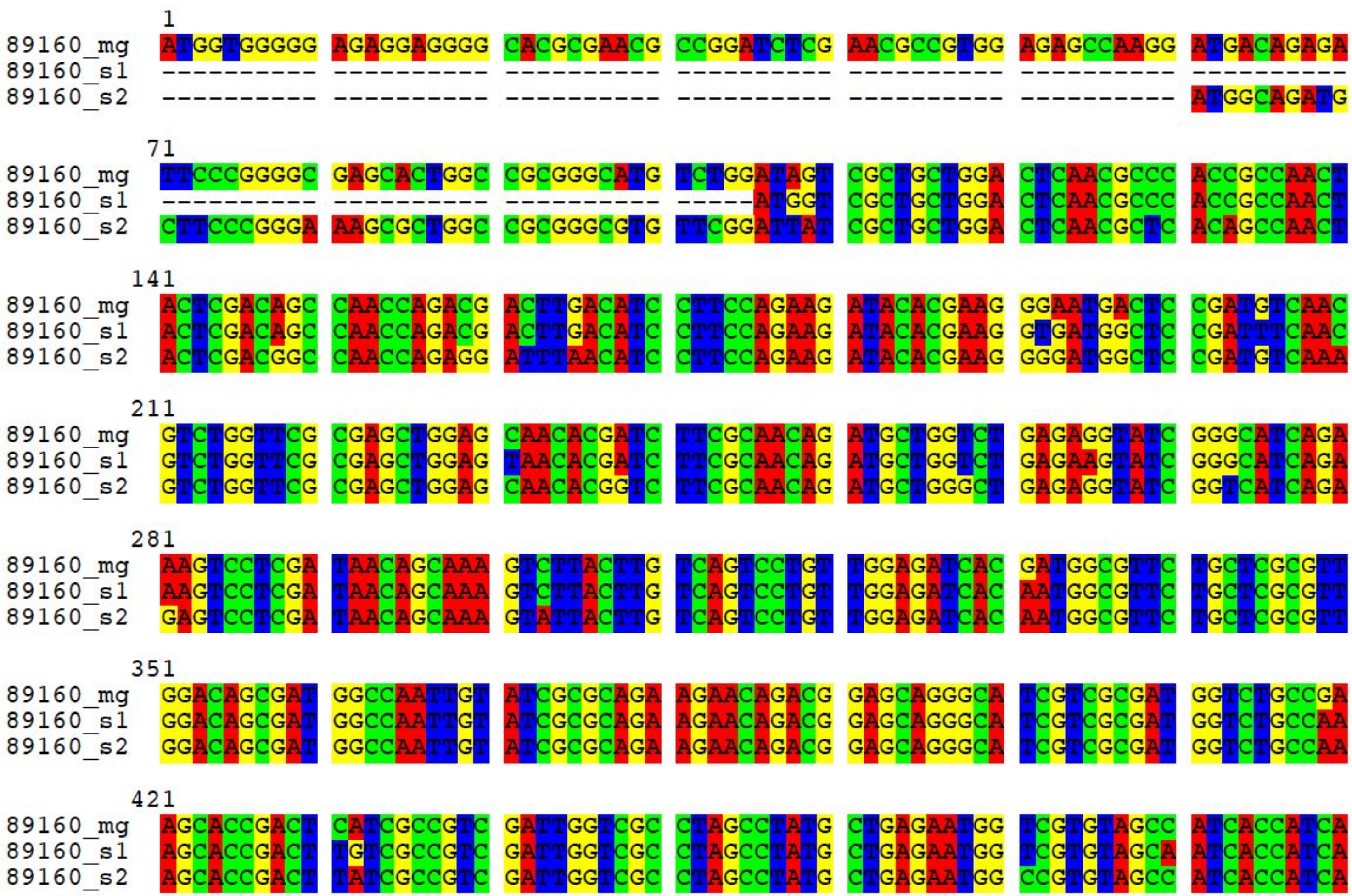
……

How can we detect positive selection?

