Phylogenomics and Population Genomics:
Inference and Applications

Population Genomics for Understanding Adaptation

Sebastian E. Ramos-Onsins
Centre of Research in Agricultural Genomics (CRAG)

CENTRE FOR RESEARCH IN AGRICULTURAL GENOMICS

Adaptation to a given environment

- Differences in individual features, such as viability or fecundity, between others, will have an effect on the next generation

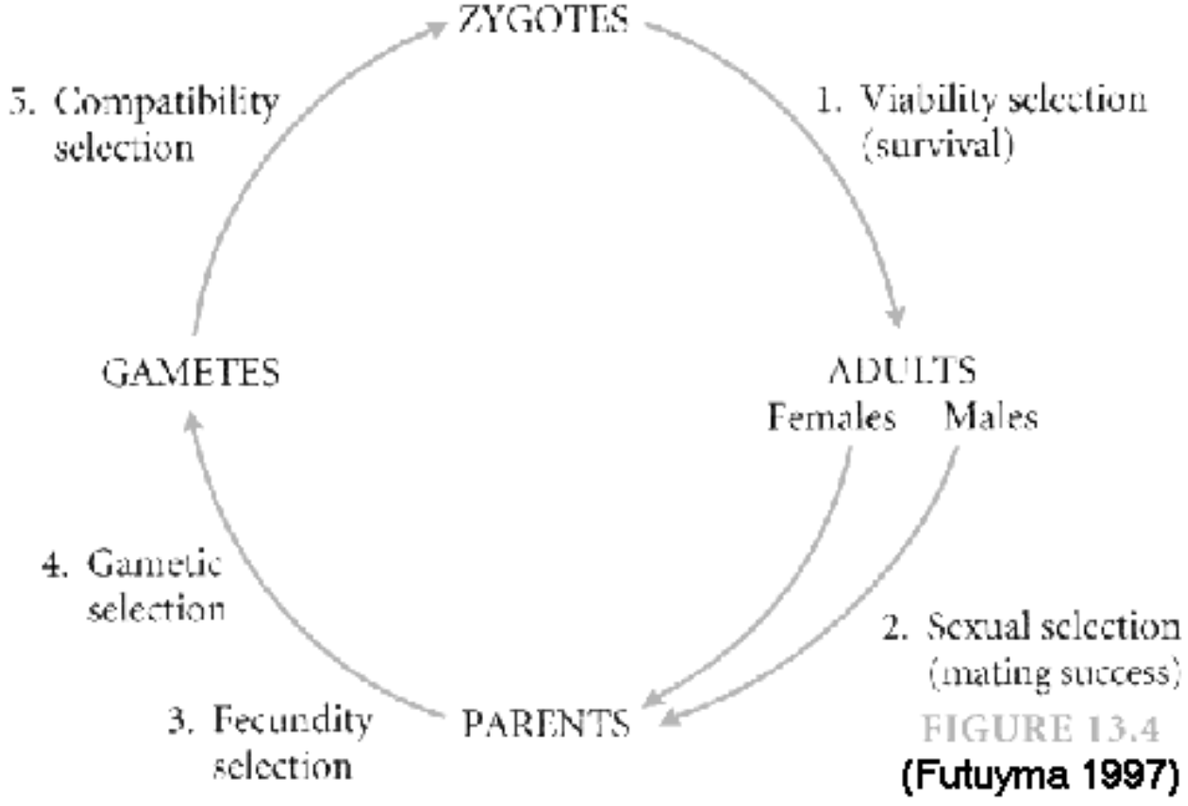
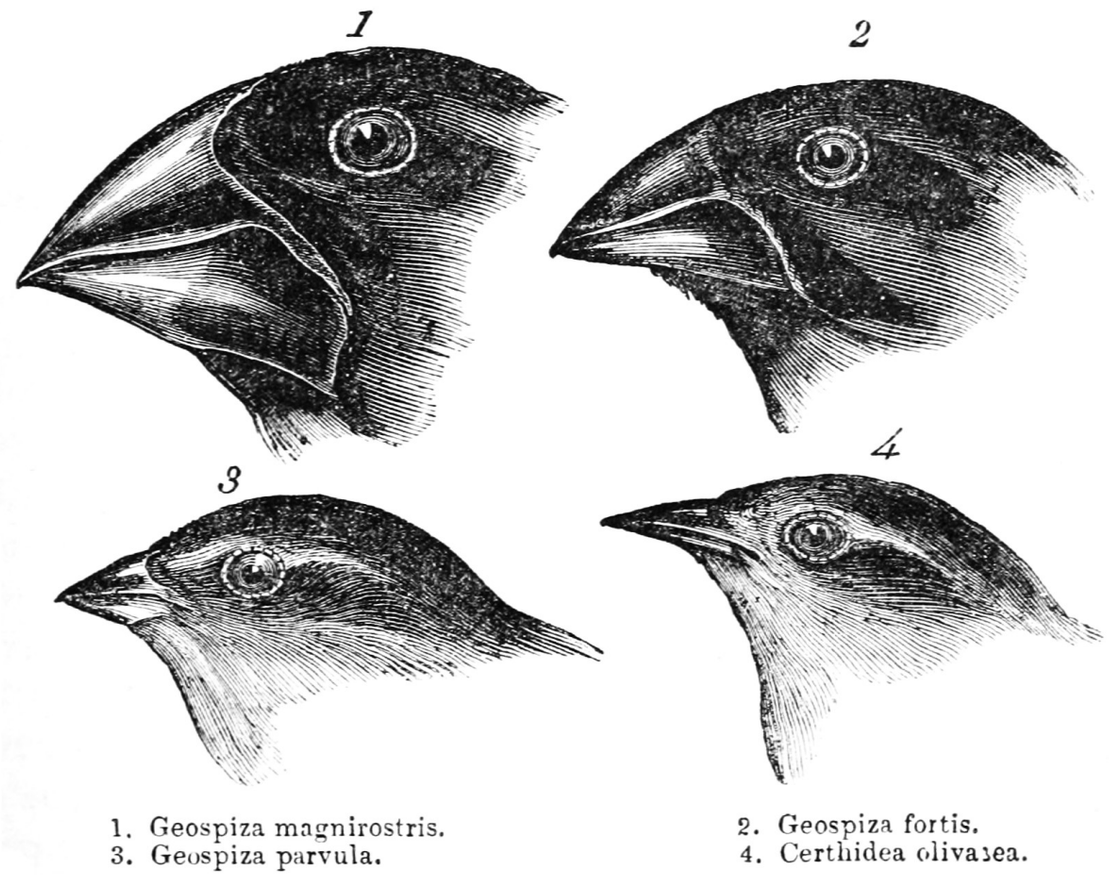


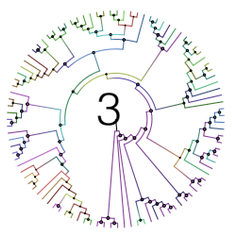
FIGURE 13.4
(Futuyma 1997)

Adaptation to a given environment

- **Fitness** involves the ability of organisms to survive and reproduce in the environment in which they find themselves (Dobzhansky, 1959).



1. Geospiza magnirostris.
2. Geospiza fortis.
3. Geospiza parvula.
4. Certhidea olivasea.



(Darwin 1845)

Adaptation to a given environment

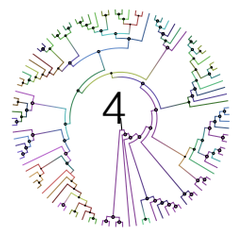
- Individuals with better features will be able to contribute more to the next generation. These individuals have higher (**relative**) **fitness**.

Industrial melanism in peppered moth (Biston betularia)

- The **mean fitness** is the average contribution of the genotypes coming from these individuals to the global pool in the next generation.

A_1A_1	A_1A_2	A_2A_2
$p^2\ \omega_{11}$	$2pq\ \omega_{12}$	$q^2\ \omega_{22}$

$$\overline{\omega} = p^2\ \omega_{11} + 2pq\ \omega_{12} + q^2\ \omega_{22}$$



Adaptation to a given environment

- Fitness allow us to predict the change of a trait under selection: the amount by which any trait changes from one generation to the next is given by the genetic covariance between the trait and relative fitness (Robertson 1966, Price 1970).

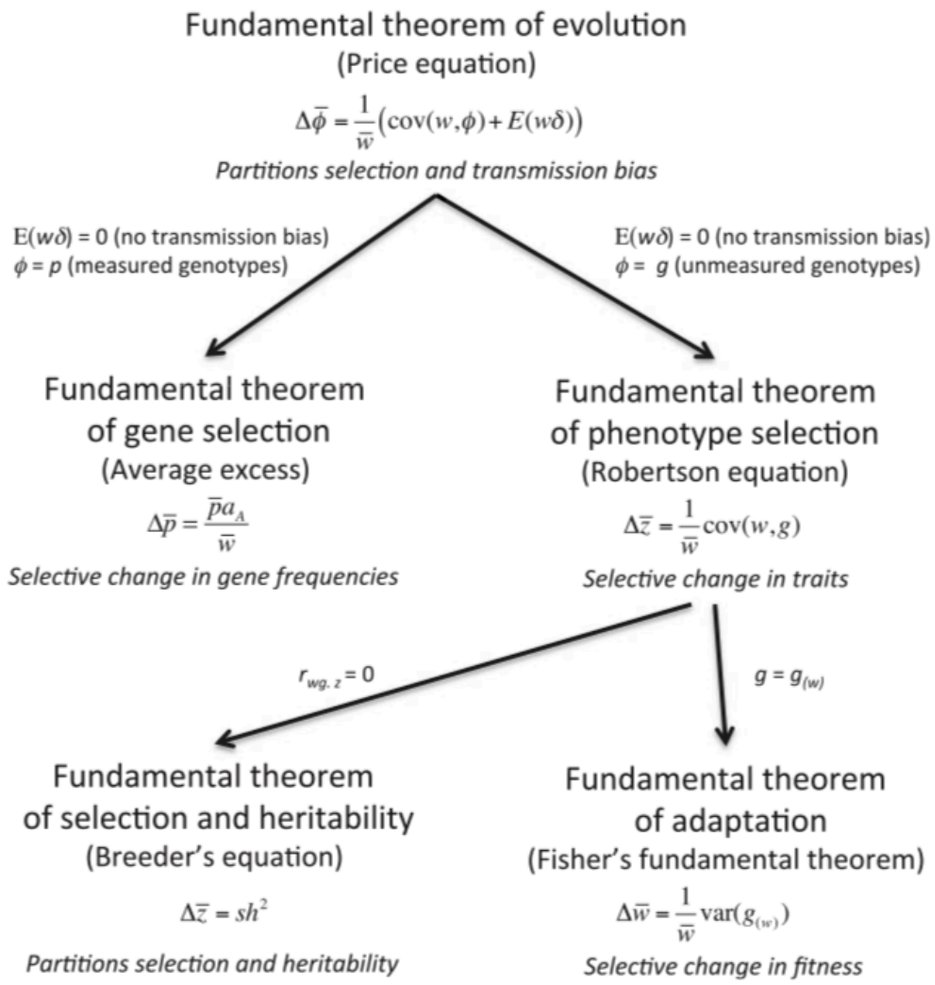


Figure 1: Fundamental theorems and their relationships. Arrows indicate derivation, with required assumptions or domain restrictions written beside them. ϕ = any trait value; δ = the change in ϕ from parent to offspring; w = fitness; p = allele frequency; a_A = average excess; g = breeding value; z = phenotype value; r = partial correlation; s = selection differential, h^2 = heritability. The i subscripts for individuals used in the text are omitted for economy.

(Queller Am. Nat. 2017)

Adaptation to a given environment

- Fitness allow us to predict the change of a trait under selection: the amount by which any trait changes from one generation to the next is given by the genetic covariance between the trait and relative fitness (Robertson 1966, Price 1970).

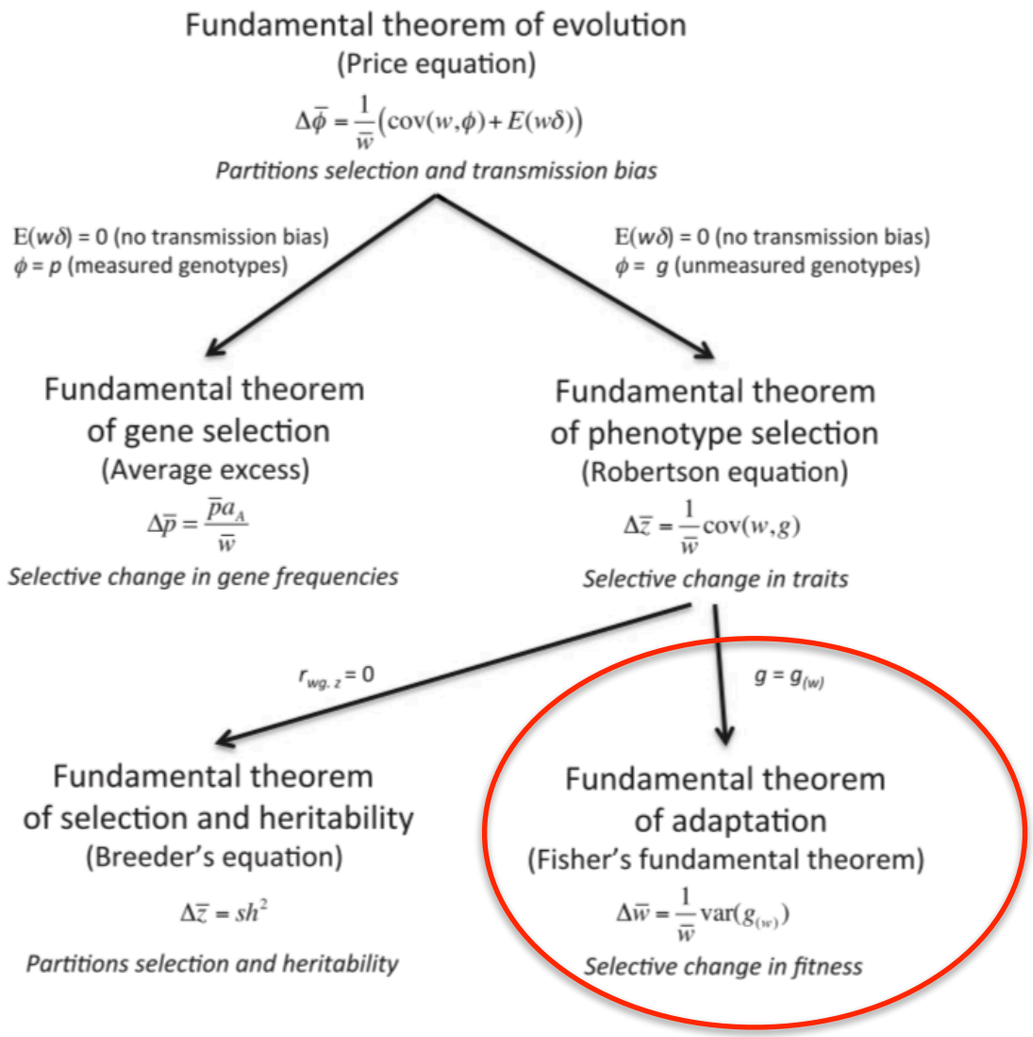


Figure 1: Fundamental theorems and their relationships. Arrows indicate derivation, with required assumptions or domain restrictions written beside them. ϕ = any trait value; δ = the change in ϕ from parent to offspring; w = fitness; p = allele frequency; a_A = average excess; g = breeding value; z = phenotype value; r = partial correlation; s = selection differential, h^2 = heritability. The i subscripts for individuals used in the text are omitted for economy.

(Queller Am. Nat. 2017)

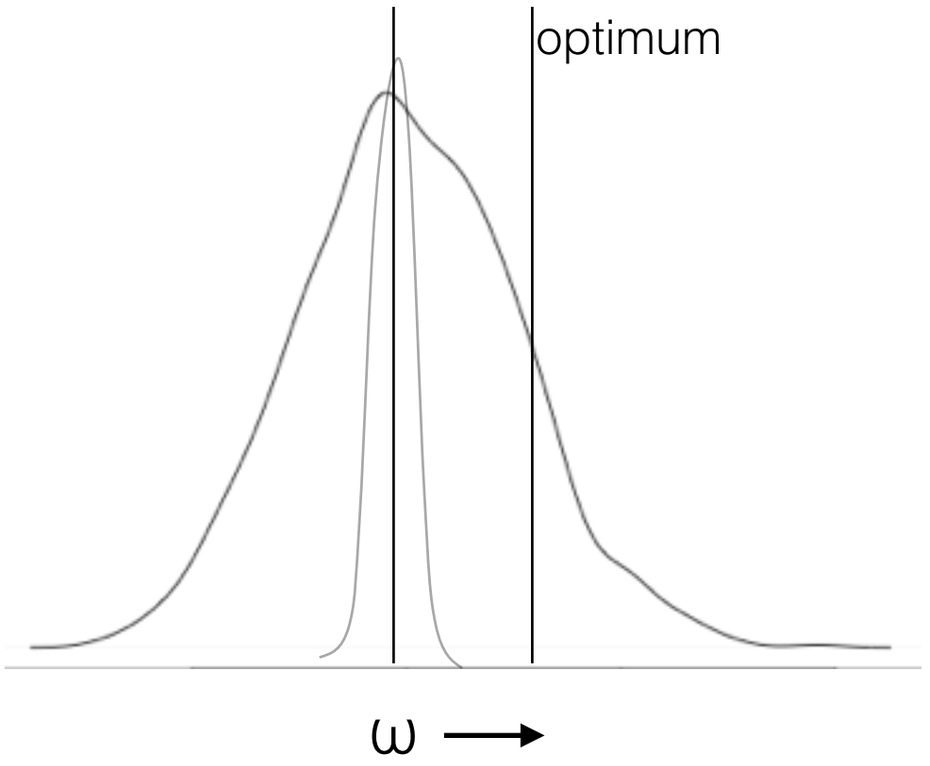
Adaptation to a given environment

- Fisher Fundamental Theorem of Natural Selection indicates that wide distributions in fitness will give more chances to increase the relative mean fitness. Suggest that natural selection (if present) is a process that (almost) always increases the mean relative fitness.

