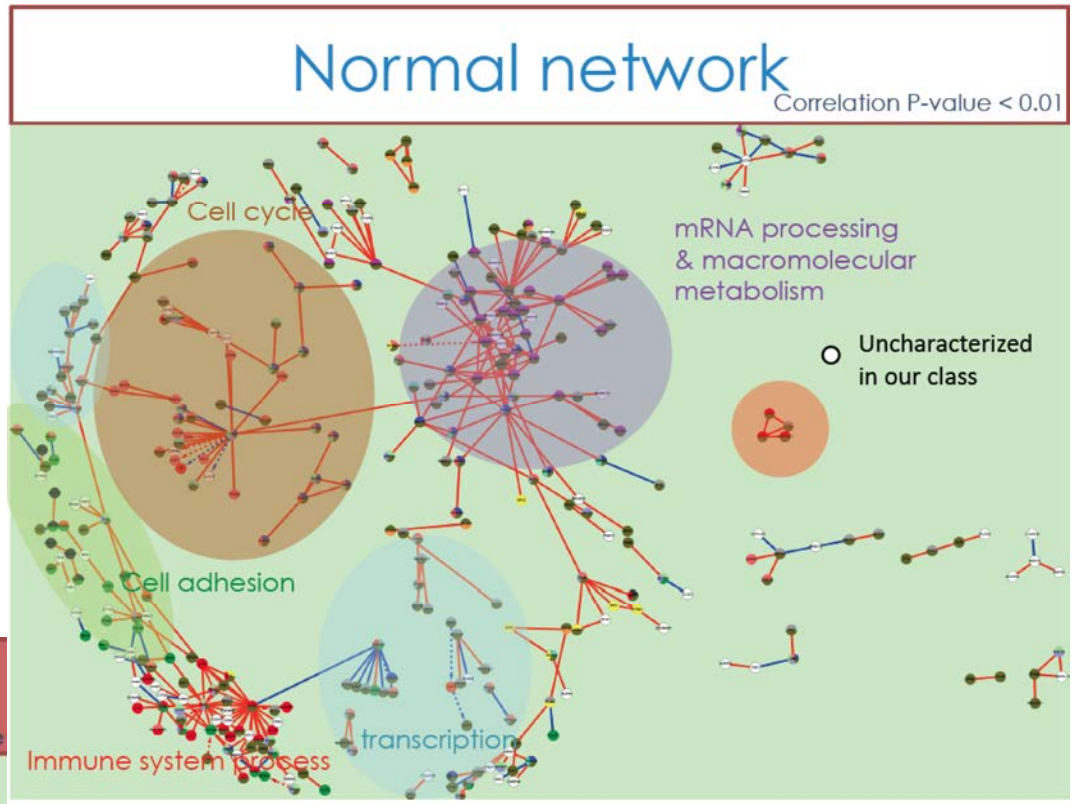
Quantitative Genome Evolution

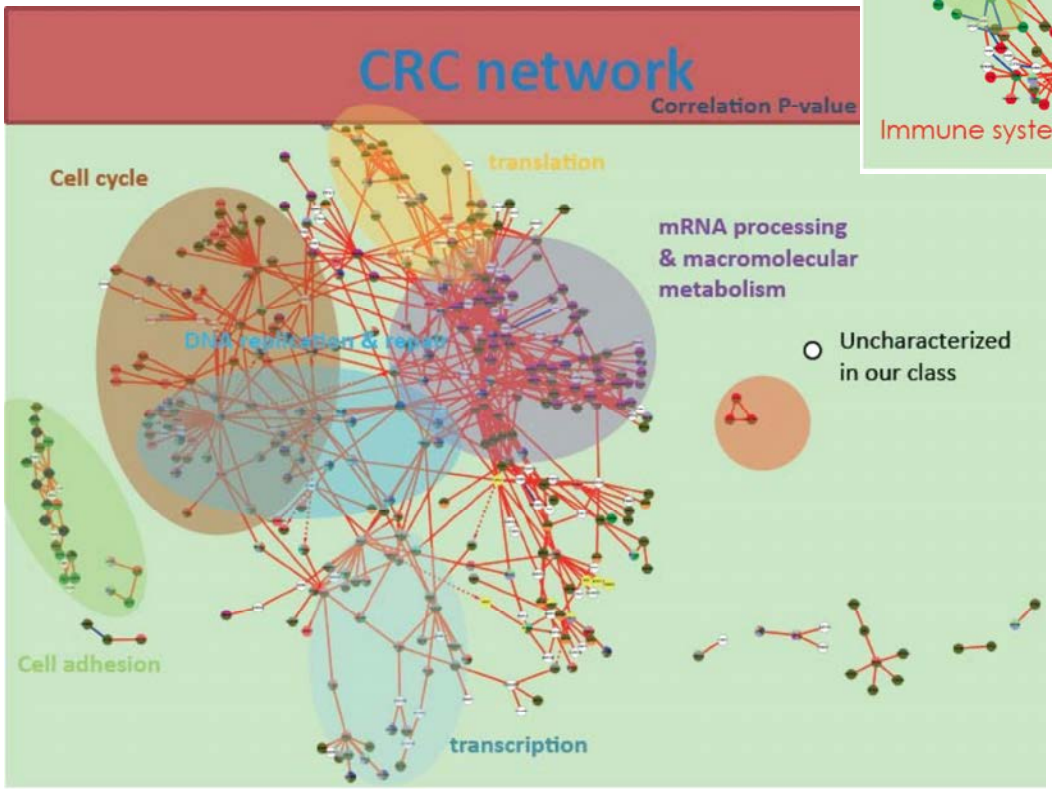
HC Lee
National Central University

Graduate Institute of Systems Biology
National Central University

Zhongli/Jhongli/Chungli, Taiwan

Enriched analysis of microarrays





Three questions I like to ask

- WHAT is the phenomenon?
  - What is important/unusual/interesting?
- HOW did it happen?
  - (Physics)
- WHY did it happen?
  - (Biology)

What we plan to discuss

How did life evolve so rapidly • random sequence • statistical properties of the genome • uniformity homogeneity and universality • order of genome • genome the information carrier • genome at the "edge of chaos" • symmetries in the genome • coding versus non-coding