the evolutionary dynamics of genes

jose cuesta

uc3m
GISC

PNAS

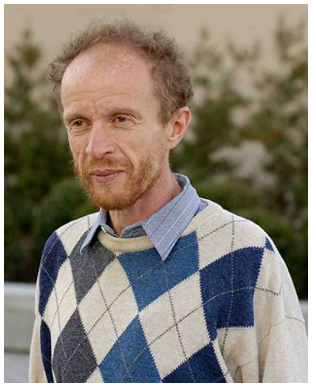
PNAS PLUS

Disentangling the effects of selection and loss bias on gene dynamics

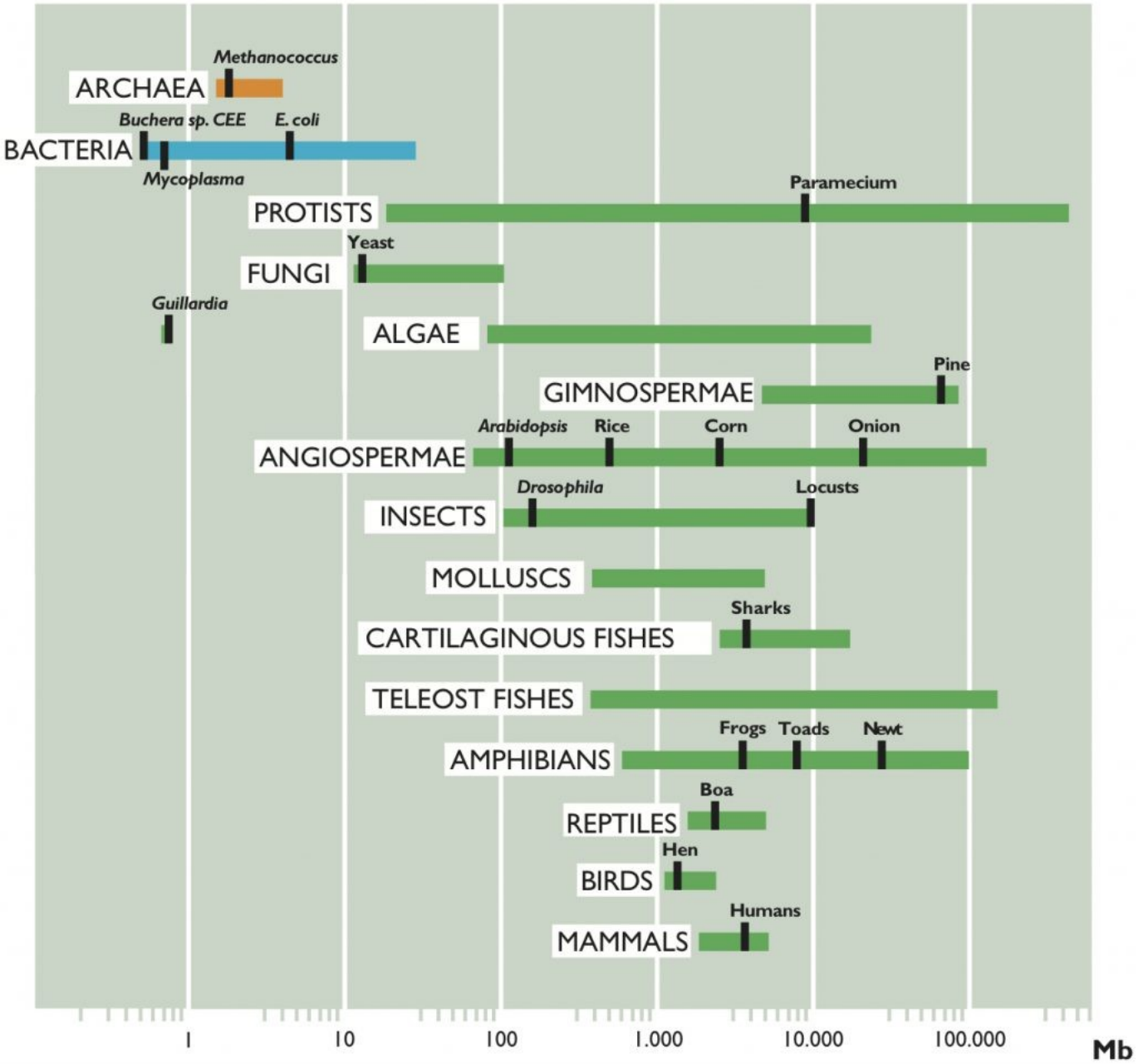
Jaime Iranzo[a], José A. Cuesta[b,c,d], Susanna Manrubia[e], Mikhail I. Katsnelson[f], and Eugene V. Koonin[a,1]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; [b]Grupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Spain; [c]Institute for Biocomputation and Physics of Complex Systems, Universidad de Zaragoza, 50018 Zaragoza, Spain; [d]Institute of Financial Big Data, Universidad Carlos III de Madrid-Banco de Santander, 28903 Getafe (Madrid), Spain, [e]Grupo Interdisciplinar de Sistemas Complejos, National Biotechnology Centre, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain; and [f]Institute for Molecules and Materials, Radboud University, Nijmegen 6525AJ, The Netherlands

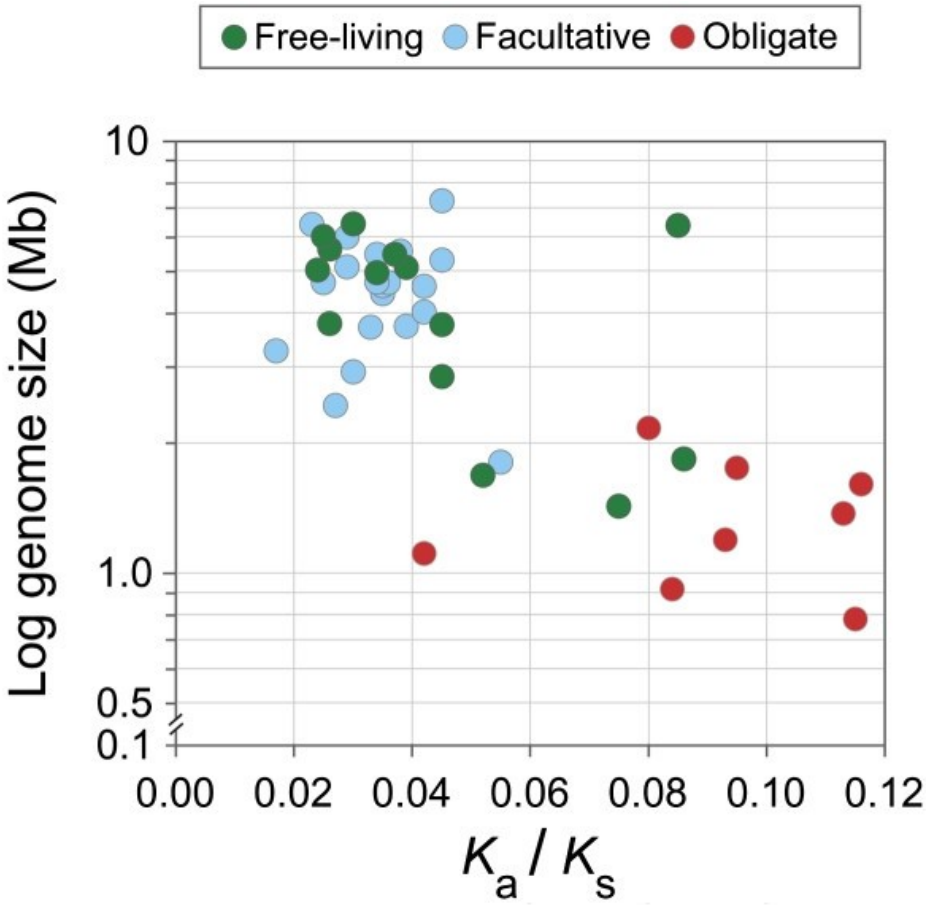
Contributed by Eugene V. Koonin, June 1, 2017 (sent for review March 24, 2017; reviewed by Sergei Maslov and Dennis Vitkup)