Fundamental theorem
of adaptation
(Fisher's fundamental theorem)

$$\Delta\bar{w} = \frac{1}{\bar{w}}\mathrm{var}(g_{(w)})$$

Selective change in fitness

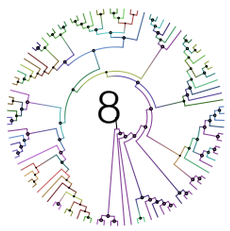


Adaptation to a given environment

- The fitness can be affected by few number of traits (**simple** environment) or alternatively by a large number of traits (a **complex** environment).

- Organisms living in simple environments are easier to achieve the theoretical optimum (maximum) fitness, while in the high multidimensional space of complex environments this optimum can be less accesible. **Larger populations give more chance to achieve optimum** (more mutations, more options).

- **Genetic load** is the relative difference between the optimal and the current fitness.

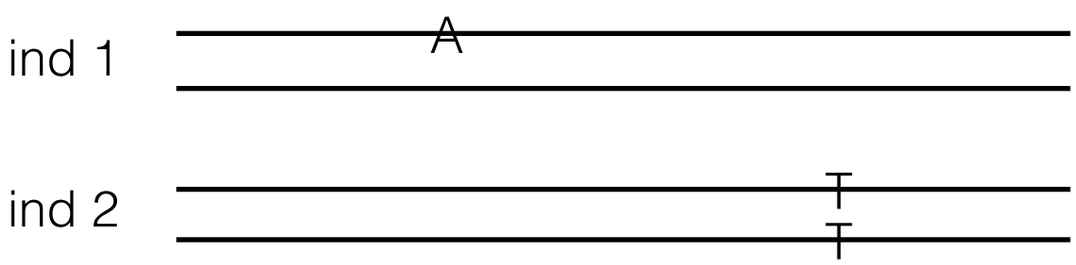


The Fitness Distribution

- The global effect of fitness and their dynamics in a population is a consequence of the the strength of the fitness and the number of positions involved, that is, of the **Distribution of Fitness Effects** (DFE).

- The fitness effect of genome positions, -positive or negative-, -strong or weak-, in a given environment can explain the capacity and the dynamics of the population to adaptation and the expected patterns of variability.

each position
has assigned a
specific **s**



Genome

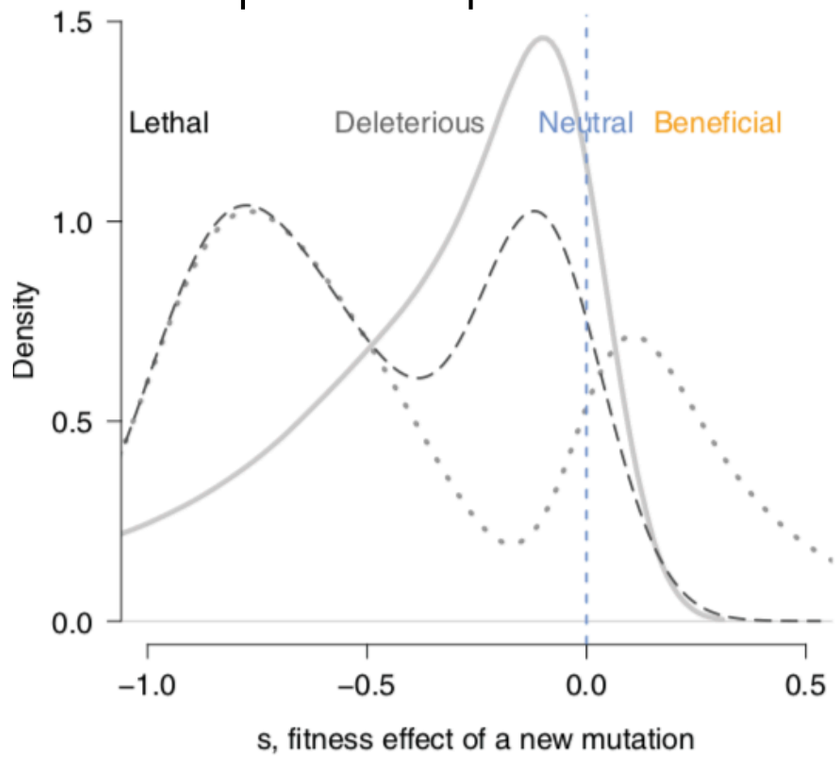


Figure 1. Hypothetical whole distributions of fitness effects.

(Bataillon & Bailey Ann NY Acad. Sc.. 2014)

The Fitness Distribution

- In case that the mean fitness of a population is close to the optimum (small genetic load), and beneficial effects are rare, it is predicted an invariant and L-shaped (exponential) distribution of beneficial effects (Orr 2003).

- In case that the mean fitness of a population is close to the optimum, most mutations are displacing from the optimum and thus are deleterious. A gamma distribution is predicted.

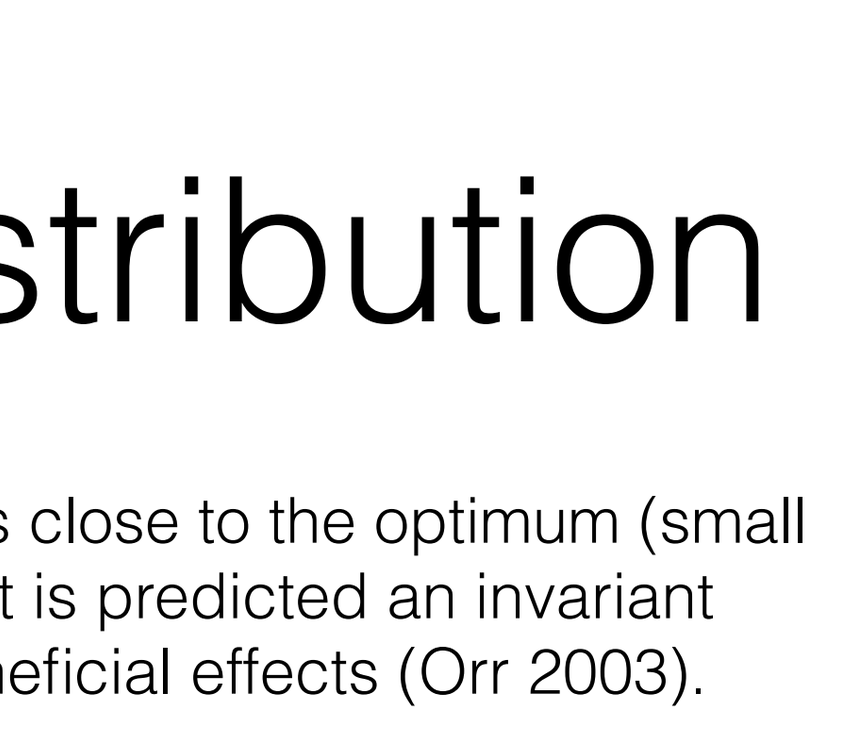


Figure 1. Hypothetical whole distributions of fitness effects.

(Bataillon & Bailey Ann NY Acad. Sc.. 2014)

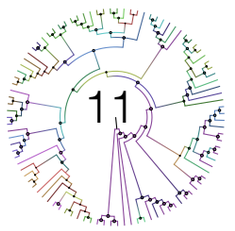
Methods to measure Fitness

- How many mutations affect the fitness and what importance each mutation has?

  - **Experimental studies measuring fitness components in contemporary populations.**

    - Estimate the effect of spontaneous mutations.
    - Mutagenesis.
    - Fitness trajectories in adapting populations
    - mutation-accumulation.

  - **Inference from Variability data**



Methods to measure Fitness

- How many mutations affect the fitness and what importance each mutation has?

  - **Experimental studies measuring fitness components in contemporary populations.**

    - **Estimate the effect of spontaneous mutations**.Usually in bacteria or viruses. Fluctuation assay: Many parallel cultures evolve on a permissive environment; then transferred to a selective environment. Count the number and the strength. Usually beneficial mutations.

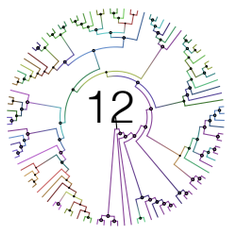
    - **Mutagenesis**: random or site-directed mutagenesis; measure beneficial and deleterious mutations.

    - **Fitness trajectories in adapting populations** (trace the changes in frequencies of adaptive strains -using neutral markers- across the time). Measure the strength of each beneficial mutation.

    - **mutation-accumulation**. Look at new mutations affecting a phenotypic trait related to fitness. measure beneficial and deleterious mutations.

  - **Inference from Variability data**

    - Use polymorphisms and divergence data from functional and non-functional regions to infer the Distribution of Fitness Effects.



Methods to measure Fitness

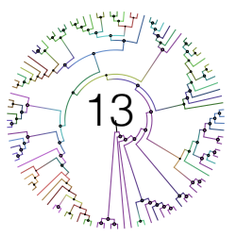
- How many mutations affect the fitness and what importance each mutation has?

  - **Experimental studies measuring fitness components in contemporary populations.**

Table 1. DFE inferred from experimental studies relying on the isolation of individual mutations

Strategy used to isolate mutations	Organism	Mutational target	Number of beneficial mutations	Mutations character-ized	DFE inferred
Resistance to antibiotic	*Pseudomonas fluorescens*	gyrA and others	18	Beneficial	Exponential
Resistance to antibiotic	*P. aeruginosa*	rpoB	15	Beneficial	Exponential
Reporter construct	*P. fluorescens*	11 genes total	100	Beneficial	Normal
Increased growth rate	ssDNA bacte-riophage Id11	WG	9	Beneficial	Weibull
Novel host growth	RNA phage ϕ6	P3 (host attachment gene)	16	Beneficial	Weibull
Site-directed mutagenesis	VSV (RNA virus)	WG	16	(A) Beneficial	(A) Γ, significantly leptokurtic
				(B) Deleterious	(B) log-normal + uniform

(Bataillon & Bailey Ann NY Acad. Sc.. 2014)

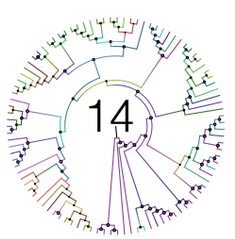


Methods to measure Fitness

- How many mutations affect the fitness and what importance each mutation has?
  - **Experimental studies measuring fitness components in contemporary populations.**

Table 2. Summaries of studies inferring mutational properties through fitness/marker trajectories over time

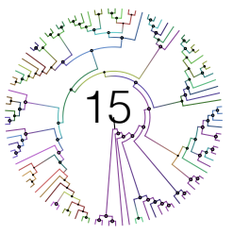
Method	Organism	Sample size (# of mutations)	DFE	Beneficial mutation rate	References
Marker frequencies	*E. coli*	66	Exponential or Γ	4×10^{-9}	74
Marker frequencies	*E. coli*	30	Peaked, unimodal	5.9×10^{-8}	44
Marker frequencies	*E. coli*	72	Exponential, uniform, and Dirac δ	2×10^{-7}	46
Marker frequencies	*E. coli*	75 and 87	Γ	10^{-5}	64



(Bataillon & Bailey Ann NY Acad. Sc.. 2014)

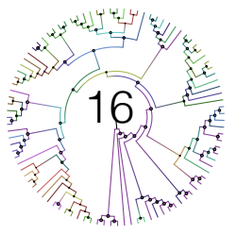
Inference from Variability data

- How many mutations affect the fitness and what importance each mutation has?



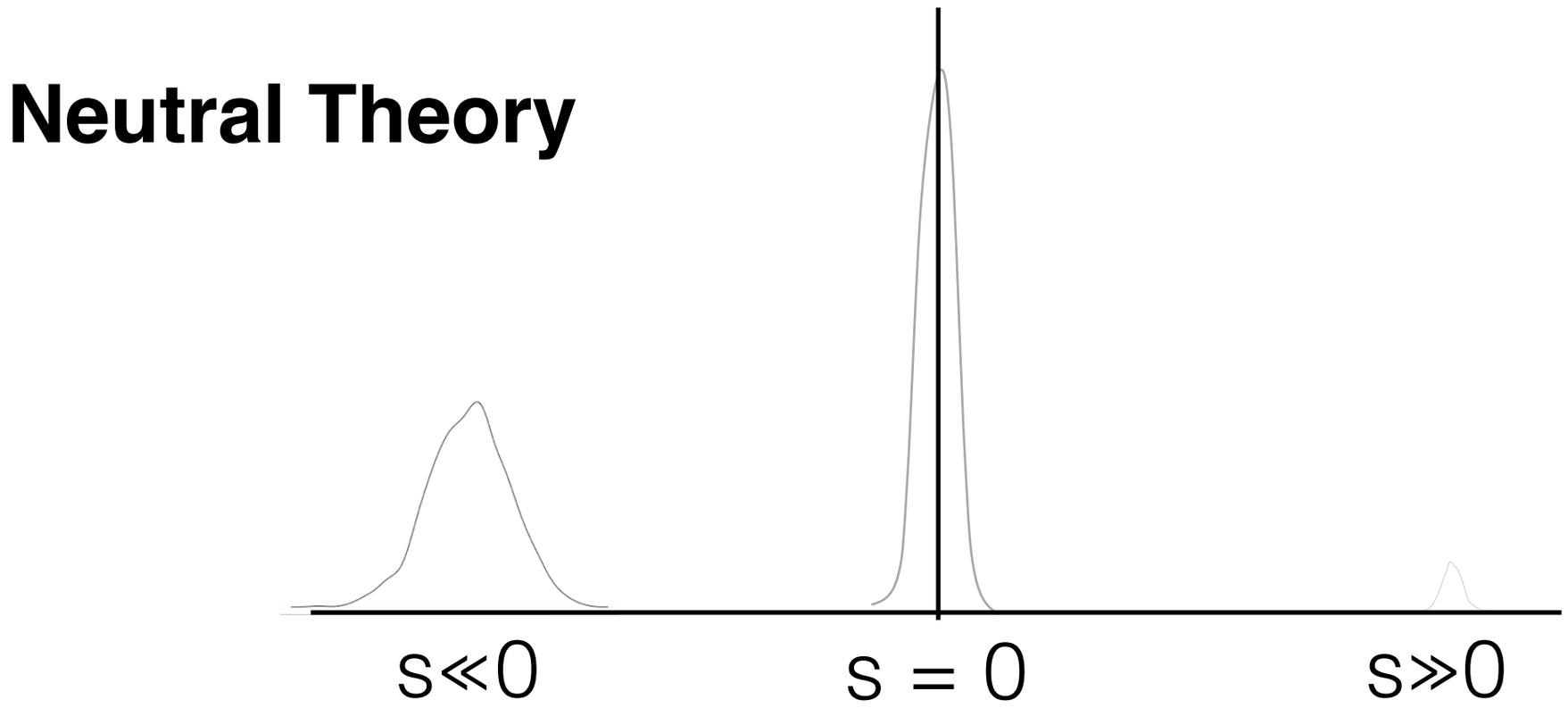
Inference from Variability data

- How many mutations affect the fitness and what importance each mutation has?
- What proportion of mutations are beneficial?



