The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2584-15

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

A Genome as a Toolbox: Cross-genome “laws” for families

Marco Cosentino Lagomarsino
Université Pierre et Marie Curie
Paris

A Genome as a Toolbox: Cross-genome “laws” for families

June 5th 2014

Spring School, Trieste

Marco Cosentino Lagomarsino

Génophysique / Genomic Physics Group

CNRS “Microorganism Genomics” UMR7238 Laboratory
Université Pierre et Marie Curie, Paris

0) Where we left yesterday ...

Data Structure – Many Species

	FUNCTION 1					FUNCTION C
	★	★	★	■		⬟
	family 1	family 2	family 3	family 4	...	family F
genome 1	5	0	2	21		5
genome 2	7	0	3	32		7
genome 3	12	2	2	23		2
...	...	...	...	...	...	...
genome G	2	4	2	24		3

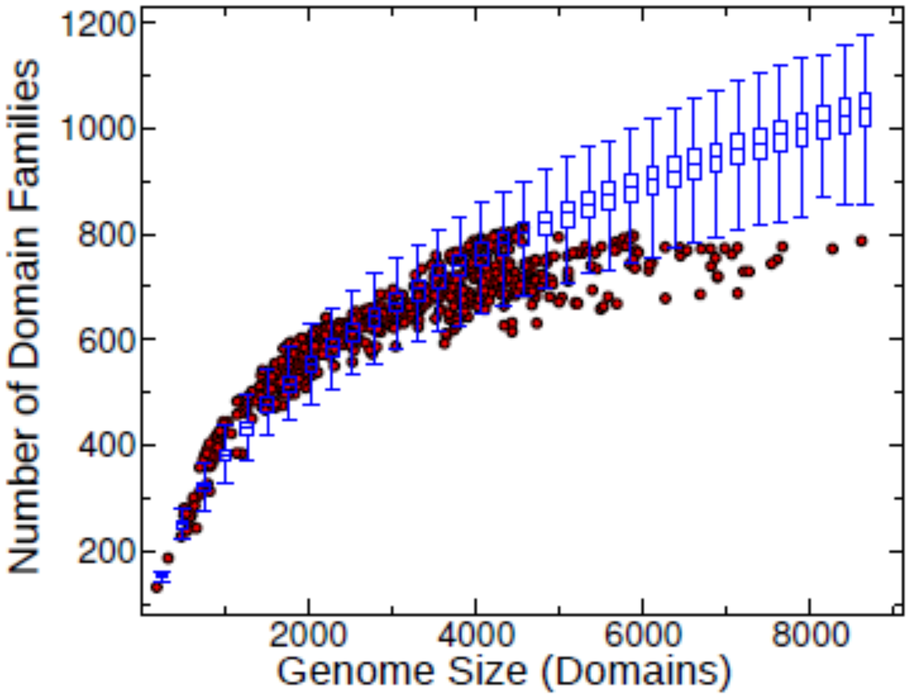
row sum

= genome "size"

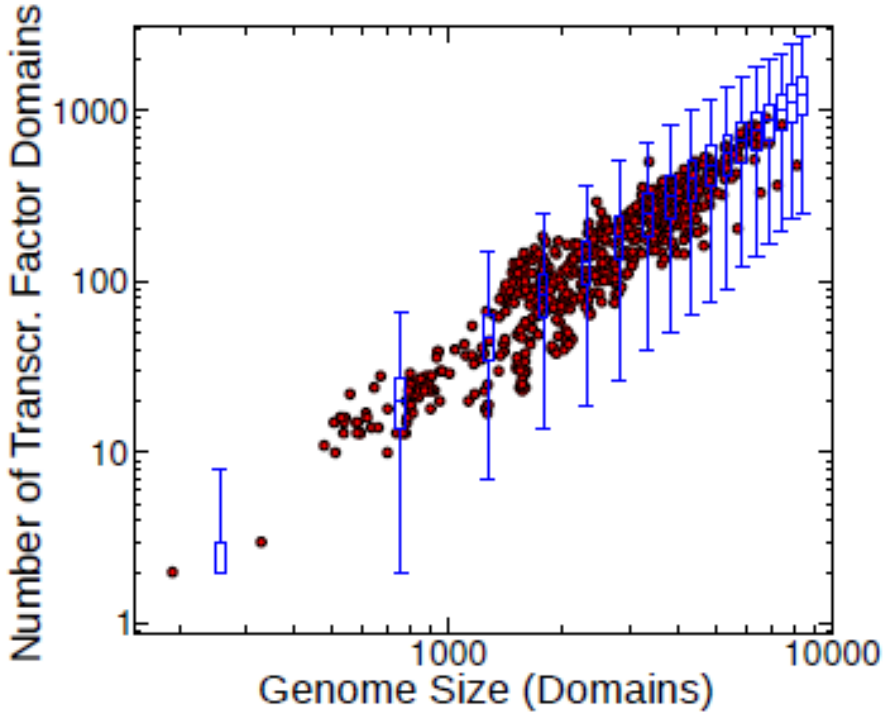
(related by phylogeny)

column sum = total family abundance

CRP with correlated duplication agrees well with empirical data (both variants)



Evolutionary Classes



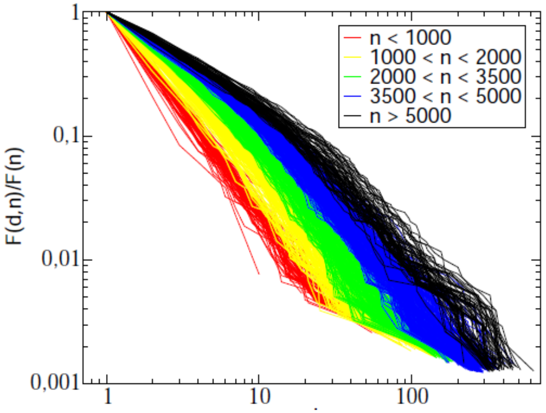
Functional Categories

Power-law with exponent

ζ ~ 1.6 for transcription factors

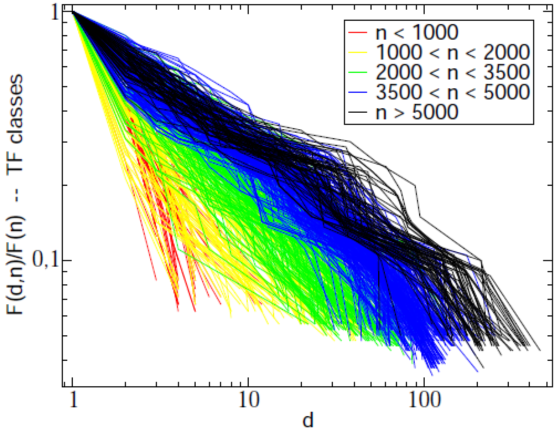
[1.6 explained as finite-size effect]

Empirical data follow the predicted trend



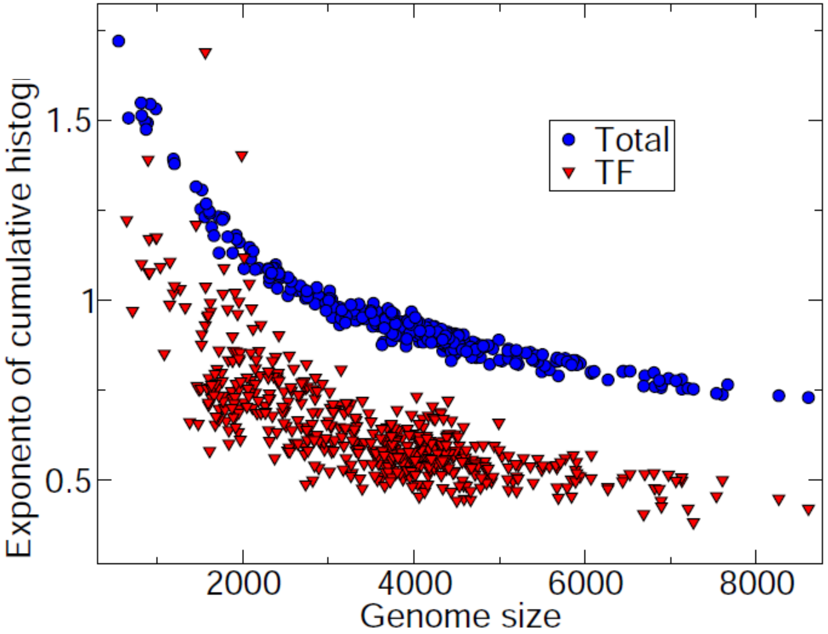
Total

TFs



$$P(d) \sim \left(\frac{1}{d}\right)^{1+\alpha}$$

$$P(d)_{TF} \sim \left(\frac{1}{d}\right)^{1+\frac{\alpha}{2}}$$

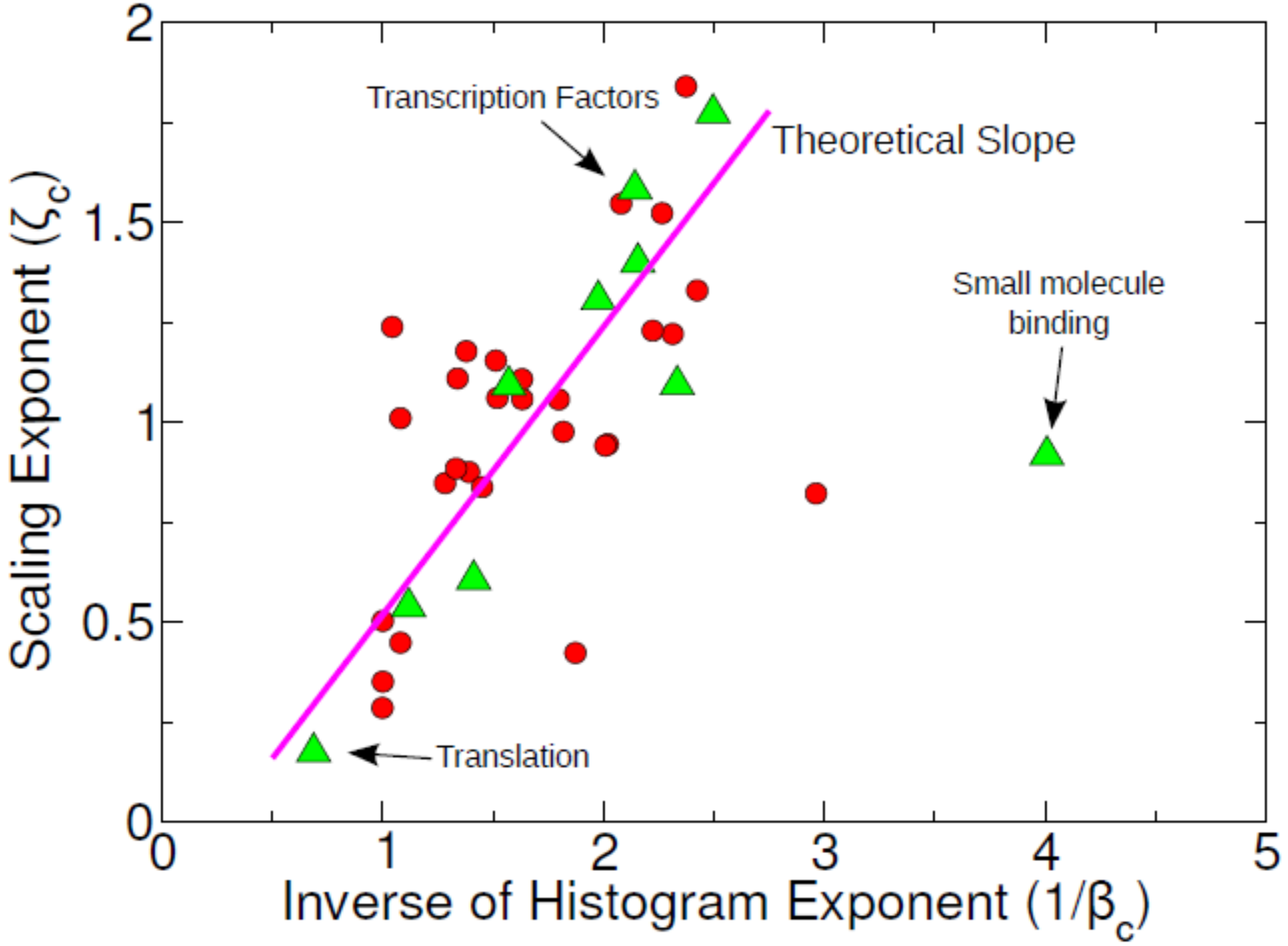


Total

TFs

Empirical data follow the predicted trend

Valid for many categories $\beta_c = \alpha/\zeta_c$



1) Cross-genome statistics:

gene-frequency distribution, the U

Underestimated Problem: observations may depend on resolution

1) At the level of philogeny

kingdoms

species

clades

strains

2) At the level of homology

Proteins

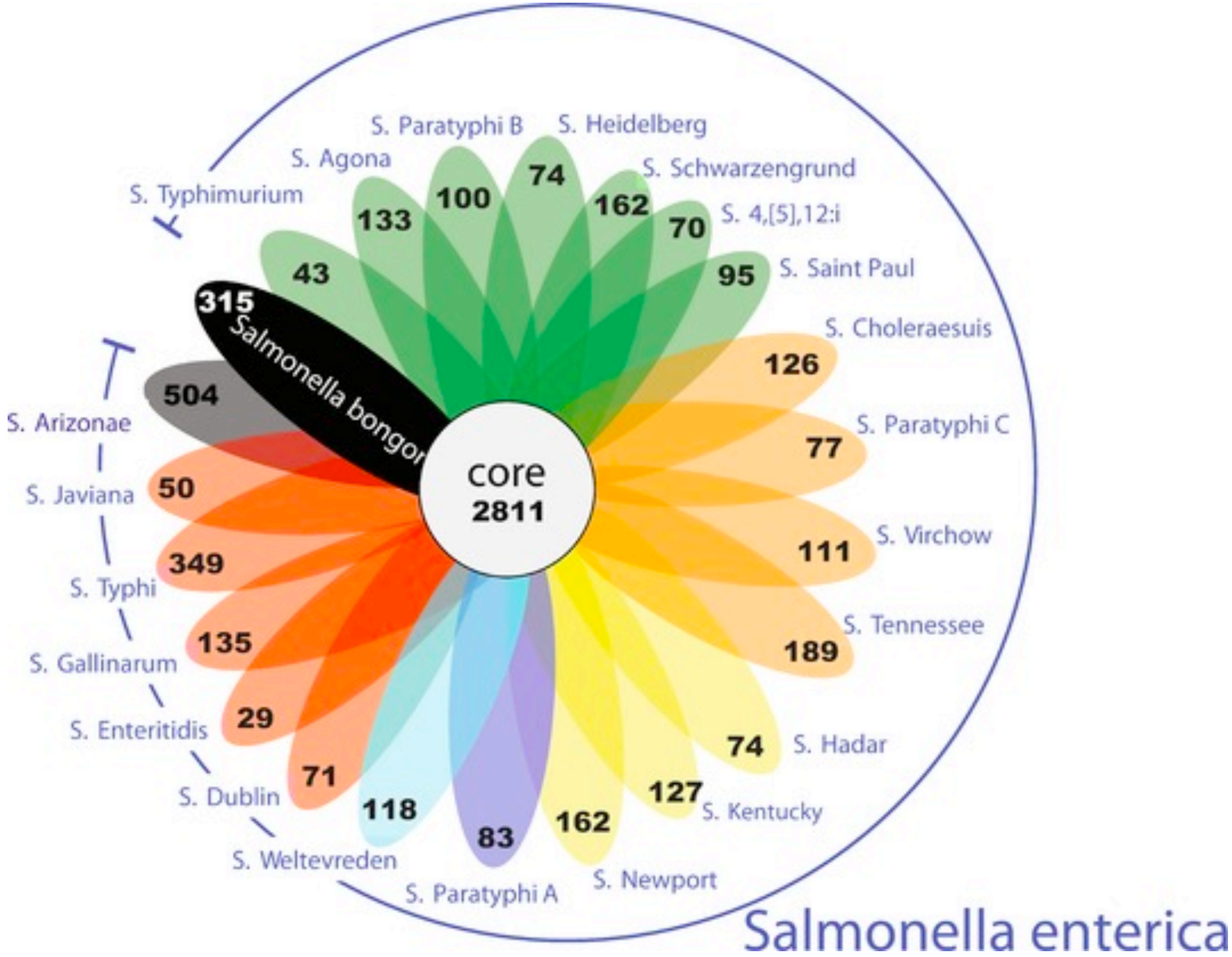
Domains

Homology criteria and thesholds

Taxonomy level

Available observations

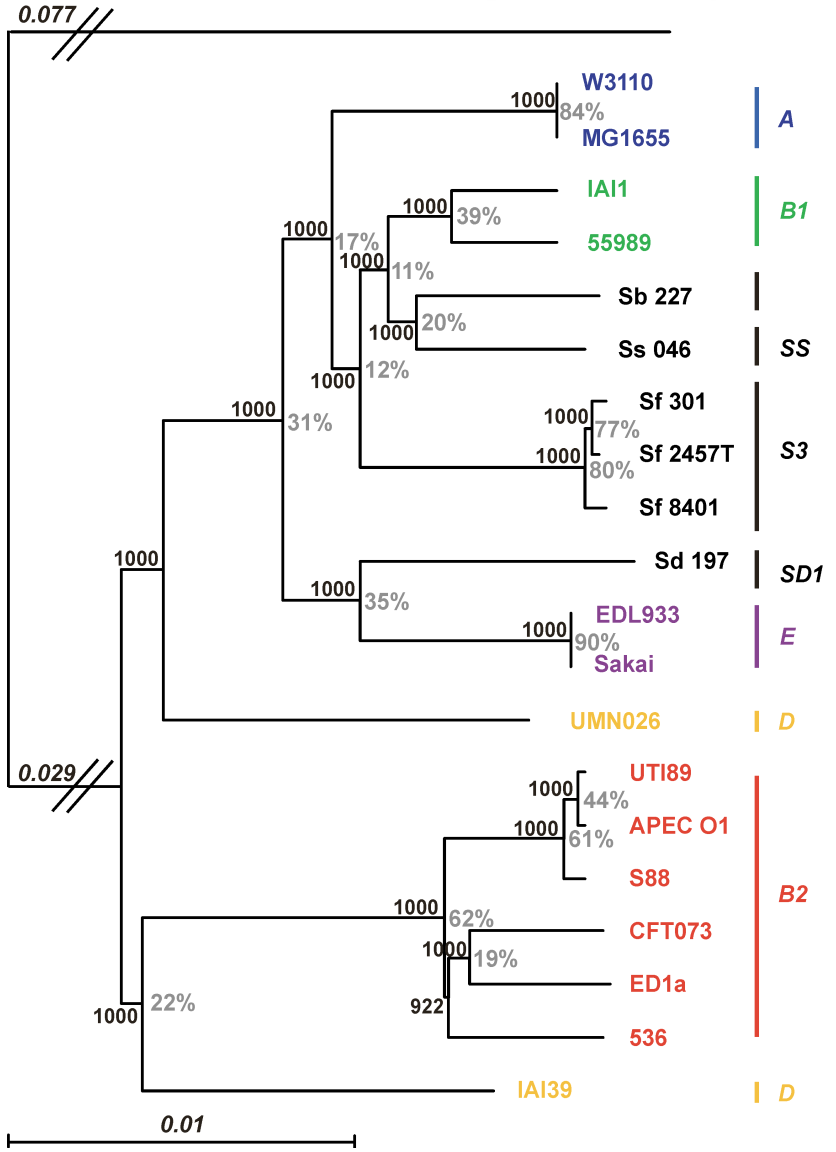
Core vs Pan genome (strain level)



Available observations

Core vs Pan genome (strain level)

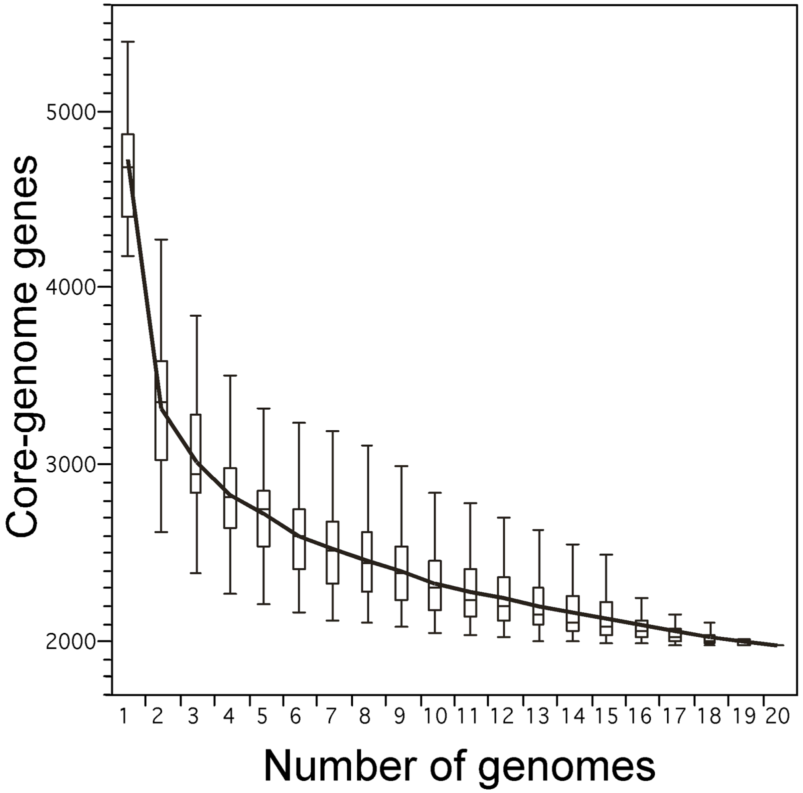
E. coli (Touchon et al, PLOS genet 2009)

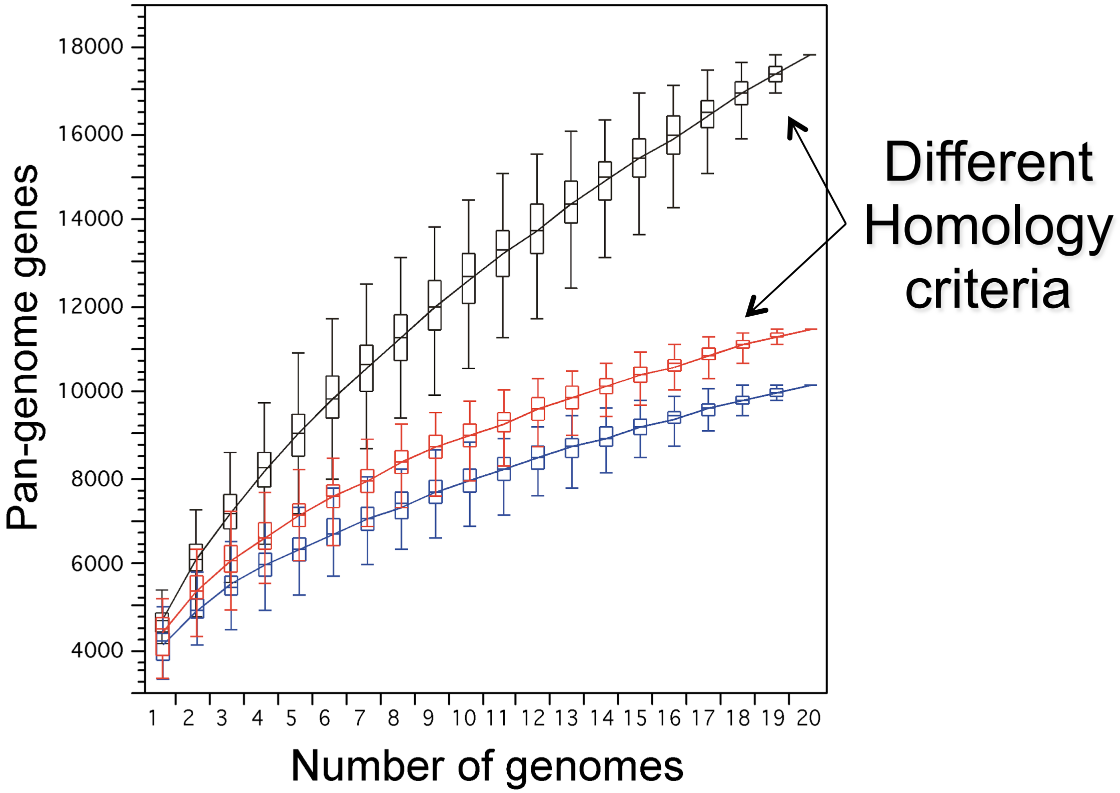


Available observations

Core vs Pan genome (strain level / gene based)