Changes in a protein sequence….

```
             1
89160_mg     MVGERRGTRT PDLERRGEPR MTEIPGASTG RGHVWIVAAG LNAHRQLLDS QPDDLTSFQK IHEGNDSDVN
89160_s1     ---------- ---------- ---------- -----MVAAG LNAHRQLLDS QPDDLTSFQK IHEGDGSDFN
89160_s2     ---------- ---------- MADASRESAG RGRVRIIAAG LNAHSQLLDG QPEDLTSFQK IHEGDGSDVK

            71
89160_mg     VWFASWSNTI FATDAGLRGI GHQKVLDNSK VLLVSPVGDH DGVLLALDSD GQLYRAEEQT EQGIVAMVCR
89160_s1     VWFASWSNTI FATDAGLRSI GHQKVLDNSK VLLVSPVGDH NGVLLALDSD GQLYRAEEQT EQGIVAMVCQ
89160_s2     VWFASWSNTV FATDAGLRGI GHQRVLDNSK VLLVSPVGDH NGVLLALDSD GQLYRAEEQT EQGIVAMVCQ

           141
89160_mg     STDSSPSIGR LAYAENGRVA ITIKQAPNGN LCHVEEFKDL ETFLRWFQDP SGDGNYPERH FMLPGRPKQL
89160_s1     STDLSPSIGR LAYAENGRVA ITIKQAPNGN LCHVEEFKDL ETFLRWFQDP SGDGNYPERH FMLPGRPKQL
89160_s2     STDLSPSIGR LAYAENGRVA ITIKQAPNGN LCHVEEFKDL EPFLRWFQDP SGDGNHPERH FMLPGRPKQL

           211
89160_mg     KAGTGIFVLL MESGQVYTWG DSRYRSLGRS VTGDGSKSAD EPAVLEALDG LHIKKVDCCG WMSAALSDDG
89160_s1     EAGTGIFVLL MESGQVYTWG DSRYRSLGRS VTGDGNKSAD EPAVLEALDG LHIKKVDCCG WMSAALSDDG
89160_s2     EAGTGIFVLL MESGQVYTWG DPRFRSLGRS VTGDGNKSAD EPAVLEALDG LHIKKVACCG WMSAALSDDG

           281
89160_mg     ALYLWGITSP SDDVKIGALA AGEDEEVALV ELPGDGSEPL DVVDVALGVE HIAVLAESGR LFVTGDKSCG
89160_s1     ALYLWGITSP SDDVKIRALA AGEDEEVALV ELPGDDSEPL DVVDVALGVE HIAVLAESGR LFVTGDNSCG
89160_s2     ALYLWGITSP SGDVIINALT AGEDEEIALV ELPGGGSEPL DVVDVALGVE HIAVLGESGR LFVTGDNSCG
```

Come from changes in the nucleotide sequence



Quantifying non-synonymous variation

- an estimate of positive selection

Synonymous mutations: neutral mutations
Non-synonymous mutations: non-neutral mutations

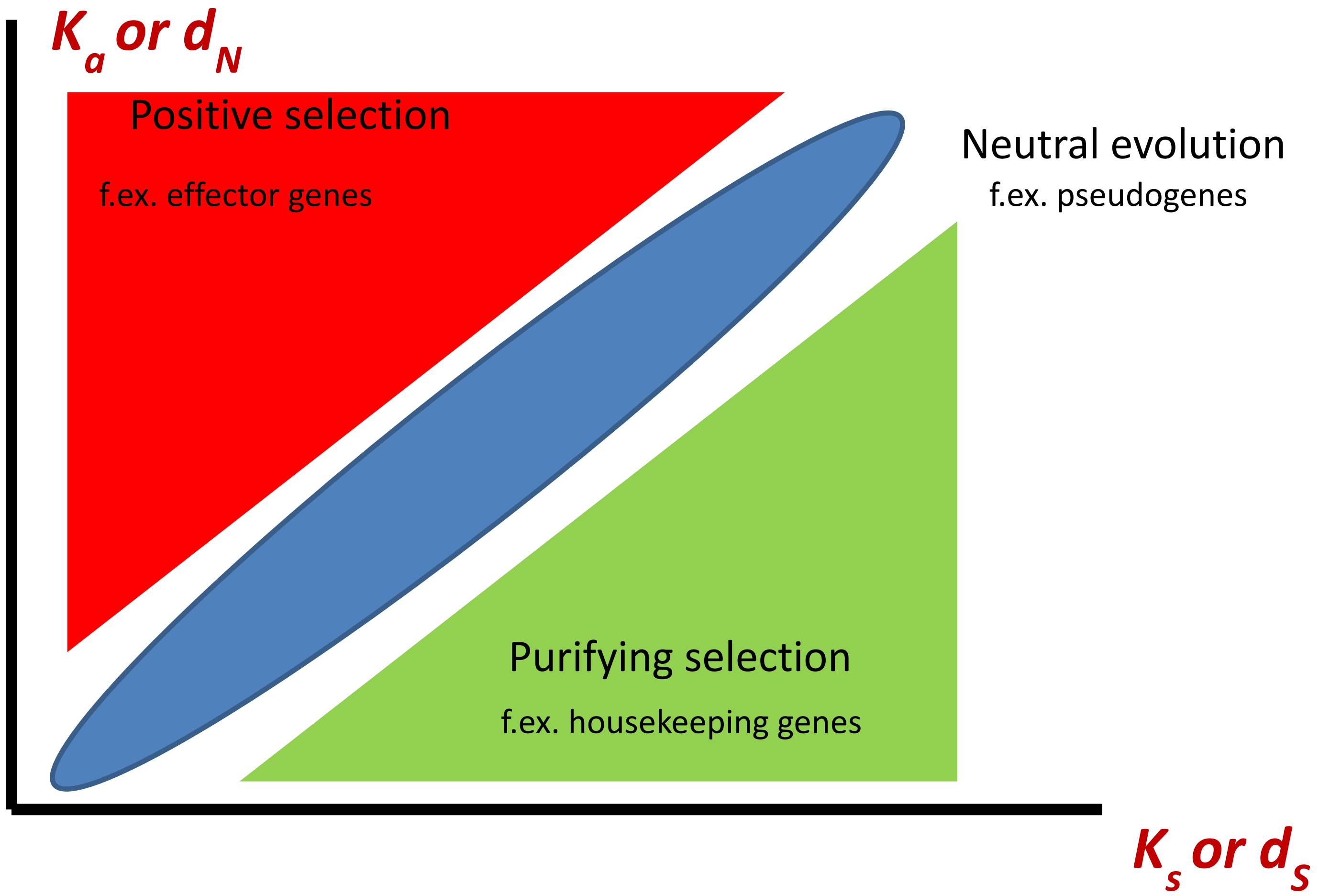
	Second Position								
	U		C		A		G		
First Position	code	Amio Acid	code	Amio Acid	code	Amio Acid	code	Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

