Family Size Statistics

David Kessler

Bar-Ilan Univ.

Nadav Shnerb (BIU)
Yosi Maruvka (BIU -> Harvard)
Robert Ricklefs (Washington U)
Gur Yaari (Hebrew Univ. -> BIU)
Soren Solomon (Hebrew Univ)

Basic Motivation

In an exponentially growing population, "families" grow exponentially (on average) - nevertheless, some go extinct

Mutations generate new families

Size of family indicates when it arose

Can learn about dynamics of growth by studying family size statistics (Manrubia, Zanette, Derrida)

History goes back to Galton and Watson - extinction of noble families in England (1880's)

What defines a family?

Any inheritable characteristic subject to mutation

- Surnames (Sociology)
- Genome (Evolutionary Biology)
- Species (Ecology)

The Model

Discrete Generations

Each individual gives rise to some number of offspring and is then removed

Number drawn from some IID distribution, with mean $\lambda \equiv 1 + \gamma$

Offspring belong to same family as parent, except for mutations, which occur in a fraction μ of births

Mutations give rise to new family

Same as birth/mutation model of Yule, with number of children now random

- In particular, we have possibility of 0 children; i.e. death

Fokker-Planck Equation

Evolution equations:

$$n_m^{t+1} = \sum_{\substack{\ell \\ p \geq m}} n_\ell P(\ell \to p) \binom{p}{m} \mu^{p-m} (1-\mu)^m; \quad m > 1$$

$$n_1^{t+1} = \sum_{\ell} n_\ell P(\ell \to p) \binom{p}{1} \mu^{p-1} (1-\mu) + \mu N(t+1)$$

In limit of small growth, mutations, get Fokker-Planck eqn. for n_m:

$$\frac{\partial n}{\partial t} = -(\gamma - \mu) \frac{\partial}{\partial m}(mn) + \frac{\sigma^2}{2} \frac{\partial^2}{\partial m^2}(mn)$$

σ^2 is the variance of the offspring distribution

Smallness of $\gamma - \mu$ allows us to truncate after 2nd derivative

Family size distribution

$$\frac{\partial n}{\partial t} = -(\gamma - \mu)\frac{\partial}{\partial m}(mn) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial m^2}(mn)$$

Same equation as derived by Manrubbia and Zanette for a Moran process, with $\sigma^2 = 2$, appropriate for geometric distribution with mean close to unity.

Solution: $$n_m = \frac{\nu R_c}{m}\Gamma\left(2+\nu\right) U\left(1+\nu, 0, R_c\frac{m}{N_o}\right)$$

With: $\nu \equiv \frac{\mu}{\gamma-\mu}$, $R_c \equiv \frac{2N_o\gamma}{\sigma^2(1+\nu)}$

Normalization fixed by total population: $\sum_m m n_m = N_o$

Power-law tail with exponent 2+ν (as in Yule)

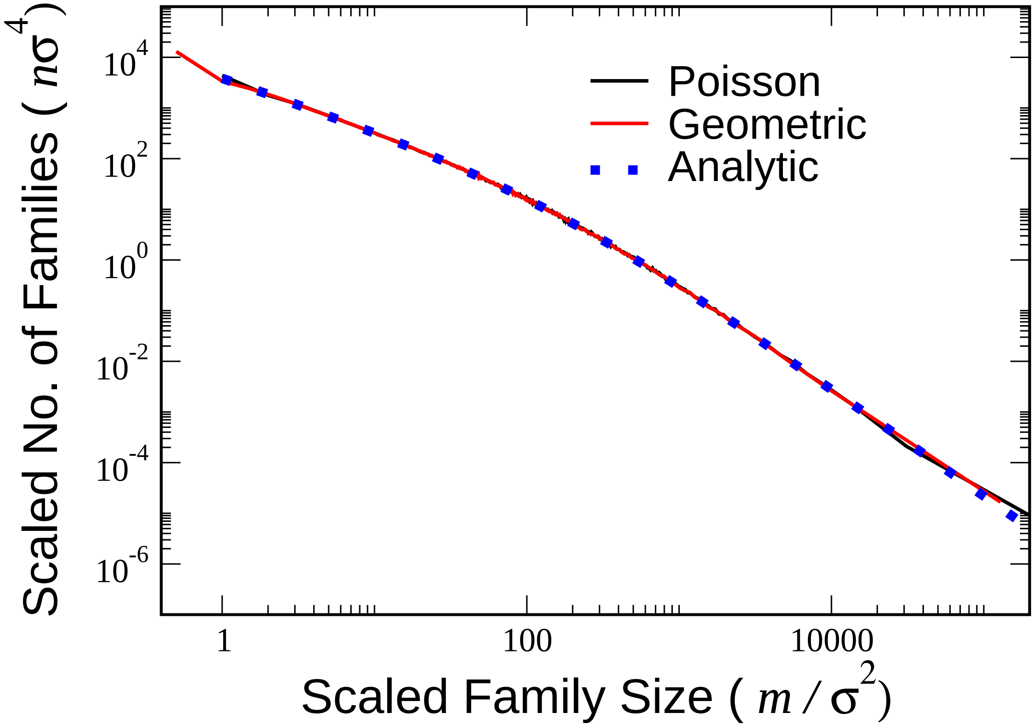
- Has shoulder at small m, fewer small families due to demographic fluctuations