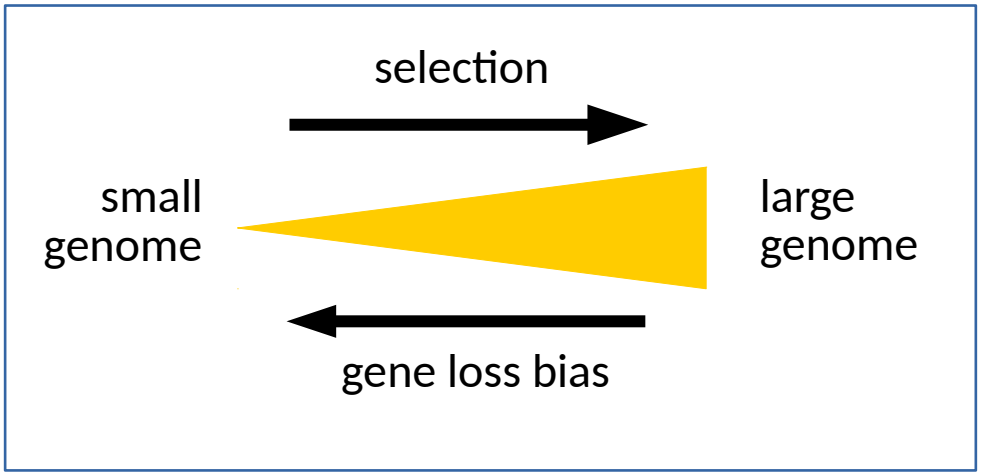


huge variation of genome sizes



selection vs bias to gene loss…

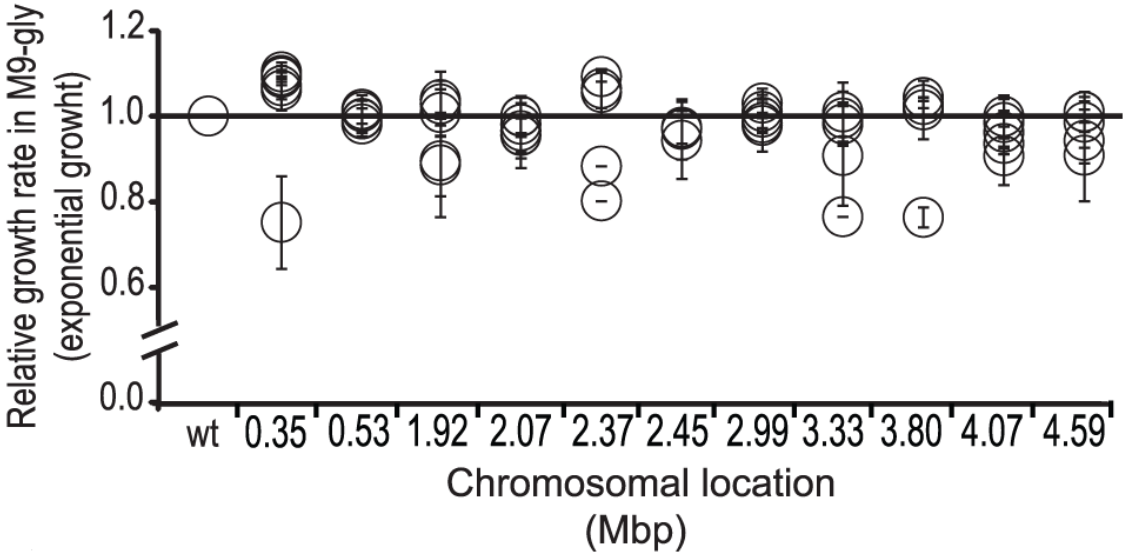




Kuo, Moran & Ochman, *Genome Res.* (2009)

… but it's complicated because…

- selection can promote gene deletions

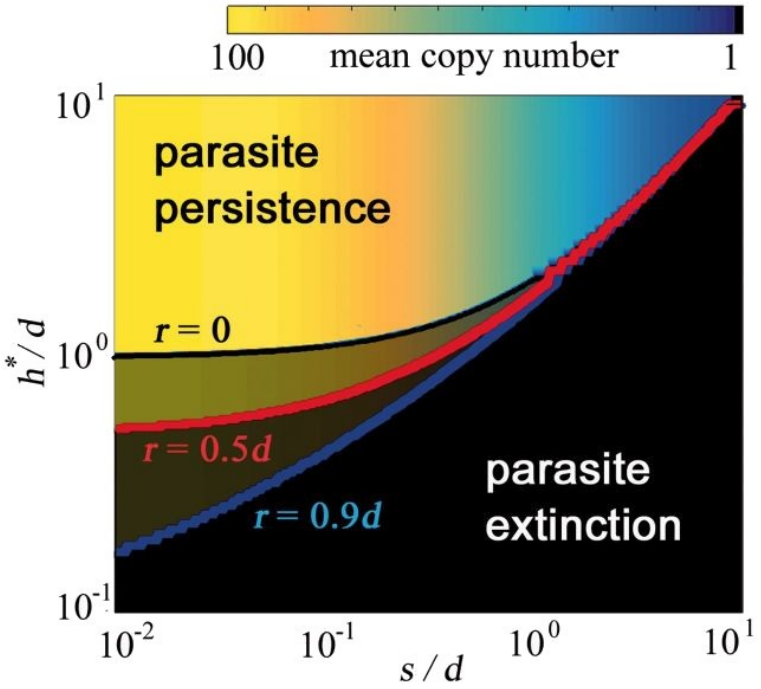


25% of large random deletions in *Salmonella enterica* are beneficial in one or more growth conditions

Koskiniemi et al., *PLoS Genet.* (2012)

… but it's complicated because…

- selection can promote gene deletions
- horizontal gene transfer (HGT) contribute to gene maintenance