E. coli (Touchon et al, PLOS genet 2009)



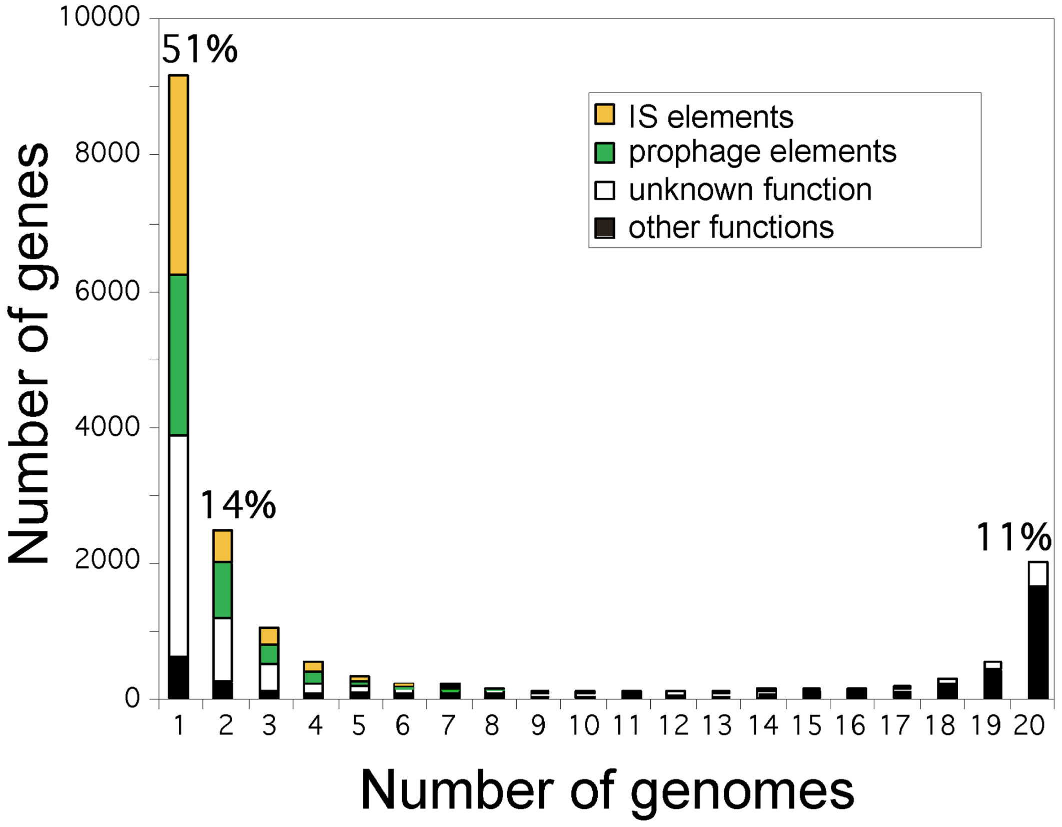


Different Homology criteria

Available observations

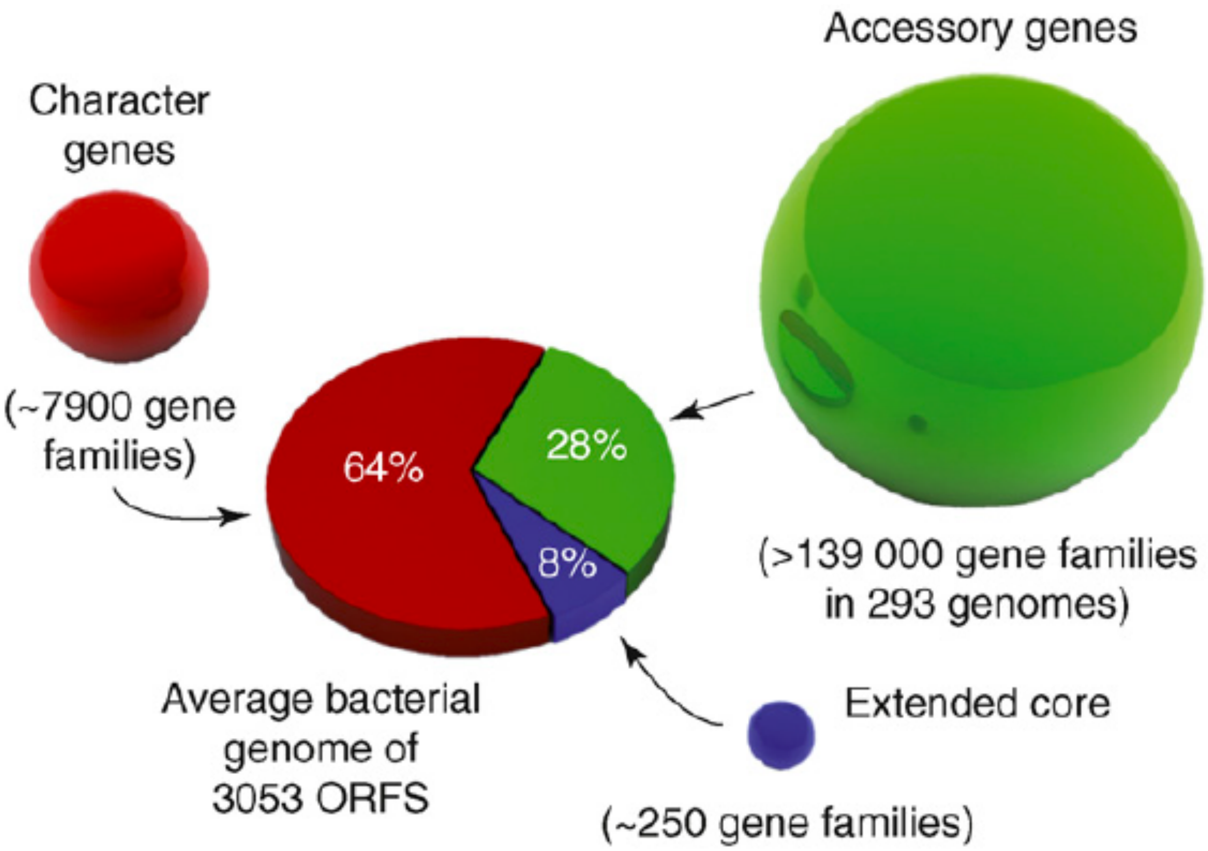
“Gene-frequency distribution”

E. coli (Touchon et al, PLOS genet 2009)



Available observations

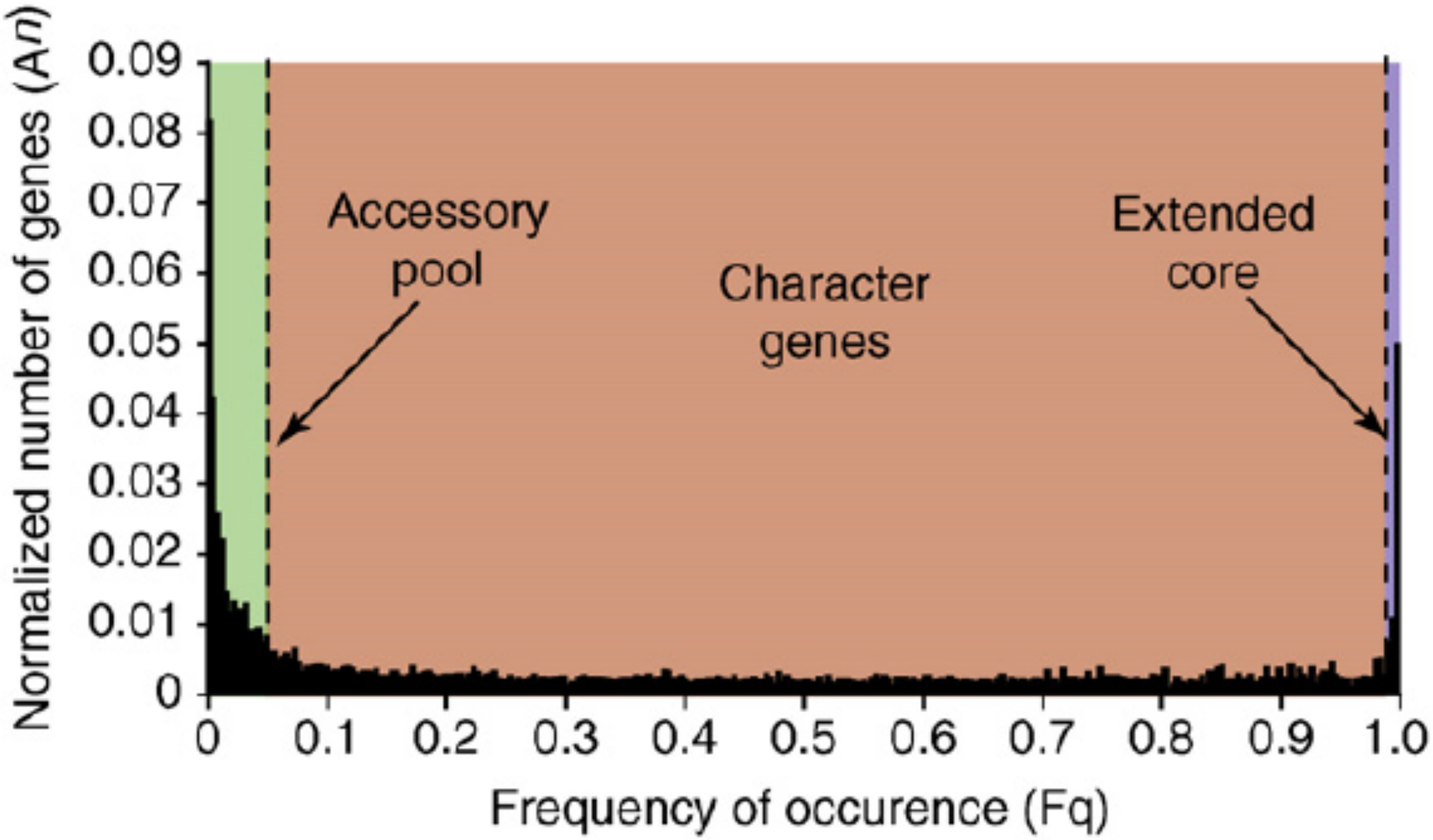
Species/gene level



~500 bacterial species (Lapierre and Gogarten TIG 2009)

Available observations

Species/gene level

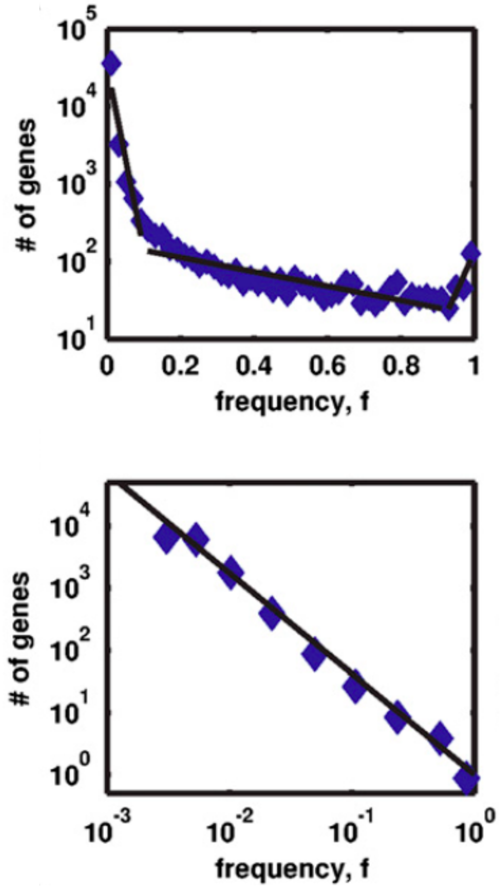


~500 bacterial species (Lapierre and Gogarten TIG 2009)

There are multiple Us!

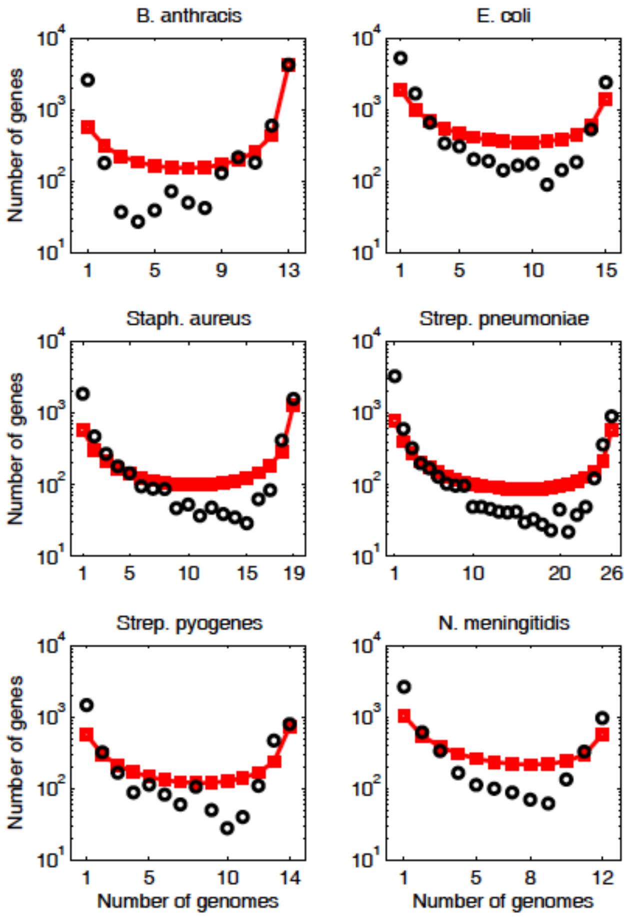
U-shaped occurrence profile at different resolutions (“gene-frequency distribution”)

species



(Pang and Maslov PNAS 2013)

strains



(Haegeman and Weitz BMC Genomics 2012)

There are multiple Us!