Generalizes Fisher-Log series for $\gamma = 0$ (no growth) case.

Universality

$$n_m = \frac{\nu R_c}{m}\Gamma\left(2+\nu\right)\, U\left(1+\nu, 0, R_c\frac{m}{N_o}\right) \qquad\qquad R_c \equiv \frac{2N_o\gamma}{\sigma^2(1+\nu)}$$

$\sigma^4 n$ is a universal function of m/σ^2 for given γ, μ



Question #2: Subsampling

Rarely have access to entire population

Especially true for genetic data, species abundances

Subsample data favors large families, small families likely to be missed entirely

Basic Observation: For a power-law distribution, sampling preserves the power-law

Full answer: For a sample of size R_o: ("red" subpopulation)

$$n_m^R = \sum_{p\geq m} n_p \frac{\binom{p}{m}\binom{N_o-p}{R_o-m}}{\binom{N_o}{R_o}} \approx \sum_{p\geq m} n_p \frac{e^{-pR_o/N_o}}{m!}\left(\frac{pR_o}{N_o}\right)^m \qquad R_o \ll N_o$$

$$\approx \nu R_c \,\mathrm{B}\left(2+\nu, m\right) s^m \, {}_2F_1\left(m, m+1; 2+\nu+m; 1-s\right)$$

where $s \equiv R_o/R_c$, normalized sampling strength

Subsampling

$$n_m^R = \nu R_c \, \mathrm{B}\,(2+\nu, m)\, s^m \, {}_2F_1\,(m, m+1; 2+\nu+m; 1-s)$$

What does this mean?

Two limits:

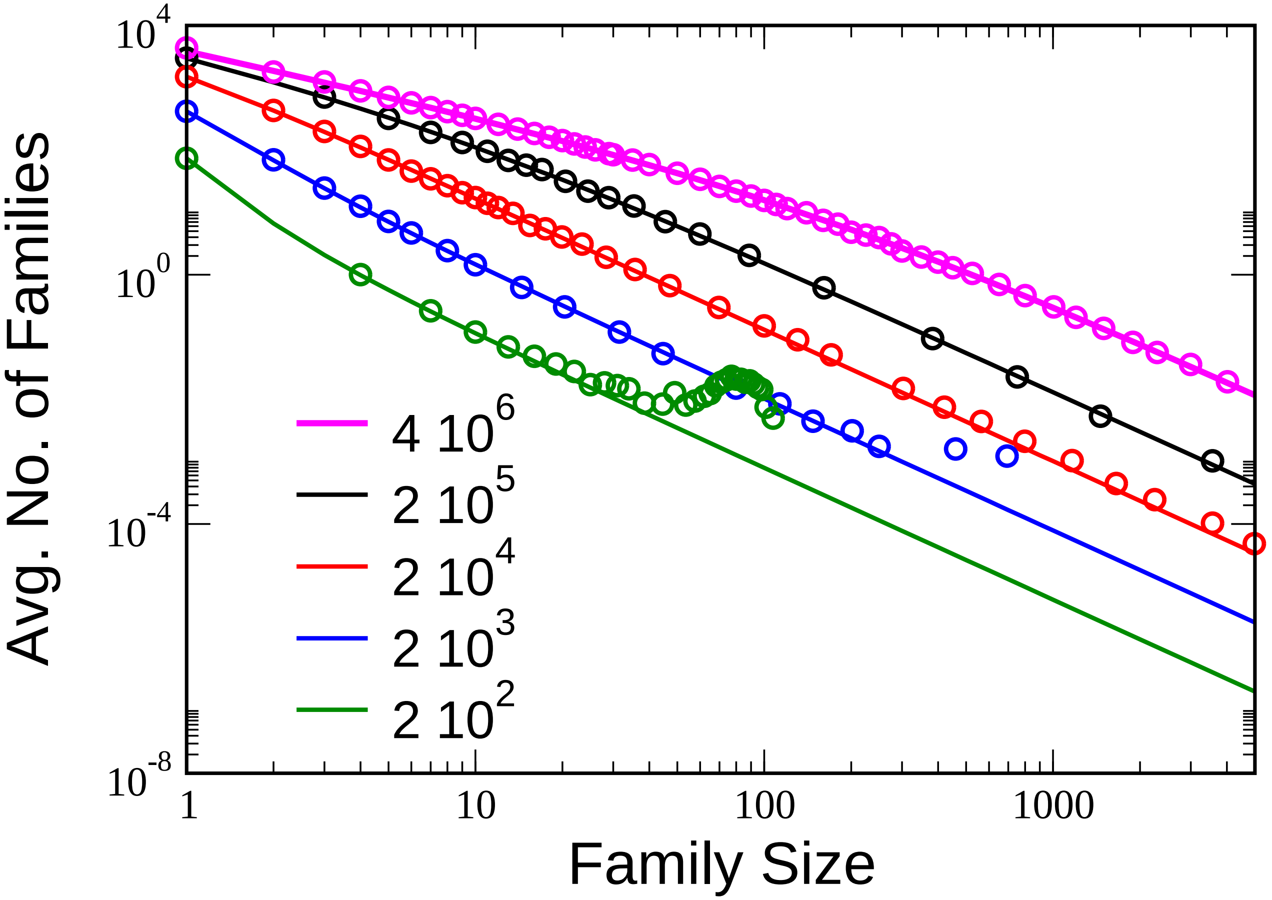
- Strong sampling: $R_o \gg R_c$

  $$n_m^R \approx \nu R_c \frac{\Gamma(2+\nu)}{m} U(1+\nu, 0, m/s)$$

  Distribution is shifted down and to the left. Left "shoulder" shrinks
- Weak sampling: $R_o \ll R_c$.

  Shoulder eliminated completely. Power-law tail starts for m over order 1.

Subsampling



Application 1: Surnames

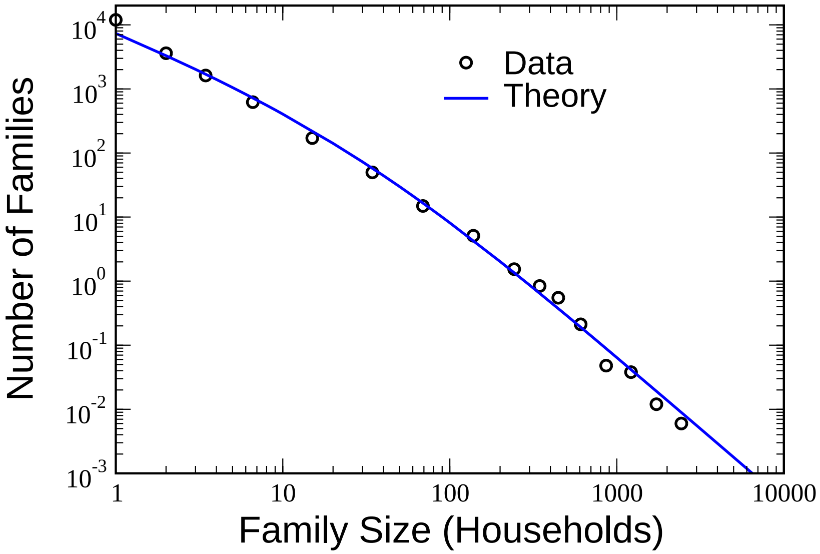
MZ looked at surnames for various cities taken from telephone books
There is better data available from censuses