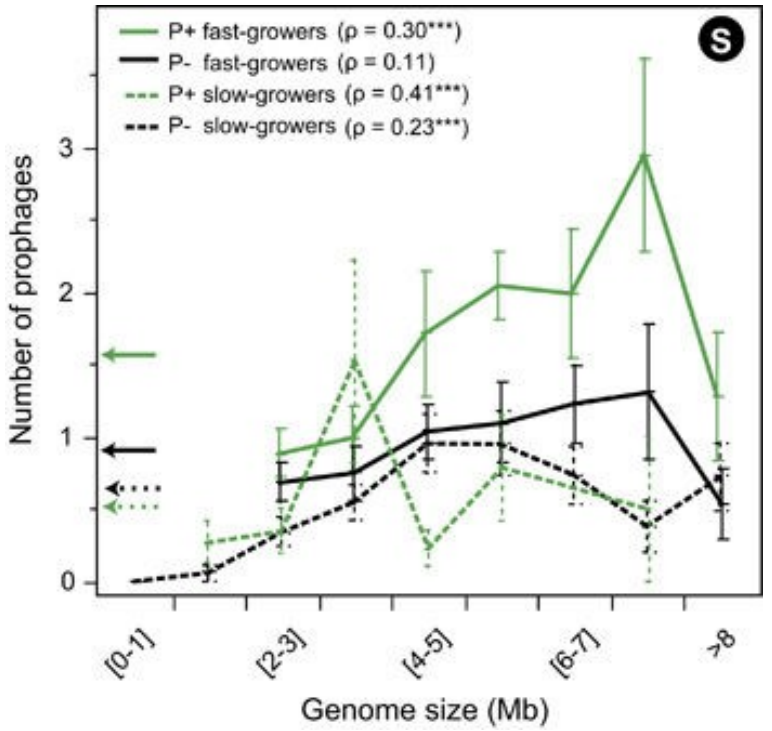


parasites persist through increased HGT despite purifying selection

Iranzo et al., *Genome Biol. Evol.* (2016)

… but it's complicated because…

- selection can promote gene deletions
- horizontal gene transfer (HGT) contribute to gene maintenance
- abundance of genetic parasites correlates positively with genome size



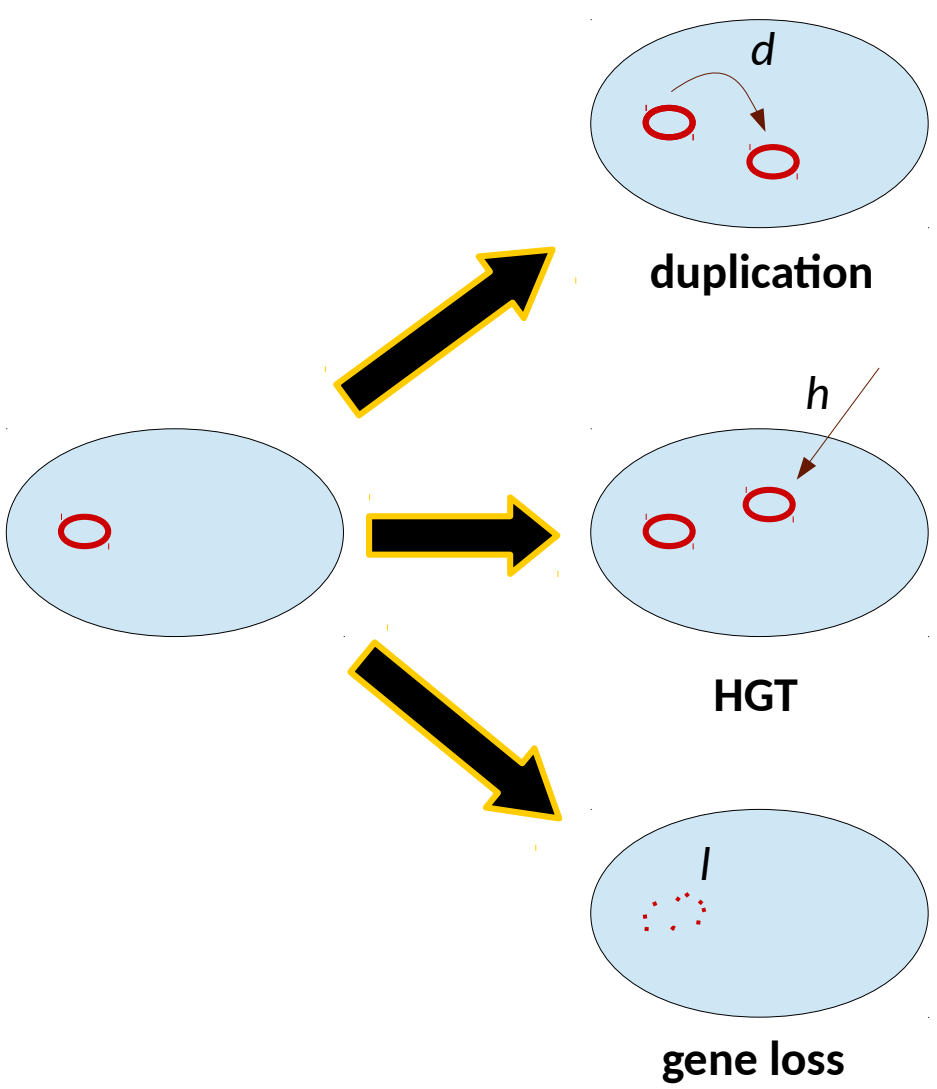
prophages are more abundant in larger genomes (supposedly subject to stronger selection)

Touchon et al., *ISME J.* (2016)

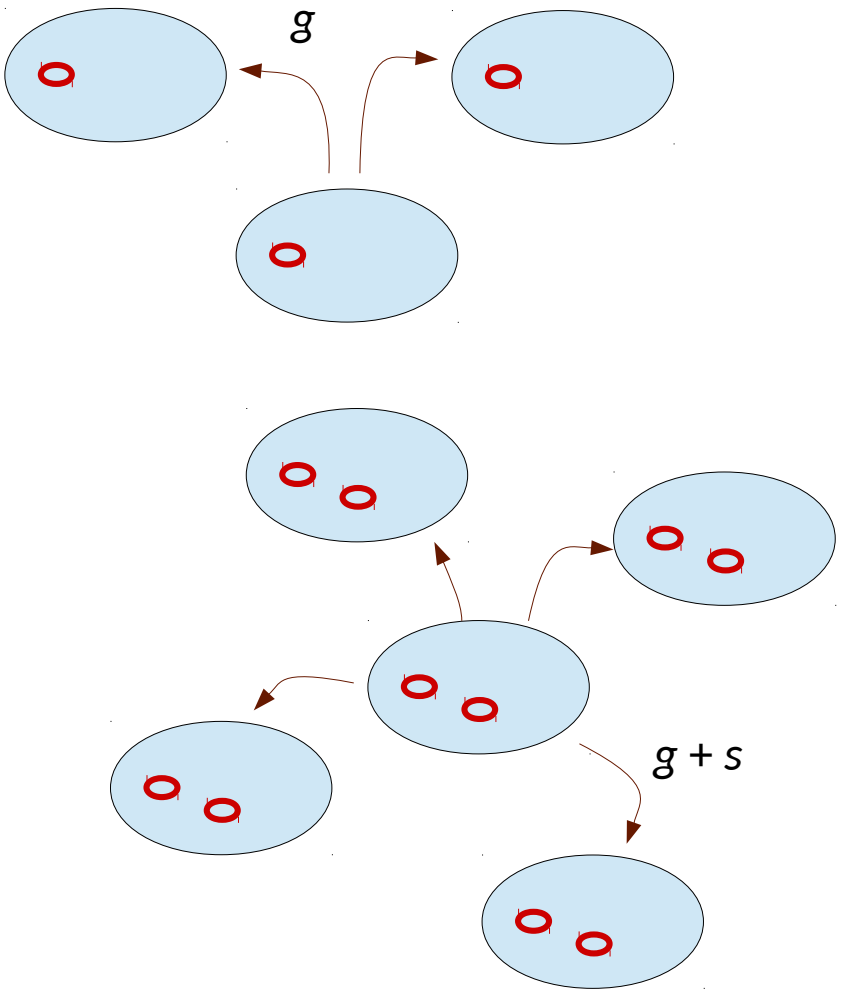
what is the interplay between selection, gene loss, HGT, etc., in a genome?