Model for strains:
neutral population dynamics with HGT

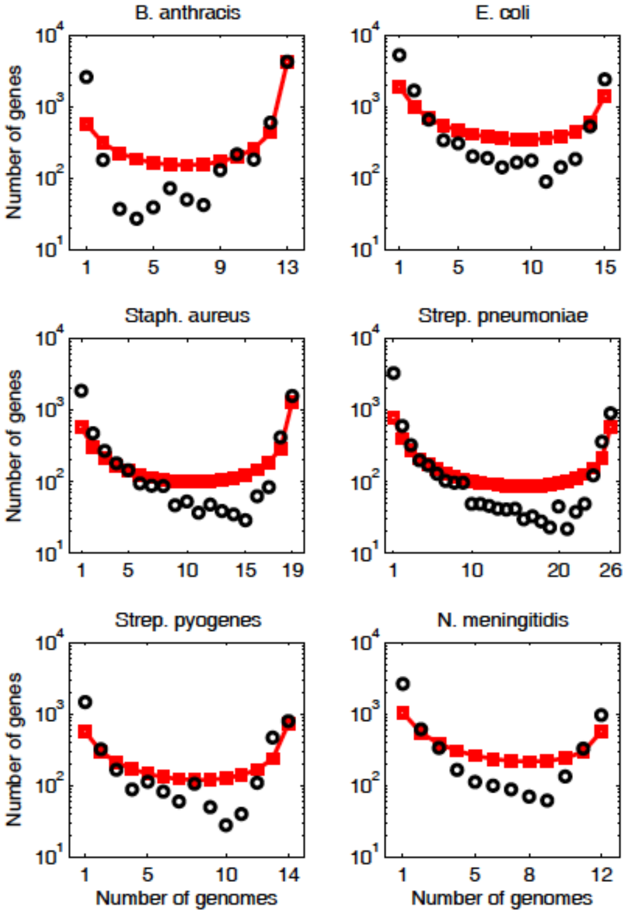
N individuals =
Gene presence/absence Boolean vectors of length M

Moran model [genetic drift]
(Polya Urn with constant population / each addition accompanied by random removal)

+

"Horizontal transfer" = innovation

strains



(Haegeman and Weitz BMC Genomics 2012)

neutral population dynamics with HGT

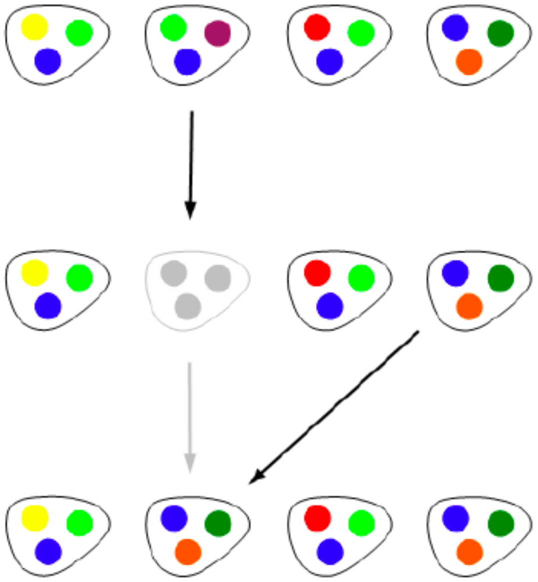
N individuals (population size) =
Gene presence/absence Boolean vectors of length *M* (genome size)

Moran model [genetic drift]
(Polya Urn with constant population / each addition accompanied by random removal)

+

"Horizontal transfer" = innovation

Moran



(Haegeman and Weitz BMC Genomics 2012)

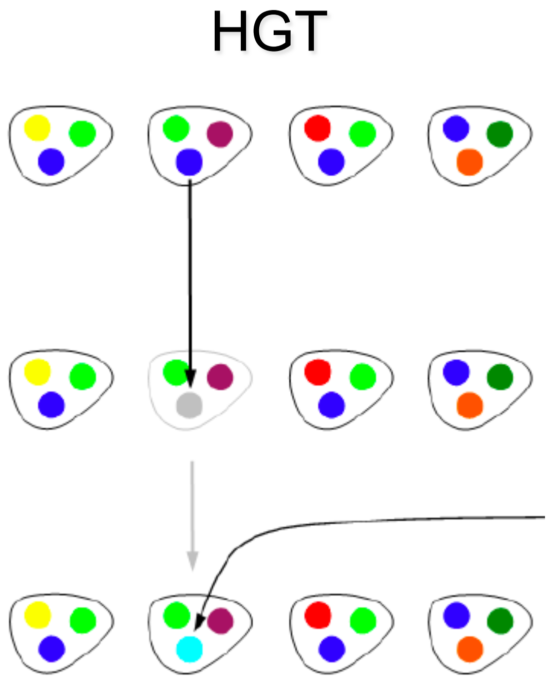
neutral population dynamics with HGT

N individuals (population size) =
Gene presence/absence Boolean vectors of length *M* (genome size)

Moran model [genetic drift]
(Polya Urn with constant population / each addition accompanied by random removal)

+

"Horizontal transfer" = innovation



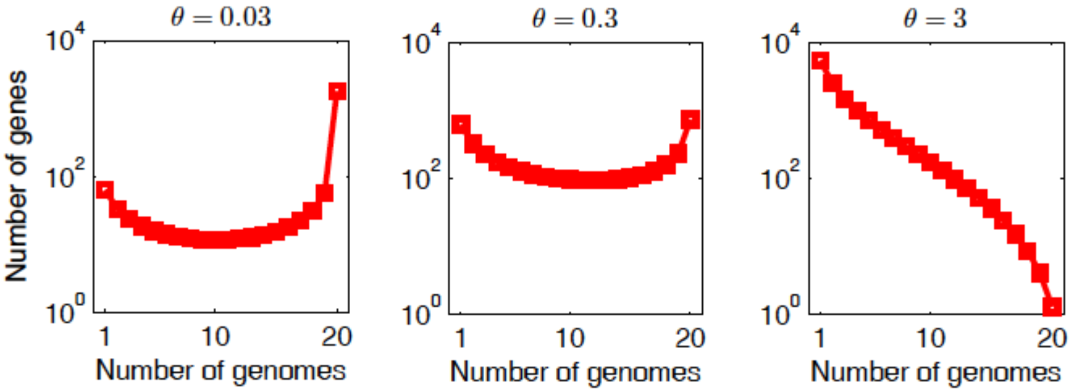
(Haegeman and Weitz BMC Genomics 2012)

neutral population dynamics with HGT

Parameters: *N, M*
reproduction rate *r*
HGT rate *s*

⇒ Combine in $\theta = Ns/Mr$

⇒ *if* θ<1 U-shape



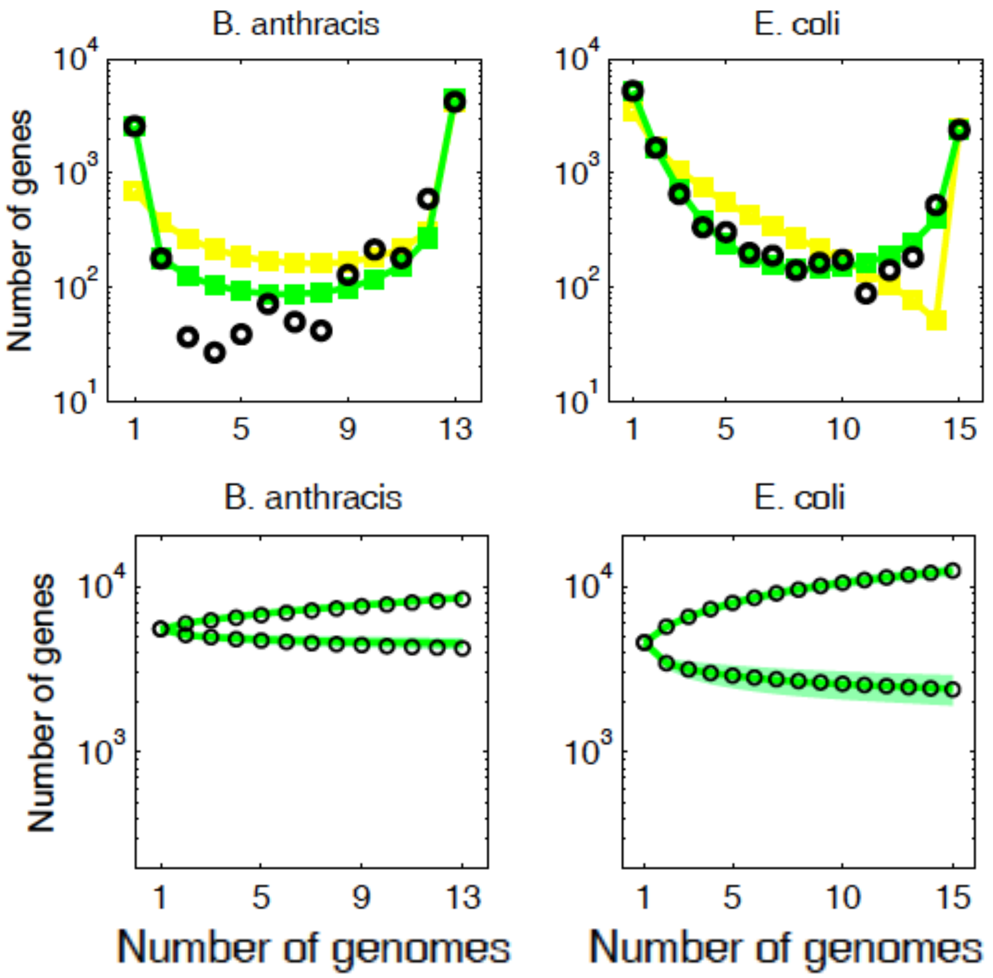
(Haegeman and Weitz BMC Genomics 2012)

neutral population dynamics with HGT

Fit θ (effective HGT rate)
for different clades

Fit pan-genome scaling
(equivalent)

Criticism:
Others with similar neutral models
Claim that thay can be rejected
evidence for selection?
(Collins&Higgs, Koonin, Baumdiecker)

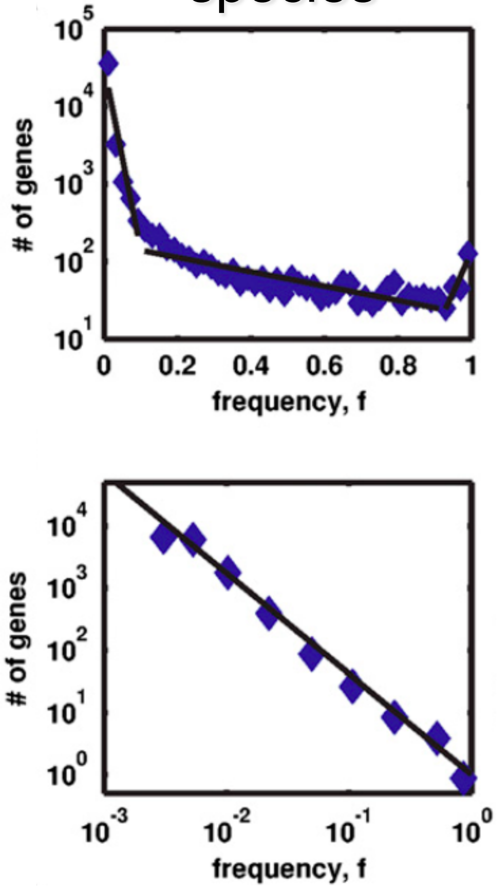


(Haegeman and Weitz BMC Genomics 2012)

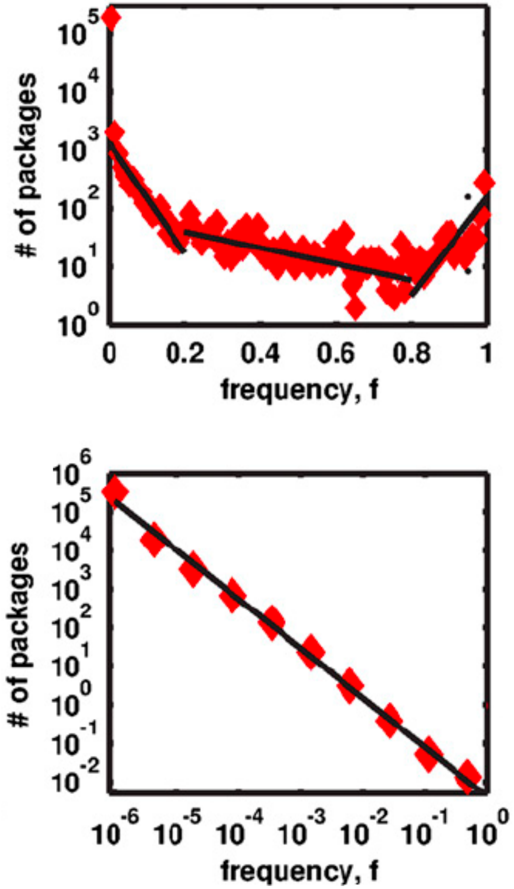
There are multiple Us!