- USA: Last three censuses, 1790
- Norway: 2008

USA Census: 1790

- Compiled in 1915 for the 125th anniversary of the Census
- Use known growth rate of England population (basically constant γ from 1086 to 1800)

USA 1790

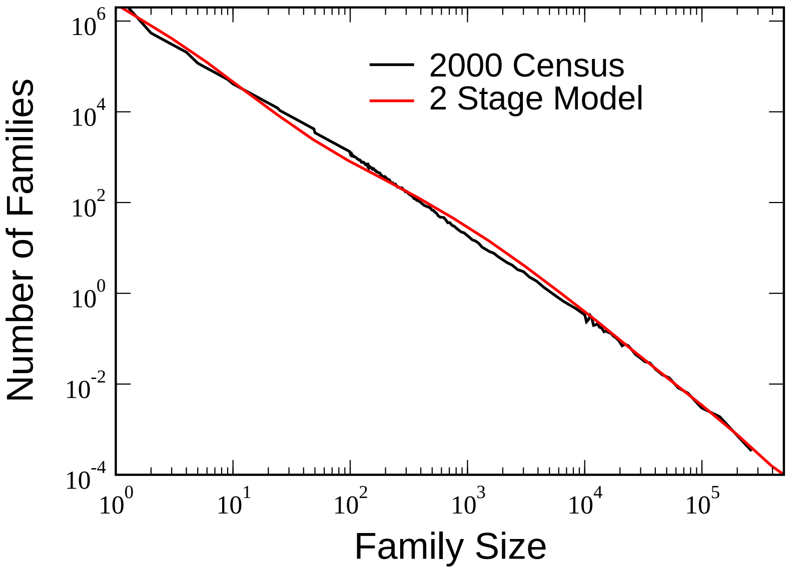


The 2000 US Census

The US (and England) growth rate from 1790-1920 is much larger (by a factor of 7.5!!) than the England 1086-1800 rate

- The US growth rate fell again after 1920

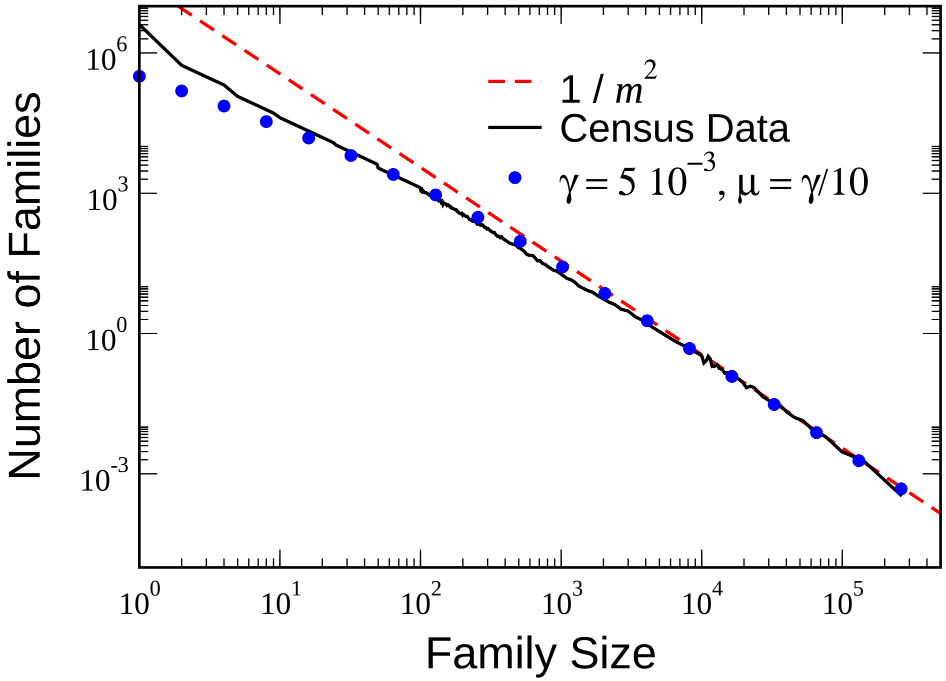
If we use a two-rate model, using the England 1986-1800 rate before 1800 and the 1790-1920 rate after, we get



