Lecture Four

Genome the Blind Self-Plagiarizer and the Mystery of "Missing" Mutations

HC Lee
National Central University

Summary of genome data

- Universality class – for fixed word length k, L_{eff} is (approximately) the same for all genomes
  - Log $L_e(k) = ak + B$
  - a, B are universal constants

- Vast majority of genomes have order index

$$\ln \phi_g = -3.49 \pm 0.65 \quad \text{, or} \quad \phi_g = 0.031^{+0.028}_{-0.015}$$

- Strong local or global, or both, inverse-symmetry

Order, Randomness, L_e and duplications

- Small L_e of genomes suggests segmental duplication
- Coarse grain homogeneity suggest random events
- Statistical similarity between coding and non-coding regions suggest random/neutral events
- Long range correlation suggest tandem events

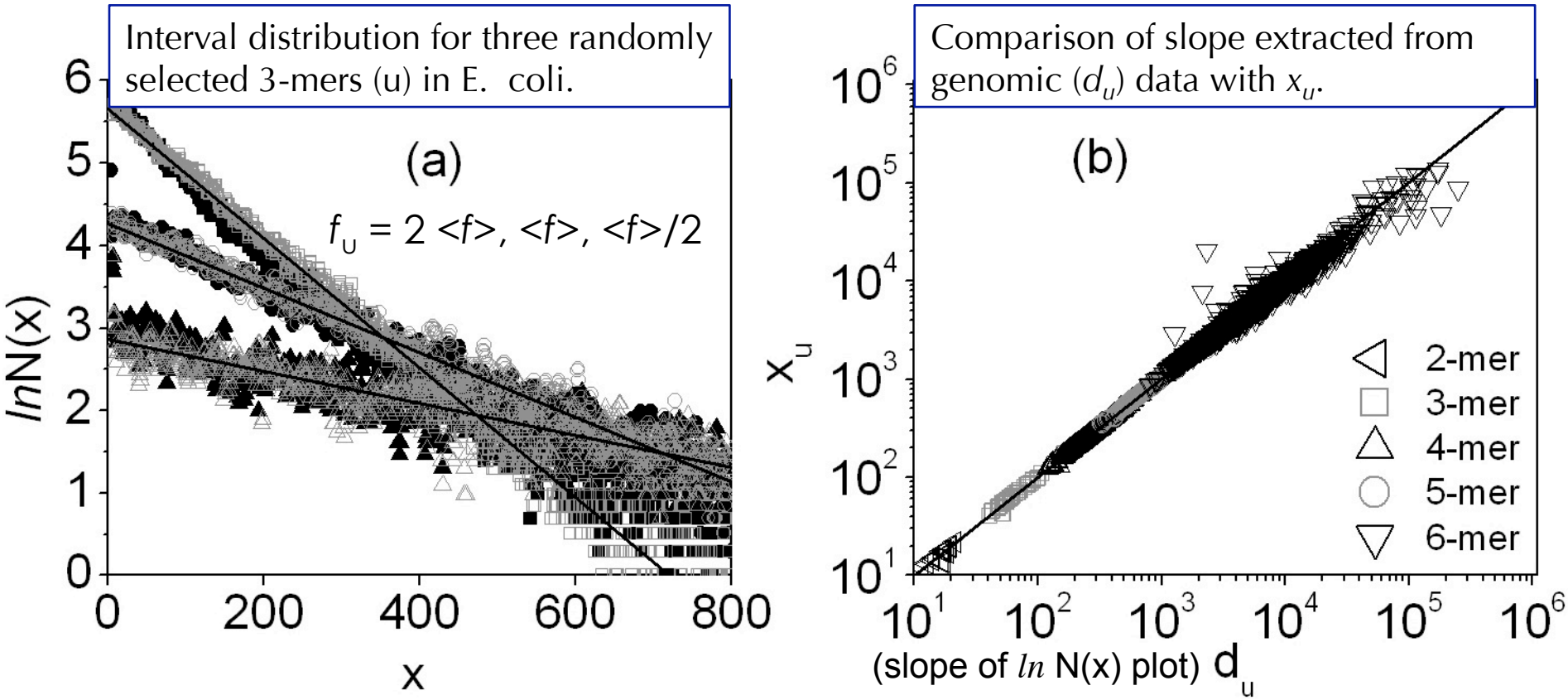
Word Intervals

- Intervals (spatial or temporal) between adjacent random uncorrelated events have an **exponential distribution**
- In a random sequence, intervals of identical words are exponential
- What is the word-interval distribution in a (non-random) genome?

k-mer interval distribution

If the SITE distribution of *k*-mers is random, then the intervals have exponential distribution

$$N^{\{ran\}}(x) = N_0 e^{-x/x_u} \qquad x_u = L/f_u$$



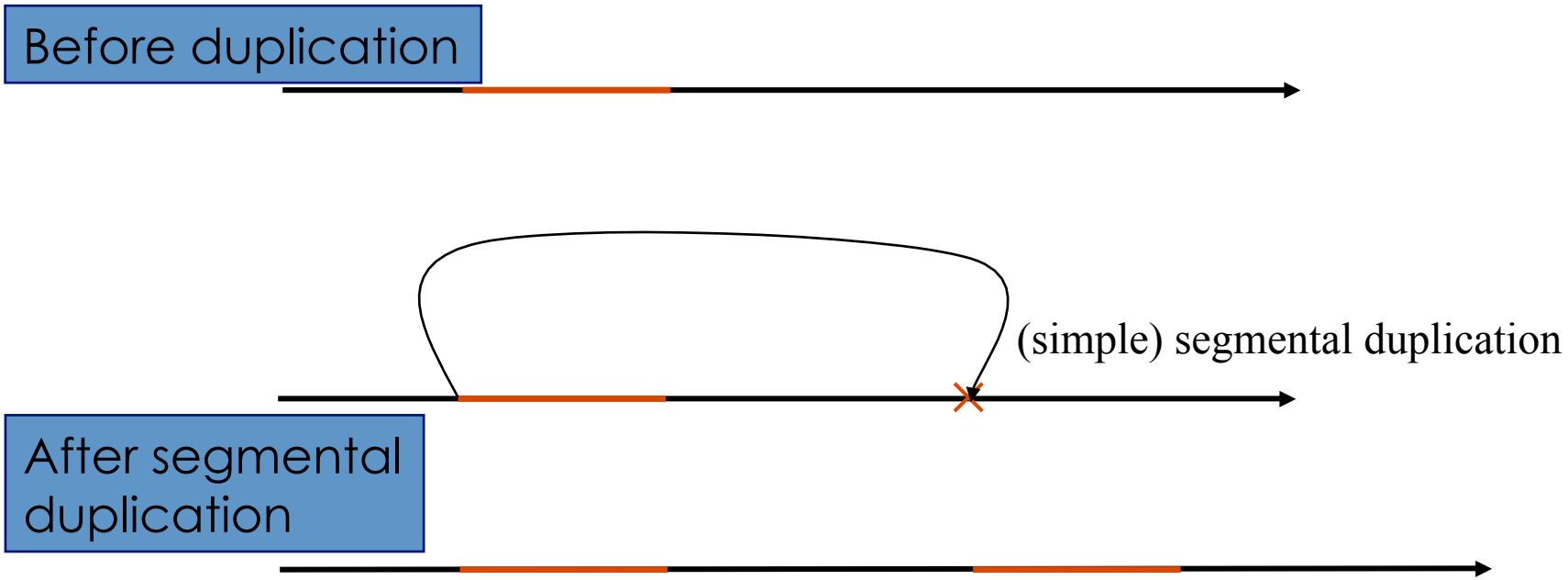
Genome Growth by Random Segmental Duplication

Susumu Ohno (1970). *Evolution by gene duplication*. Springer-Verlag.

Kimura, Motoo. 1983. *The neutral theory of molecular evolution.* Cambridge (page xi).