To measure positive selection:

Rate of synonymous mutations

Rate of non-synonymous mutations

Positive selection between species

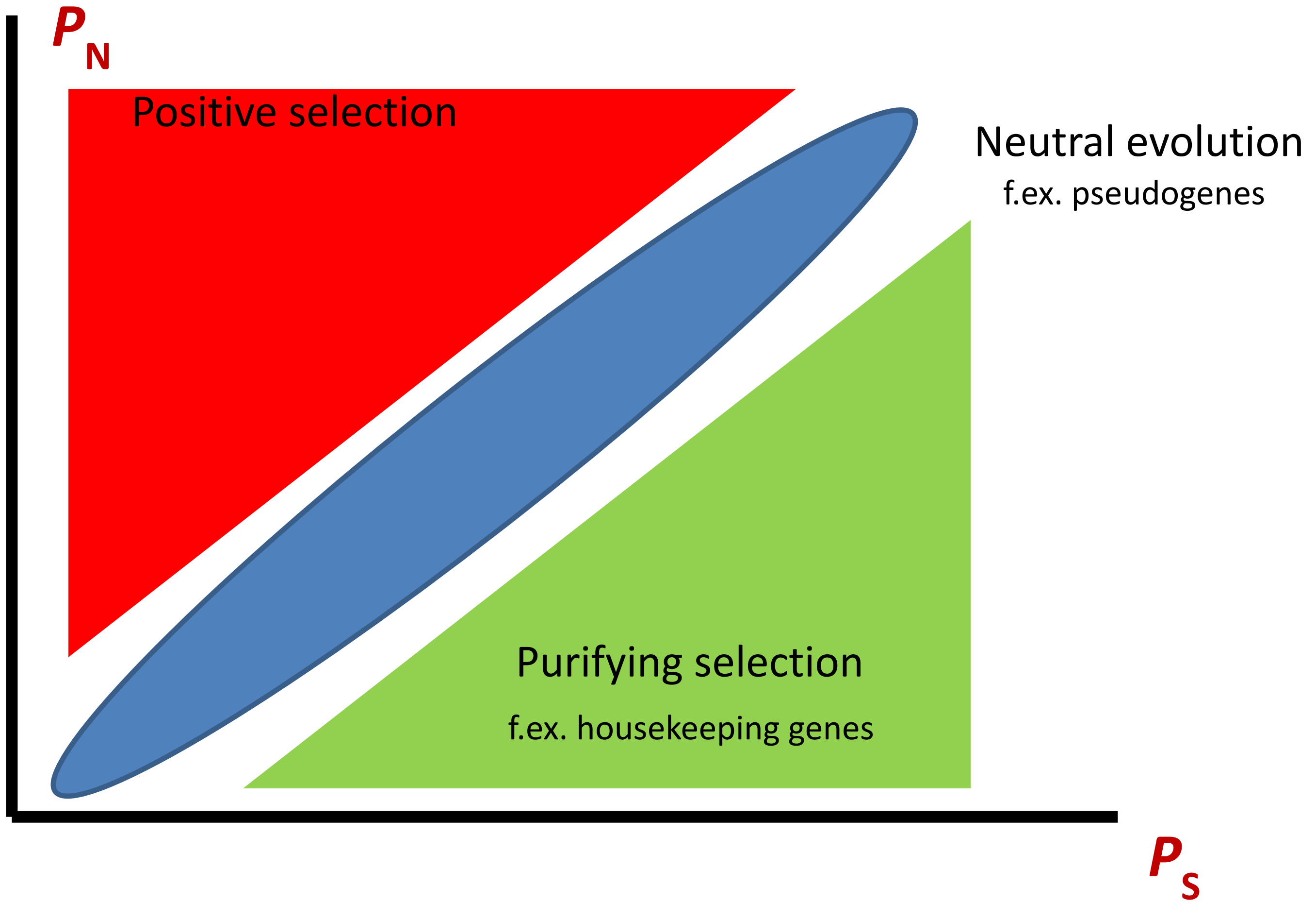


K_A or d_N: rate of non-synonymous divergence

Evolution between species

K_S or d_S: rate of synonymous divergence

Positive selection in a population