processes to model

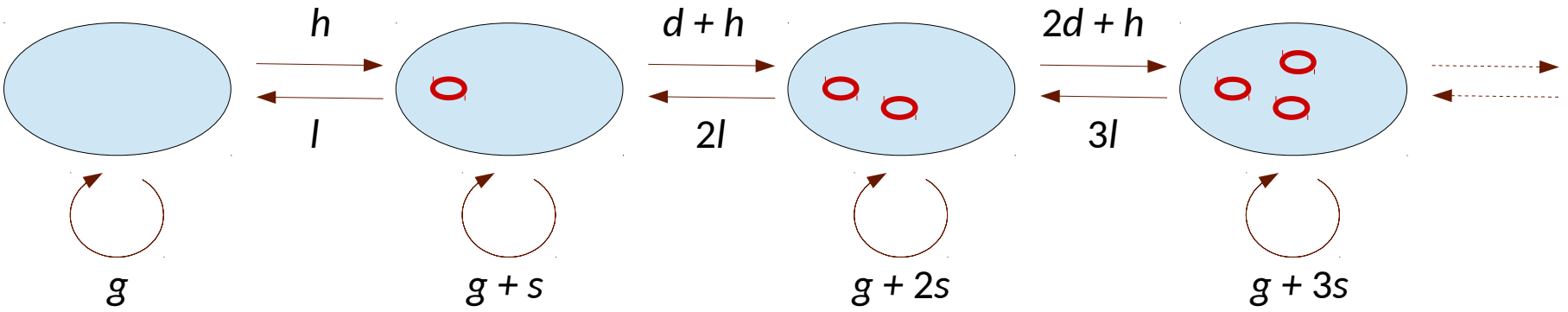
neutral processes



differential growth / selection



duplication-loss-transfer-selection model



in a genome with *k* copies:

loss rate	$k\,l$
duplication rate	$k\,d$
HGT rate	h
selection factor	$k\,s$
basal growth rate	g

n_k: number of genomes with k copies

$$n_k(t) = e^{gt} m_k(t)$$

$$m_{k<0}(t) = 0 \qquad \alpha = d - s + l$$

$$\frac{d\,m_k}{dt} = -(h + k\,\alpha)\,m_k + (k+1)\,l\,m_{k+1} + [(k-1)\,d + h]\,m_{k-1}$$

dimensionless equations

scale all rates with loss rate (l): $d \to d/l$, $h \to h/l$, $s \to s/l$, $\alpha \to \alpha/l$
scale time with l^{-1}: $t \to lt$
equivalently: $l = 1$

$$\frac{d\,m_k}{dt} = -(h + k\,\alpha)\,m_k + (k+1)\,m_{k+1} + [(k-1)\,d + h]\,m_{k-1} \qquad k \in \mathbb{Z}$$

van Kampen's shift operators:

$$\mathbf{E}\,f_k = f_{k+1} \qquad \mathbf{E}^{-1} f_k = f_{k-1} \qquad k \in \mathbb{Z}$$

$$\frac{d\,m_k}{dt} = (\mathbf{E} - 1 + s)\,k\,m_k + (\mathbf{E}^{-1} - 1)\,(d\,k + h)\,m_k$$

some useful properties

1. $\langle \mathbf{E}^{\mp 1} f_k, g_k \rangle = \langle f_k, \mathbf{E}^{\pm 1} g_k \rangle$ $\qquad \langle f_k, g_k \rangle = \sum_{k \in \mathbb{Z}} f_k^* g_k$

2. $\mathbf{E}^{\pm 1} z^k = z^{\pm 1} z^k$ $\qquad z \in \mathbb{C}$

3. $\left(z \frac{\partial}{\partial z} \right) z^k = k z^k$

generating function

$$G(z,t) \equiv \langle m_k(t), z^k \rangle = \sum_{k=0}^{\infty} m_k(t)\, z^k$$

$$\langle \frac{d\, m_k}{d\, t}, z^k \rangle = \frac{\partial G}{\partial t}$$

$$\langle (\mathbf{E} - 1 + s)\, k\, m_k, z^k \rangle = (z^{-1} - 1 + s)\, z \frac{\partial}{\partial z} \langle m_k, z^k \rangle = [1 - (s-1)\, z] \frac{\partial G}{\partial z}$$

$$\langle (\mathbf{E}^{-1} - 1)(d\, k + h)\, m_k, z^k \rangle = (z-1)\left(d\, z \frac{\partial}{\partial z} + h\right) \langle m_k, z^k \rangle = d\, z (z-1) \frac{\partial G}{\partial z} + h (z-1) G$$

$$\frac{\partial G}{\partial t} = (d\, z^2 - \alpha\, z + 1) \frac{\partial G}{\partial z} + h (z-1) G$$