Model for species: dependency networks

Bacterial species



Linux packages (!)



In both cases the Left side of the U Looks like a power law!

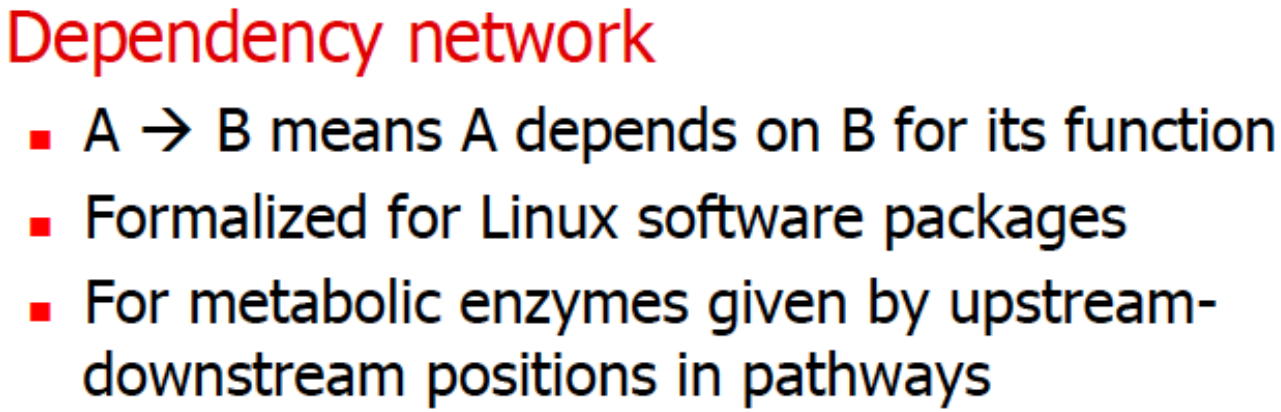
Same exponent (1.5)

(Pang and Maslov PNAS 2013)

Idea:
Occurrence = Importance

= component needed
for proper functioning of other components

= High rank in *dependency network*



Dependency network

- A → B means A depends on B for its function
- Formalized for Linux software packages
- For metabolic enzymes given by upstream-downstream positions in pathways

(Pang and Maslov PNAS 2013)

Argument

The dependency network is *feedforward*
D = mean out-degree

Poisson graph growth model
t = size of network when a package was added

A package at time *t'>t* sends link
to package added at time with
probability *t* = *D/t'*

It inherits (indirectly) its dependencies

(Pang and Maslov PNAS 2013)

Argument

Importance ~ K_{dep} (#indirect dependencies)

$$K_{dep}(t) = 1 + \int_{t+1}^{N} K_{dep}(t')D/t'$$

Implies:

$$K_{dep}(t) = (t/N)^{-D}$$

(Pang and Maslov PNAS 2013)

Argument

$$P(K_{dep} > k) = P((t/N)^{-D} > k)$$

$$= P(t < NK^{-1/D}) = \frac{NK^{-1/D}}{N}$$

Hence

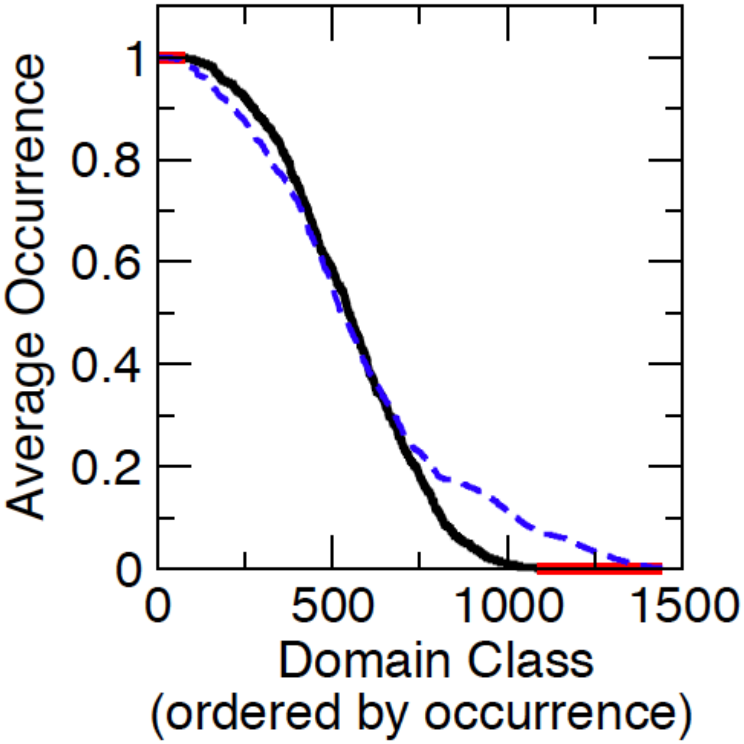
$$\boxed{P(K_{dep}) \propto \frac{1}{k^{1+\frac{1}{D}}}}$$

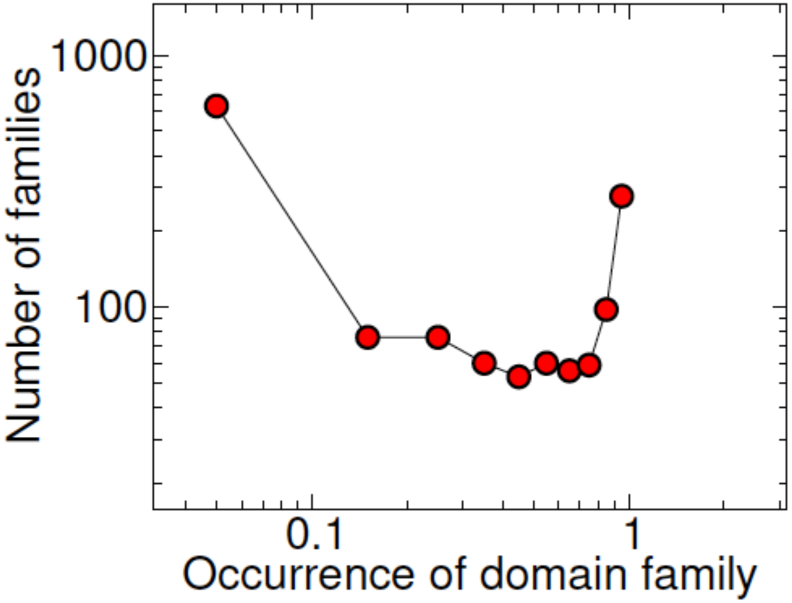
Degree can be measured:

$D_{met} = 1.7$; $D_{linux} = 2.4$

(Pang and Maslov PNAS 2013)

Side note: species/domain family level occurrence pattern is more like a U (much fewer families)





2) Cross-genome statistics: abundance fluctuations and HGT

Data Structure – Many Species

	FUNCTION 1					FUNCTION C
	★	★	★	■		⬟
	family 1	family 2	family 3	family 4	...	family F
genome 1	5	0	2	21		5
genome 2	7	0	3	32		7
genome 3	12	2	2	23		2
...	...	...	...	...	...	...
genome G	2	4	2	24		3

row sum

= genome “size”

(related by phylogeny)

column sum = total family abundance

“Moves” of gene-family dynamics

Copy-Paste

Intra species HGT + Duplication

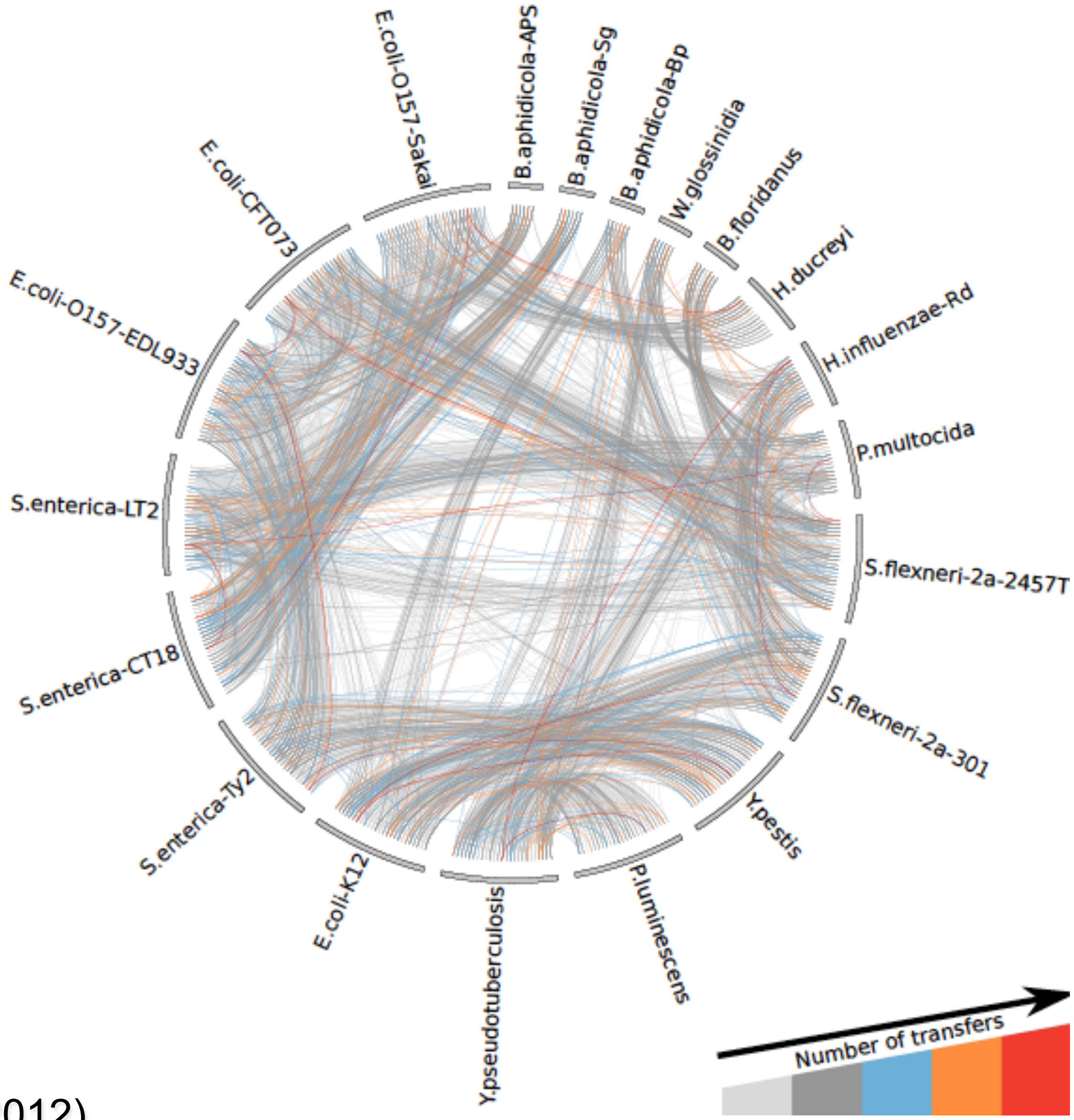
Share

Inter-species HGT

Trash

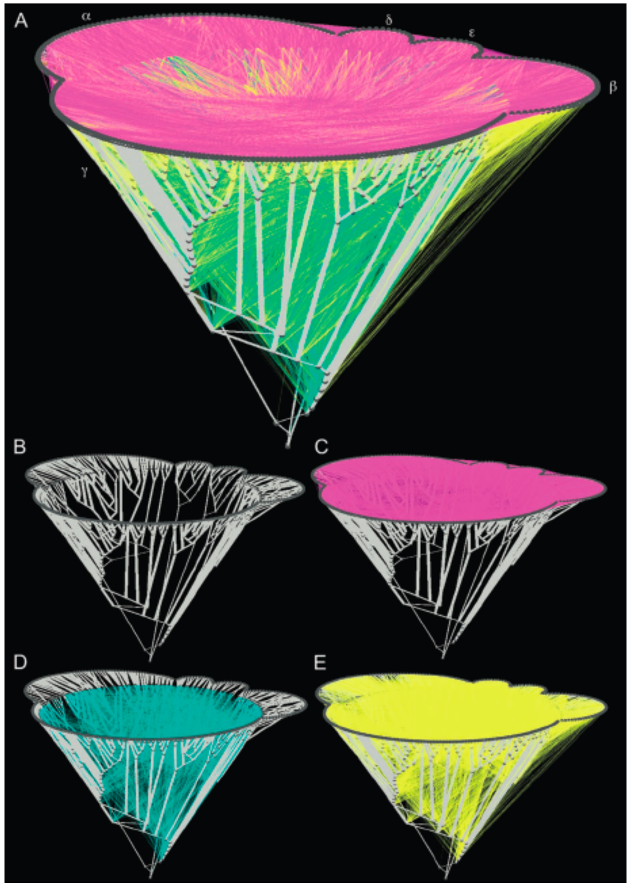
Loss

Horizontal transfer of genes is a dominant force of bacterial gene-family evolution