P_N: rate of non-synonymous polymorphisms

P_S: rate of synonymous polymorphisms

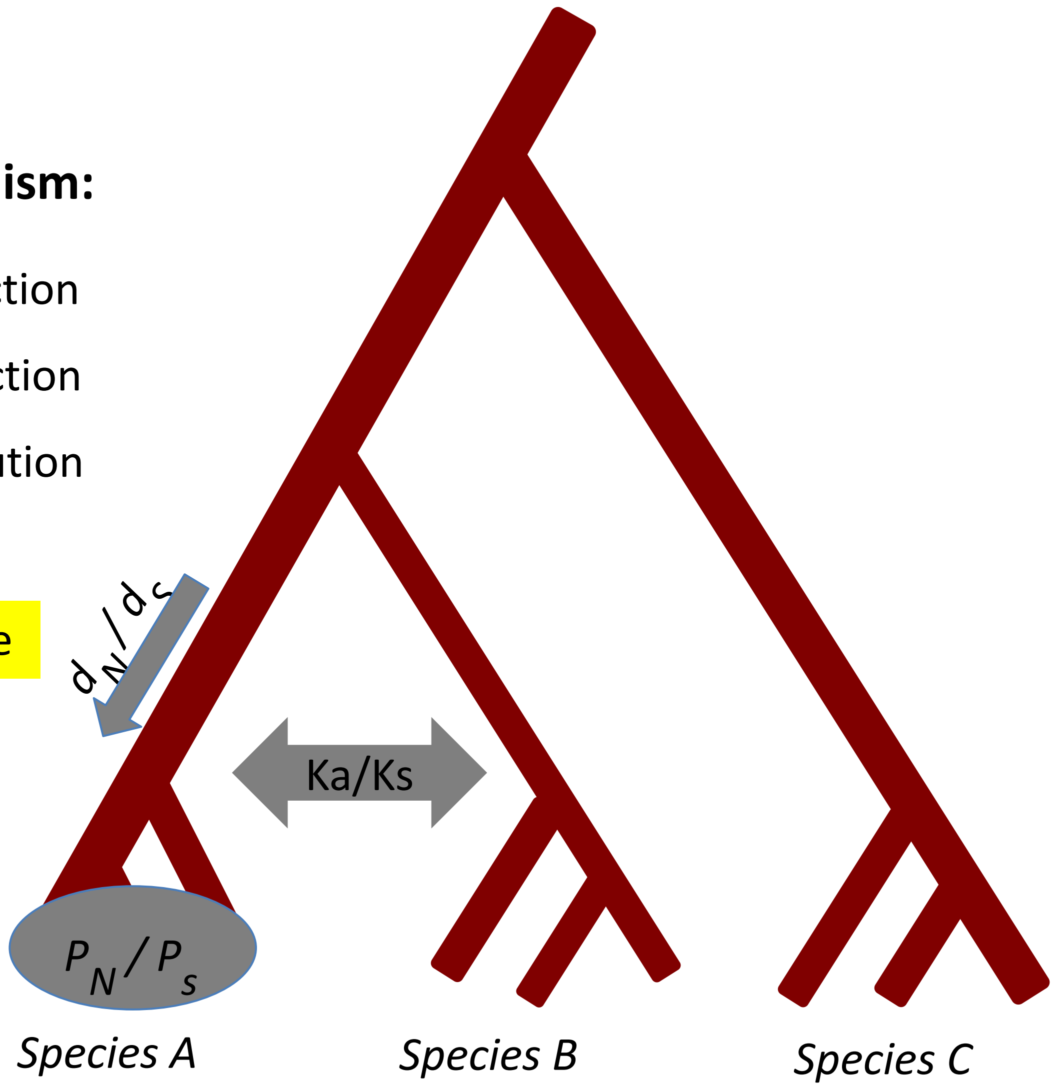
Evolution within species

Estimates of non-synonymous and synonymous polymorphisms and substitutions provide insight into the evolutionary processes

Analysing divergence and polymorphism:

- K_A / K_S ratios > 1 indicate positive selection
- K_A / K_S ratios < 1 indicate negative selection
- K_A / K_S ratios = 1 indicates neutral evolution

branch-specific estimate



K_A *and* d_N: rate of non-synonymous substitutions

K_S *and* d_S: rate of synonymous substitutions

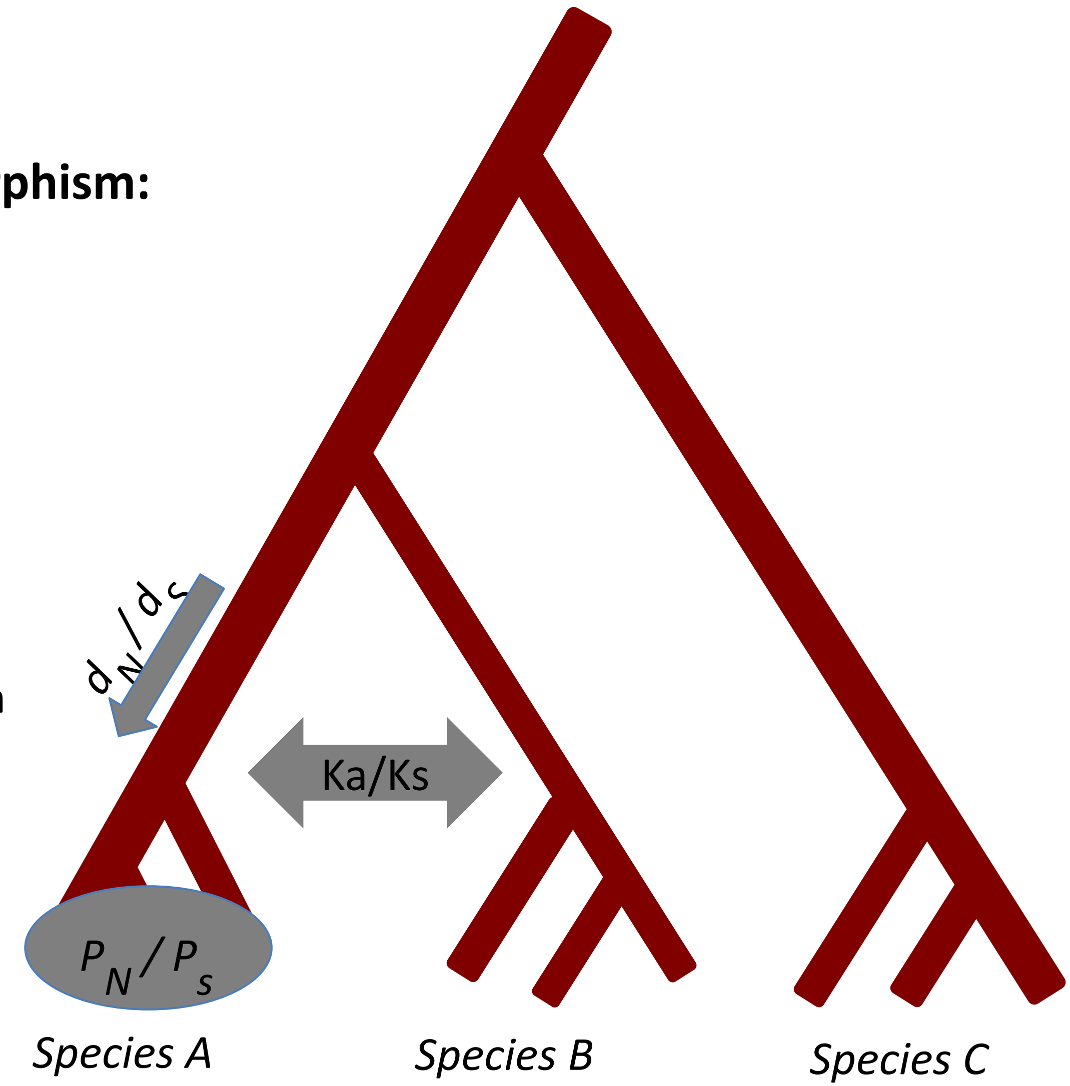
P_N: Amount of non-synonymous polymorphisms

P_S: Amount of synonymous polymorphisms

Estimates of non-synonymous and synonymous polymorphisms and substitutions provide insight into the evolutionary processes

Contrasting divergence and polymorphism:

- Ratios of K_A / K_S provide insight into the amount of non-synonymous divergence
- The branch specific d_N / d_S ratios are measures of adaptive evolution particular to one branch
- Ratios of P_N / P_S provide insight into the strength of purifying selection in the species



Basic analyses of the proportion of non-synonymous to synonymous divergence K_A/K_S

- Counts of non-synonymous mutations for each gene (N_d)

- Counts of synonymous mutations for each gene (S_d)

- Counts of potential non-synonymous sites for each gene (N)

- Counts of potential synonymous sites for each gene (S)

	Second Position								
	U		C		A		G		
First Position	code	Amio Acid	code	Amio Acid	code	Amio Acid	code	Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Non-synonymous substitution rate: $K_A = N_d / N$