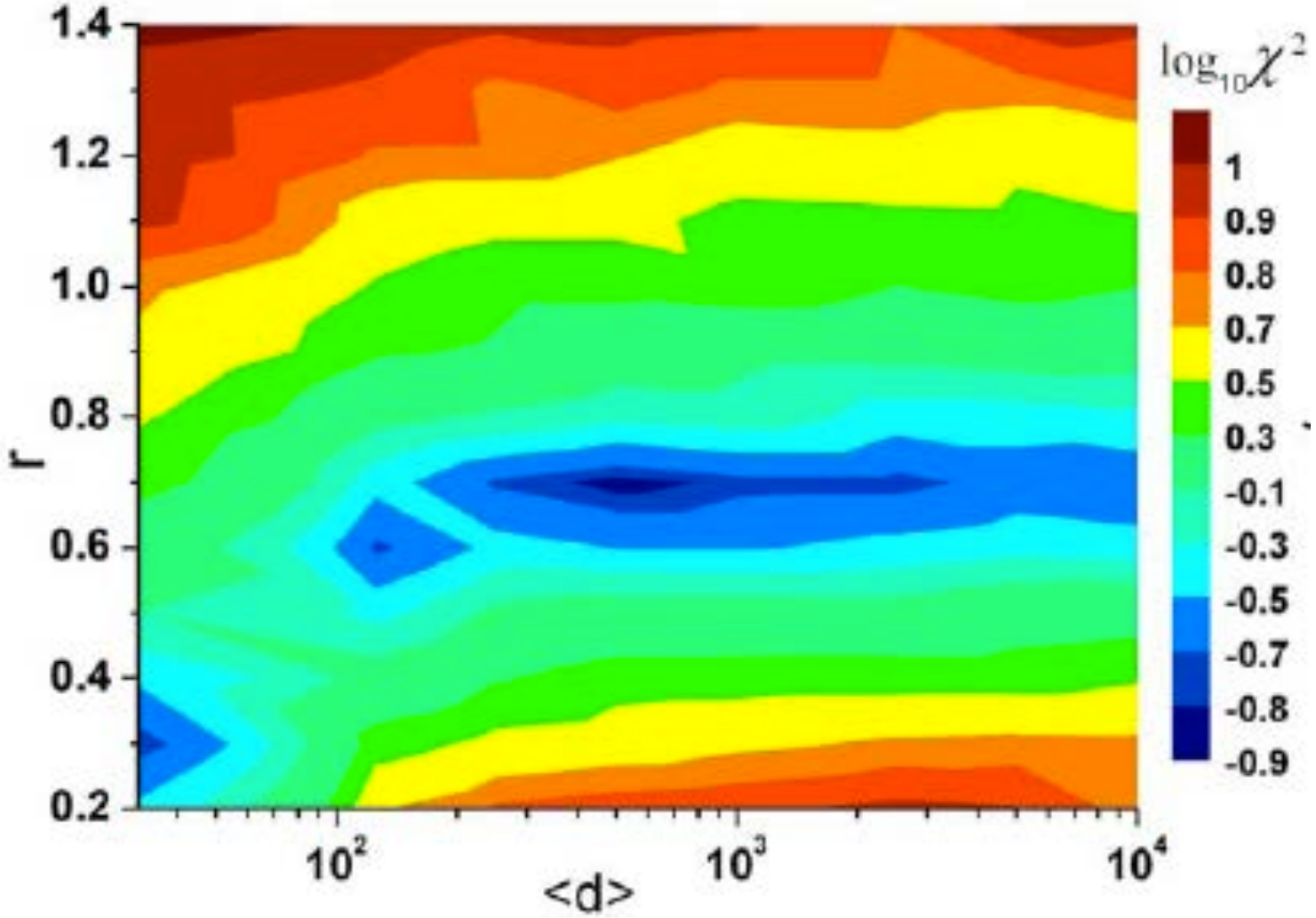
- RSD model has three parameters
  - Initial length L^0 (with chosen p)
  - Average duplicated segment length $<d>$
  - Point mutation density r (base biased to p)
- Grow protocol - maximally stochastic segmental duplication
  - Randomly selected copy site
  - Randomly selected segment length (with distribution)
  - Randomly selected insertion site

Segmental duplication



Good parameter basin
L_0=64b, <d>=120-5000b,
r = 0.65-0.75/site

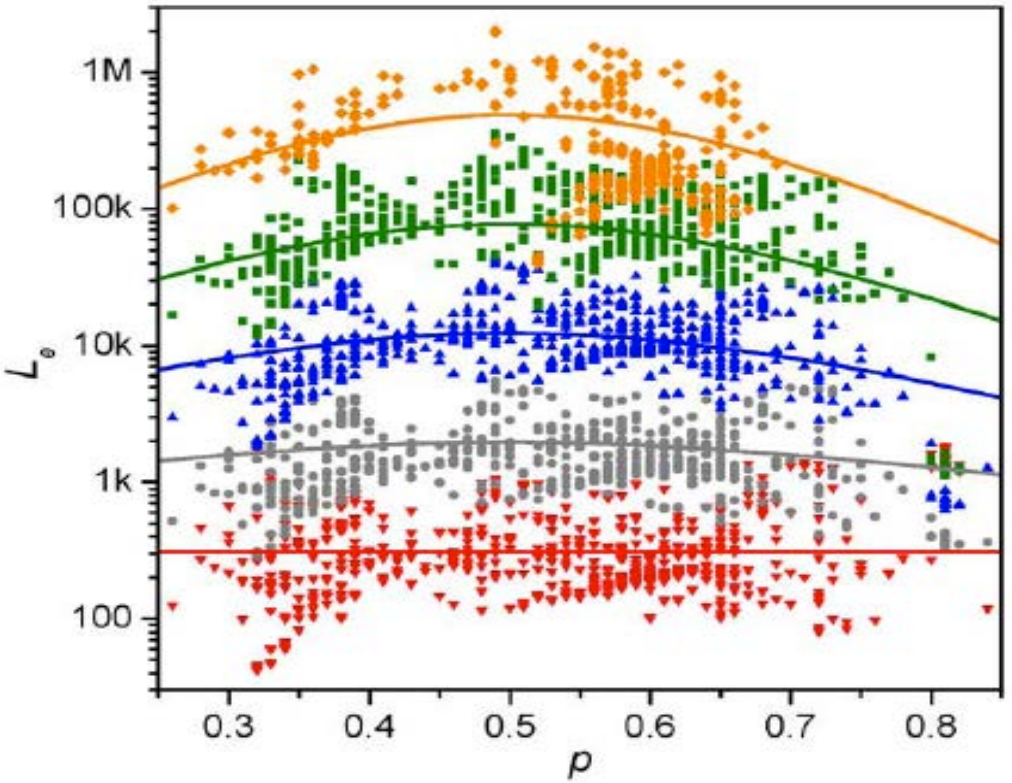
$$\chi^2_\sigma = \frac{1}{N_\sigma} \sum_{i \in \{\sigma\}} \ln^2 \left(\frac{L_{e,i}}{L_e^{\{uc\}}(k;p)} \right)$$

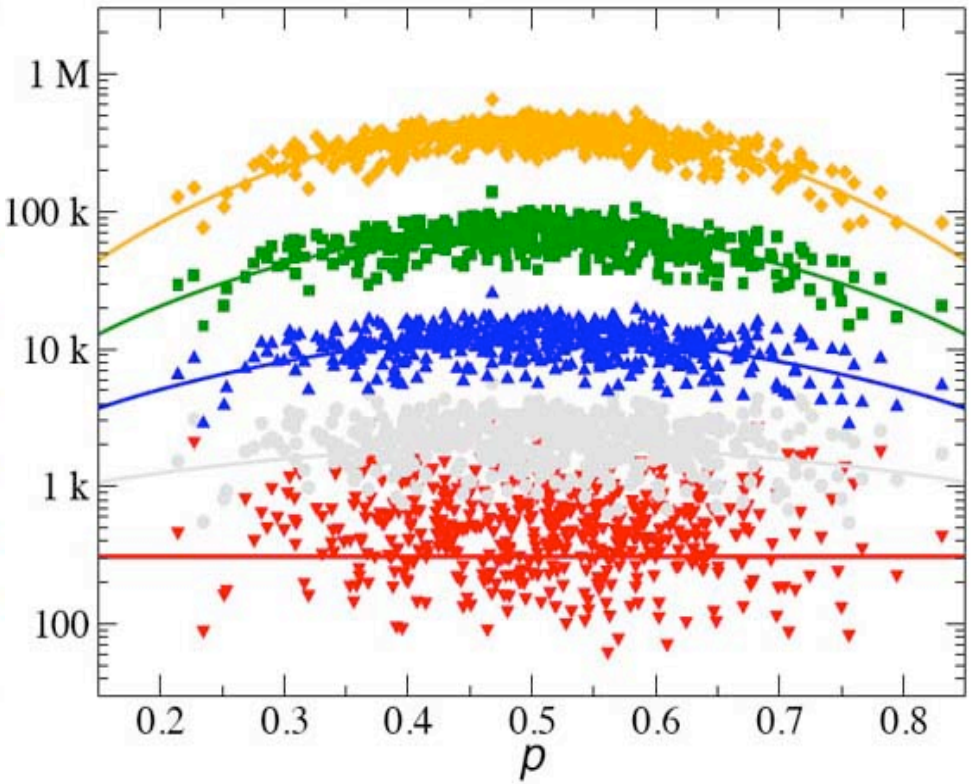


Genomic and model L_e

$$\chi^2_\sigma = \frac{1}{N_\sigma} \sum_{i \in \{\sigma\}} \ln^2 \left(\frac{L_{e,i}}{L_e^{\{uc\}}(k;p)} \right)$$