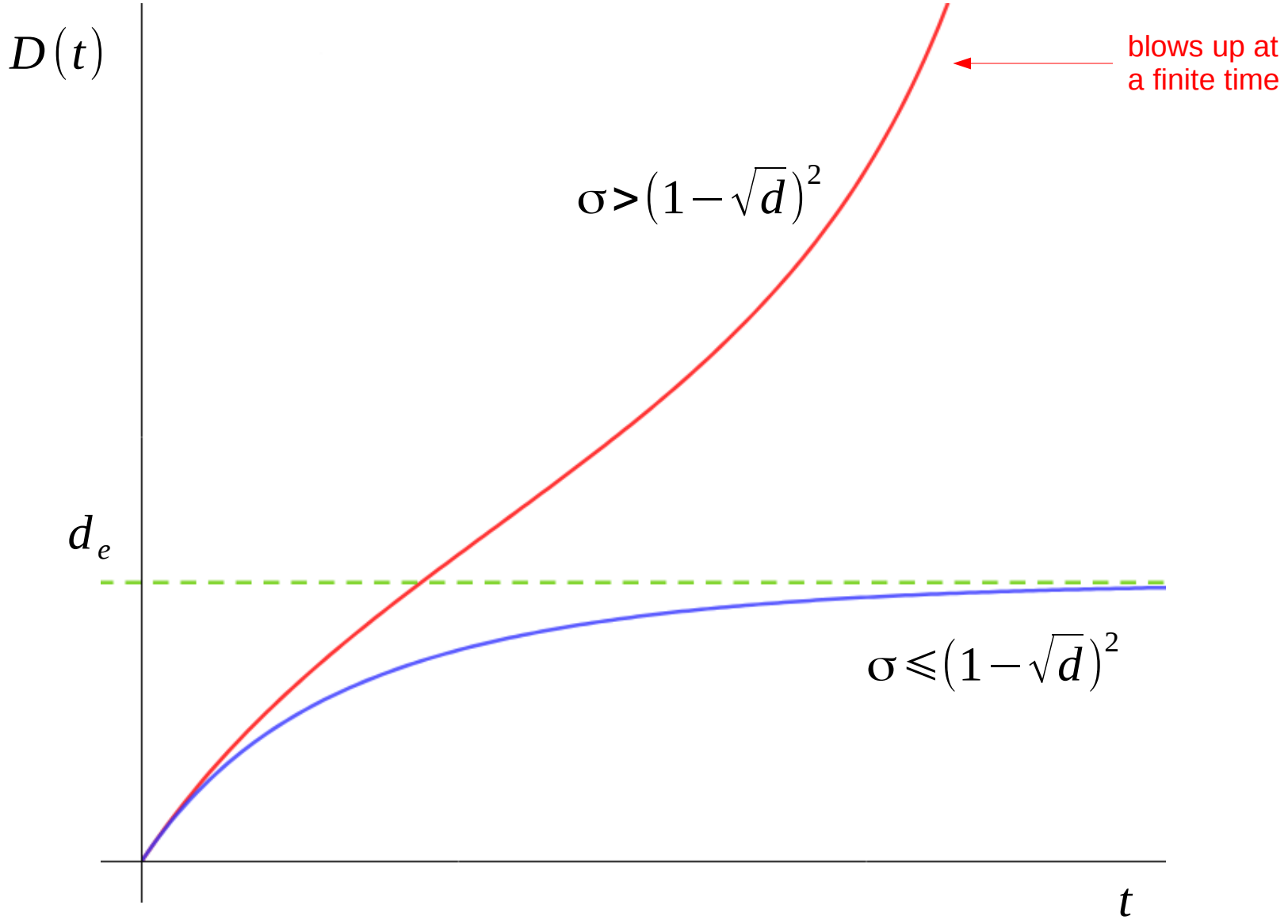
solution

$$H(z,t) \equiv \frac{G(z,t)}{G(1,t)} = \sum_{k=0}^{\infty} z^k \, p_k(t) \qquad p_k(t) = \frac{m_k(t)}{\sum_j m_j(t)}$$

$$H(z,t) = \left(\frac{1 - D(t)}{1 - z\,D(t)} \right)^{h/d}$$

$$D(t) = d_e \left(\frac{1 - \mathrm{e}^{-(d/d_e - d_e)t}}{1 - (d_e^2/d)\,\mathrm{e}^{-(d/d_e - d_e)t}} \right) \qquad d_e \equiv \frac{2d}{\alpha + \sqrt{\alpha^2 - 4d}}$$

solution


$D(t)$
blows up at
a finite time
$\sigma > (1-\sqrt{d})^2$
d_e
$\sigma \leqslant (1-\sqrt{d})^2$
t

distribution of gene copy number

$$p_k(t) = (1 - D(t))^{h/d} \frac{D(t)^k}{k!} \frac{\Gamma(k + h/d)}{\Gamma(h/d)}$$

stationary distribution (only if $\sigma \leqslant (1 - \sqrt{d})^2$):

$$p_k = (1 - d_e)^{h_e/d_e} \frac{d_e^k}{k!} \frac{\Gamma(k + h_e/d_e)}{\Gamma(h_e/d_e)} \qquad \frac{h_e}{d_e} = \frac{h}{d}$$

neutral case ($s = 0$):

$$d_e = \min(1, d)$$

d_e: effective neutral duplication rate
h_e: effective neutral HGT rate

ambiguity of the distribution

measuring the distribution of gene copy number one cannot know:

(a) if the process is stationary or transient

(b) if the process is neutral or subject to selection

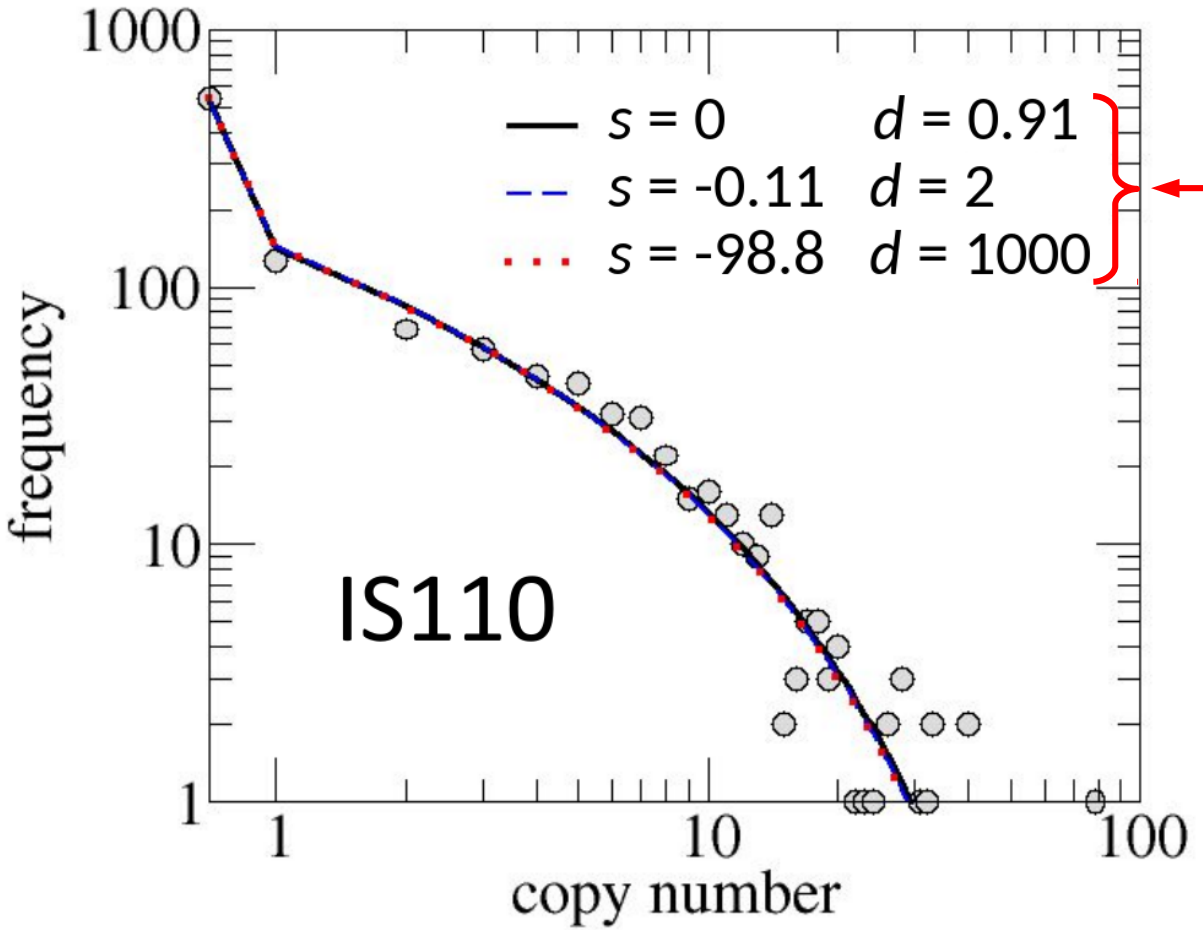
unless independent measurements of some rates are conducted

if *d* can be independently measured, selection is determined trough:

$$s = \frac{(1 - d_e)(d_e - d)}{d_e}$$

goodness of fit

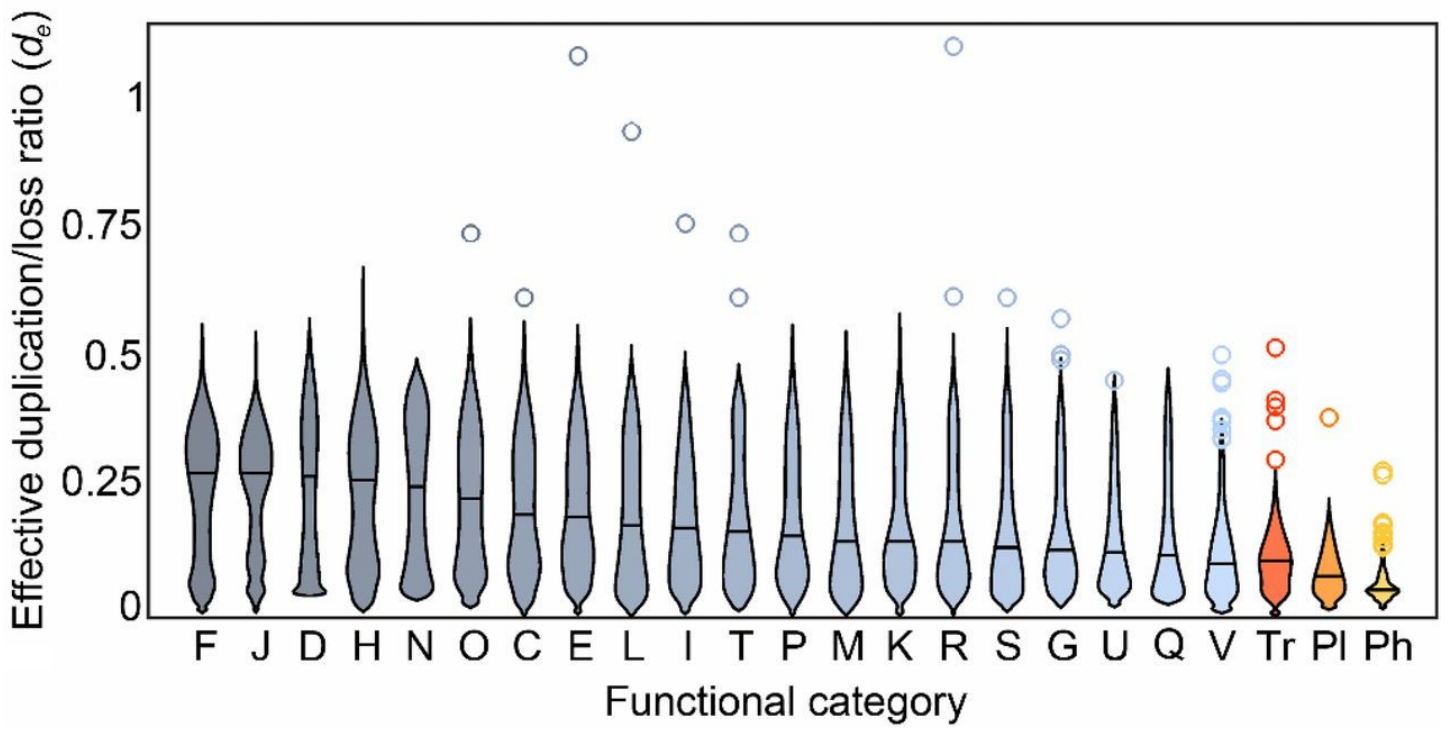
empirical distribution of copy numbers of 33 transposon families, obtained from 1811 bacterial chromosomes (typical case)



equivalent fits

Iranzo et al., *PLoS Comp. Biol.* (2014)

application to genomic data



dataset

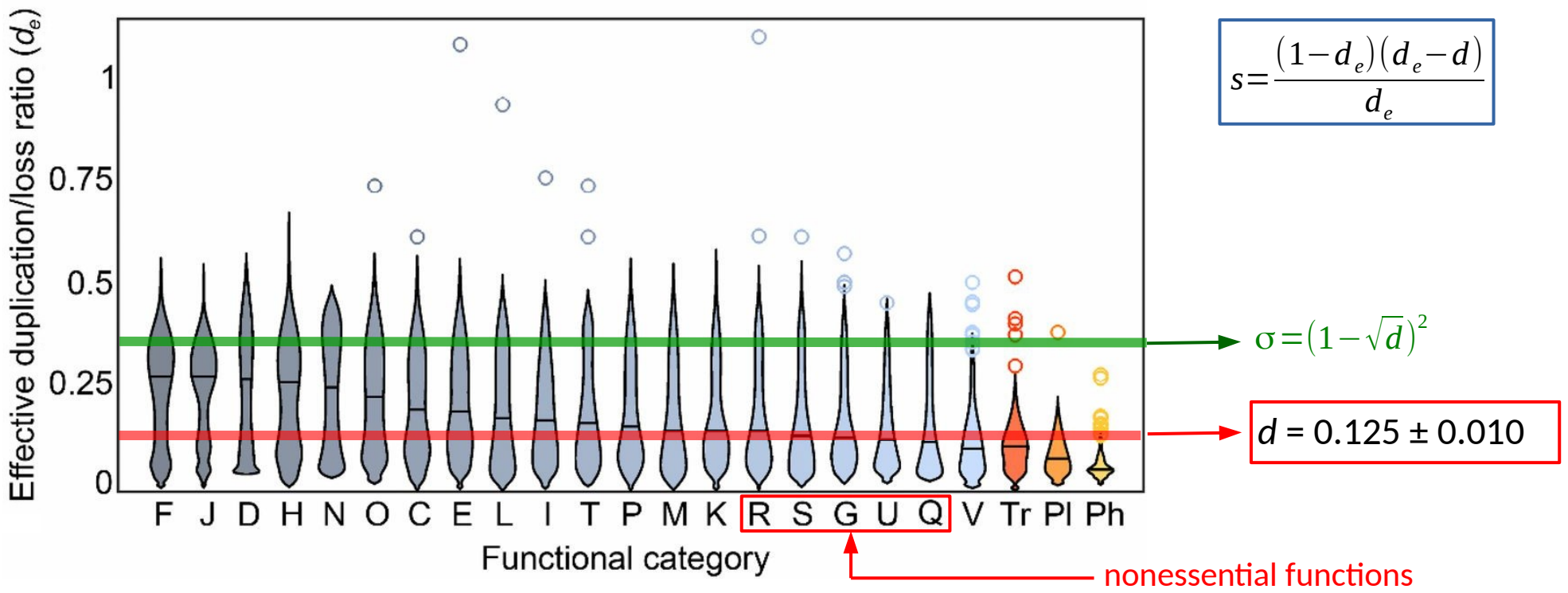
clusters of orthologs in 35 sets of closely related bacterial and archaeal genomes (678 genomes)