Synonymous substitution rate: $K_S = S_d / S$

Ratio K_A/K_S as an inidicator of evolutionary mode in each gene

Nei and Gojobori, 1986

Counts of possible synonymous sites for each gene (S)

	1	2	3	4	5
	Pro	Phe	Gly	Leu	Phe
Seq 1	CCC	UUU	GGG	UUA	UUU
Seq 2	CCC	UUC	GAG	CUA	GUA
	Pro	Phe	Ala	Leu	Val

<u>Calculate potential synonymous sites (S) for each codon</u>

A fourfold degenerate site counts as S = 1 (N = 0)

A non-degenerate site counts as S = 0 (N = 1)

A two fold degenerate site counts as S = 1/3 (N = 2/3)

1. Proline S = 0 + 0 + 1 = 1
2. Phenylalanine S = 0 + 0 + 1/3 = 1/3
3. For Glycine S = 0 + 0 + 1 = 1, for Alanine S = 0 + 0 + 1

   Take the average: S=1
4. Leucine for UUA, S = 1/3 + 0 + 1/3 = 2/3

   for CUA, S = 1/3 + 0 + 1 = 4/3

   Take the average of these: S = 1 for codon 4

5 Phenylalanine for UUU, S = 1/3

   for guanine, S = 1

   Take average: S = 2/3

For whole sequence, S = 1 + 1/3 + 1 + 1 + 2/3 = 4

N = total number of sites: S = 15 - 4 = 11

	Second Position								
	U		C		A		G		
First Position	code	Amio Acid	code	Amio Acid	code	Amio Acid	code	Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Counts of synonymous changes

	1	2	3	4	5
	Pro	Phe	Gly	Leu	Phe
Seq 1	CCC	UUU	GGG	UUA	UUU
Seq 2	CCC	UUC	GAG	CUA	GUA
	Pro	Phe	Ala	Leu	Val

<u>Calculate S_d and N_d for each codon.</u>

1. $S_d = 0$, $N_d = 0$
2. $S_d = 1$, $N_d = 0$
3. $S_d = 0$, $N_d = 1$
4. $S_d = 1$, $N_d = 0$
5. this could happen in two ways

UUU --> GUU --> GUA

$N_d = 1$ $S_d = 1$ Route 1: $S_d = 1$, $N_d = 1$

UUU --> UUA --> GUA

$N_d = 1$ $N_d = 1$ Route 2: $S_d = 0$, $N_d = 2$

Take average of these two:

$S_d = 0.5$, $N_d = 1.5$

First Position	U code	U Amio Acid	C code	C Amio Acid	A code	A Amio Acid	G code	G Amio Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

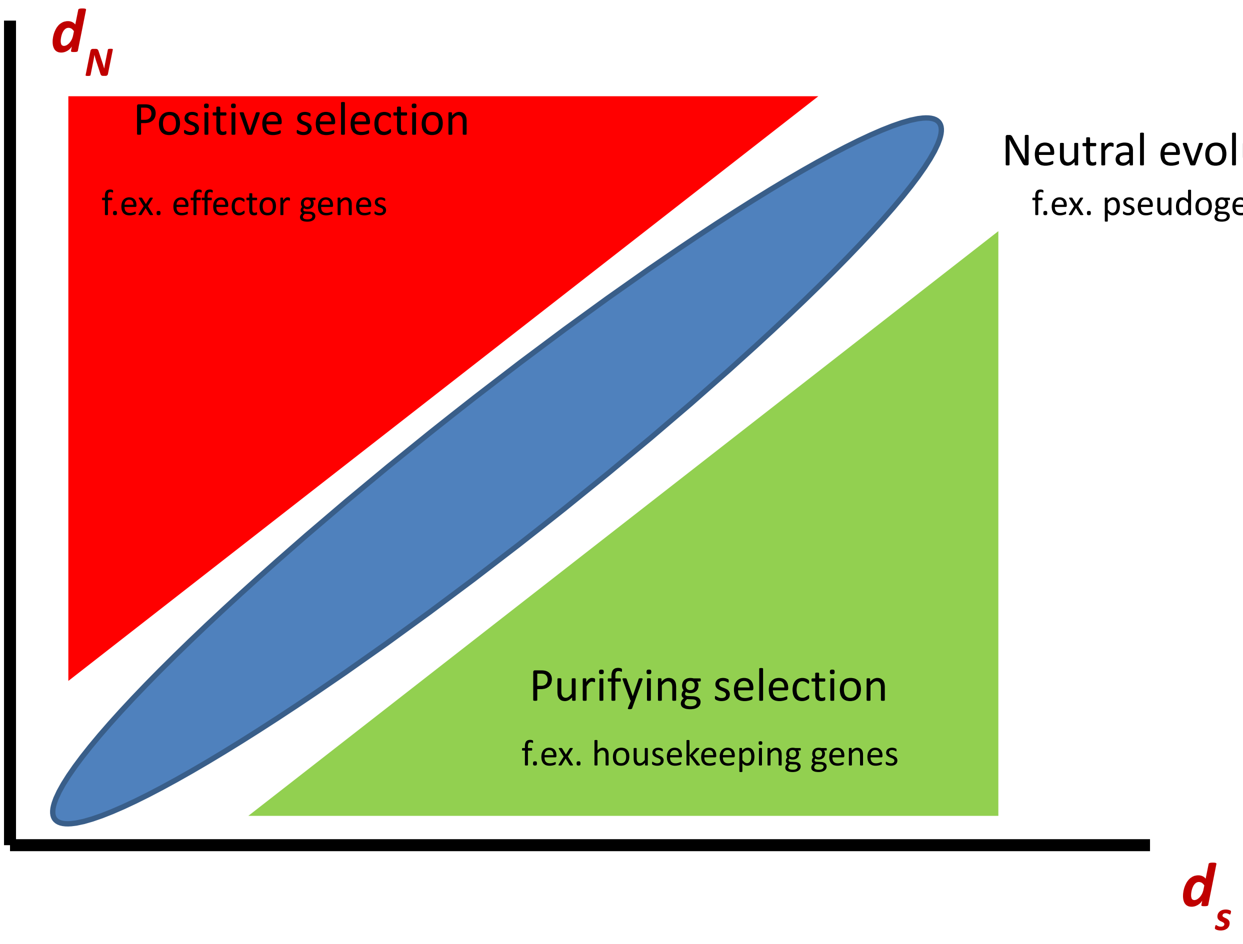
(Second Position: U, C, A, G)