"Best" parameter set:
L_0=64b, <d>=1000b,
r = 0.73/site

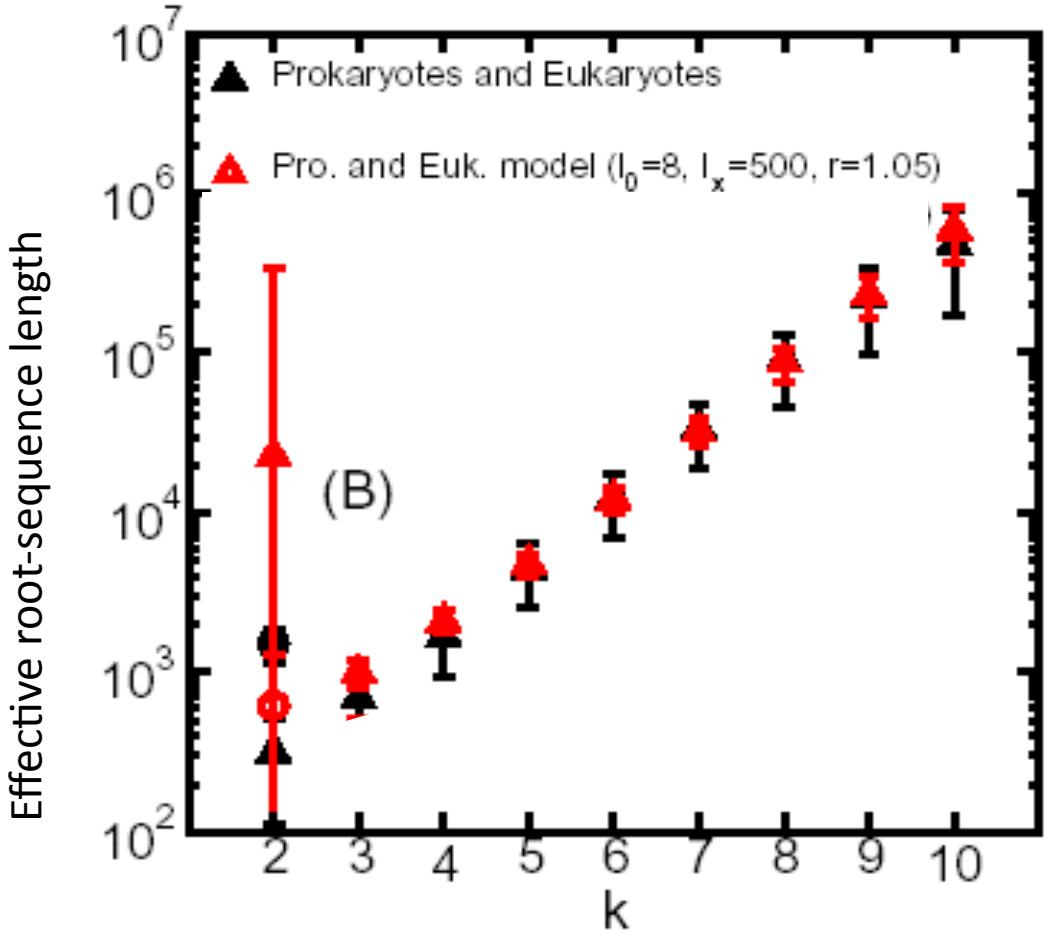




Genome χ^2=0.43

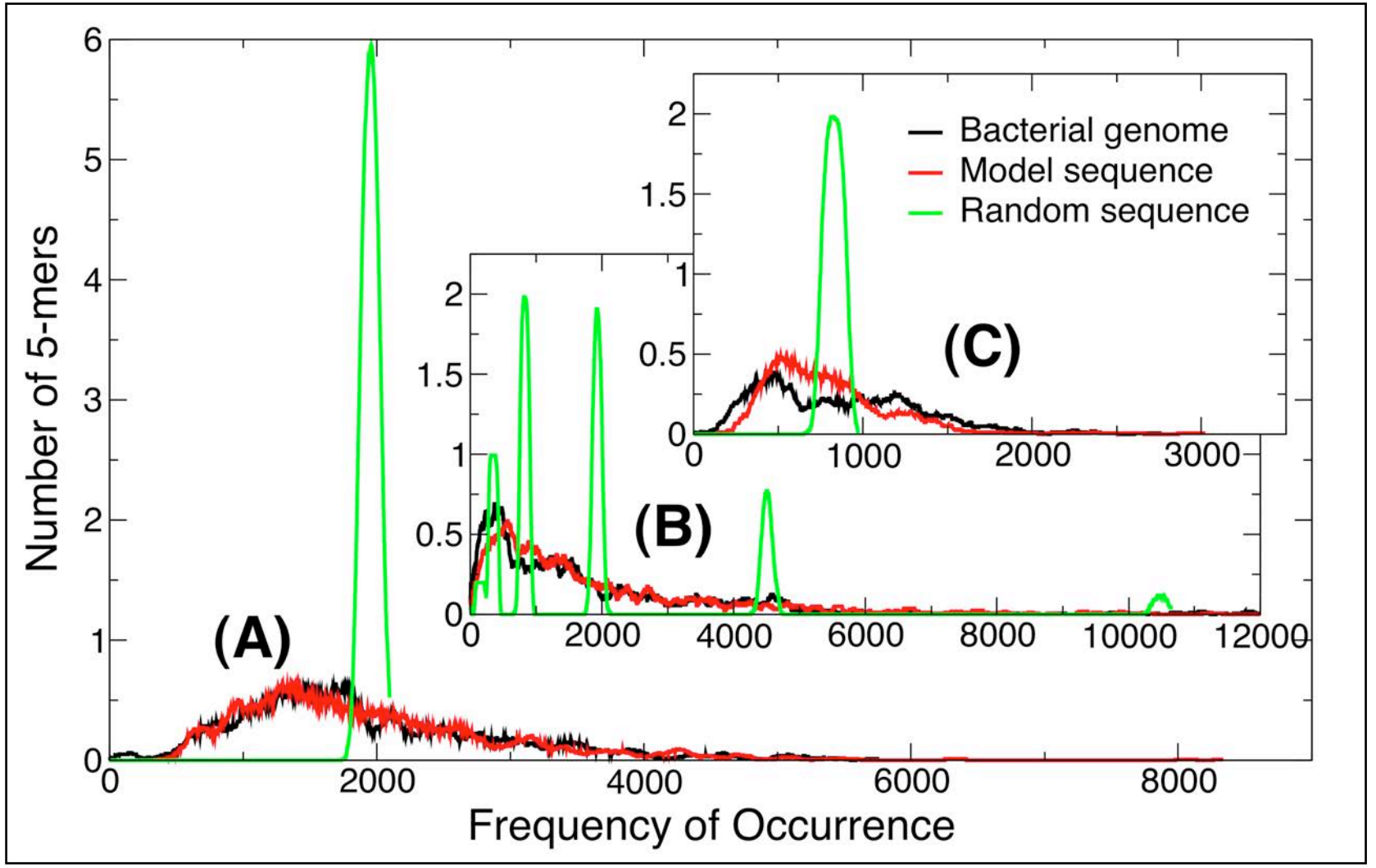
Model χ^2=0.18

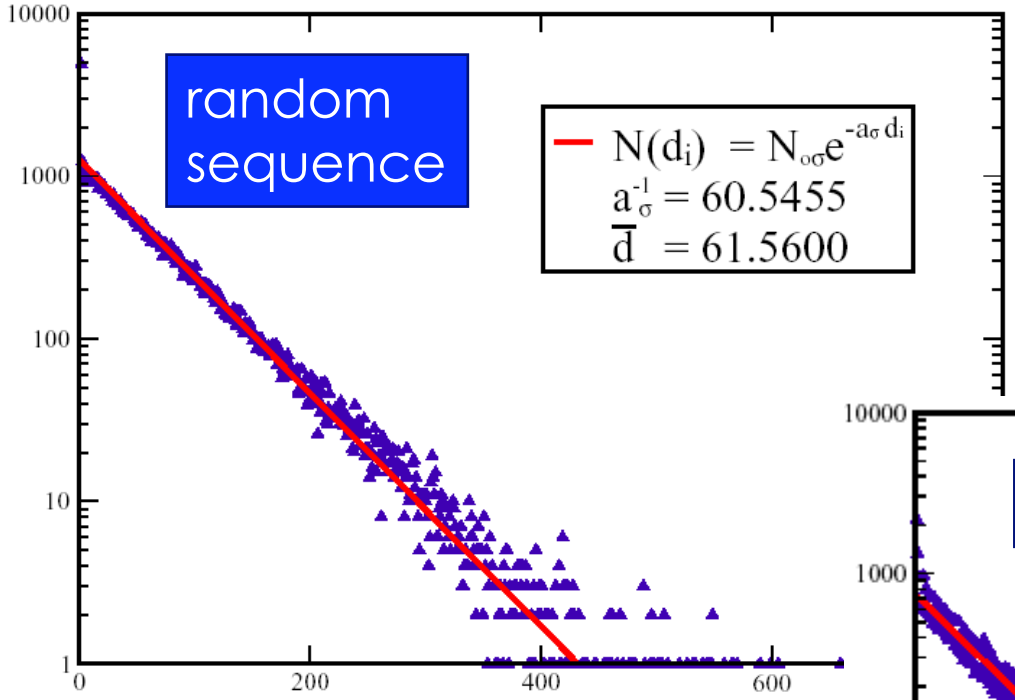
RSD model with three universal parameters generates artificial genomes with universal L_e