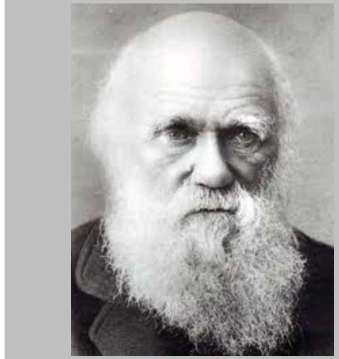
(Grassi et al MGE 2012)

A tree or a network, or both?



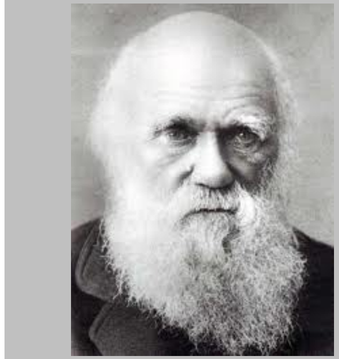
A null “collisional” model

(bearded scientists)



…

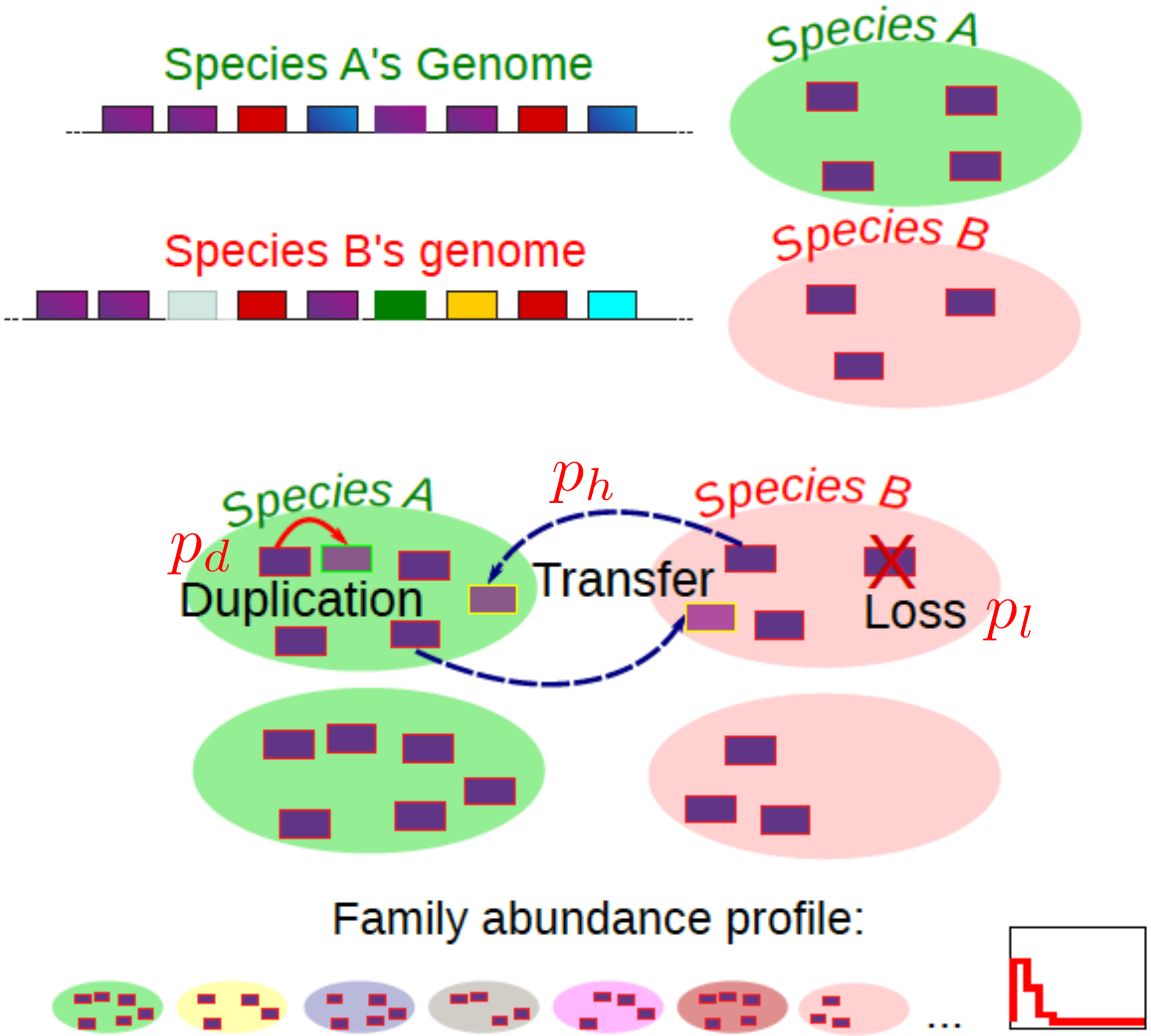
A null “collisional” model



...

Boltzmann-like
“collisional” model
Between *species*
(no population)

Model Ingredients



Model Ingredients

$$\begin{cases} V_j(\tau+1) = V_j(\tau) + DL[V_j(\tau)] + H[V_i(\tau)] \\ V_i(\tau+1) = V_i(\tau) + DL[V_i(\tau)] + H[V_j(\tau)] \end{cases}$$

p_h

Species *i* samples species *j* for horizontal transfer, and itself for “duplication”/loss

p_d p_l

Assumptions:

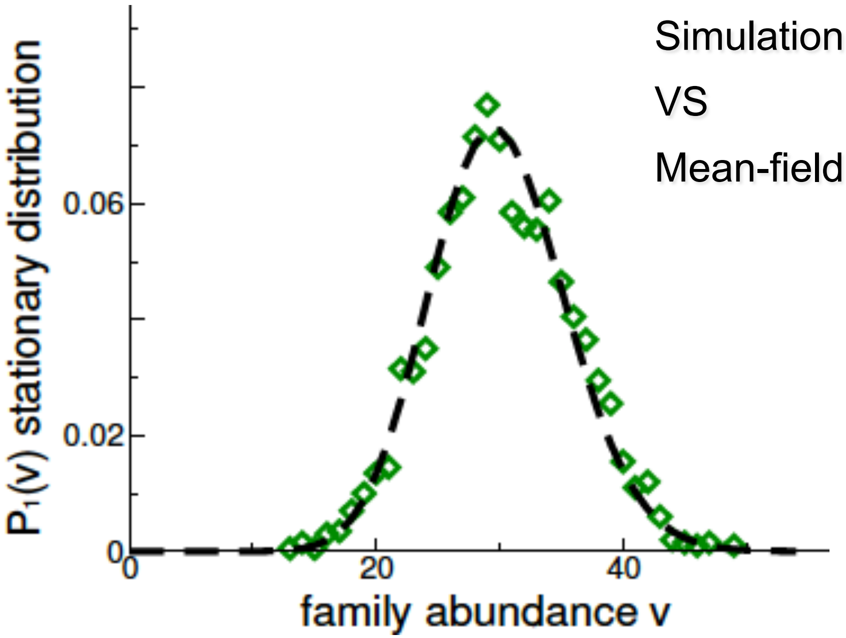
(I) Independence of families

(II) *Mean* abundance conserved

$p_h + p_d = p_l$
$$\left\langle \sum_{i=1}^{N} V_i(\tau) \right\rangle = \left\langle \sum_{i=1}^{N} V_i(0) \right\rangle$$

(III) What matters is steady state

Model Predictions



Simulation

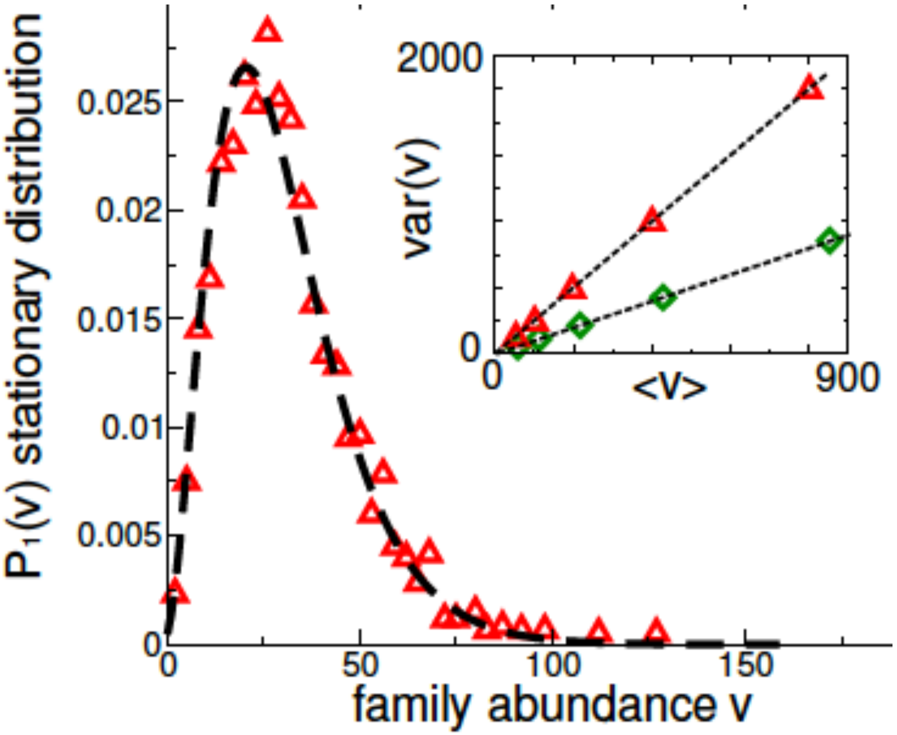
VS

Mean-field calculation

HGT / loss → Poisson abundance profile

$p_d = 0$

Model Predictions



Simulation

VS

Mean-field calculation

$p_d > 0$

HGT + *duplication* / loss

→ increasingly dispersed abundance profile

The model is tractable analytically

Mean-field theory:

$$\frac{\partial f(v,t)}{\partial t} = \mathrm{Prob}\big(V_1 + DL[V_1] + H[V_2] = v\big) - f(t,v)$$