No single growth-rate model fits the data

The US Census Puzzle

Using Best Fit Growth Rate

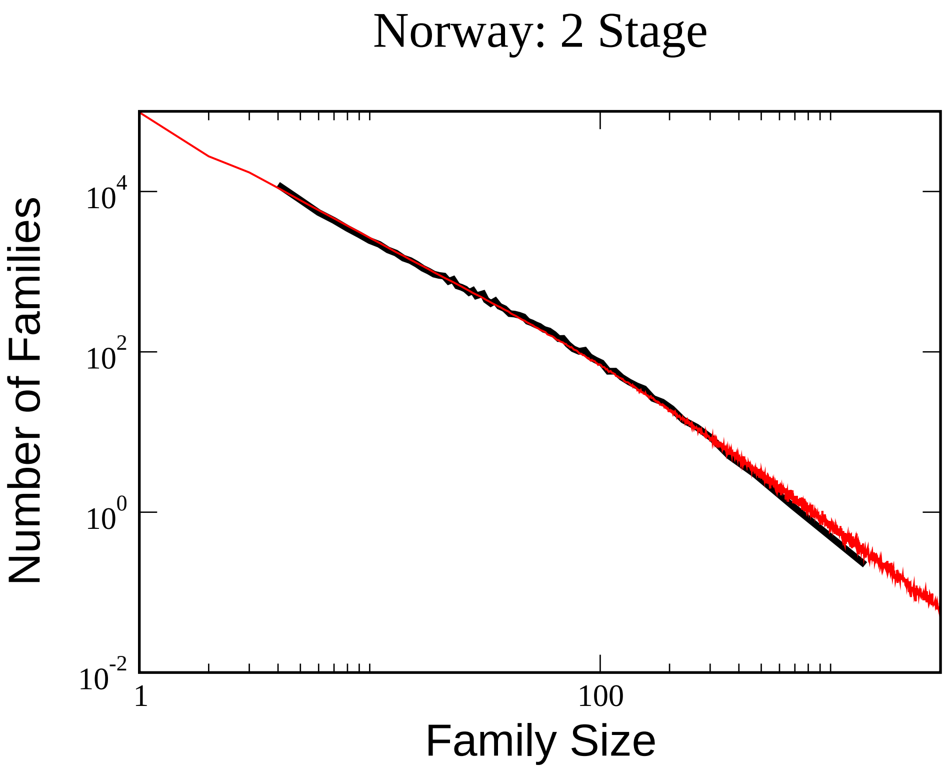


Discrepancy of a factor of 10^2 in γ!!!!!!

Norway 2 Stage Model

Same holds for Norway

$\gamma = 0.05$ before 1800, $\gamma = 0.2$ afterwards.



Application 2: Haplotype Statistics

Inherited characteristic is noncoding mtDNA sequence

Not enough data (yet) to look at full family statistics

Concentrate on number of haplotypes (different sequences)

- This is only a single number - does not fix model parameters
- Consider number of haplotypes as function of sequence length: $\mu(\ell) = \mu_1 \ell$
- This is a nontrivial function - starts out linearly and saturates

Number of Haplotypes (Family Surnames) follows from Kummer distribution

$$F = \begin{cases} \frac{\nu R_c}{2+\nu}\left[{}_2F_1\left(1,1;3+\nu;1\right) - (1-s)\,{}_2F_1\left(1,1;3+\nu;1-s\right)\right] & \mu < \gamma \\ R_c s\,{}_2F_1(1,1;2+\nu;-s) & \mu > \gamma \end{cases}$$

$$s \equiv R_0/R_c\,; \qquad \nu \equiv \frac{\min(\mu(\ell),\gamma)}{|\gamma-\mu(\ell)|}\,; \qquad R_c \equiv \frac{2N_o|\gamma-\mu(\ell)|}{\sigma^2}$$