parameter estimates

COUNT: software for phylogenetic analysis

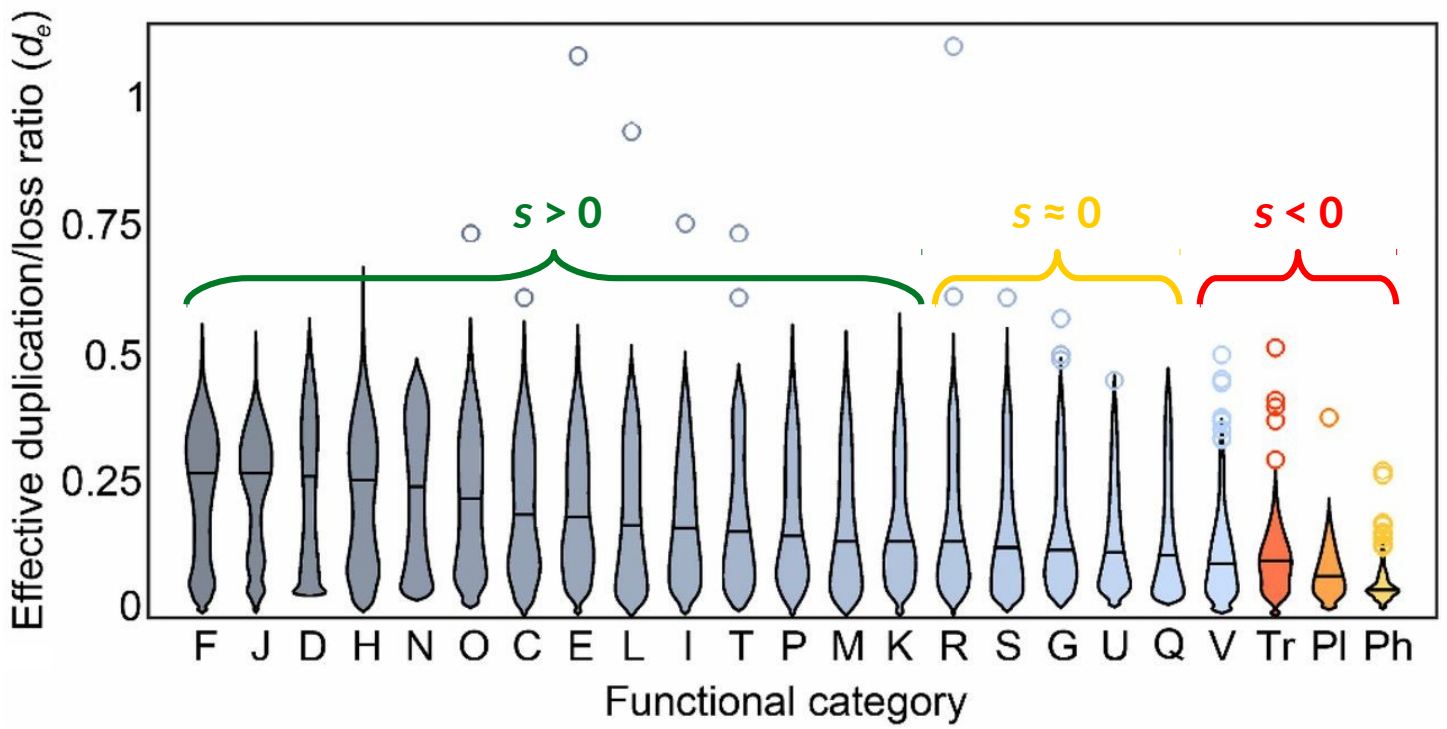
C: energy production/conversion
D: cell division
E: amino acid metabolism and transport
F: nucleotide metabolism and transport
G: carbohydrate metabolism and transport
H: coenzyme metabolism
I: lipid metabolism
J: translation
K: transcription
L: replication and repair
M: membrane and cell wall structure and biogenesis
N: secretion and motility
O: posttranslational modification, protein turnover & chaperone functions
P: inorganic ion transport and metabolism
Q: biosynthesis, transport, and catabolism of secondary metabolites
R: general functional prediction only (typically, of biochemical activity)
S: function unknown
T: signal transduction
U: intracellular trafficking and secretion
V: defense mechanisms
Tr: transposon
Pl: conjugative plasmid
Ph: prophage or phage-related

application to genomic data



C: energy production/conversion
D: cell division
E: amino acid metabolism and transport
F: nucleotide metabolism and transport
G: carbohydrate metabolism and transport
H: coenzyme metabolism
I: lipid metabolism
J: translation
K: transcription
L: replication and repair
M: membrane and cell wall structure and biogenesis
N: secretion and motility
O: posttranslational modification, protein turnover & chaperone functions
P: inorganic ion transport and metabolism
Q: biosynthesis, transport, and catabolism of secondary metabolites
R: general functional prediction only (typically, of biochemical activity)
S: function unknown
T: signal transduction
U: intracellular trafficking and secretion
V: defense mechanisms
Tr: transposon
Pl: conjugative plasmid
Ph: prophage or phage-related

application to genomic data



$$s = \frac{(1-d_e)(d_e-d)}{d_e}$$

l = (0.5 – 4) x 10^{-8} losses per gene per generation

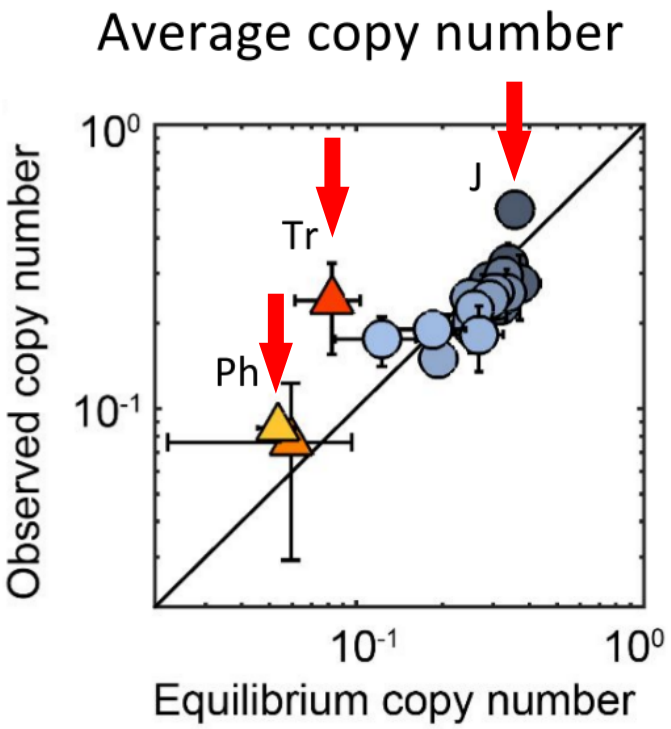
Nilsson et al., *PNAS* (2005)
Sung et al., *G3* (2016)

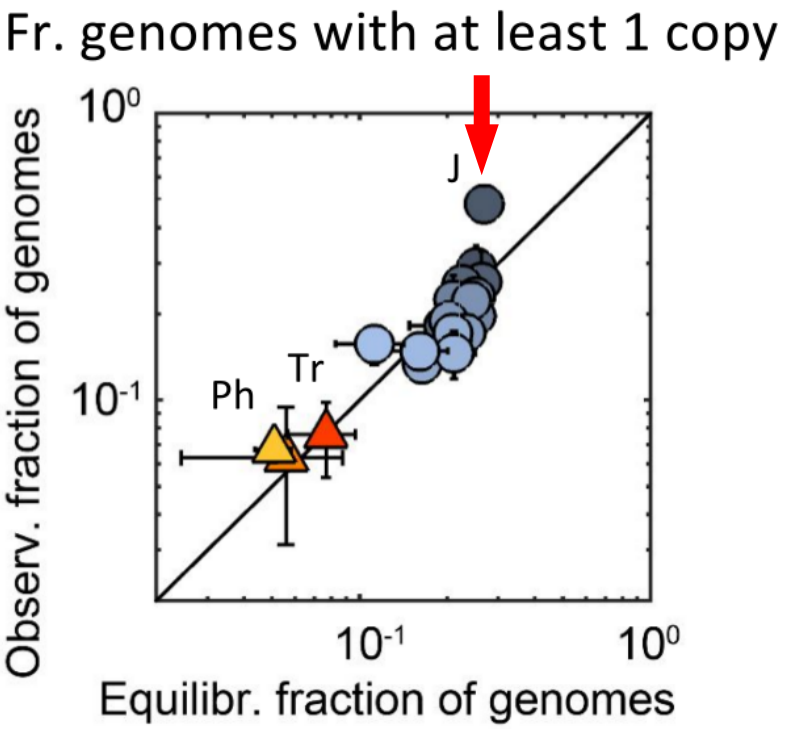
C: energy production/conversion
D: cell division
E: amino acid metabolism and transport
F: nucleotide metabolism and transport
G: carbohydrate metabolism and transport
H: coenzyme metabolism
I: lipid metabolism
J: translation
K: transcription
L: replication and repair
M: membrane and cell wall structure and biogenesis
N: secretion and motility
O: posttranslational modification, protein turnover & chaperone functions
P: inorganic ion transport and metabolism
Q: biosynthesis, transport, and catabolism of secondary metabolites
R: general functional prediction only (typically, of biochemical activity)
S: function unknown
T: signal transduction
U: intracellular trafficking and secretion
V: defense mechanisms
Tr: transposon
Pl: conjugative plasmid
Ph: prophage or phage-related