$$V_1, V_2 \sim f$$

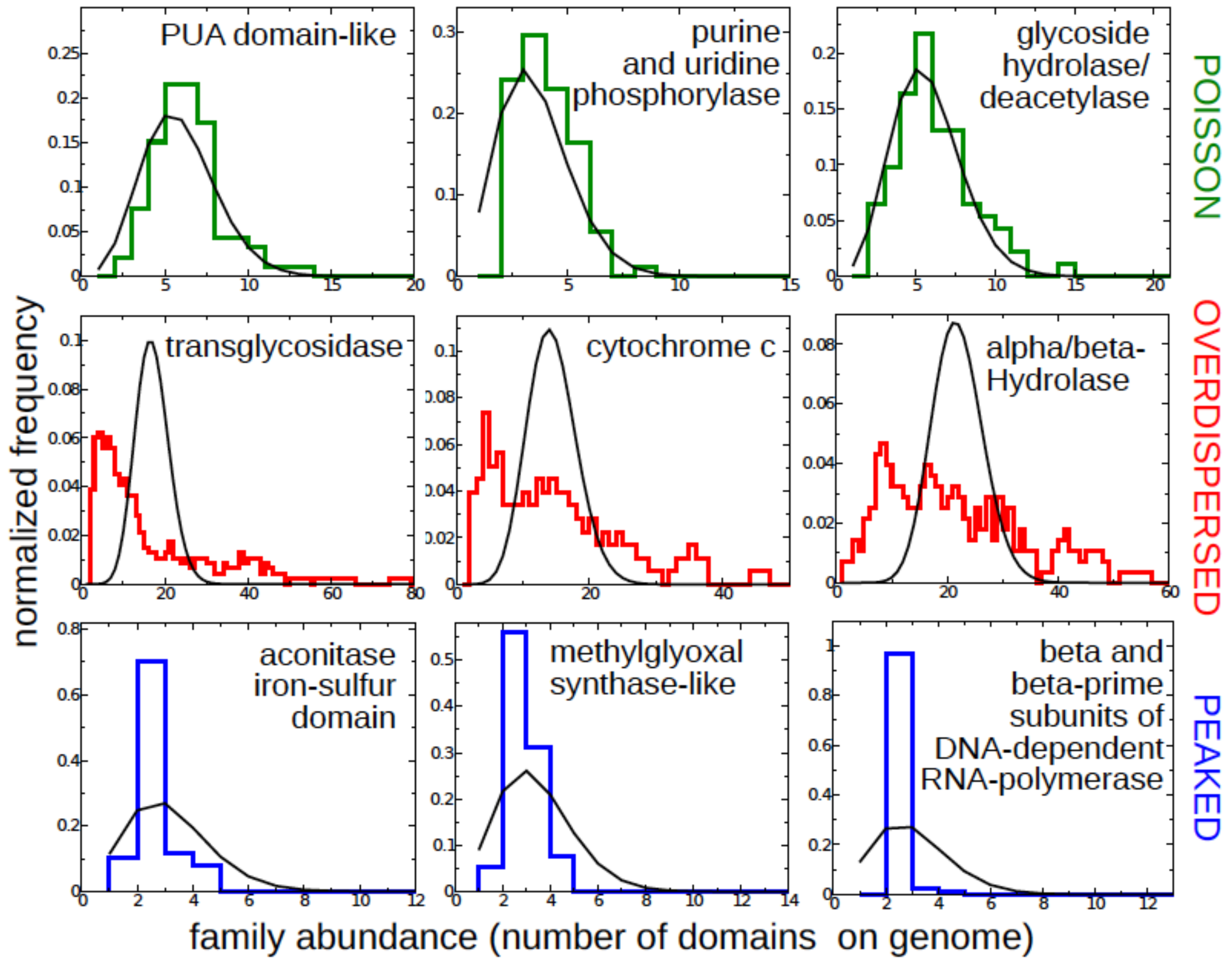
Equations for moments using generating function

Self-consistent argument for $p_d = 0$
leading to Poisson

For $p_d > 0$ approximate solution
(negative binomial)

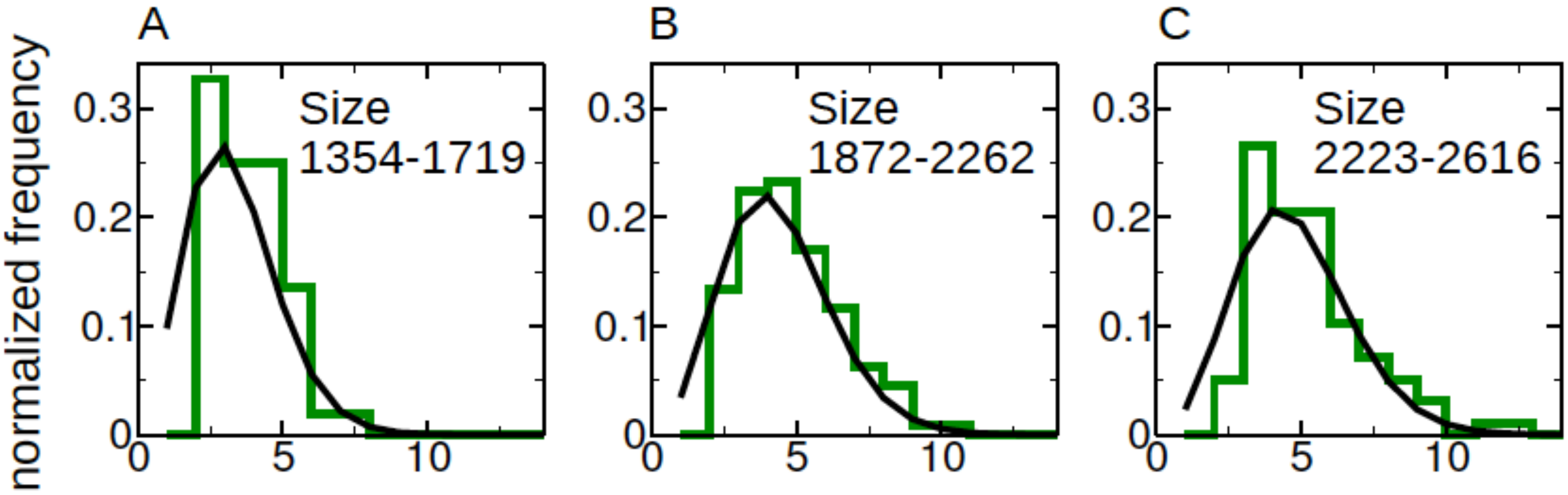
Empirical data on abundance fluctuations (domain families)

Empirical family abundance profiles



(binned by genome size in domains)

Family abundance profiles are robust for different ranges of genome size



55424: FAD/NAD-linked reductases, dimerisation (C-terminal) domain

Order Parameters

"*Qf*"

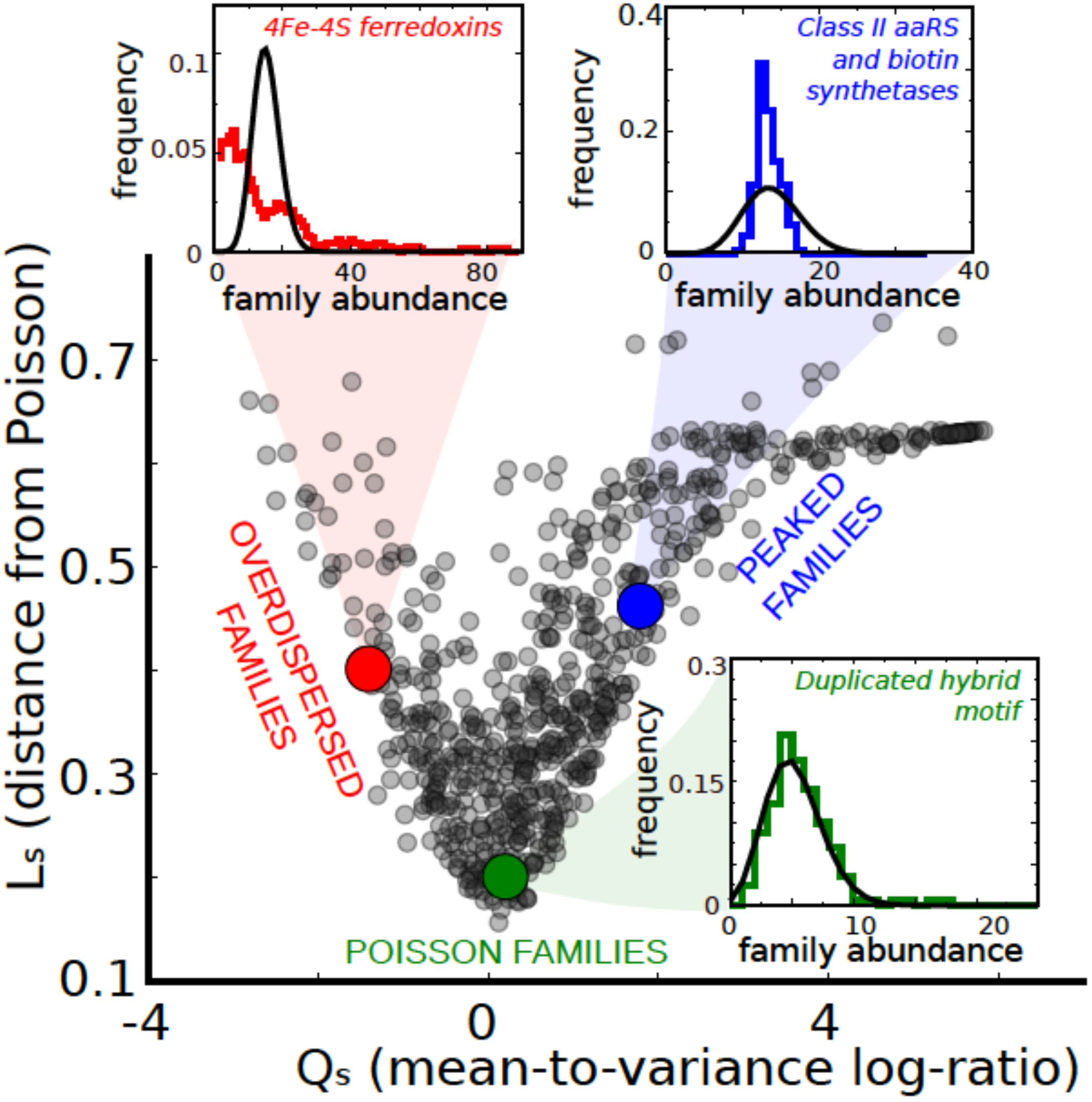
Average mean-to-variance log ratio of the family abundance histograms across bins of genome size

"*Lf*"

Average L1 distance with Poisson distribution

(both weighted on sampling)

Classification of families by abundance profiles



Abundance profiles and functions

Enrichment Tests:

Peaked abundance profile families

Are enriched for translation & RNA processing

Poisson abundance profile families

Are enriched for metabolism

Overdispersed abundance profile families

Are enriched for DNA-binding (TF) & signal transduction

Horizontal transfer candidate data

HGT-DB

(S. Garcia-Vallve et al NAR 2003)

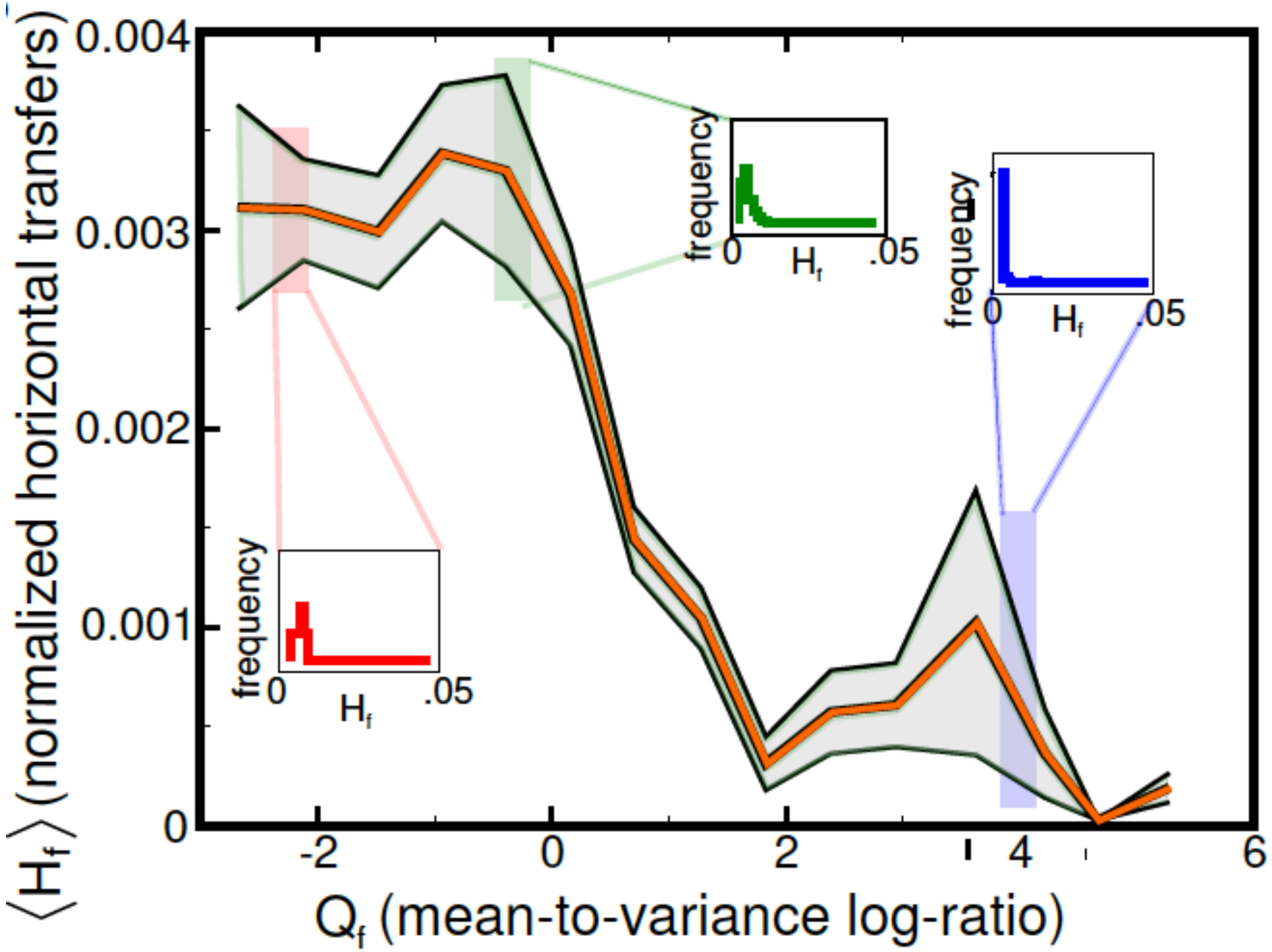
DarkHorse HGT Candidate Resource

(Podell et al Genome Biol 2007)

And other data (Treangen & Rocha, Abby et al, ...)