Total $S_d = 2.5$ Total $N_d = 2.5$

$S_d / S = 2.5/4 = 0.625$ $N_d / N = 2.5/11 = 0.227$

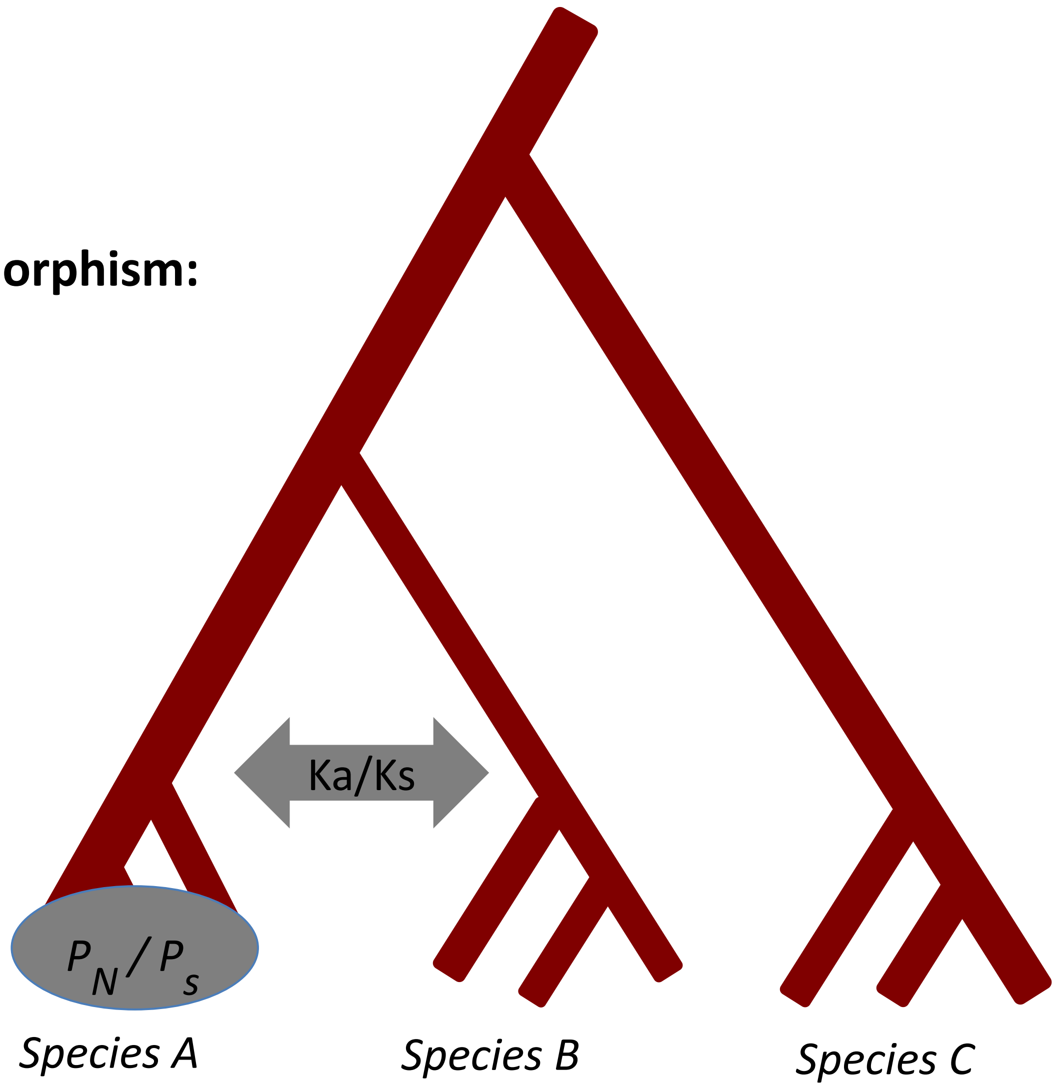
dN/dS = 0.363

Positive selection between species



When positive selection is related to species divergence

Contrasting divergence and polymorphism:



McDonald Kreitman (MK) test to contrast within and between species variation

Adaptive protein evolution at the *Adh* locus in *Drosophila*

John H. McDonald & Martin Kreitman

NATURE · VOL 351 · 20 JUNE 1991

Drosophila dataset alcohol dehydrogenase

		D. melanogaster												*D. simulans*						*D. yakuba*													
	Con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	a	b	c	d	e	f	g	h	i	j	k	l		
781	G	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Repl.	Fixed
789	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
808	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	Repl.	Fixed
816	G	T	T	T	T	-	-	-	-	-	-	-	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
834	T	-	-	-	-	-	-	-	-	-	-	-	-	C	C	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
859	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	Repl.	Fixed
867	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	A	G	G	G	G	G	G	Syn.	2 Poly.
870	C	T	T	T	T	T	T	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Fixed
950	G	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
974	G	-	-	-	-	-	-	-	-	-	-	-	-	T	-	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
983	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
1019	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	Syn.	Poly.
1031	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	Syn.	Poly.
1034	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	-	-	C	-	C	C	Syn.	Poly.
1043	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	Syn.	Poly.
1068	C	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
1089	C	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-	-	Repl.	Fixed
1101	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	A	A	A	A	A	A	A	A	Repl.	Fixed
1127	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
1131	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	Syn.	Poly.
1160	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
1175	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
1178	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	Syn.	Poly.