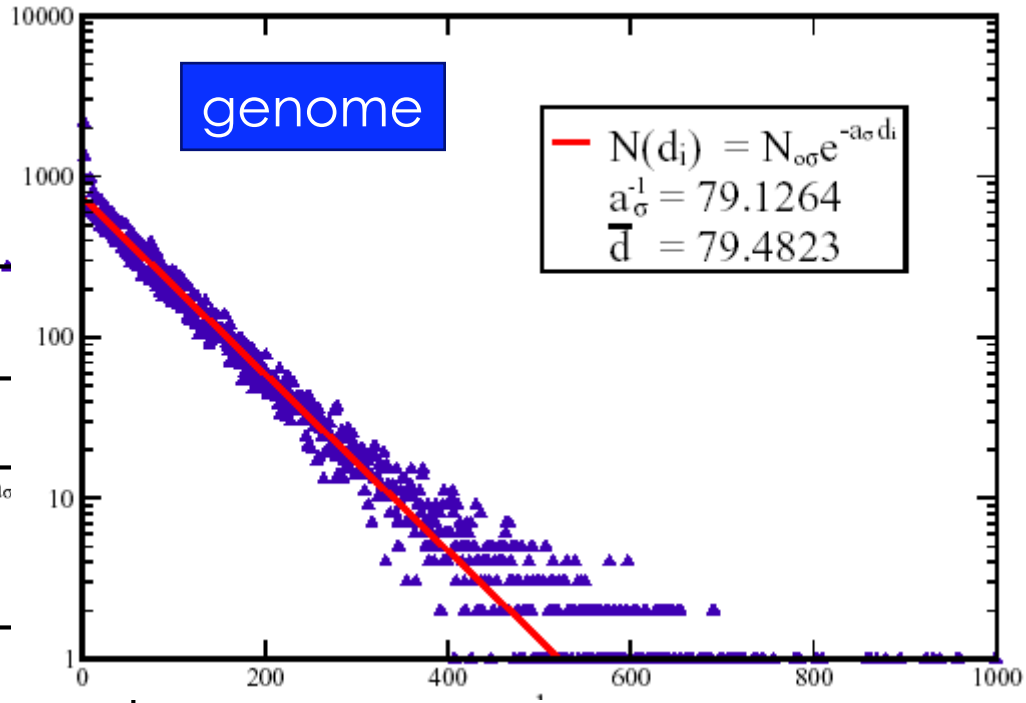
Red symbols : from 278 genome-matching model sequences

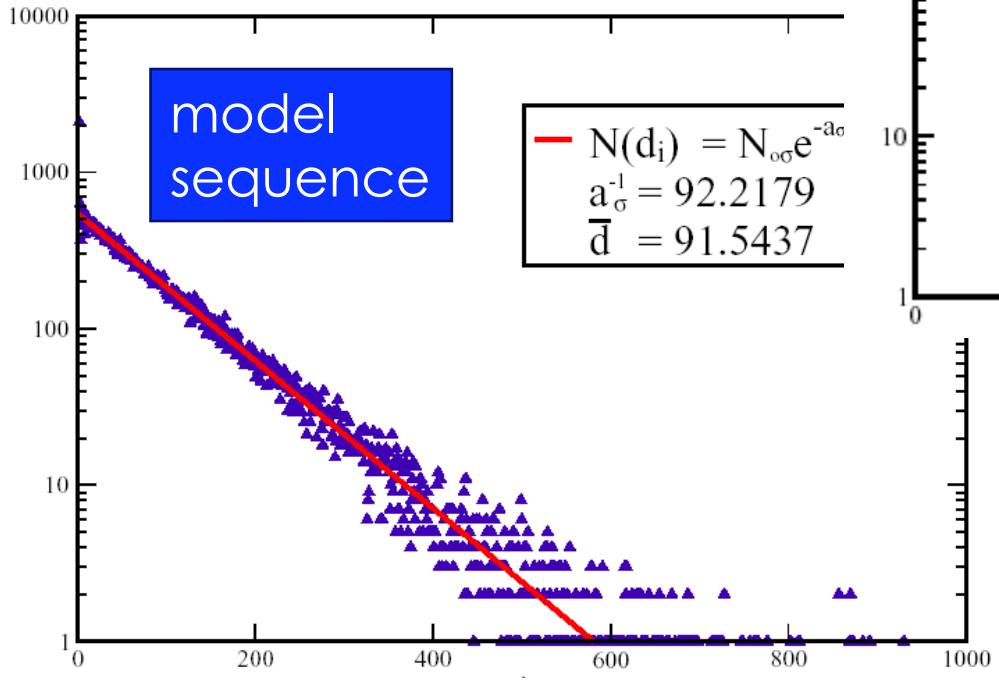
Distribution of frequency of occurrence





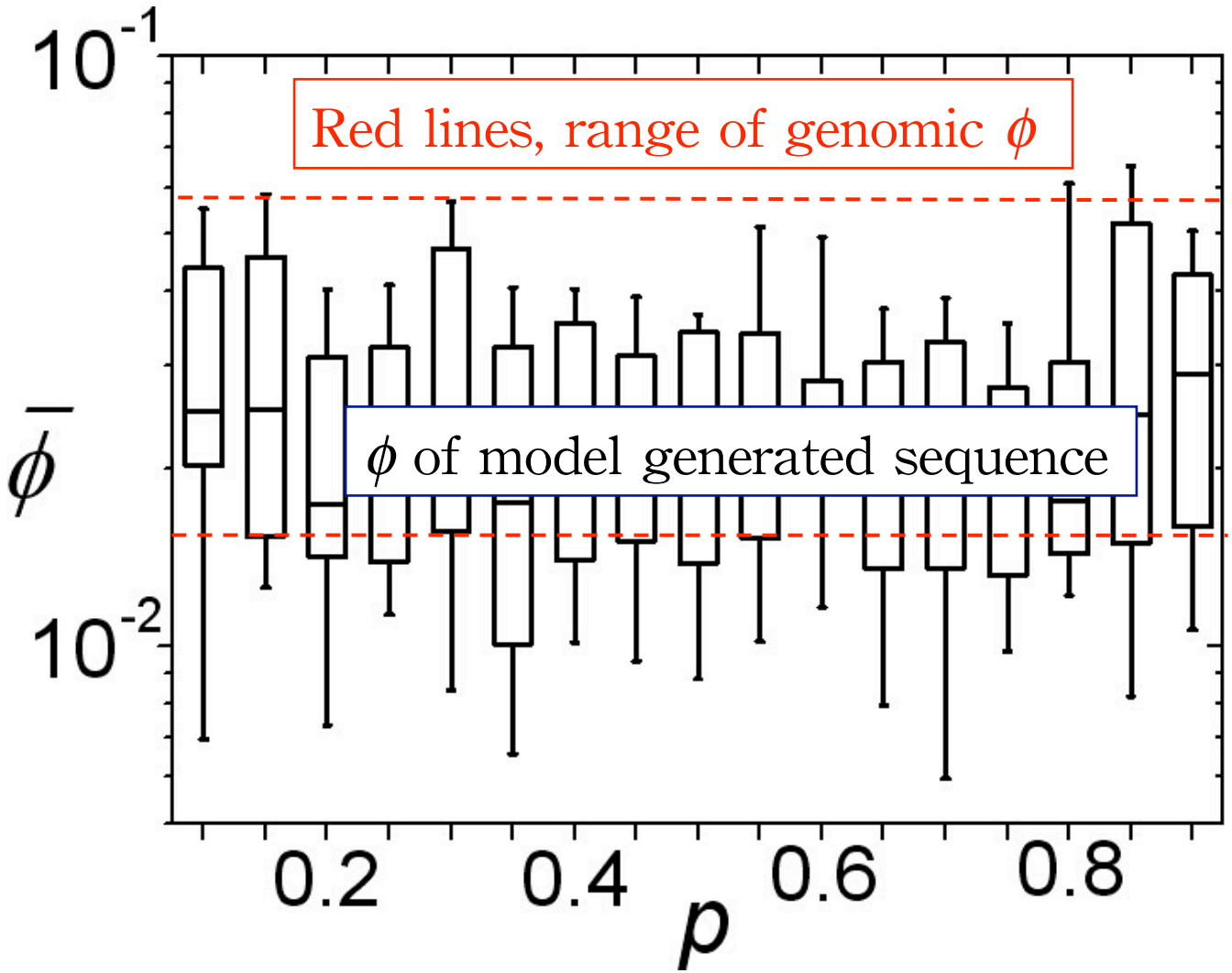
Interval distribution is exponential in random sequence as expected.
But also in genome!