Inference from Variability data

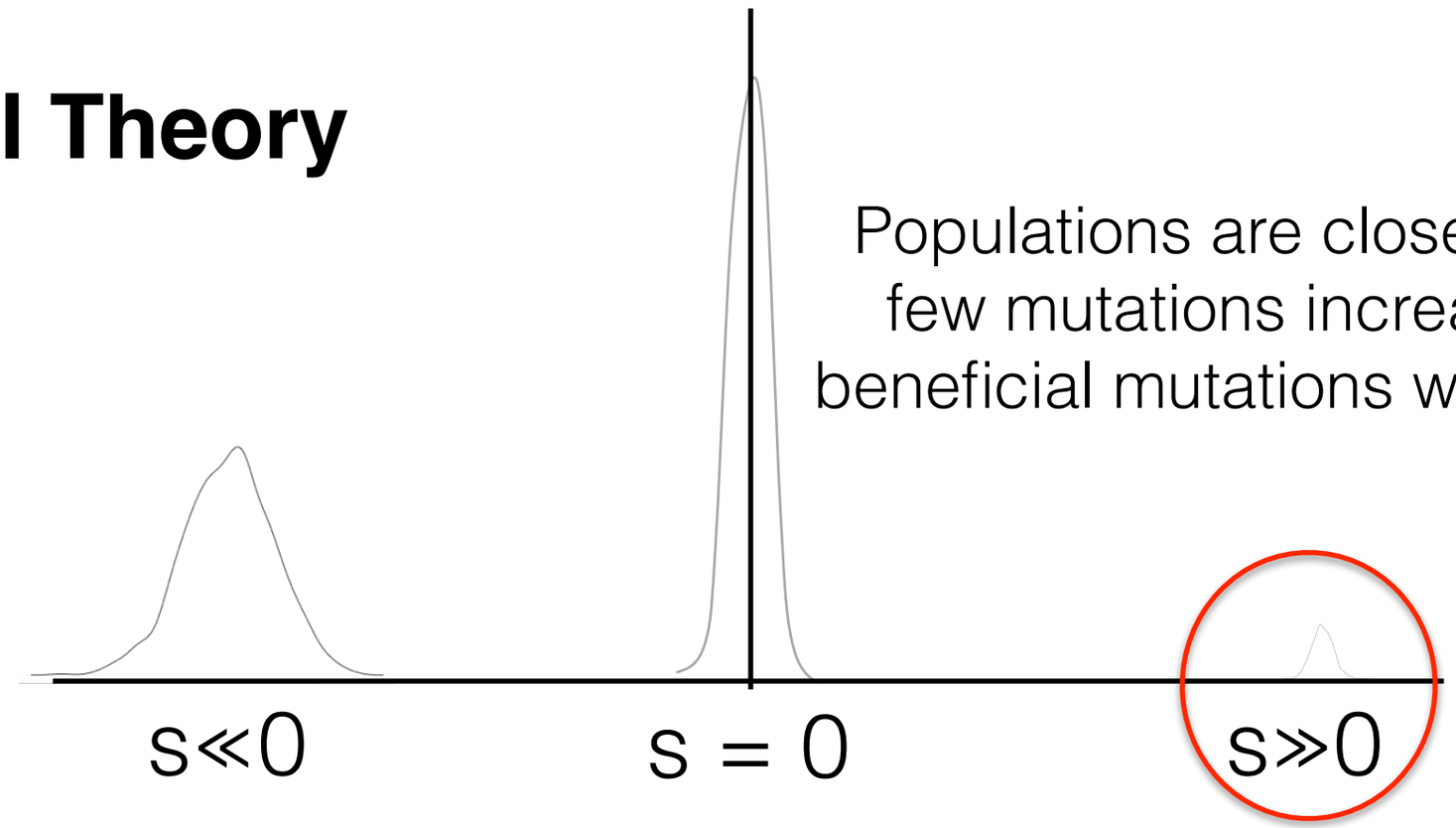
- How many mutations affect the fitness and what importance each mutation has?
- What proportion of mutations are beneficial?



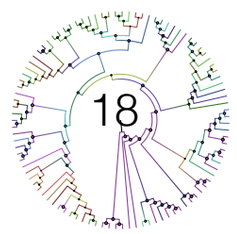
Inference from Variability data

- How many mutations affect the fitness and what importance each mutation has?
- What proportion of mutations are beneficial?

Neutral Theory



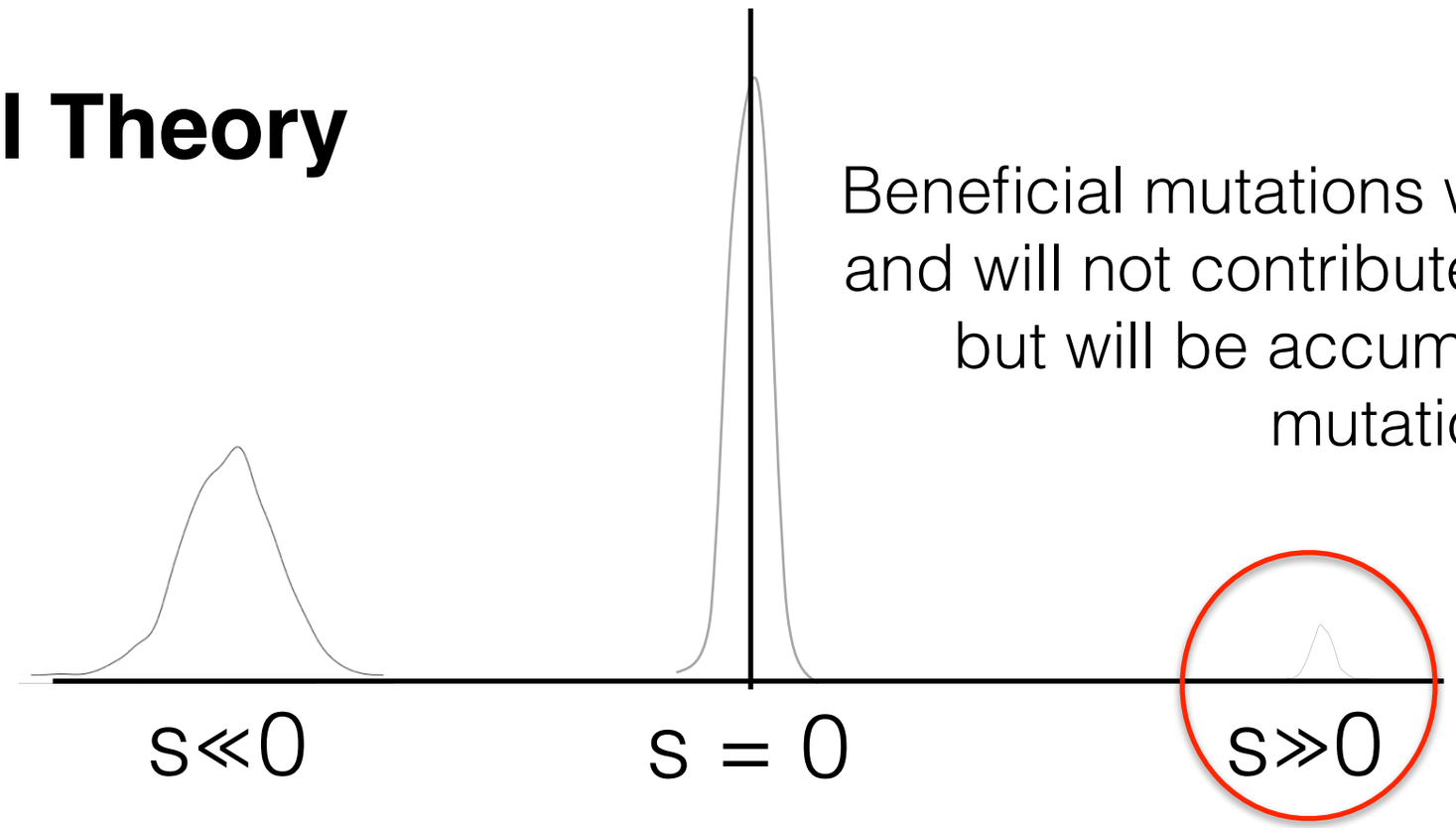
Populations are close to optimum. Only few mutations increase fitness. Weak beneficial mutations will hardly contribute.



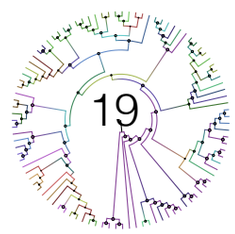
Inference from Variability data

- How many mutations affect the fitness and what importance each mutation has?
- What proportion of mutations are beneficial?

Neutral Theory



Beneficial mutations will be rapidly fixed and will not contribute to polymorphism, but will be accumulated as fixed mutations.

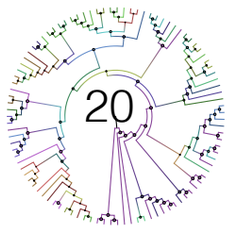


Inference from Variability data

- How many mutations affect the fitness and what importance each mutation has?

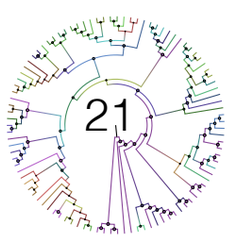
- What proportion of mutations are beneficial?

- The MacDonald and Kreitman framework (1992)



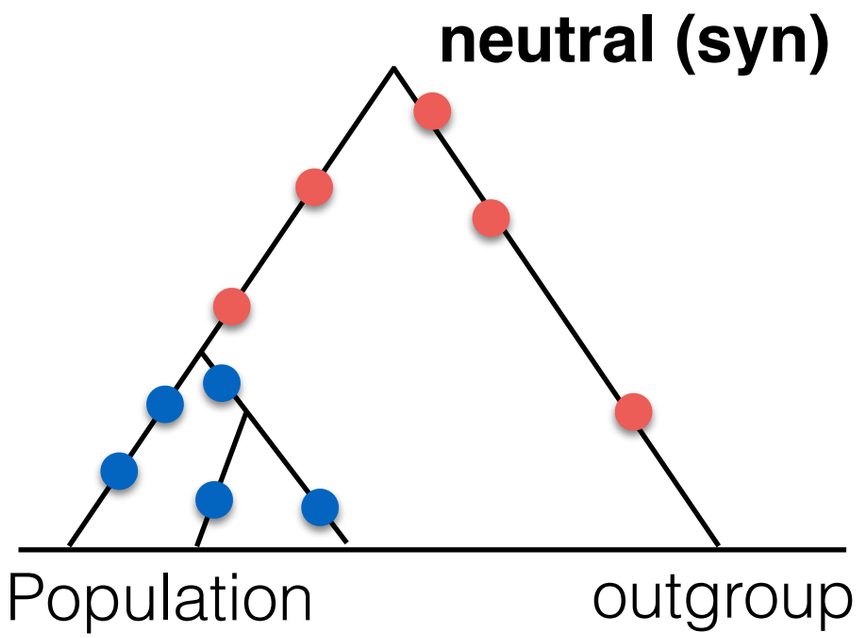
Inference from Variability data

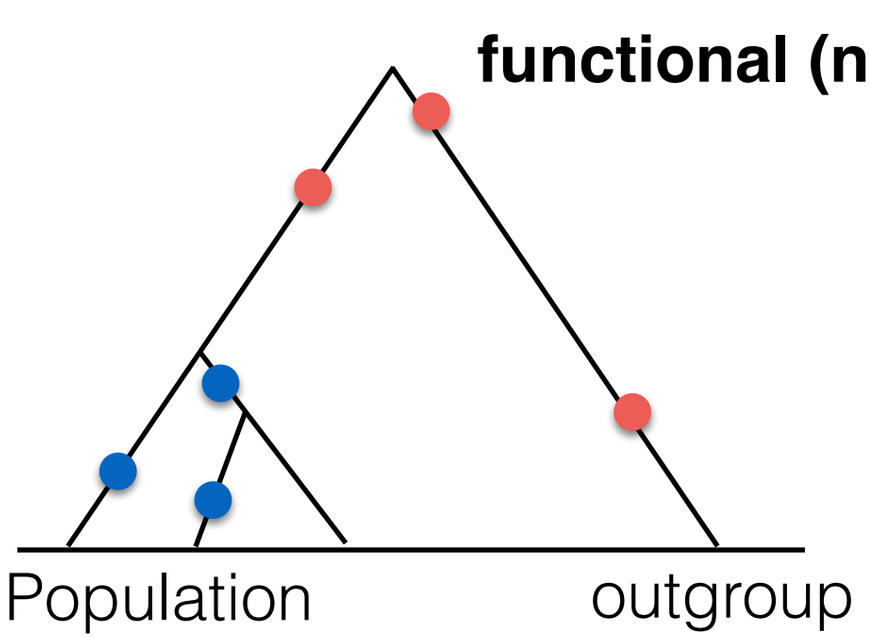
- The MacDonald and Kreitman framework (1992)
  - Separate the positions into neutral and others susceptible of selection (functional) Using coding regions: synonymous and non-synonymous.
  - Consider the polymorphism and the divergence with a close related species.



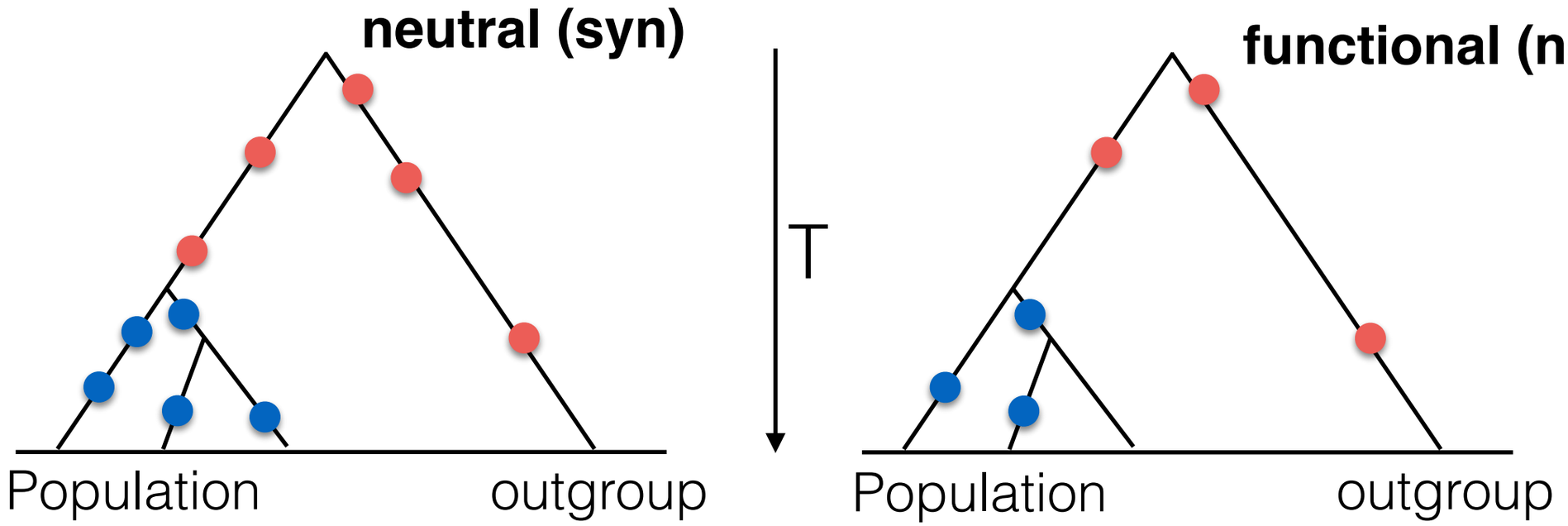
Inference from Variability data

- The MacDonald and Kreitman framework (1992)
  - Separate the positions into neutral and others susceptible of selection (functional) Using coding regions: synonymous and non-synonymous.
  - Consider the polymorphism and the divergence with a close related species.