Recurrent Mutations

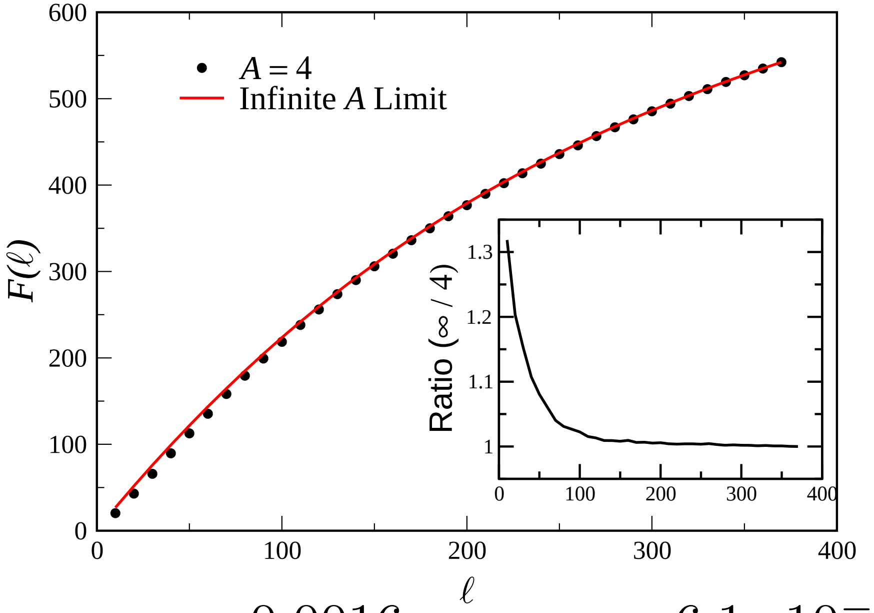
Theory assumes that each mutation gives rise to a new haplotype: NO RECURRENT MUTATIONS

- Equivalent to infinite allele model

In real life, there are only four choices for each nucleotide

Less variety in short sequences

Simulated Data

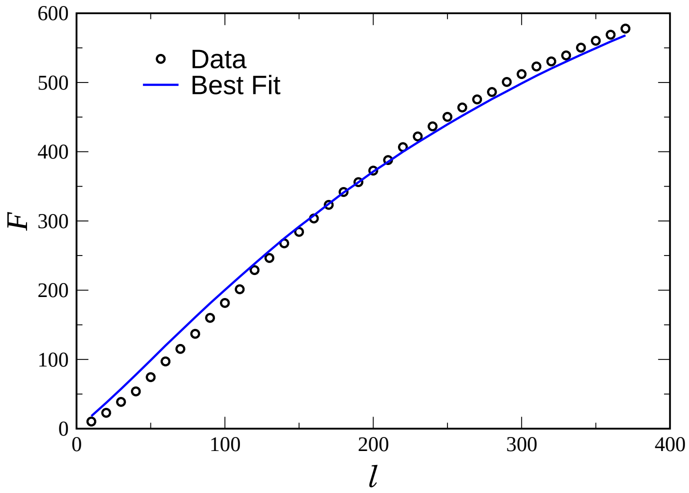


$N_o = 1.4 \cdot 10^5, \qquad \gamma = 0.0016, \qquad \mu_1 = 6.1 \cdot 10^{-6}, \qquad R_o = 1000$

The Real Test: Data from China

1212 sequences from China of the HVR1 region of mtDNA, $L = 377$

China



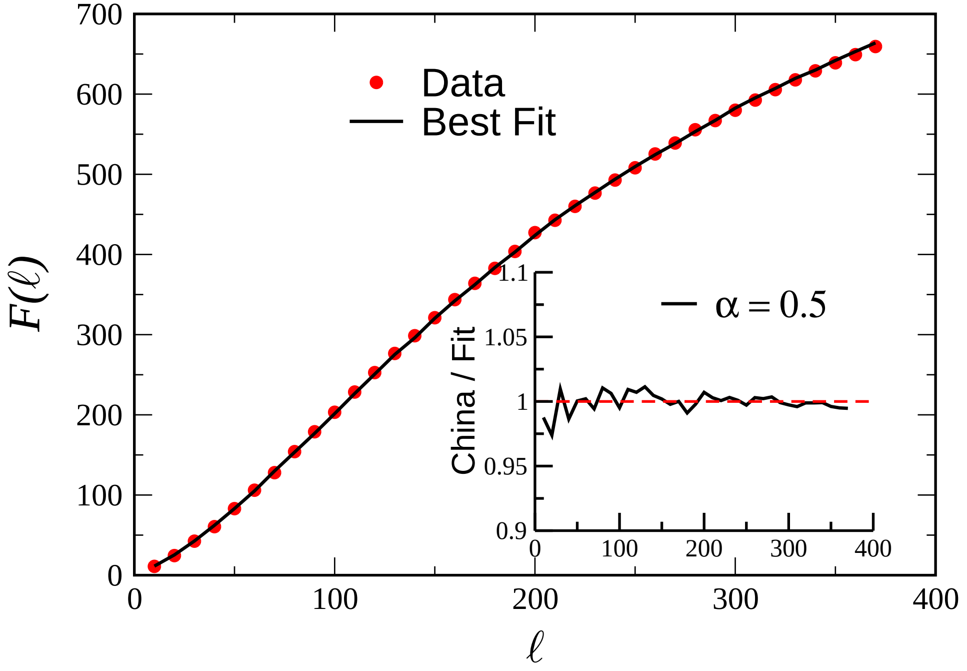
Small ℓ is qualitatively wrong, convex!