And in the model sequence (not surprising, because growth mechanism is **maximally stochastic**).

Artificial sequence with genomic L_e generated in RSD model has genomic ϕ

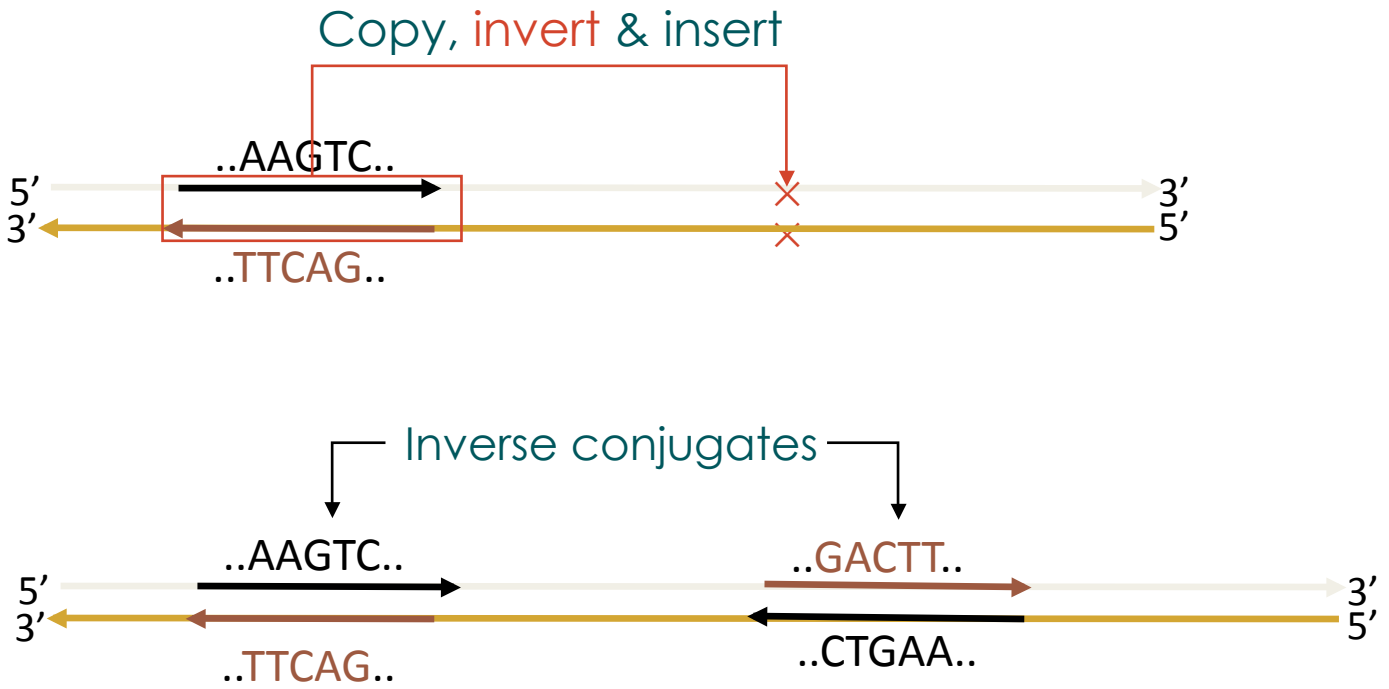


Inverse symmetry

- Had $<\chi_{\mathrm{inv}}> \sim 0.073 +/- 0.066$
- Consider Type D (IS everywhere). Then percentage of chromosome generated by ISD (from mean-field approximation) is

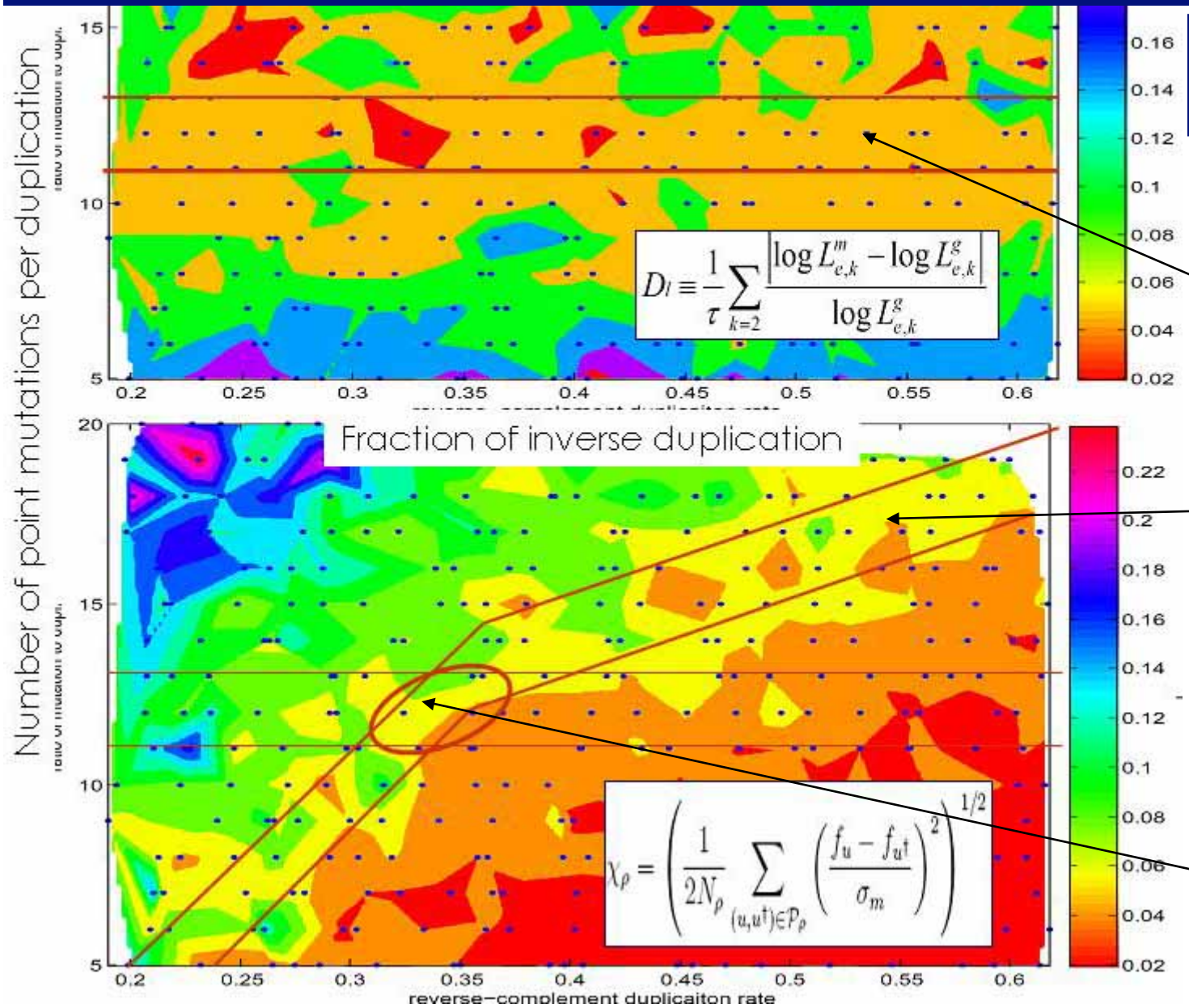
$$2\delta v_{inv} \sim 0.25 +/- 0.15$$

Inverse segmental duplication (ISD) generates IS



Absence of similar mechanism for generating reverse/ complement symmetries may explain their absence