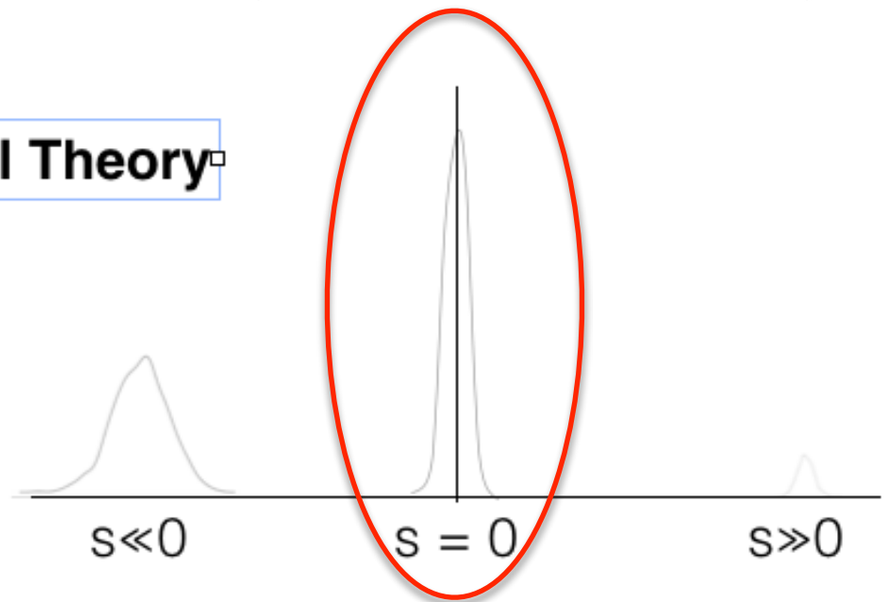




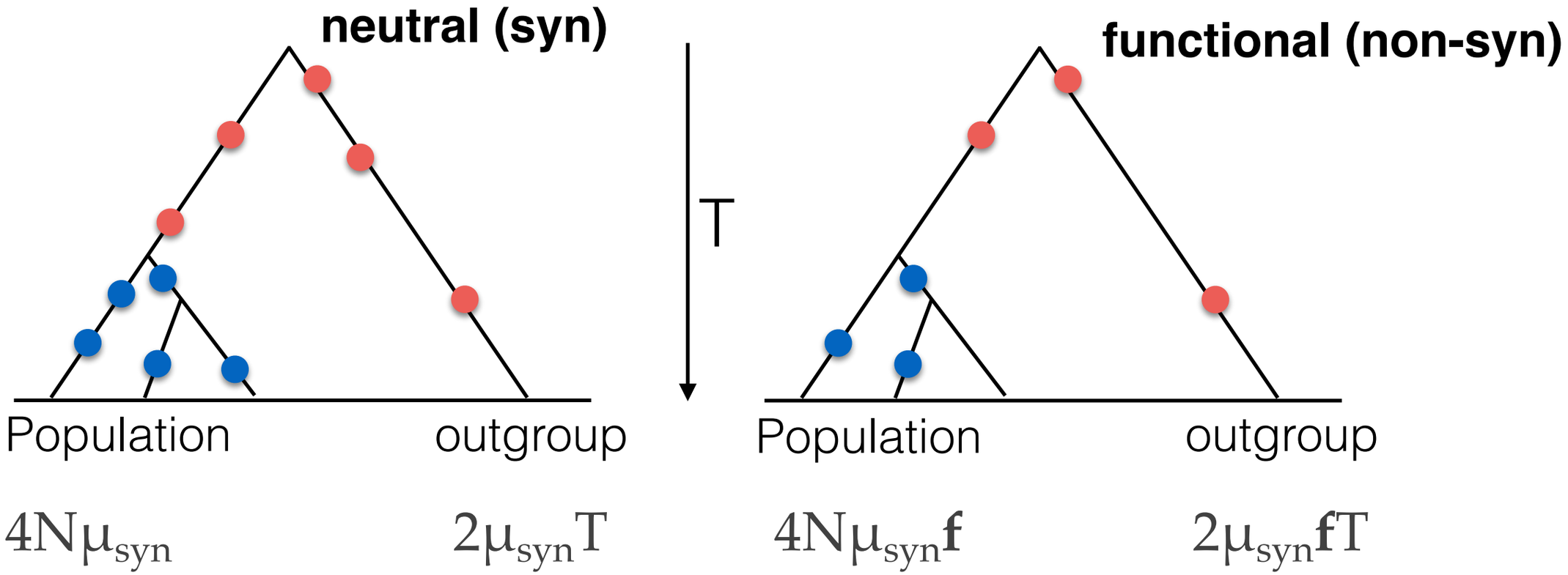
Inference from Variability data

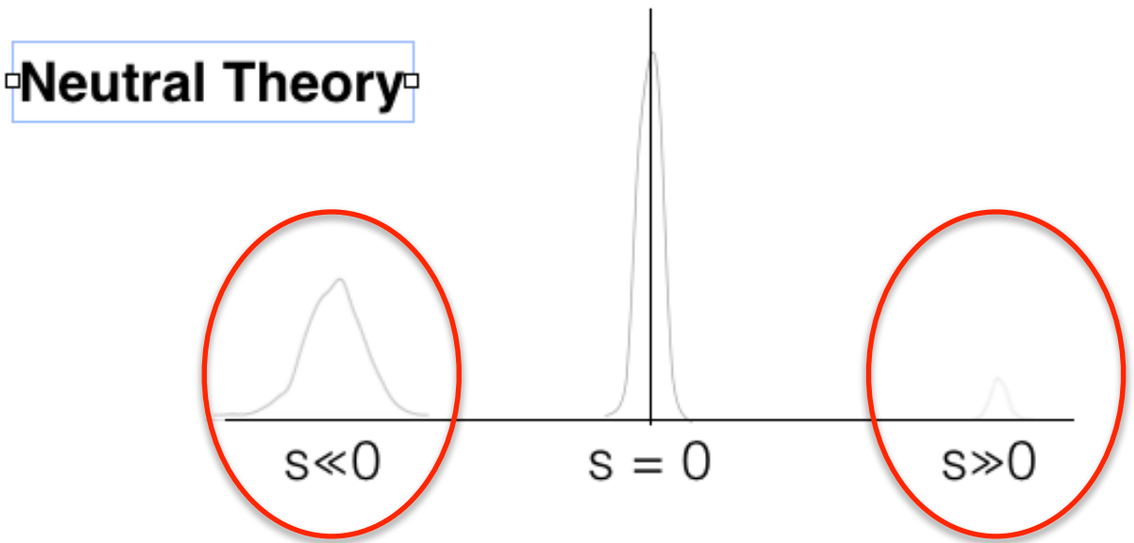


Neutral Theory

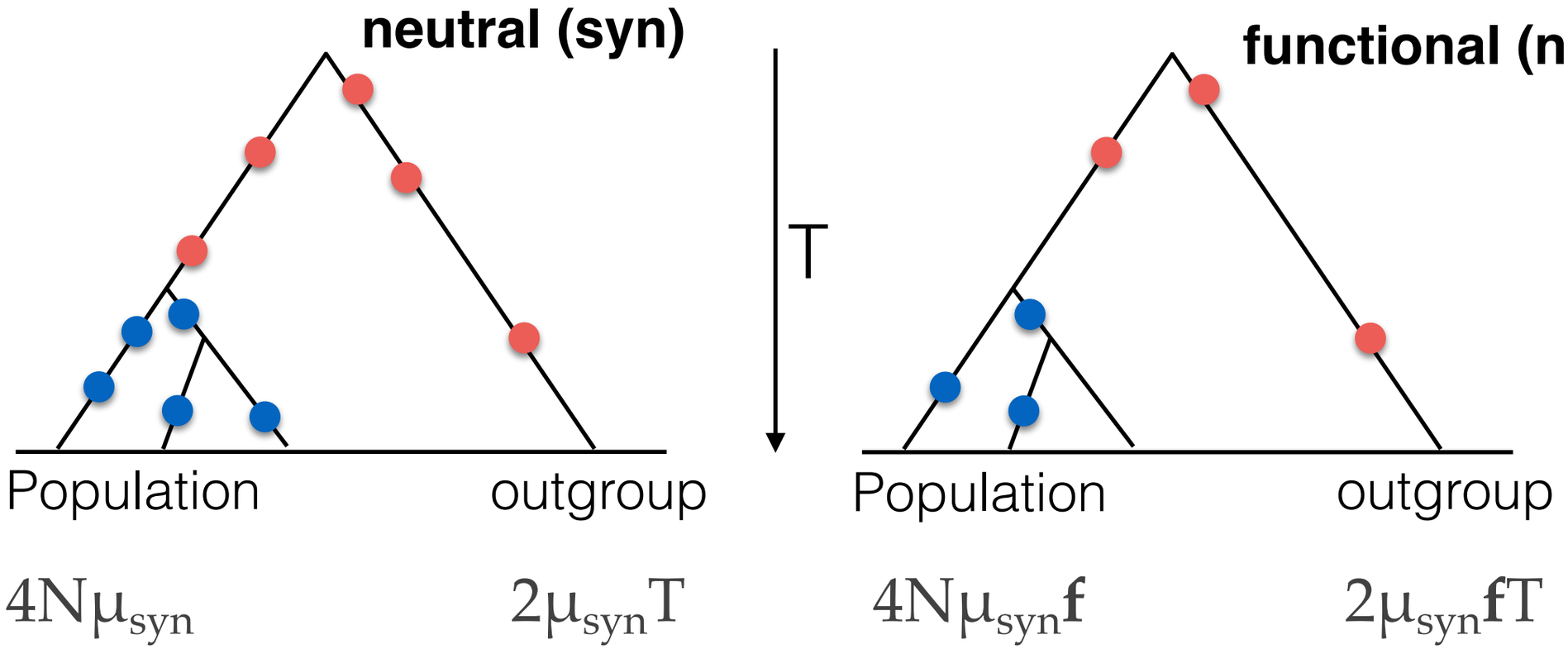


Inference from Variability data

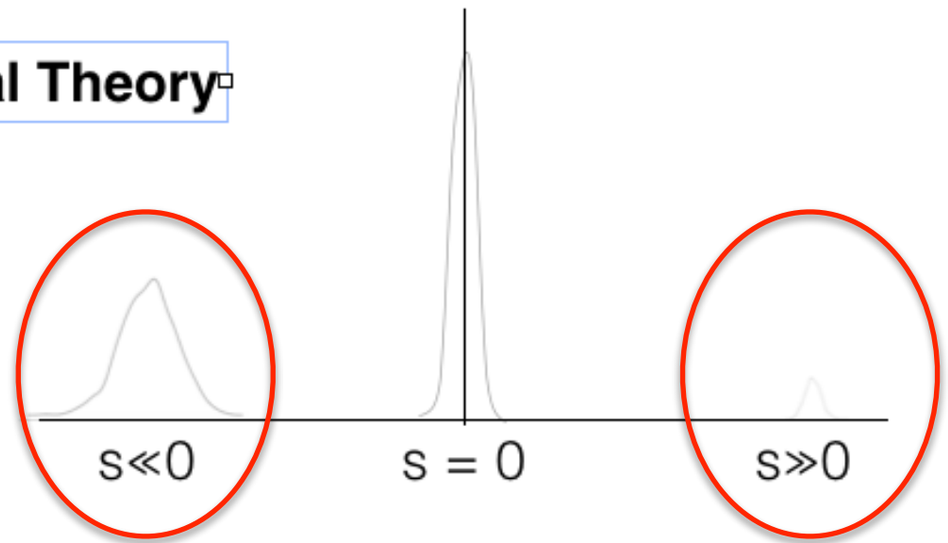




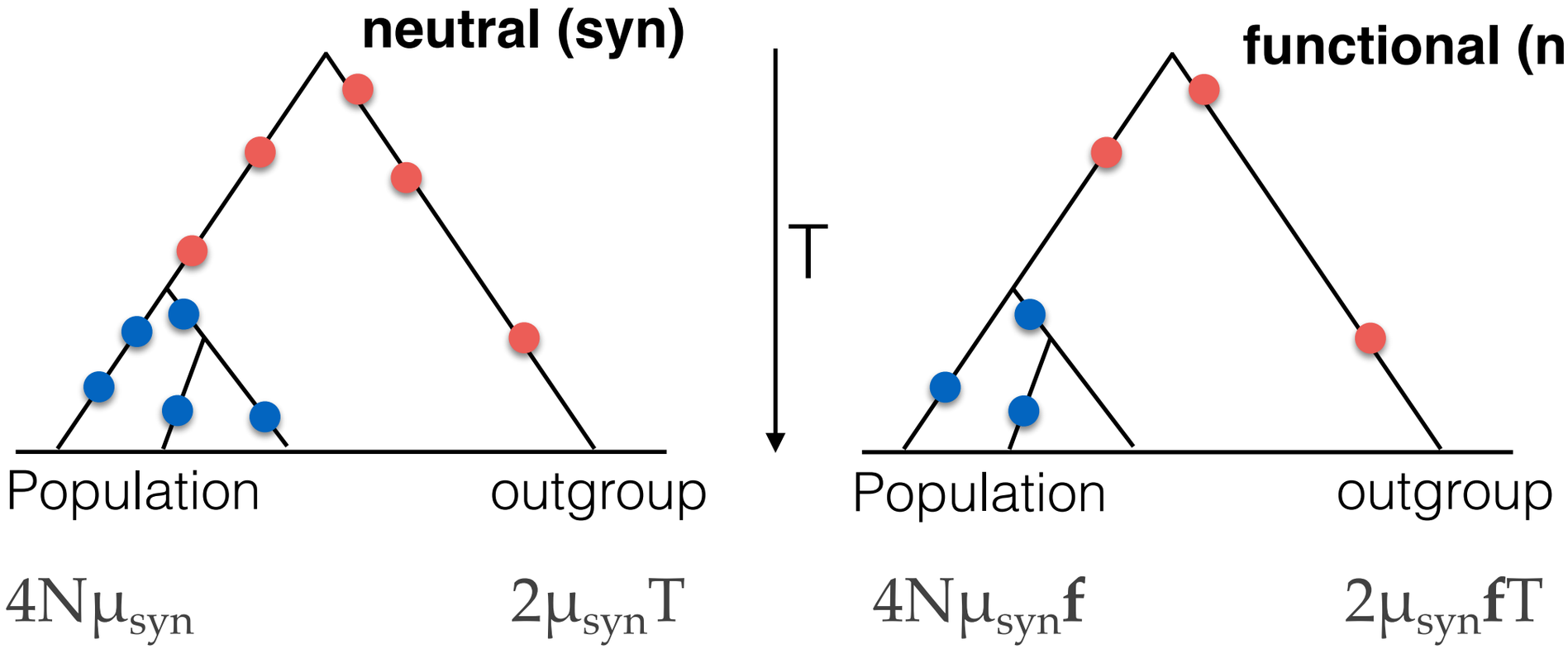
Inference from Variability data



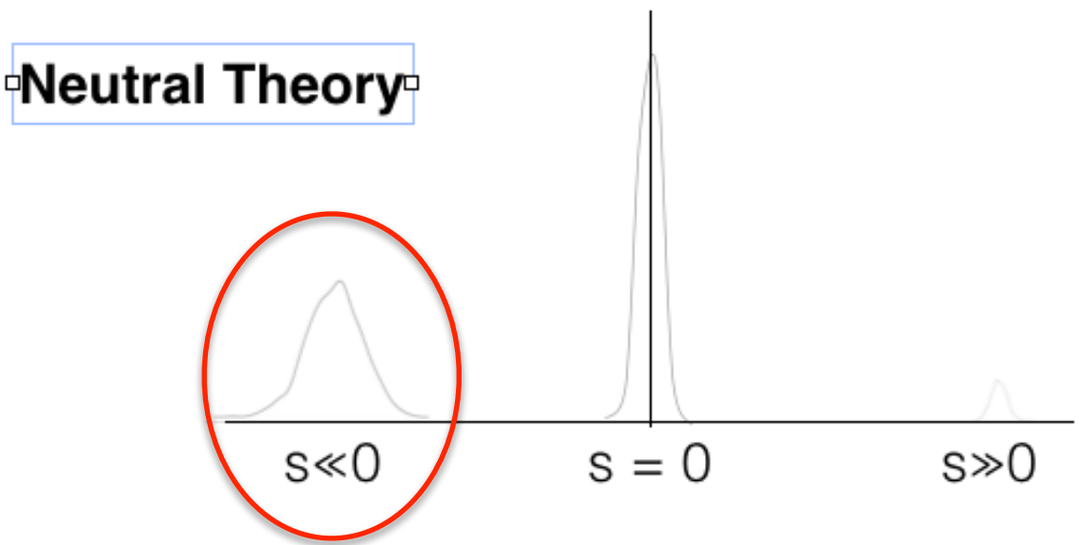
But only the beneficial mutations contribute to divergence!



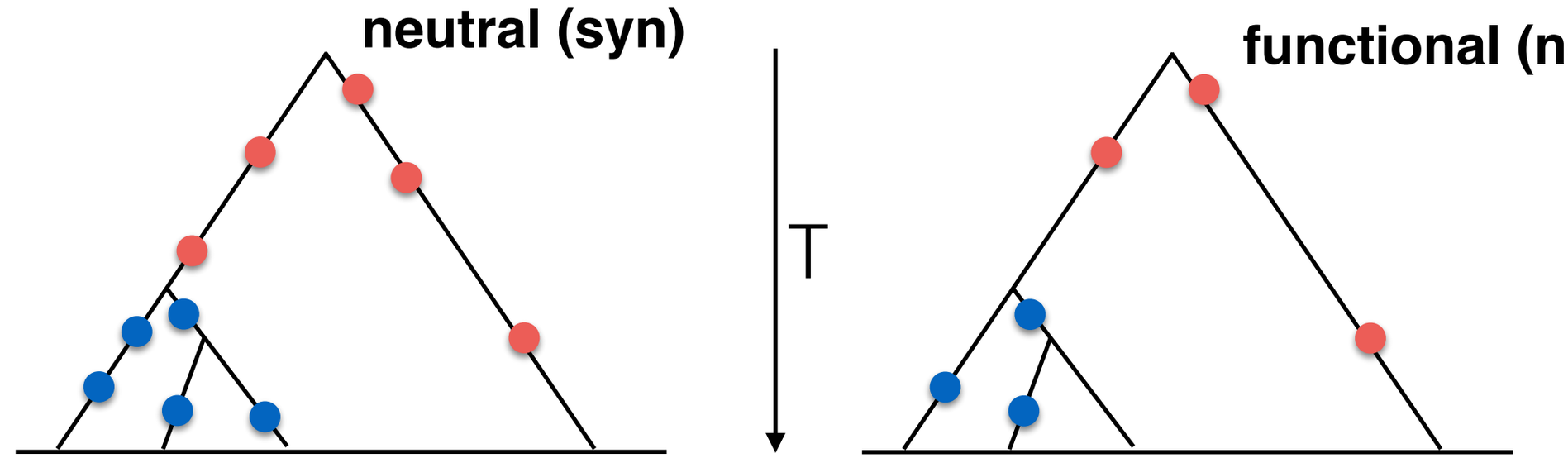
Inference from Variability data



If we do not consider beneficial mutations:



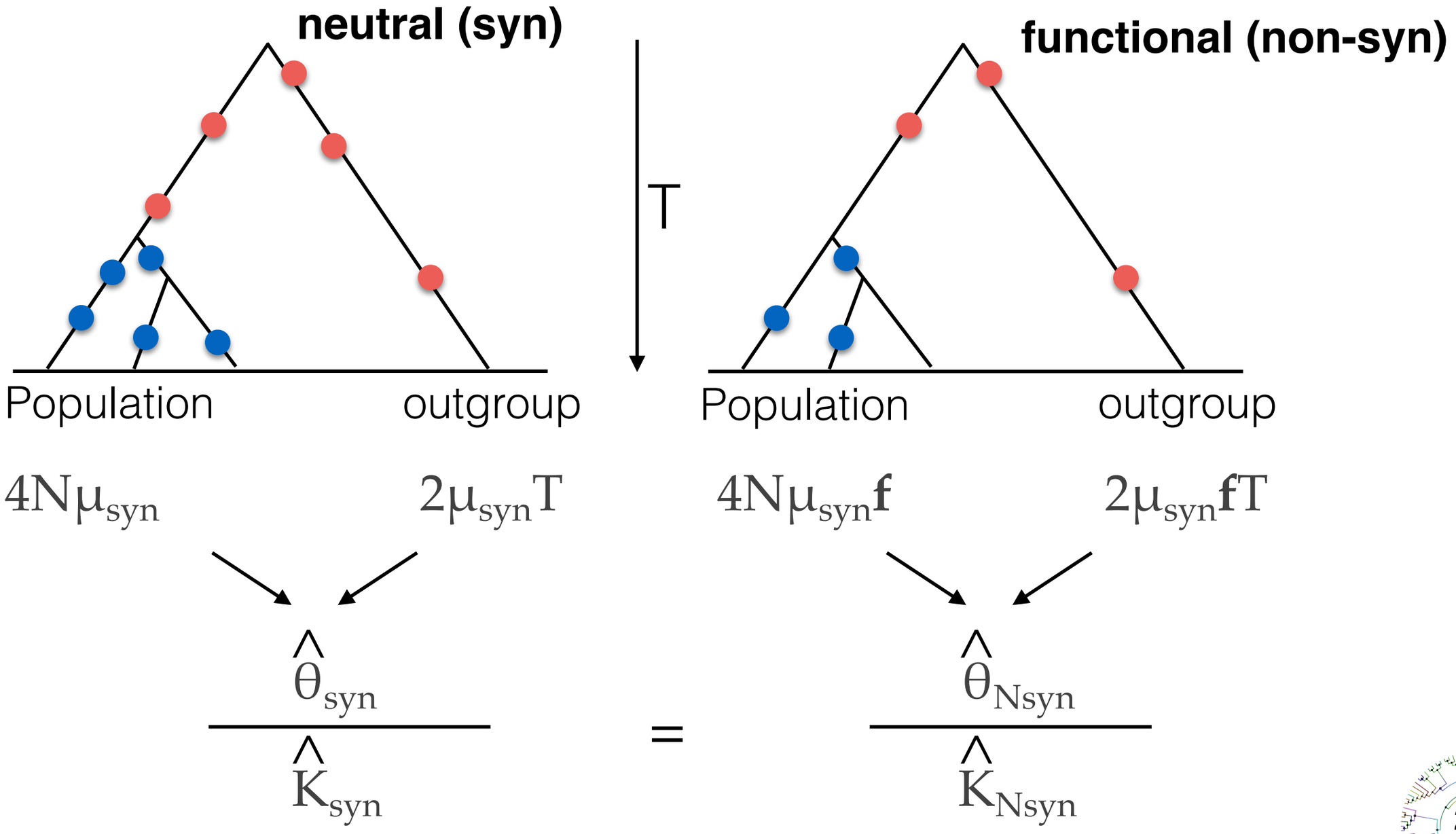
Inference from Variability data



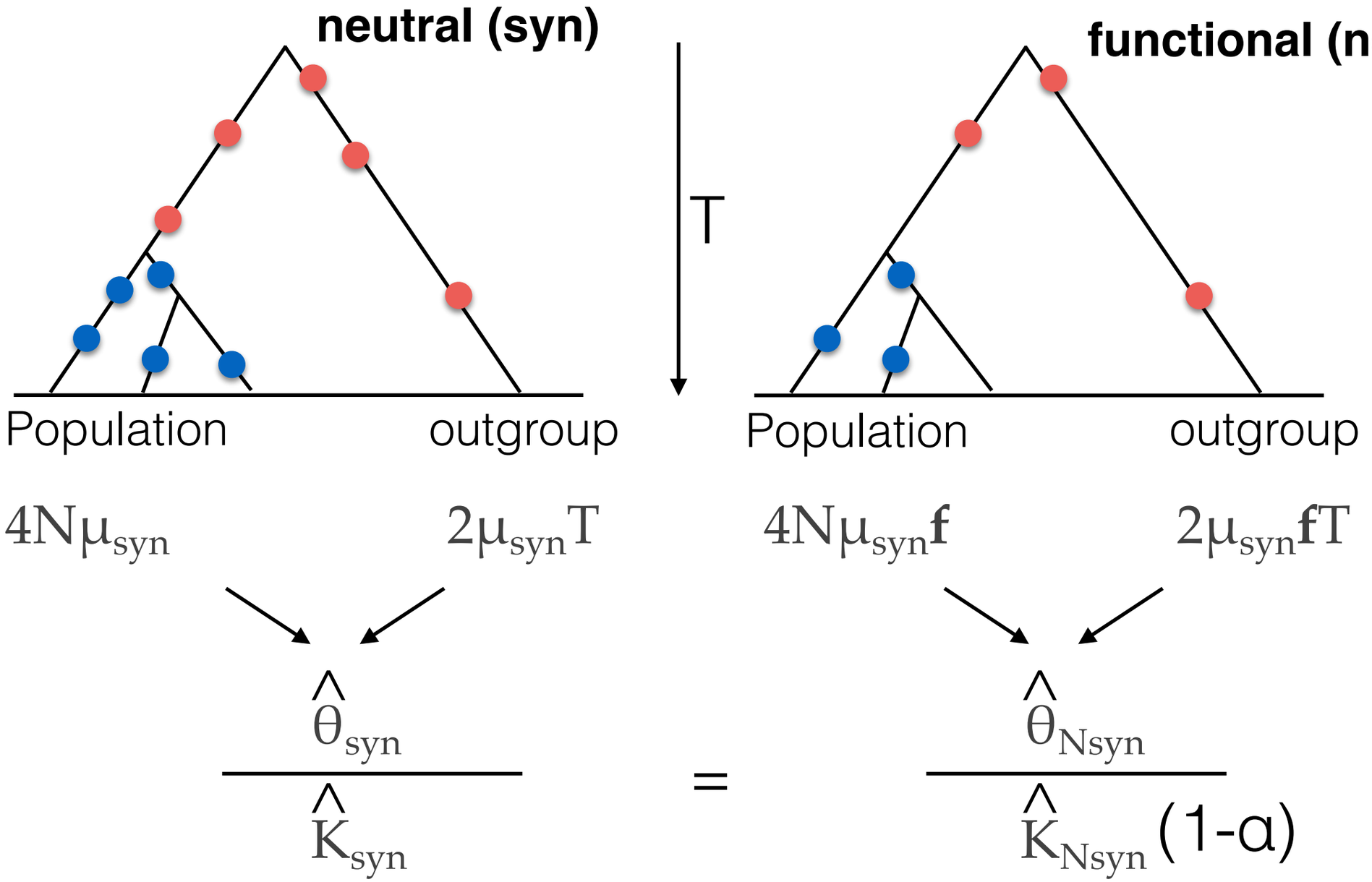
$4N\mu_{syn}$ $2\mu_{syn}T$ $4N\mu_{syn}\mathbf{f}$ $2\mu_{syn}\mathbf{f}T$

$$\frac{4N\mu_{syn}}{2\mu_{syn}T} = \frac{4N\mu_{syn}\mathbf{f}}{2\mu_{syn}T\mathbf{f}}$$

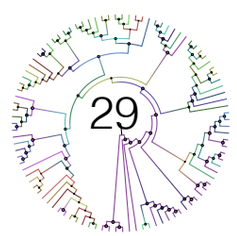
Inference from Variability data



Inference from Variability data



Considering beneficial mutations affect divergence but not polymorphism:

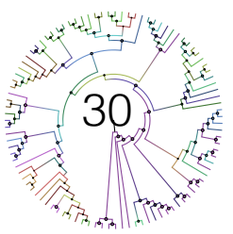


Inference from Variability data

$$\frac{\hat{\theta}_{syn}}{\hat{K}_{syn}} = \frac{\hat{\theta}_{Nsyn}}{\hat{K}_{Nsyn}(1-\alpha)}$$

α is the proportion of beneficial (adaptive) substitutions.

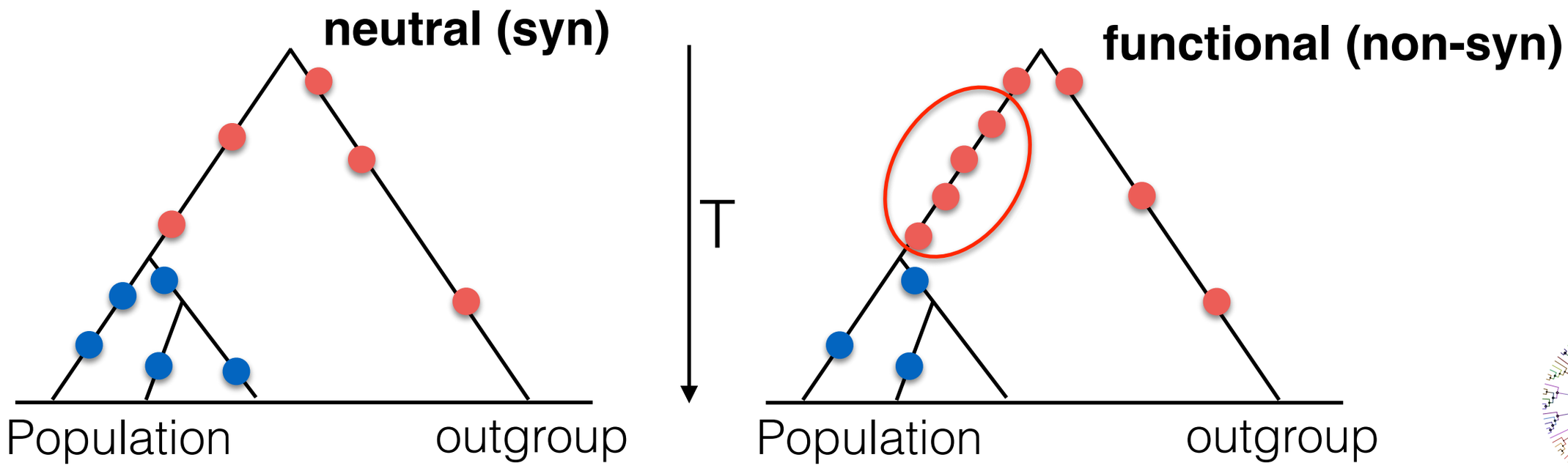
$$\hat{\alpha} = 1 - \frac{\hat{\theta}_{Nsyn}\ \hat{K}_{syn}}{\hat{\theta}_{syn}\ \hat{K}_{Nsyn}}$$



Inference from Variability data

α is the proportion of beneficial (adaptive) substitutions.

$$\hat{\alpha} = 1 - \frac{\hat{\theta}_{Nsyn}\ \hat{K}_{syn}}{\hat{\theta}_{syn}\ \hat{K}_{Nsyn}}$$



The proportion of Adaptive Substitutions

$$\overline{\alpha} = 1 - \frac{\overline{D}_s}{\overline{D}_n}\overline{\left(\frac{P_n}{P_s + 1}\right)}$$

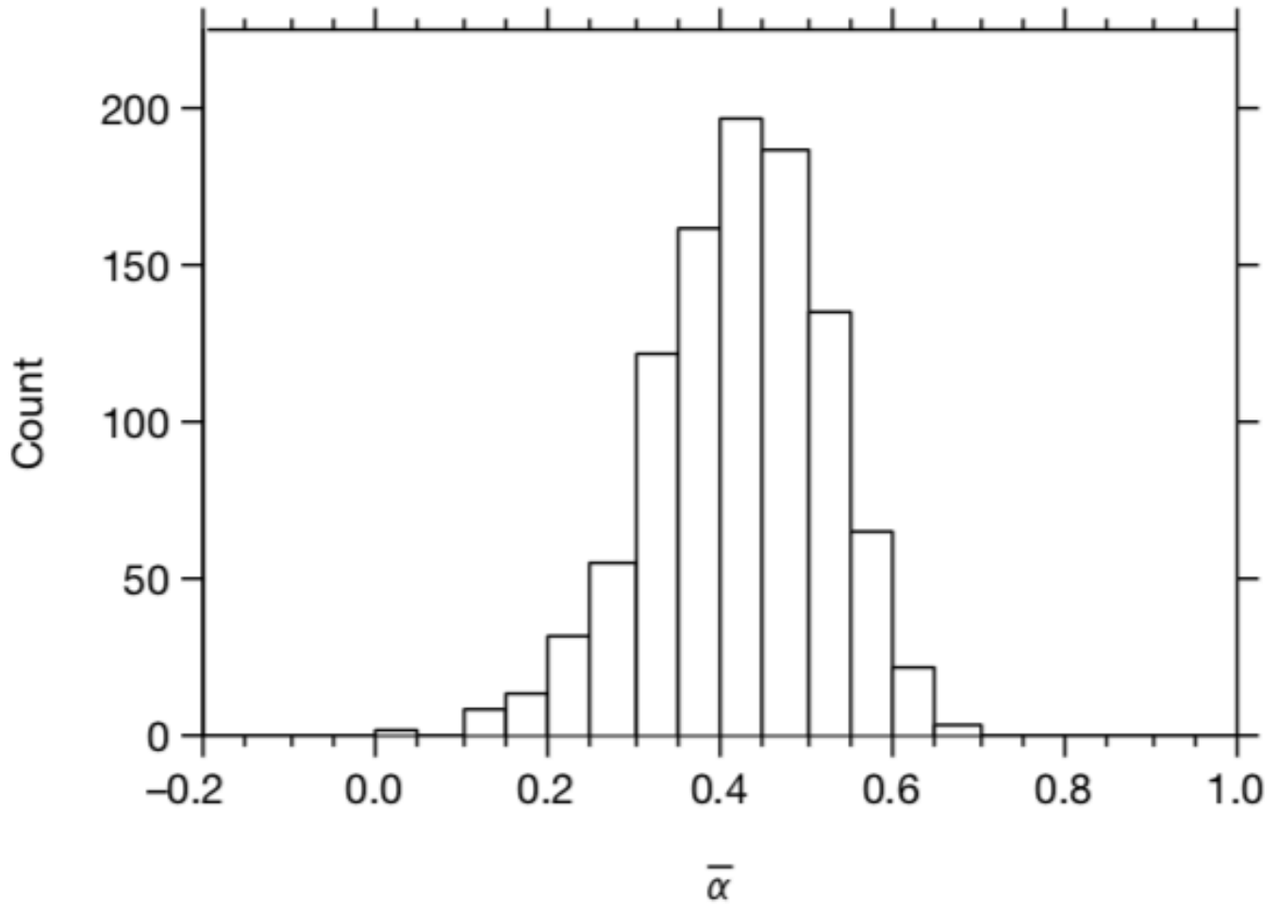


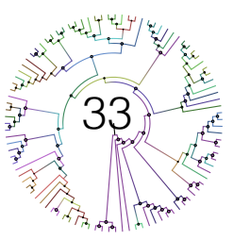
Figure 1 The distribution of 1,000 bootstrap values of $\overline{\alpha}$ for the divergence between *Drosophila simulans* and *D. yakuba* for genes in which $P_s > 5$. $\overline{\alpha}$ is the average proportion of amino-acid substitutions driven by positive selection.