comparison of long-term dynamics

$$\langle k \rangle = \left(\frac{\partial}{\partial z} H(z,t) \right)_{z=1} = \frac{h}{d} \frac{D(t)}{1-D(t)}$$

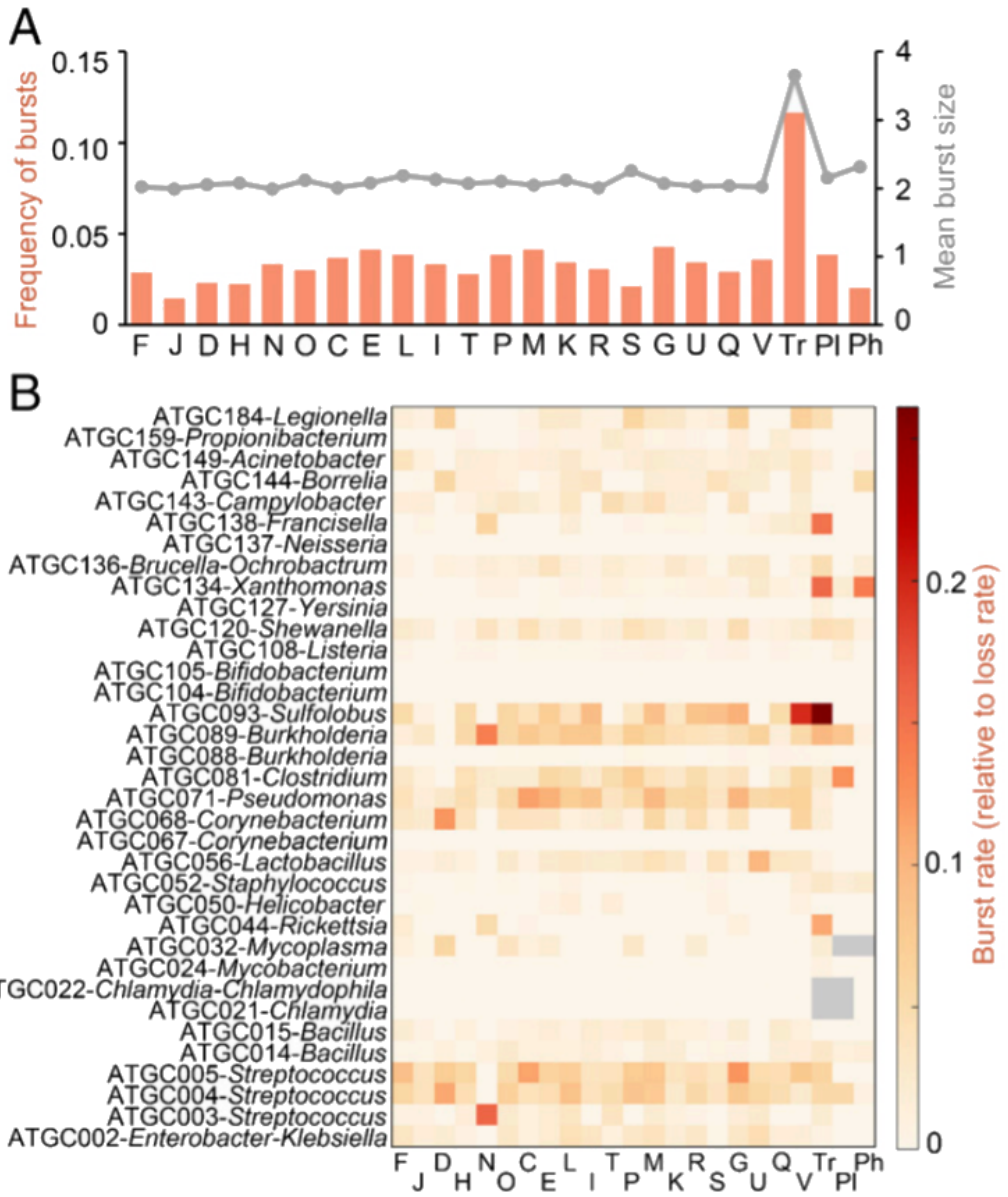
$$p_{k>0}(t) = 1 - p_0(t) = 1 - [1 - D(t)]^{h/d}$$





category J: translation (selection acts to keep just one copy)
category Tr: transposons (indicative of recent bursts)

bursts of activity



burst detection

- families with $d_e > 1$
- transposons have bursts of size ~ 4

modification of the model

- occasional bursts at a rate ϕ
- bursts reset copy number to K
- new mean copy number:

$$[\![k]\!] = \int_0^\infty \left[\left(\frac{h}{d} + K \right) \frac{D(t)}{1 - D(t)} + K \frac{d - \alpha D(t)}{d + (1 - \alpha) D(t)} \right] \phi e^{-\phi t} dt$$

bayesian estimate of parameters

ϕ = 0.04 (one burst every 25 losses)

K = 4.2 (experimental: K = 3.9)

conclusions

- mathematical modeling + phylogenetic analysis allows to quantify selection in different genes
- abundance distribution does not distinguish neutral and non-neutral evolution
- the time-dependent solution allows to disentangle different effects
- genes of key informational or metabolic pathways are subject to positive selection
- transposons and especially prophages are deleterious
- transposons experience intermitent bursts
- anti-parasite defenses are as costly as some genetic parasites

nonlinear selection

selection rate: $\frac{s\,k}{1+k/\Omega} \approx s\,k\left(1-\frac{k}{\Omega}\right)$ $\qquad \Omega \gg k$

$$\frac{d\,m_k}{dt} = -\left(h + k\,\alpha + \epsilon\,k^2\right)m_k + \left(k+1\right)m_{k+1} + \left[\left(k-1\right)d + h\right]m_{k-1} \qquad k \in \mathbb{Z} \qquad \epsilon = \frac{s}{\Omega}$$

$$\frac{\partial G}{\partial t} = \left(d\,z^2 - \alpha\,z + 1\right)\frac{\partial G}{\partial z} + h\left(z-1\right)G - \epsilon\,z\,\frac{\partial}{\partial z}\left(z\,\frac{\partial G}{\partial z}\right)$$