Percentage of ISD (for Type D) ~ 30%



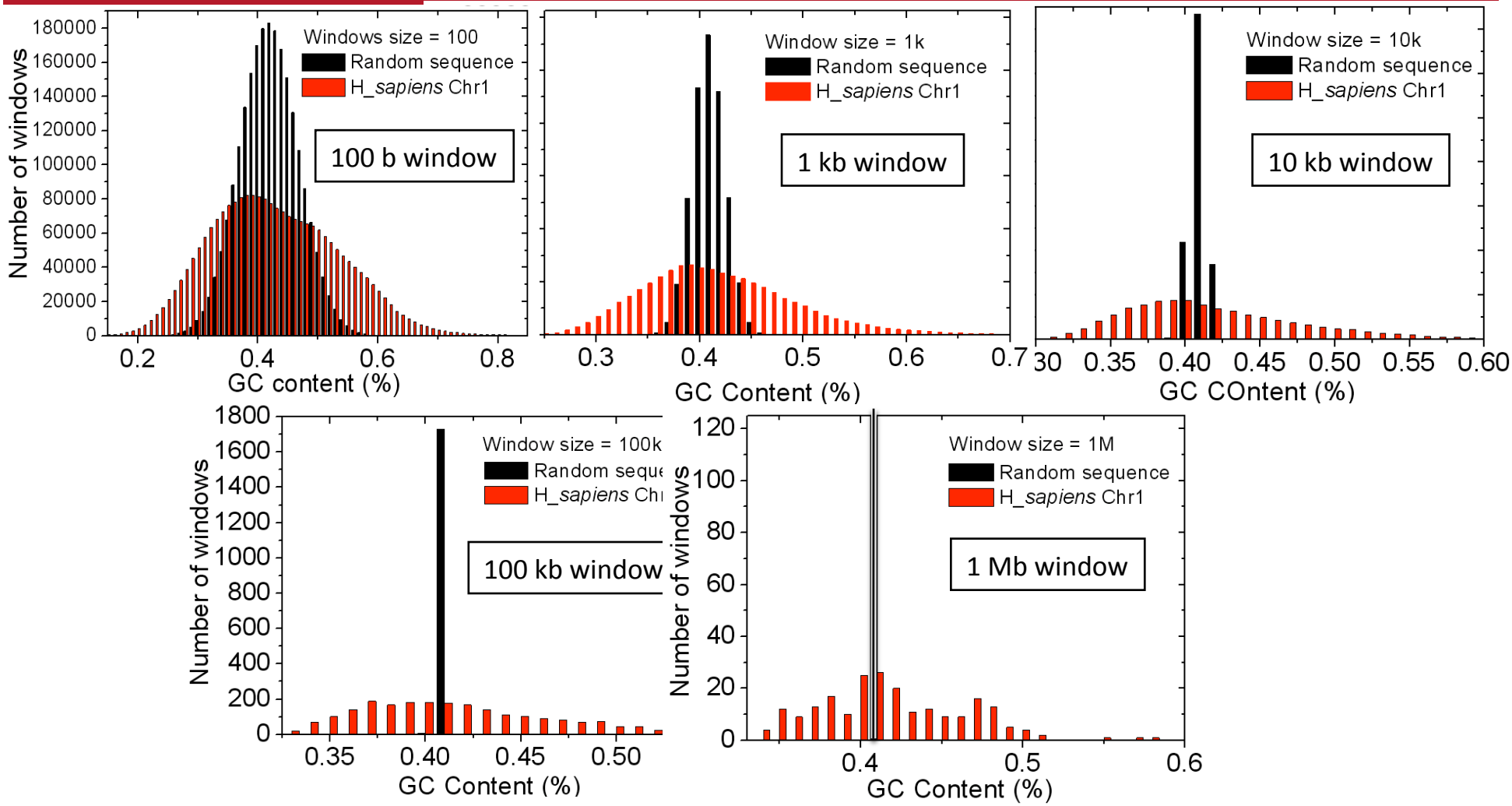
Best model parameter

Region Permitted by L_e

Region permitted by IS (χ~0.05)

Region Permitted by L_e & IS

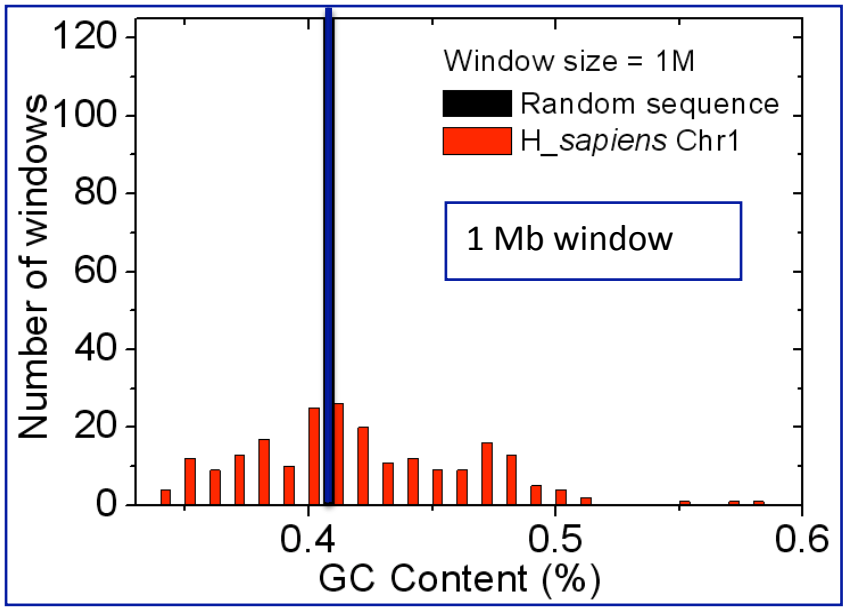
For GC-content histograms: sample size = window size



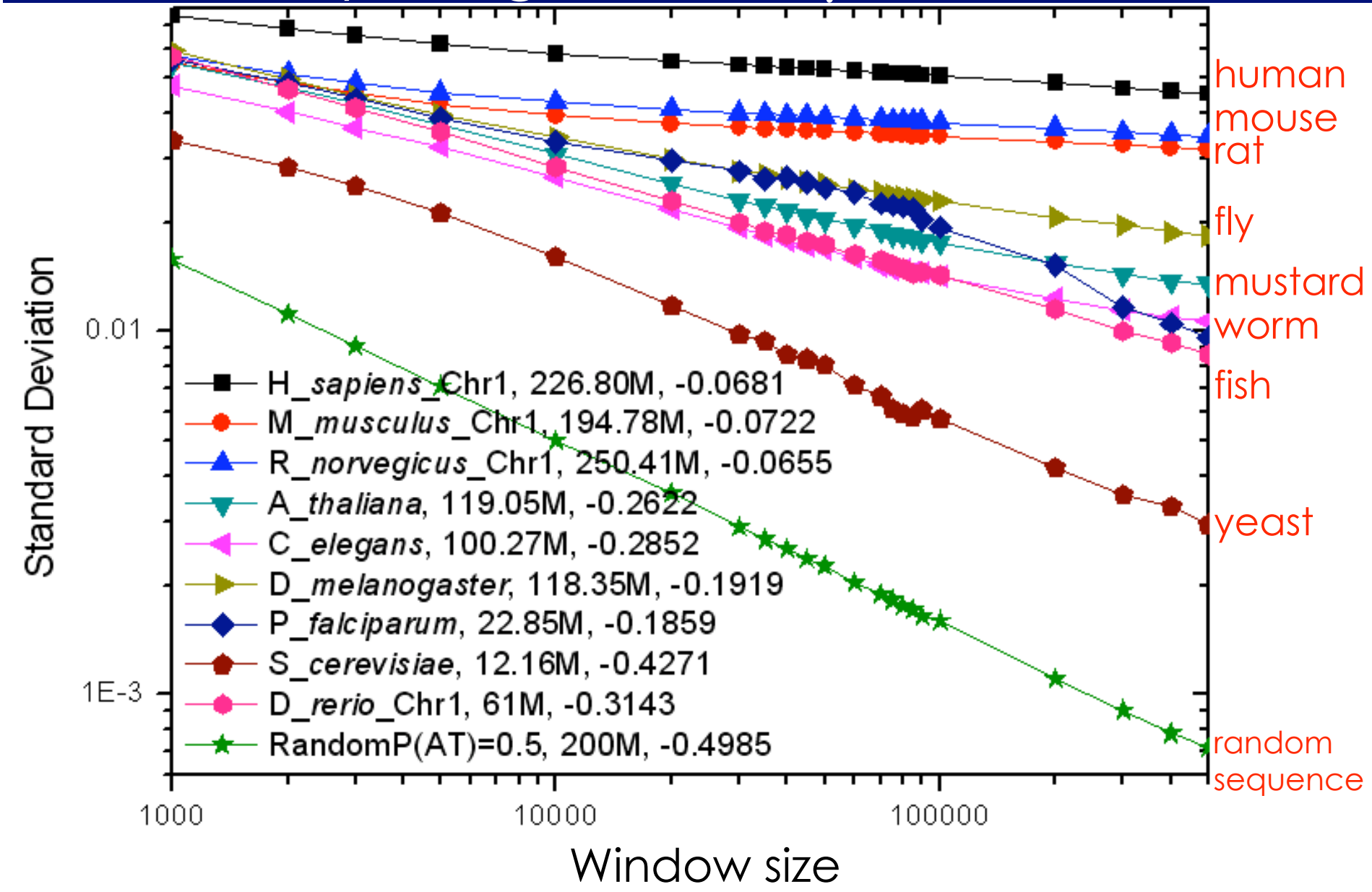
Variation of CG-content in Human genome does not obey central limit theorem