(Smith & eyre-Walker Nature 2002)

Expectations from other (more realistic) Scenarios

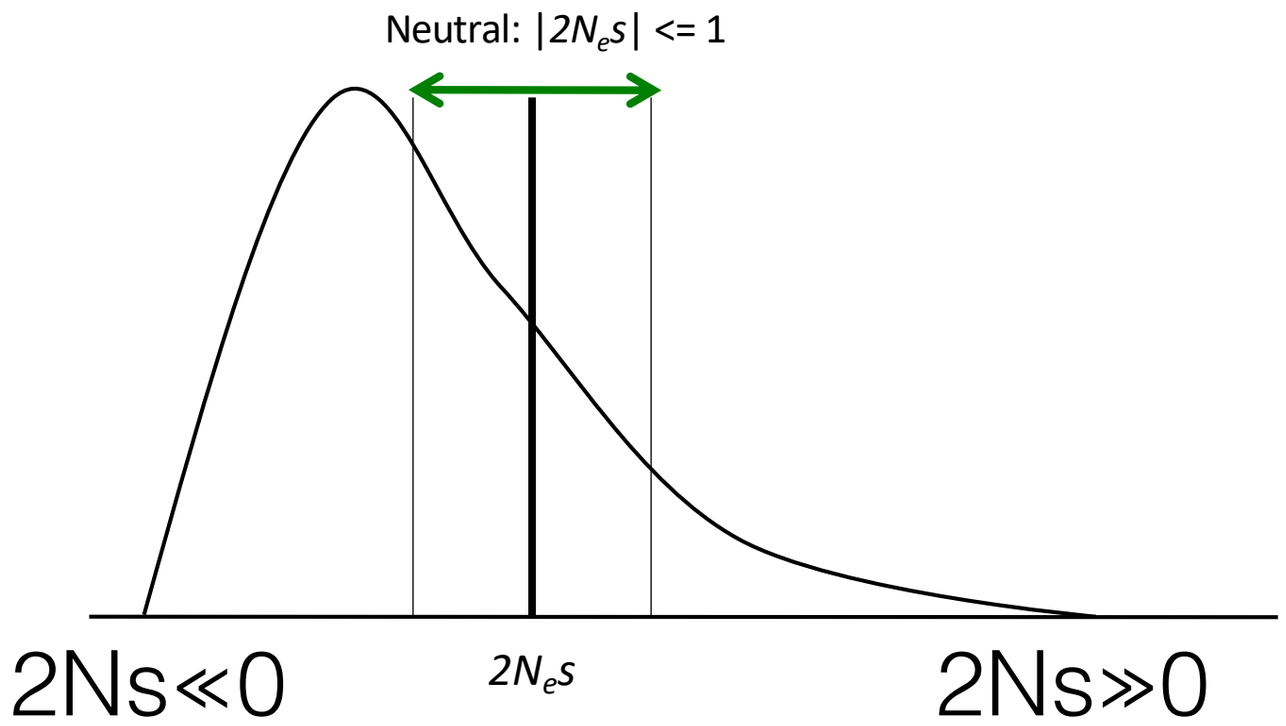
- Not all polymorphisms are neutral. The effect of segregating deleterious mutations has been observed across the variability of the genomes.

- Many mutations at functional positions are segregating at low frequency but do not arrive to fixation in the proportions expected by SNM.



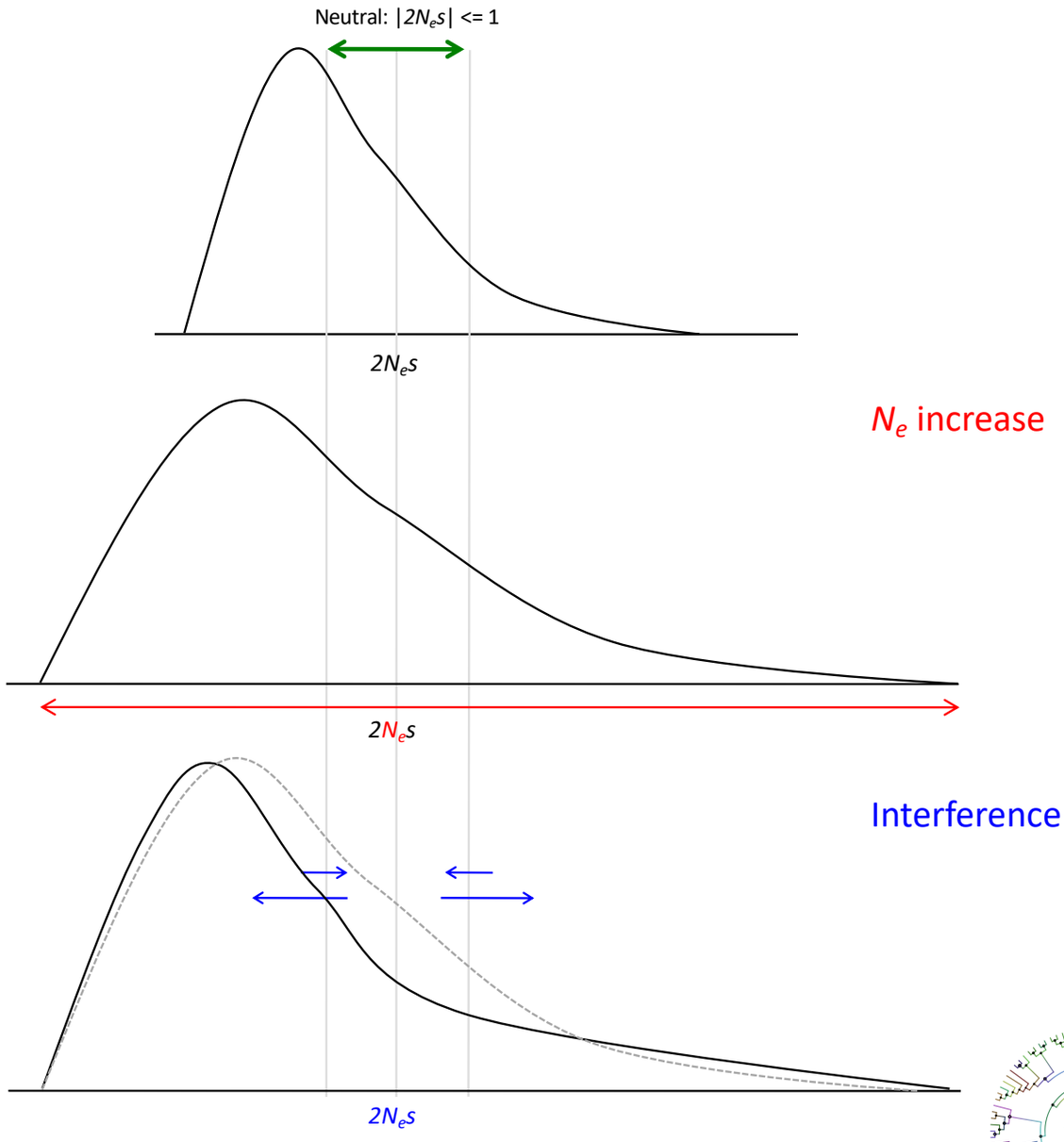
Expectations from Different Scenarios

- Otha's nearly-neutral theory define the selective effects in a more wide distribution. These selective effects are dependent on *Ne s.*



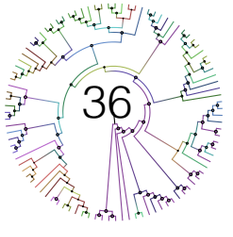
Expectations from Different Scenarios

- Demographic changes may affect the probability of fixation because they change the effective population size.
- Environmental changes may modify the selective effect of the variants.
- The distribution of the selective variants on the genome and their linkage relationship may also modify the general effect of selection on the individuals.



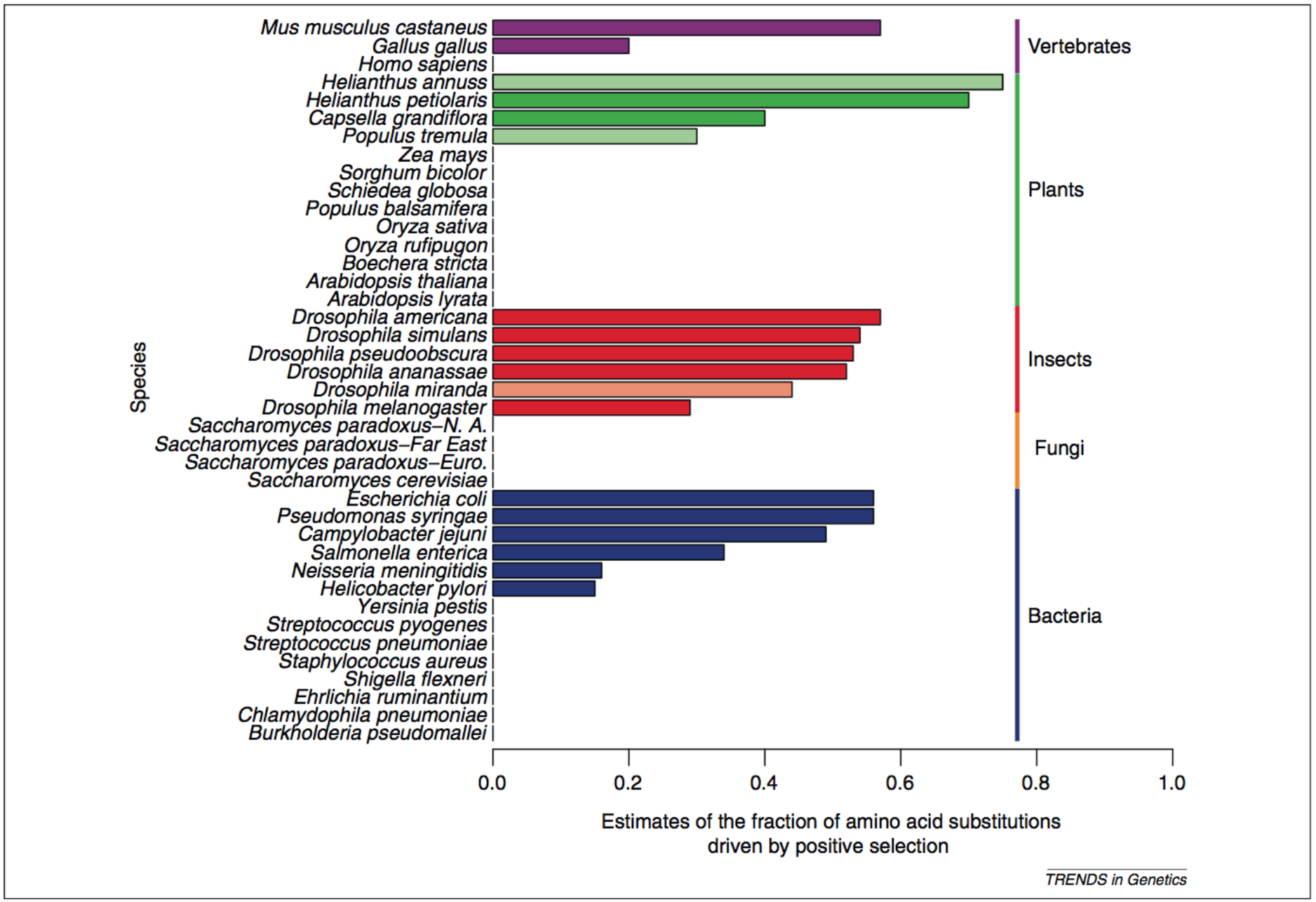
MKT modified framework

- Approaches mostly used to estimate α:
  - Exclude the low frequency variants to eliminate the effect of segregating deleterious mutations.
  - Estimate the asymptotic value of α at high frequencies.

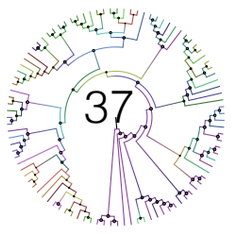


MKT modified framework

- Approaches mostly used to estimate α:

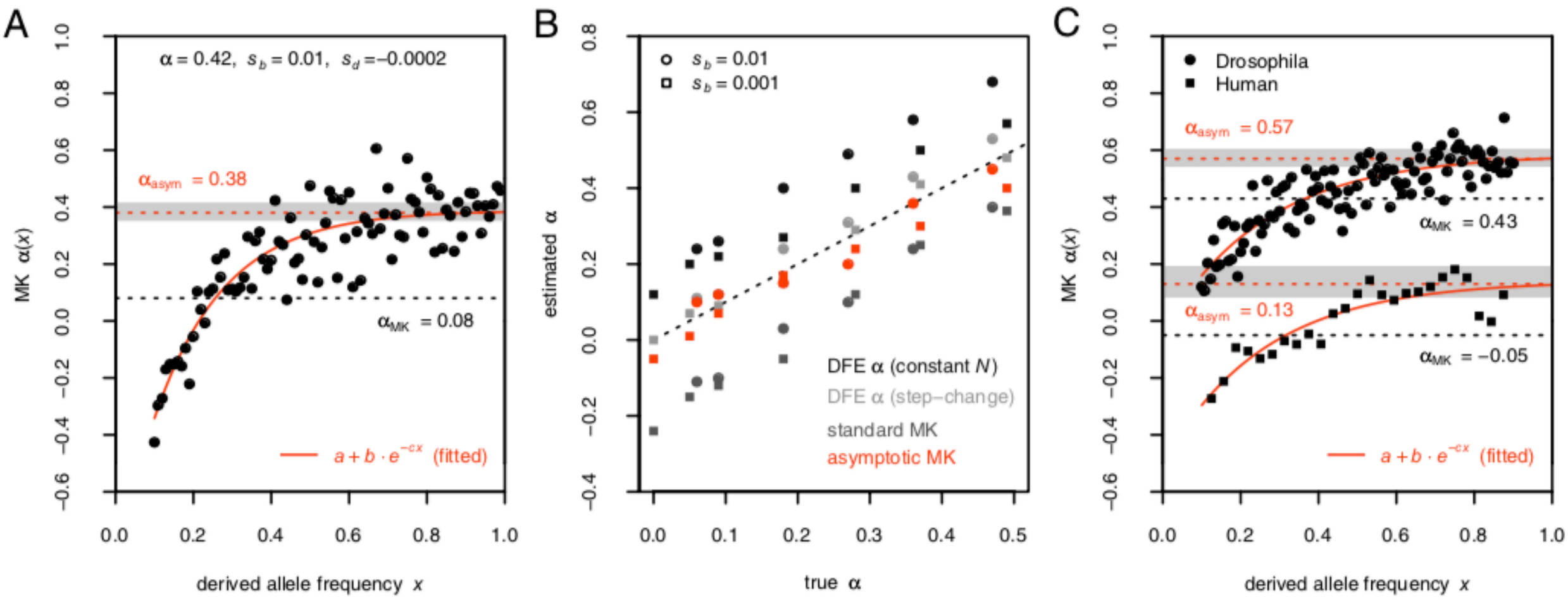


(Fay, TIG 2011)

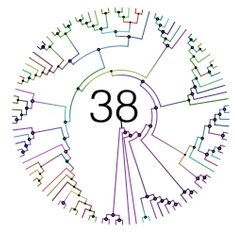


MKT modified framework

- Approaches mostly used to estimate α:



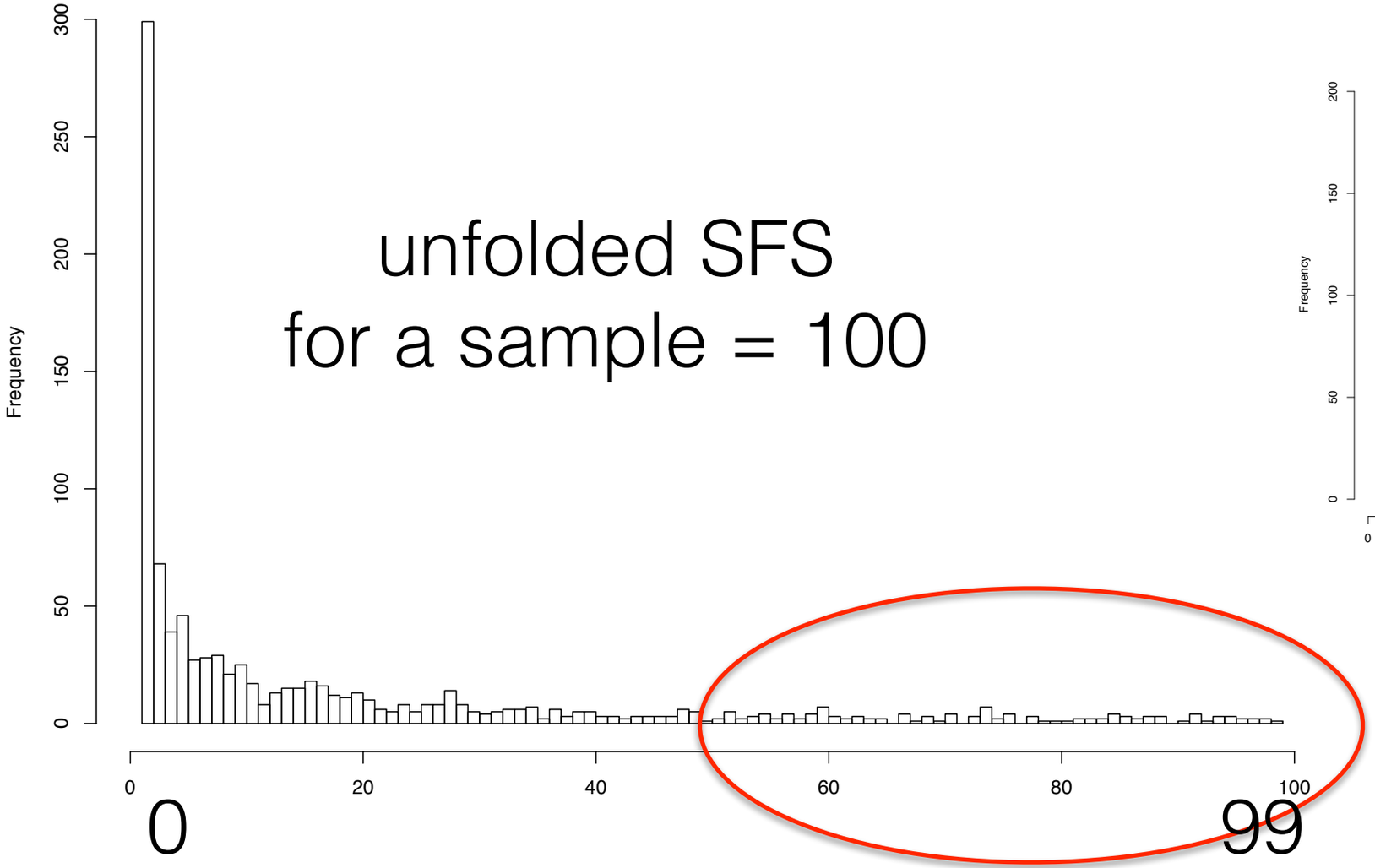
(Messer and Petrov PNAS 2013)