Result of "Hot Spots" in Genome: Loci with high mutation rate

- Mutation rates on various loci are distributed according to a Gamma Distribution: $f(\mu_1) = \frac{(\alpha/\bar{\mu}_1)^{\alpha}}{\Gamma(\alpha)}\mu_1^{\alpha-1}e^{-\alpha\mu_1/\bar{\mu}_1}$
- 2 parameters, the average mutation rate, $\bar{\mu}_1$, and α, $0.1 \leq \alpha \leq 1.1$

More recurrent mutations at these loci $\Rightarrow$ impacts short sequences

Best Fit



Previous estimates for α: 0.44-0.6 Wakeley, 1993; 0.28-0.39 Excoffier & Yang, 1999

Application 3: Species in Genera

Consider species (within a kingdom, say) as individuals. Reproduction ⇒ Speciation. Family ⇒ Genus.

Generally, new species belongs to some genus as originating species

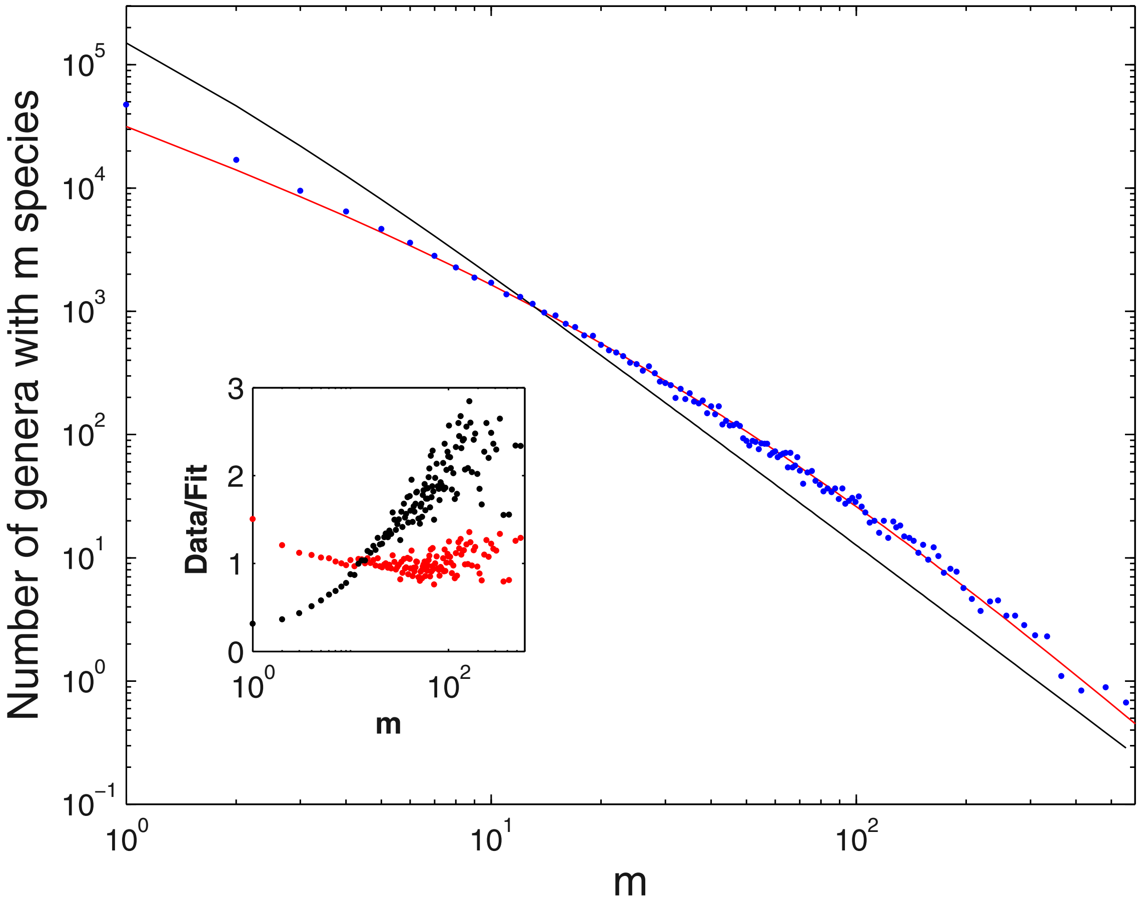
Occasionally, new species starts new genus ⇒ Mutation