Abundance profiles and horizontal transfers

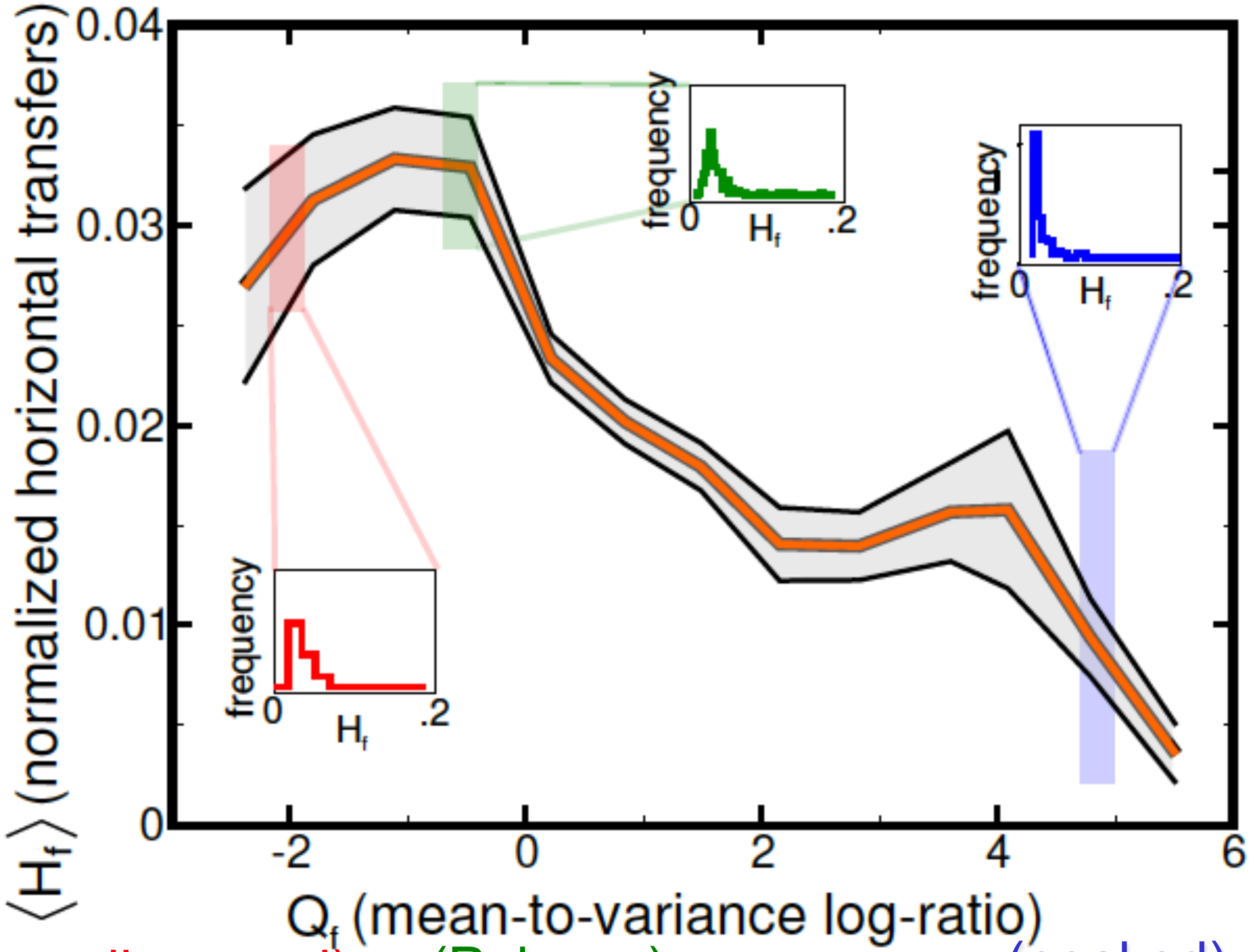
DarkHorse HGT Candidate Resource



(overdispersed) ← (Poisson) → (peaked)

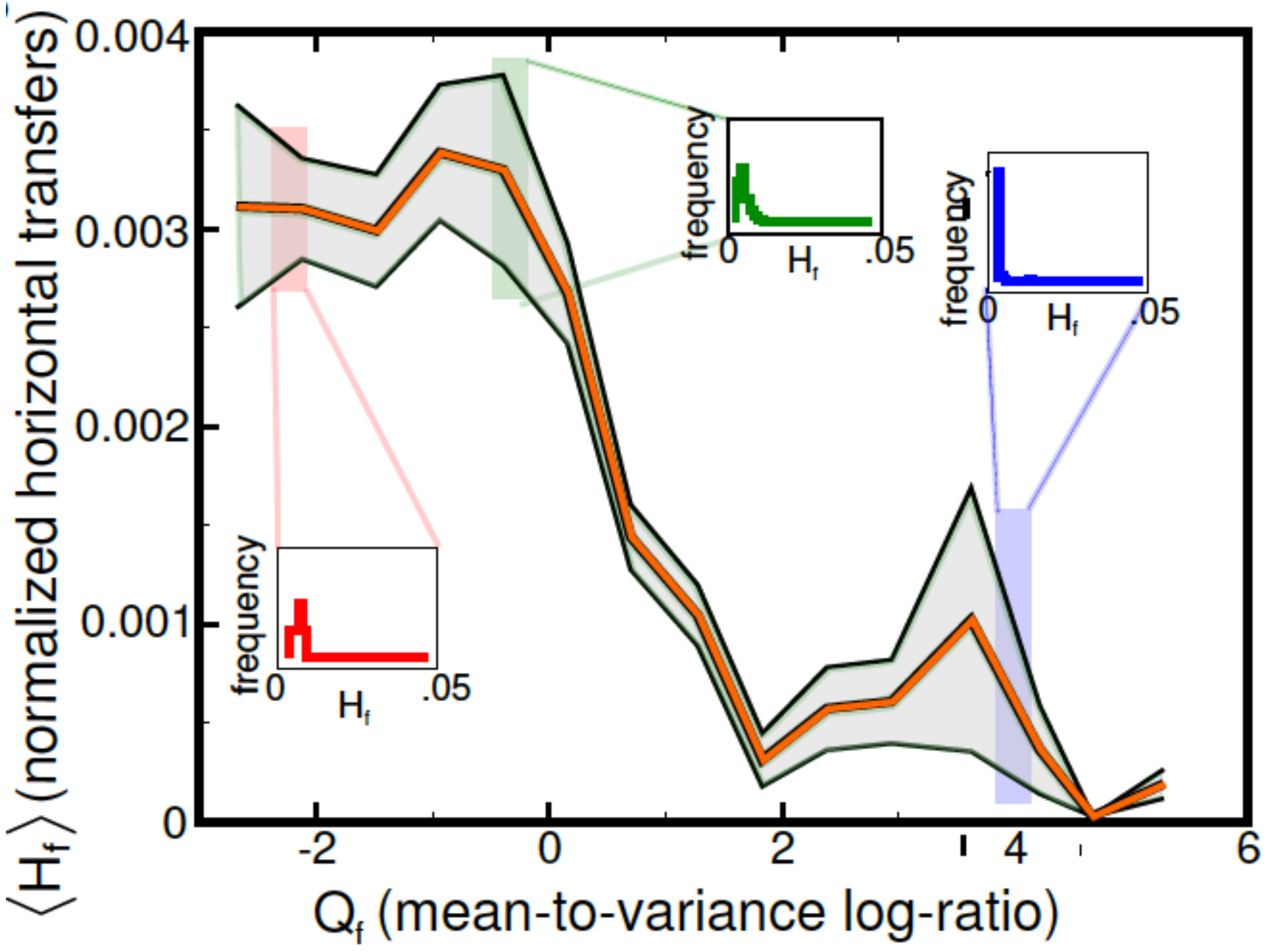
Abundance profiles and horizontal transfers

HGT-DB



(overdispersed) (Poisson) (peaked)

Abundance profiles and horizontal transfers



Fluidity ←

(“phylogenetic network”)

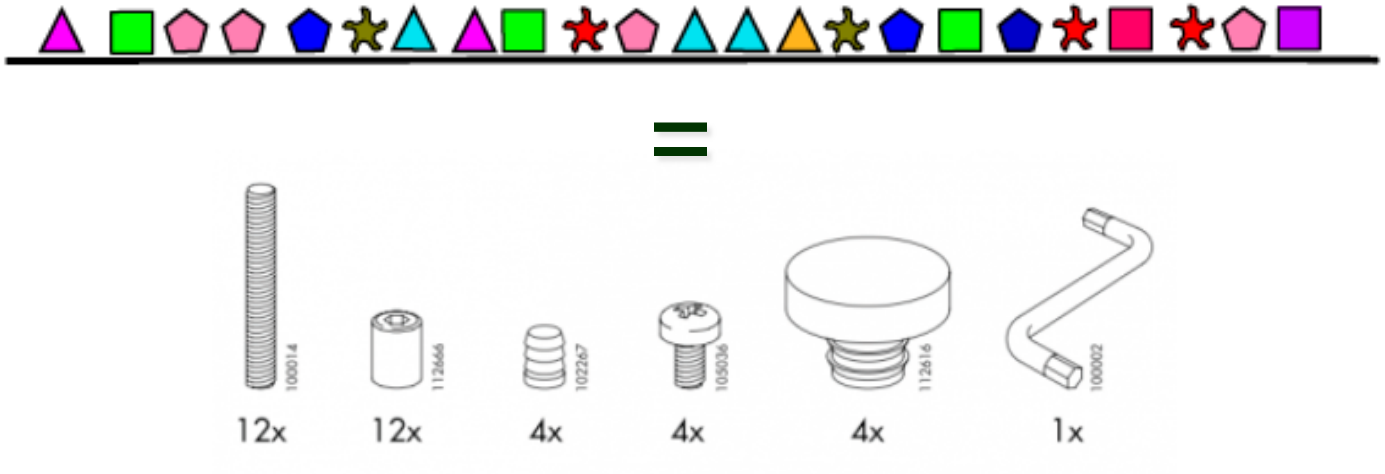
→ Stability

(“phylogenetic tree”)

Conclusions

- Population models for strain-level gene occurrence distribution
- Species-level gene occurrence distribution and dependency networks
- Heuristic value of “collisional” model
- There is a link between abundance fluctuations and HGT
- Differential genome fluidity for different functional classes of genes

Thank you!


=
100014
12x
112666
12x
102267
4x
105036
4x
112616
4x
100002
1x

Normal Boltzmann Eq.

$$\frac{\partial f}{\partial t} + \frac{\mathbf{p}}{m} \cdot \nabla f + \mathbf{F} \cdot \frac{\partial f}{\partial \mathbf{p}} = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}}$$

$$dN = f(\mathbf{r}, \mathbf{p}, t)\, d^3\mathbf{r}\, d^3\mathbf{p}$$

$$\left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} = \iint gI(g, \Omega)[f(\mathbf{p'}_A, t) f(\mathbf{p'}_B, t) - f(\mathbf{p}_A, t) f(\mathbf{p}_B, t)]\, d\Omega\, d^3\mathbf{p}_A.$$