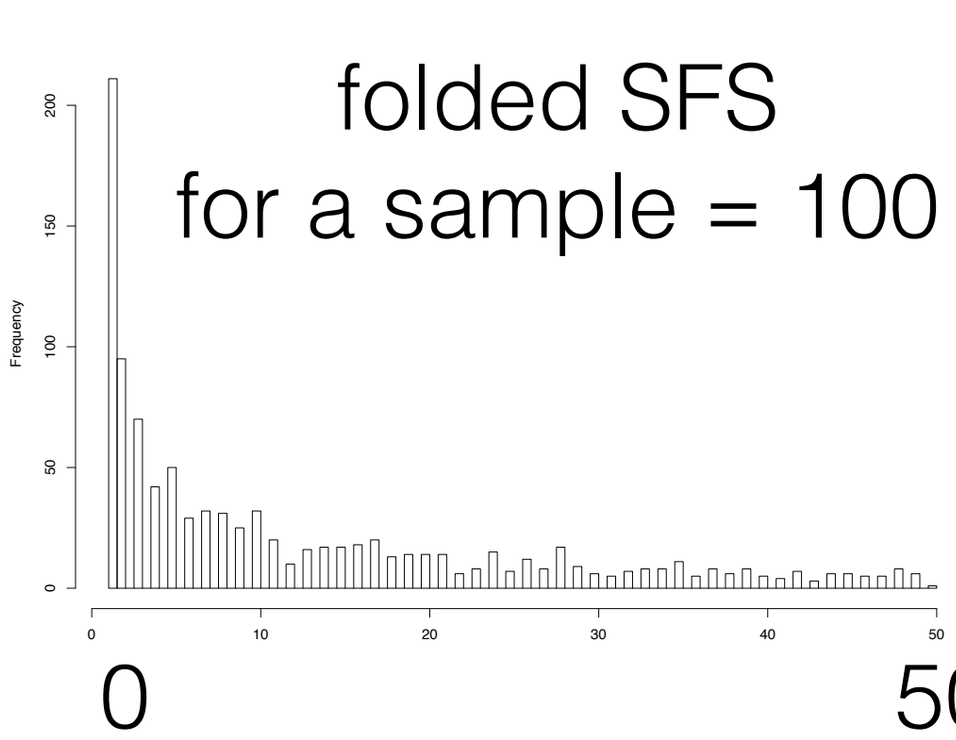


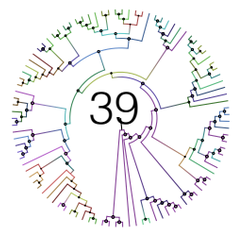
Inference of DFE from Genomics Variability

- Usual Assumptions

  - Independence between positions. Use the **Site Frequency Spectrum** and the Divergence as Observations.







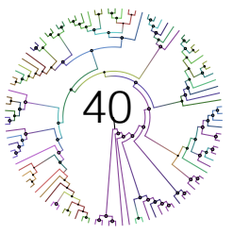
Inference of DFE from Genomics Variability

- Usual Assumptions

  - Independence between positions. Use the **Site Frequency Spectrum** and the Divergence as Observations.

  - Few or no demographic processes (no migration, no population structure, few or no changes in Ne). But see the use of nuisance parameters.

  - Effect of Directional Selection on new mutations.

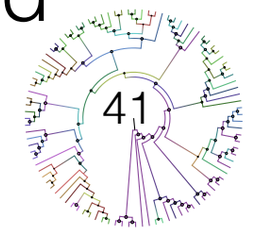


Inference of DFE from Genomics Variability

- From observations of **Site Frequency Spectrum** (SFS) of functional and neutral positions:

  - **No divergence** (no outgroup species reference): the **SFS is folded** (it is not possible to know if the variant is derived or ancestral). Assuming beneficial mutations are no contributing to polymorphisms

  - **Include divergence. SFS is unfolded.** Must correct for possible multiple hits. The proportions of adaptive substitutions plus the complete DFE may be estimated

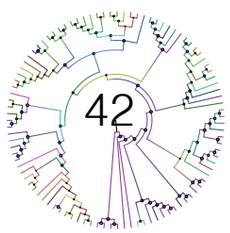


Inference of DFE from Genomics Variability

Table 3. Summary of studies inferring the distribution of scaled fitness effects, *Ns*, of nonsynonymous mutations

Organism	Method/dem model	$-1 < Ns < 0$	$-10 < Ns < 1$	$-10 < Ns$	Distribution(s) fitted	References
Human	Diffusion + complex demography	0.27	0.30	0.43	Mix of normal exponential/ neutral	57
Human	EWK2009	0.35	0.09	0.56	Γ	56
Mus musculus castaneus	K&K	0.19	0	0.81	LN, Γ, β, Spikes	59
Pan troglodytes	EWK2009	0.09	0.06	0.74	Γ	76
D. melanogaster	EWK2009	0.06	0.07	0.87	Γ	56
Saccharomyces cerevisae		0.25	0.25	0.5	Γ	63
S. paradoxus		0.2	0.2	0.6		
Angiosperms	EWK2009	0.1–0.35	0.05–0.15	0.7–0.8	Γ	77
Medicago truncatula	EWK2009	20–35	12–15	50–65	Γ	78

NOTE: EWK200956: diffusion based, simple demographic model fitted featuring a possible step change from population size *N*1 to population size *N*2 at some time *t* in the past (*N*1, *N*2, and *t* become "nuisance parameters" estimated alongside DFE and the fraction of favorable mutations).
K&K: discrete W–F matrix based, demographic model identical to EWK2009.
Dem, demographic; LN, log normal; spikes, spikes at different *Ns* class values.



(Bataillon & Bailey Ann NY Acad. Sc.. 2014)

Inference of DFE from Genomics Variability

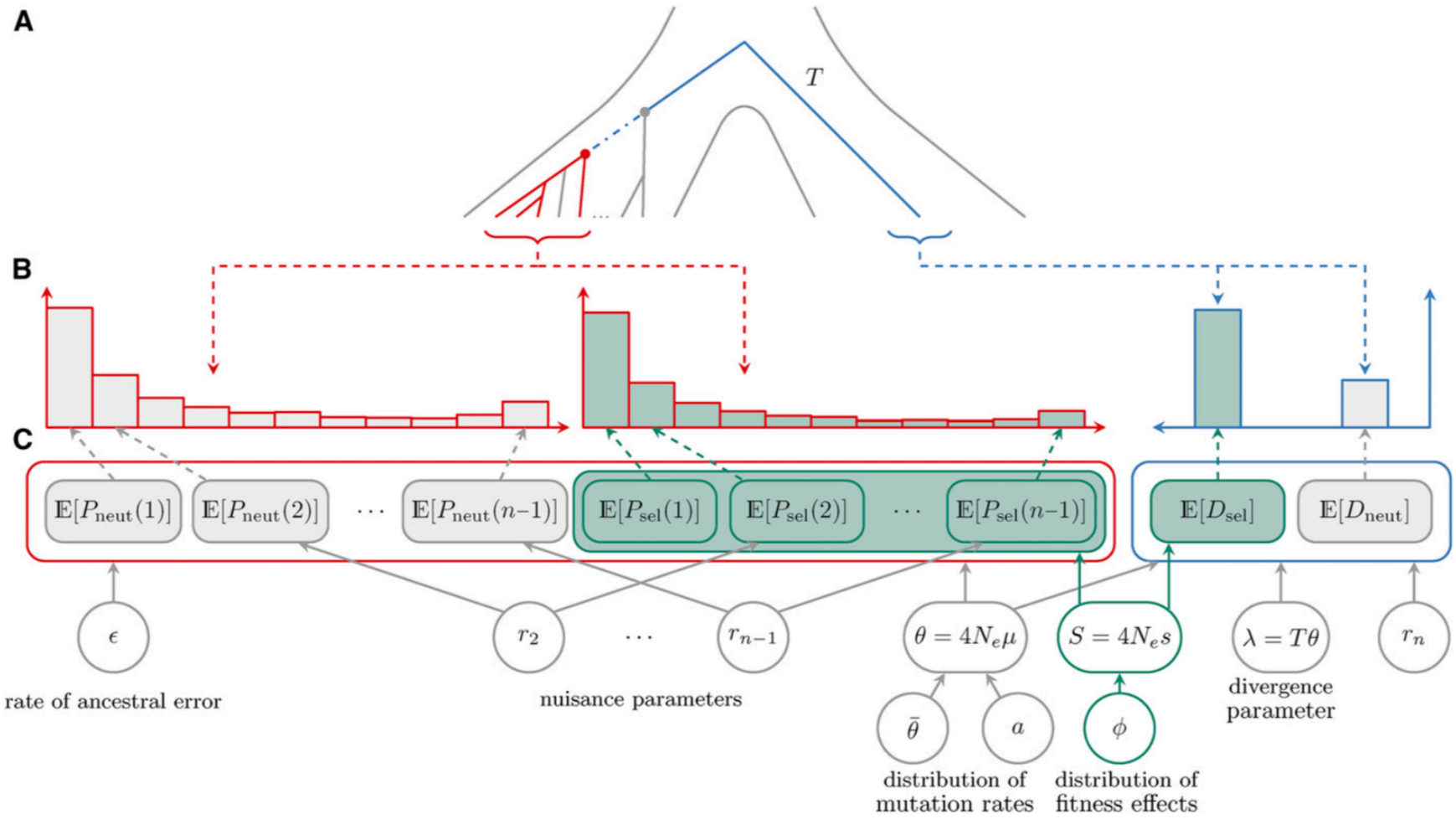
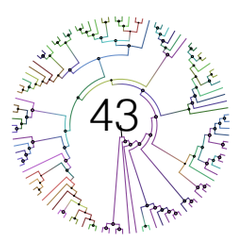


Figure 1 Schematic of data and the hierarchical model assumed in the inference method. Throughout the figure, gray and blue fill indicates sites that are assumed to be evolving neutrally or potentially under selection, respectively; while red and blue outlines indicates polymorphism and divergence data (expectations), respectively. (A) The history and coalescent tree of two populations: the ingroup (on the left side), for which polymorphism data are collected, and the outgroup (on the right side), for which divergence counts are obtained. A total of n sequences are sampled from the ingroup (marked in red), with the MRCA marked with a red circle. The MRCA of the whole ingroup population is marked with a gray circle. From the outgroup we typically have access to one sequence (marked in blue). The total evolutionary time between the MRCA of the sample (red circle) and the sampled outgroup sequence can be divided into the time from the MRCA of the sample to the MRCA of the whole ingroup population (blue dot–dash line) and T, the time from the ingroup MRCA to the sampled outgroup sequence (blue solid line). (B) Observed SFS and divergence counts [$p_z(i)$ and d_z, with $z \in \{\text{neut}, \text{sel}\}$ and $1 \le i < n$]. (C) Expected counts [$\mathbb{E}[P_z(i)]$ and $\mathbb{E}[D_z]$, with $z \in \{\text{neut}, \text{sel}\}$ and $1 \le i < n$], model parameters, and relations between parameters, expectations, and data. The dashed gray and blue ↓'s connect observed counts from (B) with matching expected counts from (C).



(Tataru et al.Genetics 2017)

Inference of DFE from Genomics Variability

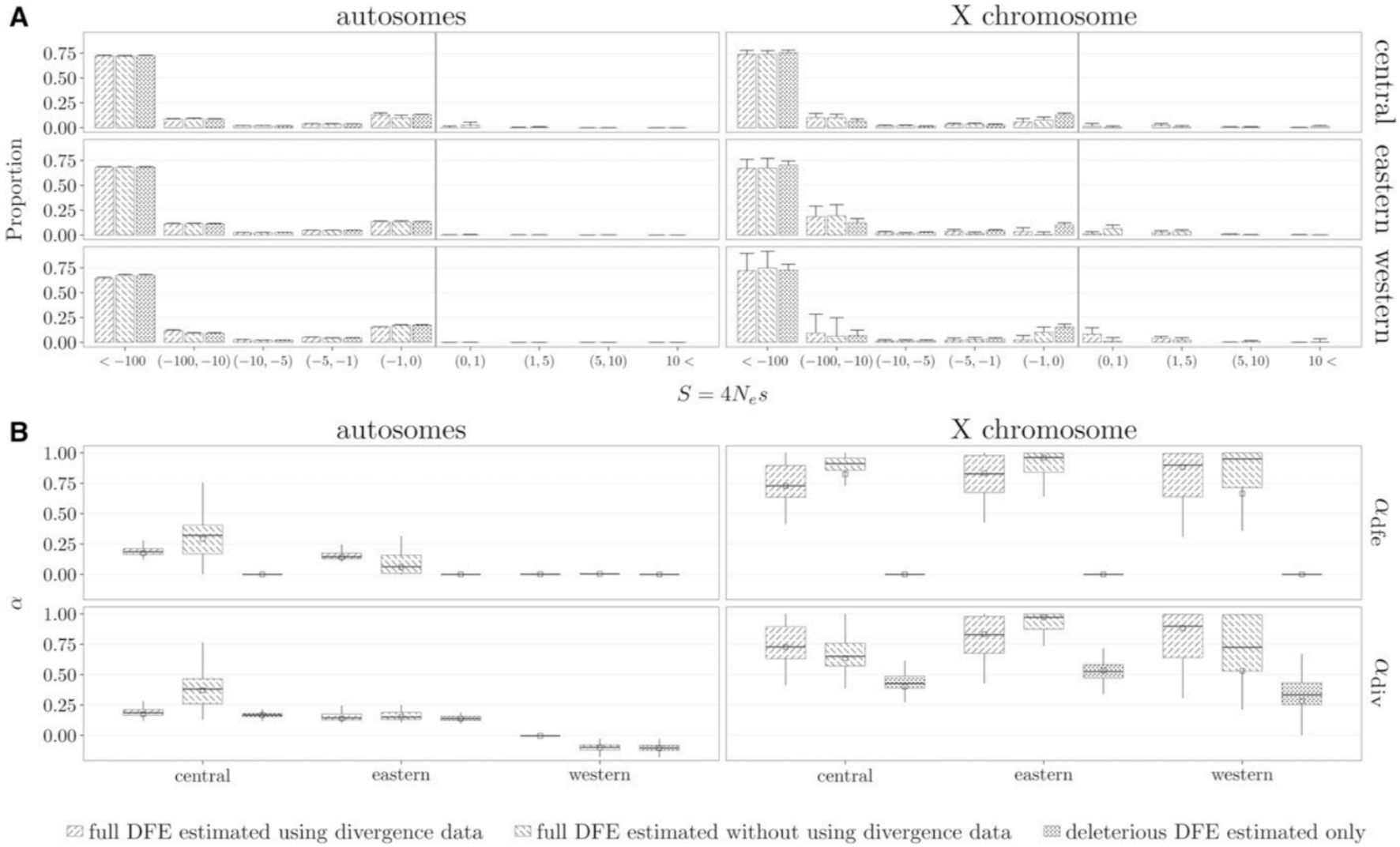
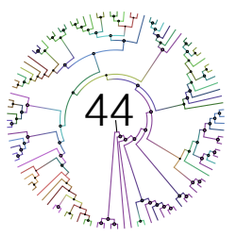


Figure 8 Inference of DFE and α (proportion of beneficial substitutions) on three chimpanzee subspecies. A full DFE was inferred from both polymorphism and divergence data, while also both a full DFE and deleterious DFE were inferred from polymorphism data only. (A) Inferred discretized DFE. The error bars indicate 1 SD obtained from the inferred discretized DFEs from 100 bootstrap data sets. (B) Box plot of inferred α_{dfe} and α_{div} from the 100 bootstrap data sets. The values of α inferred on the original data sets are given as empty squares. Note that when inferring a deleterious DFE only, α_{dfe} is zero.



(Tataru et al.Genetics 2017)

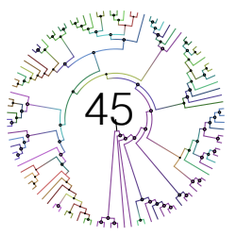
Relationship between traits and fitness

- Traits can be unlinked with fitness (neutral traits) while others may be strongly associated.

- Traits can be determined by a single gene or alternatively by a large number of genes (Quantitative Trait Loci, QTL).