RSD model does not generate long range correlation

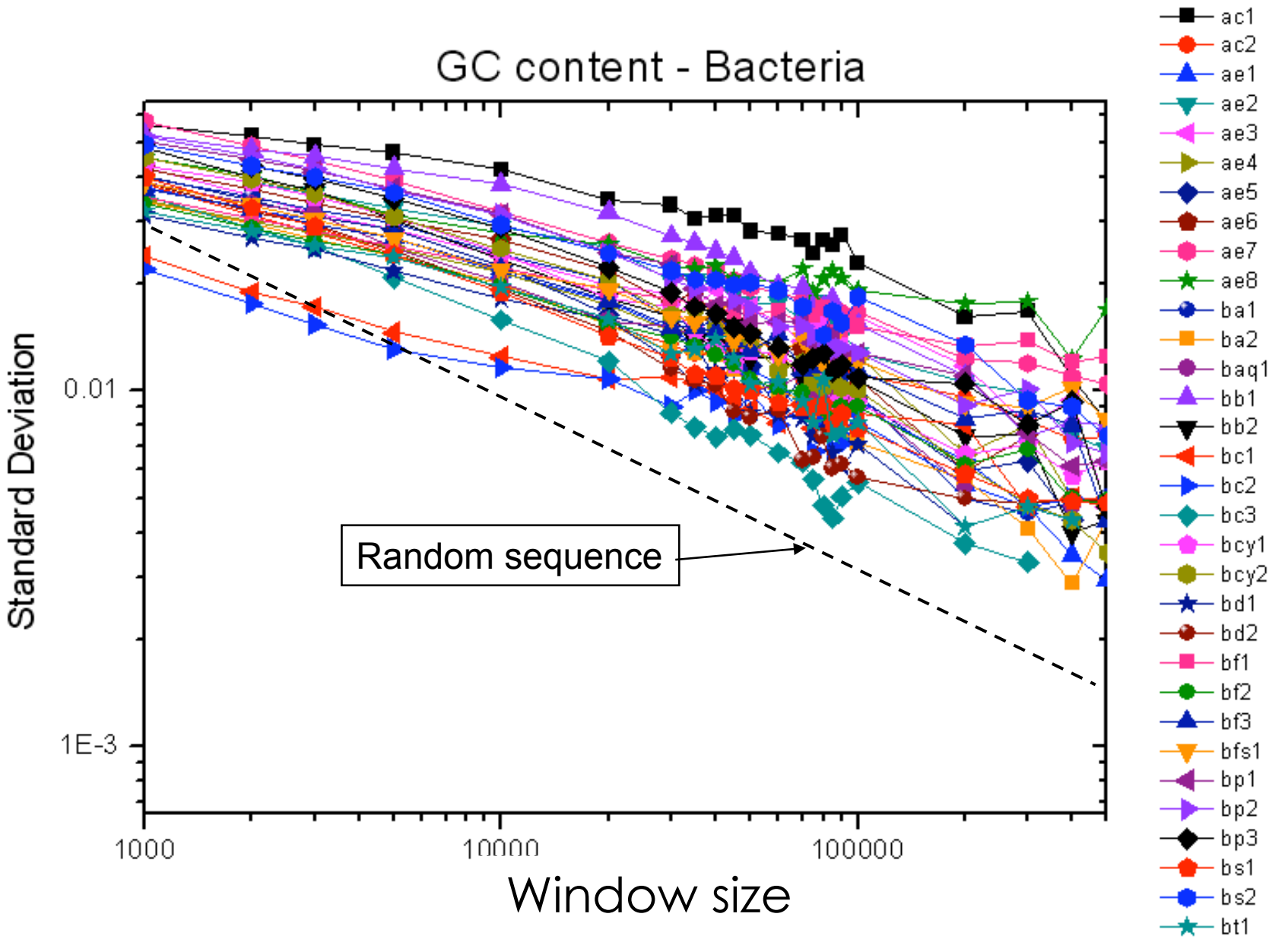
- RSD starts with very short sequence with GC content q (= 1-p)
- Duplicated segment length ~ 1000 b
- Segments randomly placed, therefore on scales greater than ~ 1000 b model sequence is a random system
- Cannot have correlation at lengths greater than 1000 b



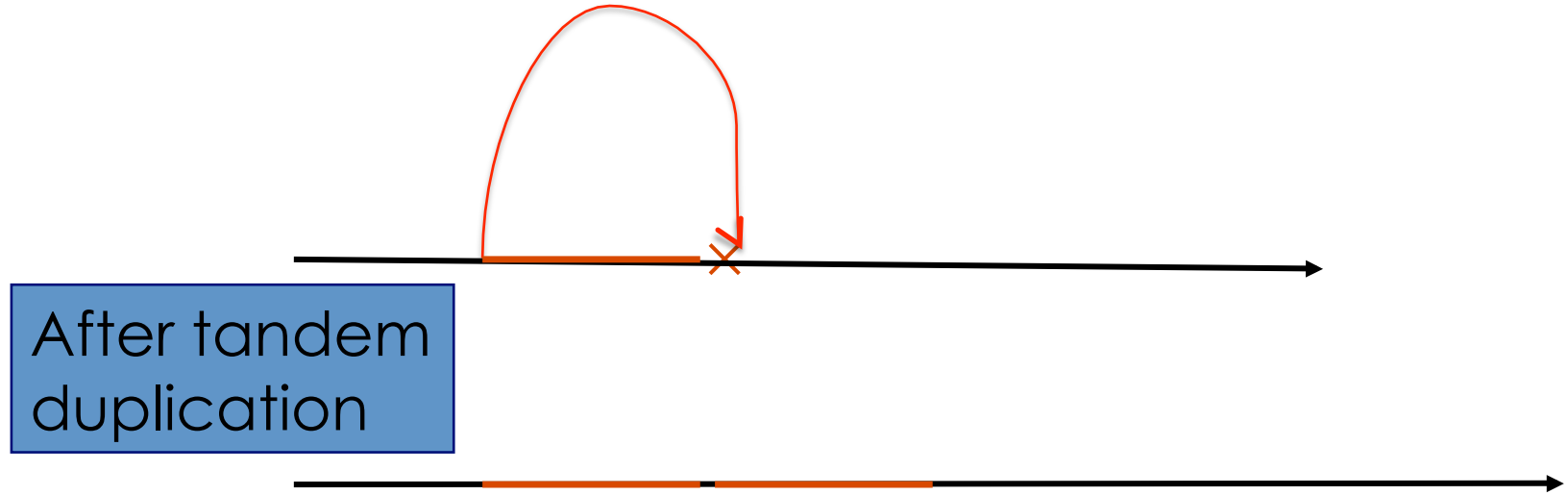
Power law is universal for complete eukaryotic genomes: γ=−0.06 to −0.42