Then, our theory should be appropriate for numbers of species within different genera

Question originally posed by Yule (1925): Yule's model had speciation & mutation, no extinction of species

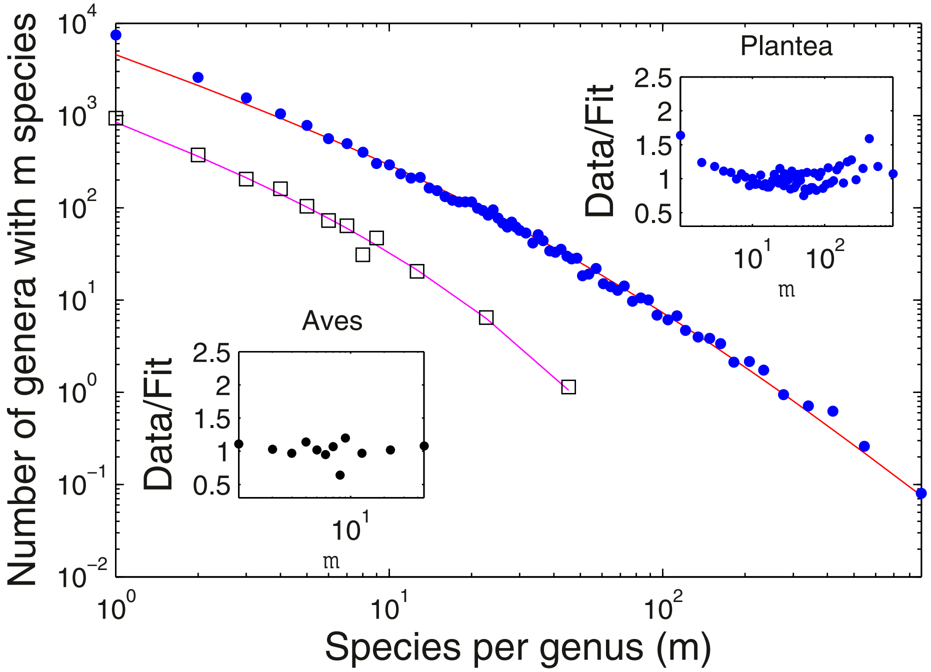
Species in Genera: The Data

Animalia Kingdom



Red is our model, Black is Yule’s model

Species in Genera: More Data



From growth rates, can infer Time to Most Recent Common Ancestor for the various kingdoms. These are roughly consistent ($\pm 30\%$) with fossil & DNA estimates (except for Diplopoda (factor of 7 low).

Controversy: Do Genera exist?

No rigorous criteria for grouping species into genera

Many biologists therefore claim that genera are purely human artifacts with no biological reality

This data would suggest otherwise

Summary

Universality for slow growth, mutation

"Critical Sampling" $R_c = \frac{2\gamma}{\sigma^2(1+\nu)} N_o$ for disappearance of shoulder

Simplest models have to be modified to fit reality

- Varying growth rate in census data
- Recurrent mutations, variable mutation rate in genome data
- Failures are always more instructive than success

Genera may actual be real

References

- J. Theor. Bio., 262, 245-256 (2010)
- PLoS One, 6, e26480 (2011)
- J. Stat. Phys., 142, 1302-1316 (2011)
- PNAS, 110, E2460-E2469 (2013)