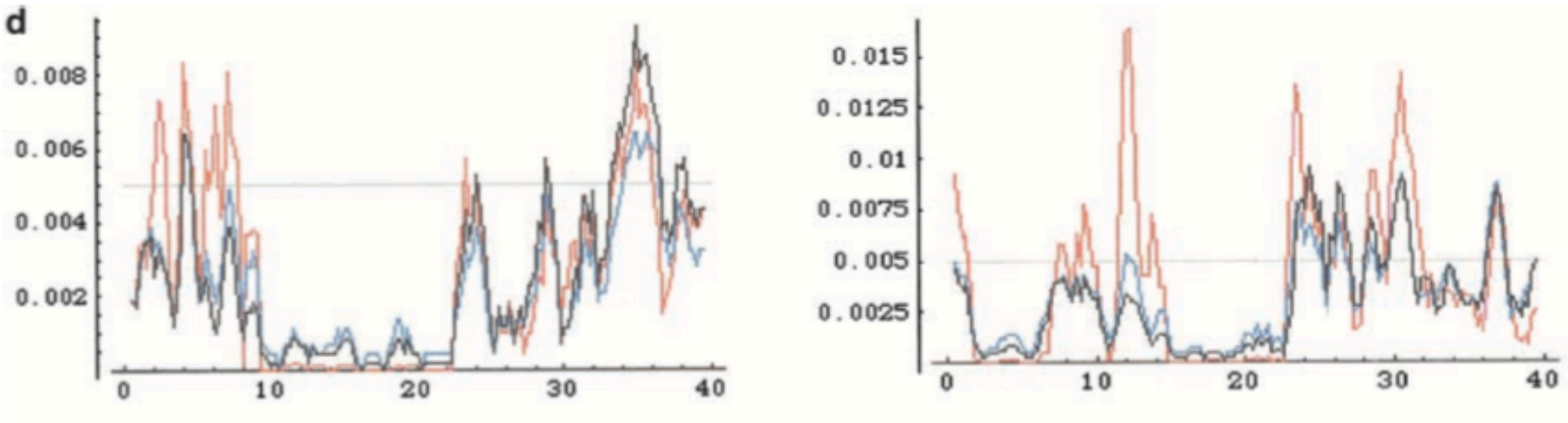
- Traits can change their association with fitness in function of the environment.

- Loci can affect single trait or alternatively affect a large number of traits (pleiotropic scenario).



Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.
- Traits affecting severely the fitness that are determined by a single gene would show a selective sweep pattern.



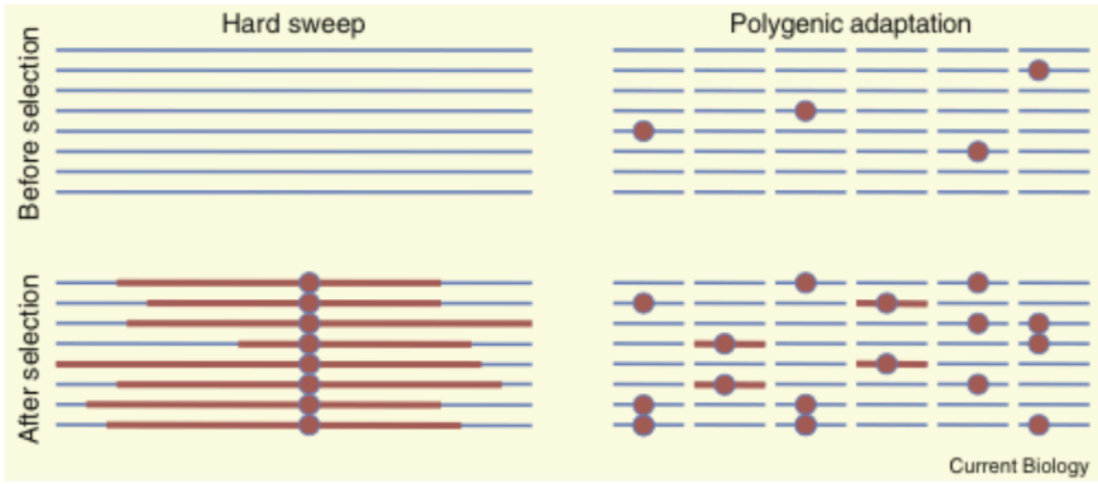
(Kim and Stephan, Genetics 2002)

Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.

- Instead, **a trait** affecting fitness that is determined by **many genes** may only show a slight increase in frequencies in all them (**infinitesimal effect**).

- In this case, it is very important to determine **what loci are affecting the trait** (for example, using GWAS methods).

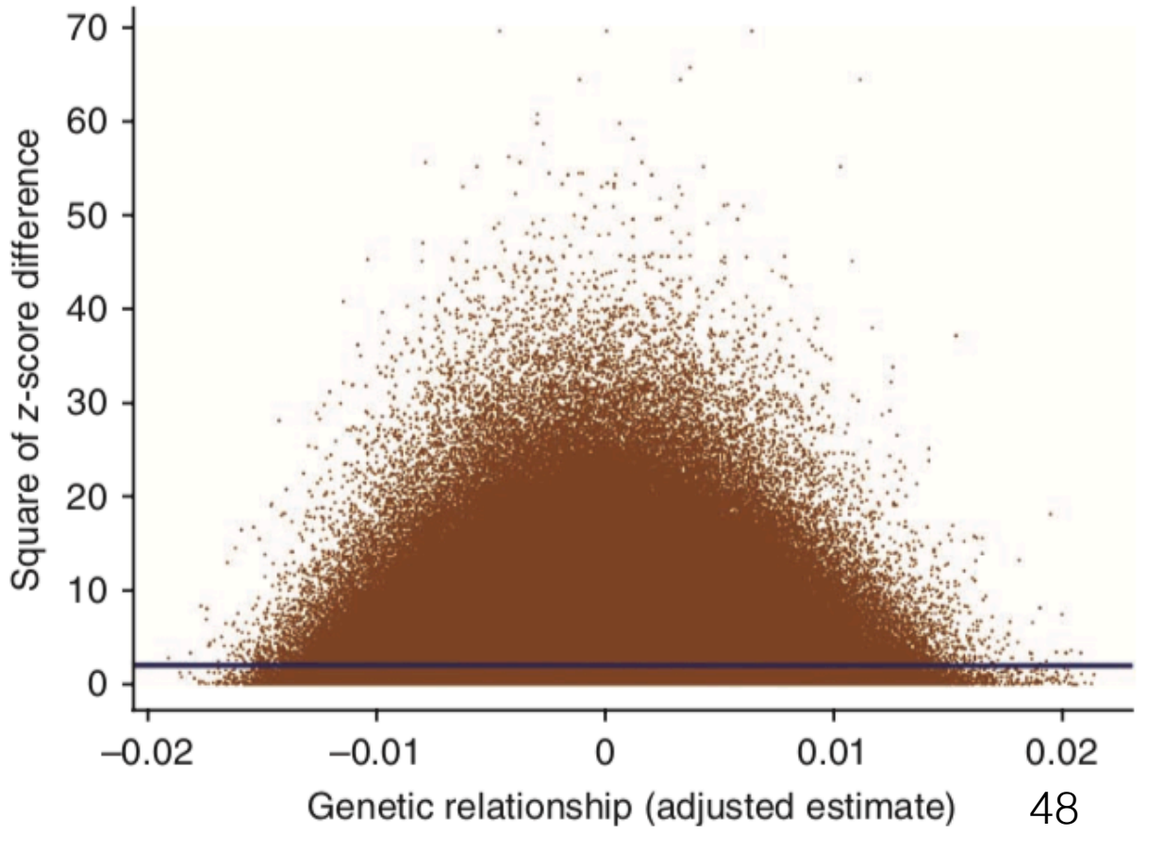


(Pritchard et al. Curr Biology 2010)

Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.

- If **many traits are affecting the fitness**, the patterns of variability may be also hard to distinguish from neutrality as well, because many genes would be implicated.



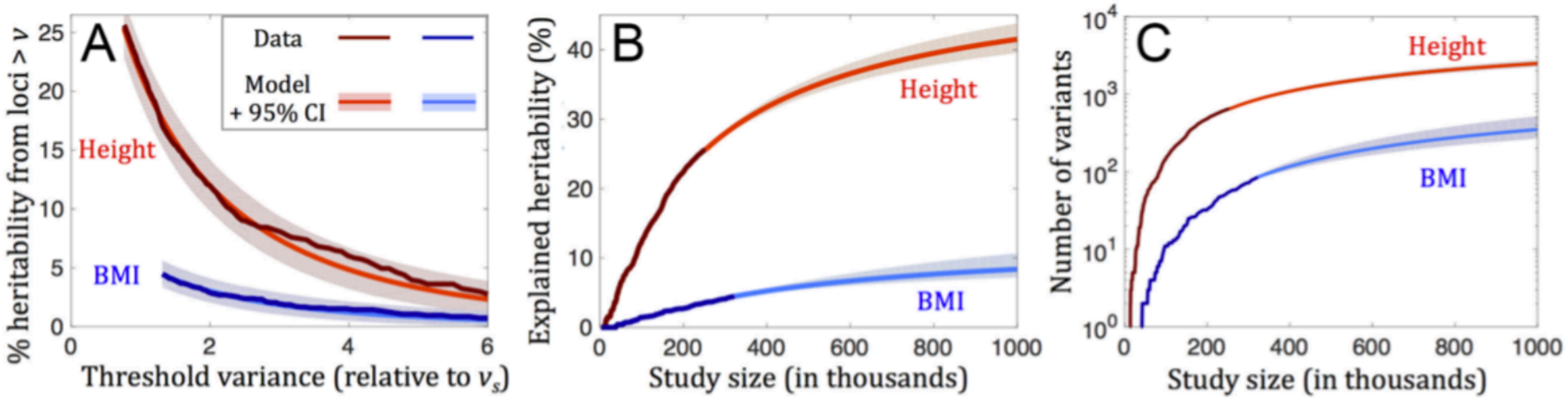
(Yang et al. -Visscher-Nat. Genet. 2010)

Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.

- If **stabilising selection** for a number of traits is maintaining the variability in **many** loci -instead of an equilibrium between mutation-selection-drift-

- Estimation of the effect of the loci implicated and their number may explain the observed patterns of association under this model.

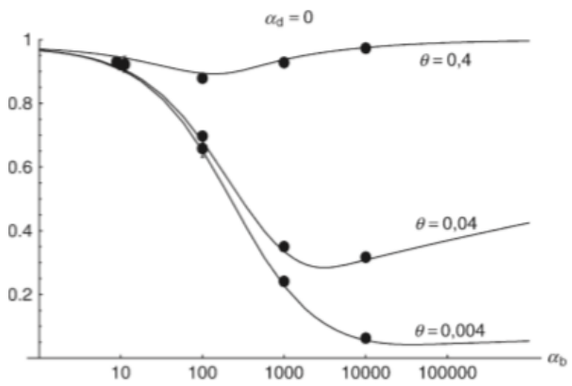


(Sella et al. PLoS Biol. 2018)

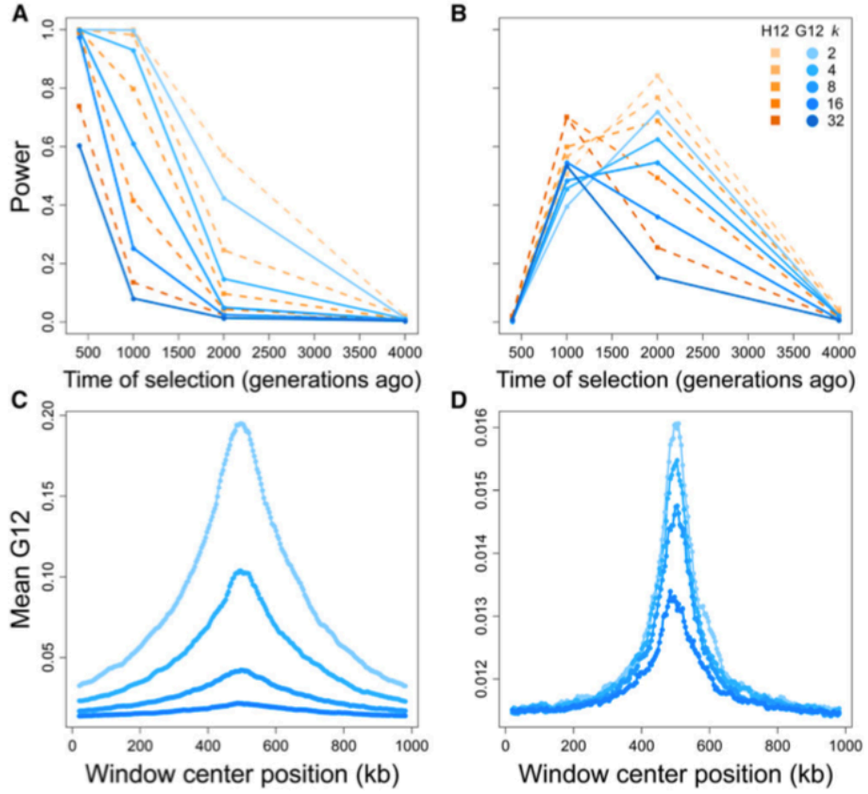
Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.

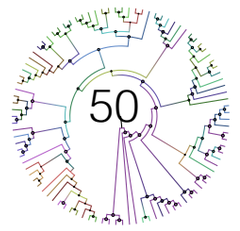
- **Changes in the environment** can change the traits - and consequently the loci- that are implicated in fitness, producing new genomic patterns. **Soft selective sweeps** may be common.



(Permission and Pennings. Genetics 2015)



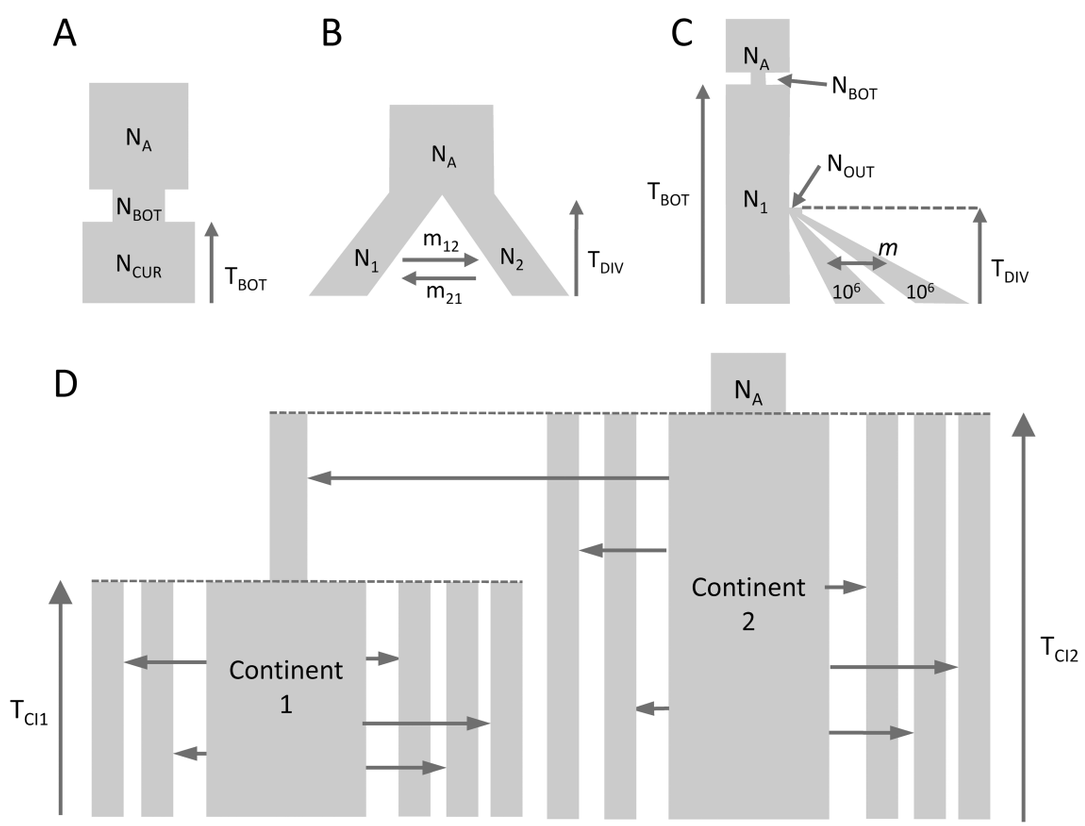
(Harris et al. Genetics 2018)

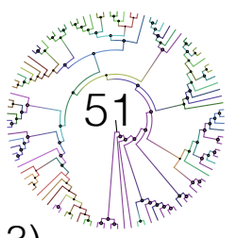


Relationship between traits and fitness

- The relation of the genes that control the trait and the relation of these traits with the fitness determine the pattern of variability at the genome.

- **Demographic Changes** can modify the genetic architecture by the modification of the number of implicated loci and the effect of each locus to fitness.





(Excoffier et al. PLoS Genet. 2013)

Review Lectures

- Fitness and its role in evolutionary genetics. H. Allen Orr. Nature Reviews Genetics. Volume 10, August 2009. pp 531-539.

- Effects of new mutations on fitness: insights from models and data. Thomas Bataillon and Susan F. Bailey. Ann. N.Y. Acad. Sci. pp. 76-92. ISSN 0077-8923.

- Introduction in: A population genetic interpretation of GWAS findings for human quantitative traits. Simons YB, Bullaughey K, Hudson RR, Sella G. (2018) PLoS Biol 16(3): e2002985.

- Selective Sweeps, Wolfgang Stephan. Genetics January 1, 2019 vol. 211 no. 1 5-13;

- Sònia Casillas and Antonio Barbadilla. Molecular Population Genetics. Genetics, Vol. 205, 1003–1035 March 2017.