And for bacteria: γ=-0.16 to -0.45



Tandem duplication

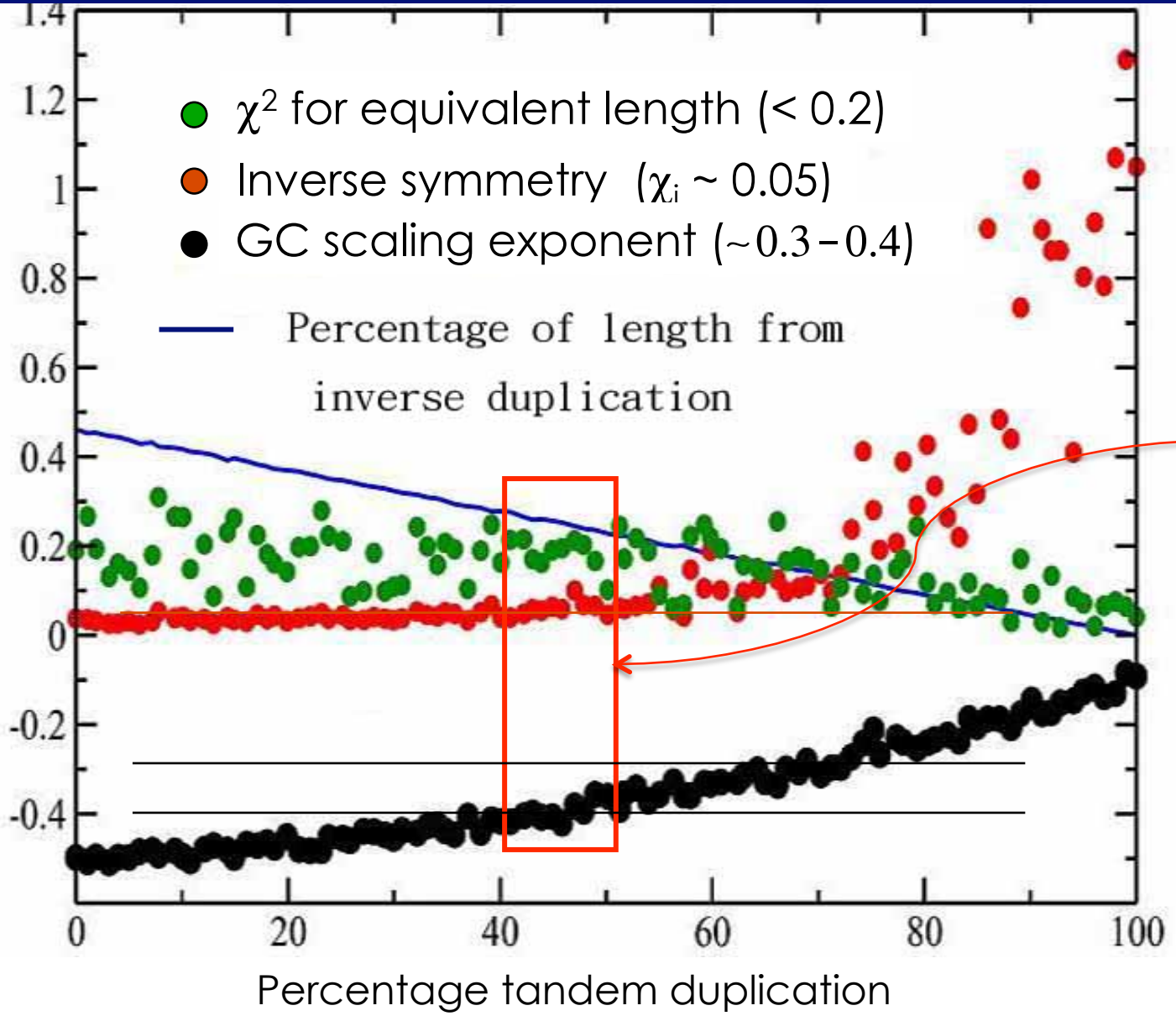


(Refer to extensive work by Lessig, (Peter) Ardnt group and early work by WT Li)

Can increase correlation length.

Proximal duplication probably biologically more viable

RSD+ Model: RSD plus inverse & tandem duplications



25-40% inverse duplication

40-50% tandem duplication

Genome the blind self-plagiarizer

- Mostly random, hence "blind"
- Random process is robust
- Self copying is the fastest and most efficient way to accumulation hard-to-come-by information, even if most of the time junk is copied
- The RNA world. If genome really started the RSD+ process when it was < 300 b, then RNA world. Copying machinery was composed of (proto) ribozymes that could be very small

Many biological phenomena explained by RSD+ model

- Preponderance of **homologous gene families** in all genomes
- Genome is full of **non-coding repeats**
- Transposons and operons
- Large-scale genome **re-arrangments**
- Rapid **speed of evolution** – RSD+ may be genome's way to "beat" the **2nd law of thermo-dynamics**
- Diversity of species
- Growth by random self-copying likely is the result of **natural selection**
- …

Some data on rates from human

- Data
  - Estimated silent site substitute rates for plants and animals range from 1 to 16 (/site/By) (Li97)
  - Humans: r_S ~2 (Lynch00) or 1 (Liu03) /site/By .
  - Animal gene duplication rate ~ 0.01 (0.002 to 0.02) per gene per Mya (Lynch00)
  - Human (coding region ~ 3% of genome) translates to 3.9/Mb/Mya.
  - Human retrotransposition event rate ~ 2.8/Mb/Mya (Liu03)
- Estimate rates for human

  $$\mu \sim 1\text{-}2 \text{ /site/Bya}, \quad r_D \sim 2.8\text{-}3.9 \text{/Mb/Mya}$$
- Human genome grew 15-20% last 50 My (Liu03)
- References
  - Lynch & Conery Science 290 (2000)
  - Liu (& Eichler) *et al*. Genome Res. 13 (2003)

Human genome growth rate

- Suppose per length per time growth rate is λ

$$\Delta L = \lambda L \Delta t$$

- Solution is $L_1/L_2 = e^{\lambda(t_1 - t_2)}$.
- t_1 ~ 3.4 Bya ago (bacteria-archaea+eukarya diverge time), t_2 current time, L_1 ~ 0.1 Mb, L_2 ~ 150 Mb (average human chromosome length)

$$\lambda = \ln(L_2/L_1)/(t_2 - t_1) \approx 2.2/By \quad \text{(human)}$$

- Implies human genome grew 13% in last 50 Mya (cf. 15-19%, Liu et al. 2003).

Human genome mutation rates

- If growth is mainly by SD, with rate r_D, then
- <d> (d bar) is average SD length

$$r_D = \lambda/\bar{d} = \begin{cases} 4.4/\text{Mb/My} & (\bar{d} = 0.5\ \text{kb}), \\ 2.2/\text{Mb/My} & (\bar{d} = 1.0\ \text{kb}). \end{cases}$$

(cf. r_D ~2.8-3.9/Mb/Mya, Liu, Lynch et al).

- Mutation rate $\mu \approx r\lambda/2$ (note factor ½), r is mutation density in RSD model

$$\mu_{hs} \approx 0.73 \times 2.2/2/\text{site/By} = 0.80 \times 10^{-3}/\text{site/By}$$

(cf. μ ~ 1/Mb/Bya (Liu (2003), Lynch (2000))

E. coli looks OK, too

- Suppose E. coli got to its current size 0.4 Bya after bacteria-archaea+eukarya diverge time (~ 3 Bya ago) then

$$\lambda_{ecoli} = \ln(4.6/0.1)/(0.4) \approx 9.6/By$$

$$\mu_{ecoli} \approx 0.73 \times 9.6/2/\text{site/By} = 3.5 \times 10^{-3}/\text{site/By}$$

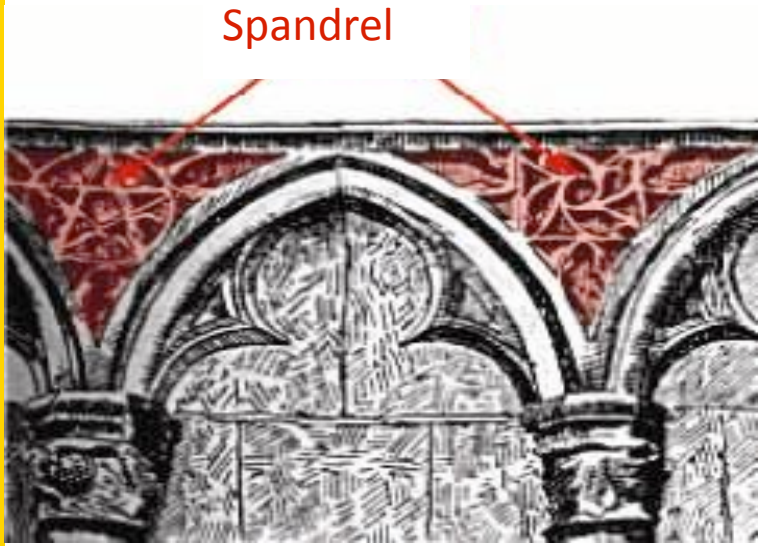
- μ_{ecoli} is about same as its current rate.

Trouble with bacteria

- Trouble. Suppose *E coli* sustained a low mutation rate of 1/site/Bya in the last 3 Bya, then accumulated mutation density of 3/site.
- Recall universal equivalent mutation density of one-half of $\mu_c = \frac{1}{4} \ln L$, or μ_{eq}=1.8/site (not a rate).
  - In RSD, 0.73/site from direct mutation, other 1.1/site from SD events.
- Would not allow many more, certainly not additional 3/site.
- There exist unknown mechanism that protects genome, especially its non-coding parts, from mutation events.

"Spandrels"

- Spandrels
  - In **architecture**. The roughly triangular space between an arch, a wall and the ceiling
  - In **evolution**. Major category of important evolutionary features that were originally side effects and did not arise as adaptations *(Gould and Lewontin 1979)*



- RSD events to genome are what the construction of arches, walls and ceilings (yielding spandrels) to a cathedral
- Genes and other codes are the décorations in the spandrels

Gradualism vs. Punctuated Equilibrium

- **Phyletic gradualism** - Evolution generally occurs uniformly and by the steady and gradual transformation of whole lineages (*Wikipedia*)
- **Punctuated Equilibria** - Model for discontinuous tempos of change (in) the process of speciation and the deployment of species in geological time (*Wikipedia*)
  - SJ Gould & N Eldredge (1977) *Paleobiology* **3** (2): 115-151. (p. 145); Eldredge & Gould (1972). "Punctuated equilibria: an alternative to phyletic gradualism" In TJM Schopf, ed., *Models in Paleobiology*. San Francisco: Freeman Cooper. pp. 82-115.
- RSD plus whole genome duplication provides a basis to accommodate a wide range (with power-law distribution) of tempos of change

End of Lecture Four

End of Lee Tutorials

People

- Dr. Li-Ching Hsieh, (WH Li Lab) Academia Sinica, Genome Research Center
- Dr. Wen-Lang Fan (WH Li Lab) Academia Sinica, Genome Research Center
- Dr. Yuan-Da Chen, He-Hsin Hospital Cancer Research Center
- PhD students (now PhDs) at Lee Lab
  - Hong-Da Chen
  - Sing-Guan Kong