1184	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	G	G	G	G	G	G	G	Syn.	Fixed
1190	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	Syn.	Poly.
1196	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	-	T	T	T	-	T	-	-	Syn.	Poly.
1199	C	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
1202	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
1203	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.
1229	T	-	-	C	C	C	C	C	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Syn.	Poly.

Repl: Nonsynonymous, Syn: Synonymous
Fixed: Substitution, Poly: Polymorphisms

MK test contrasts within and between species synonymous and non-synonymous differences

The proportion of **non-synonymous fixed** differences between species much higher than the proportion of **non-synonymous polymorphisms**

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

A *G*-test of independence (with the Williams correction for continuity)[1] was used to test the null hypothesis, that the proportion of replacement substitutions is independent of whether the substitutions are fixed or polymorphic. $G = 7.43$, $P = 0.006$.

Contingency table can be tested by a G-test

Conclusion from MK-test:

Adh locus in Drosophila has accumulated adaptive mutations (been under positive selection) when the Drosophila species diverged

One problem with the "counting methods"

Sometimes the signal of selection is not very strong

Positive selection on one or few particular codons or in one particular branch

```
           1
94534_S1   MANNTKVQDR TSTNKPRPRK RQRGATRADK PLDTSQPSIF ELIPEEQVDG AIAYYYDHPE QLPYRLSGAH
94534_Mg   MANNTKVQGR TSTNKPRPRK RQRGATRADK PLDTSQPSIF ELVPEEQVDG AIAYYYDHPE QLPYRLSGAH
94534_S2   MANNTKVQDR TSANKPRPRK RQRGATRADK PLDTSQPSIF ELIPEEQVDG AIAYYYDHPE QLPYRLSGAH

          71
94534_S1   QRASQRPLRI MLKFDEKIFE RKVQLRPIPD CELEGVDPFS LDMAFETEDD DKDSAASTEP GG--GSGDTI
94534_Mg   QRASQRPLRI MLKFDEKIFE RKVQLRPIPD CELEGVDPFS LDPVFETKDD DTDSAASTEP GG--GSGDTI
94534_S2   QRASQRPLRI MLKFDEKIFE RKVQLRPIPD CELEGVDPFS LDTADEAEDD GTNTDSSSES GNVRENSASI

         141
94534_S1   YCGQESLVVG DLLHGPRDYG HILAHGQTEQ CKT*
94534_Mg   YCGQESLVMG DLLHGPRDYG YILAHGQTEQ CKT*
94534_S2   YCGQEFLIMG DLLHGPRDYG YMNPPGQIKQ CKT*
```

 Evolutionary model to detect selection in particular codons or branches