regions • how genome grew • segmental & whole genome duplication • genome the blind self-plagiarizer • genome in a state of fixed-point • the nearly universal cumulative point mutation density • age of the genome • the mystery of the "missing" mutations • the phyletic gradualism versus punctuated equilibria debate

Motivation: Understand large scale growth and evolution of genome; how did it evolve so rapidly

Materials: Complete genome of prokaryotes and eukaryote

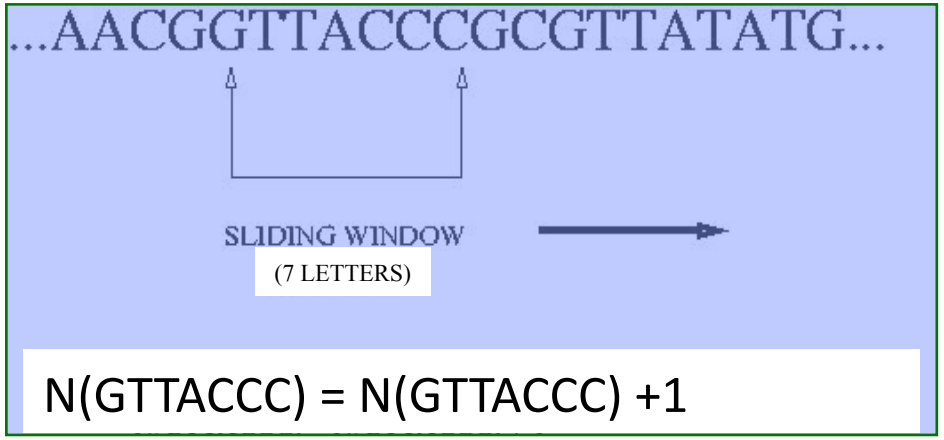
Method: Variety of statistical analysis of frequency of oligonucleotides (k-mers)

Sequence alignment versus statistical analysis

Feature	Seq alignment	Stats analysis
Nature of quantity measured	relative	absolute
Sequence length	< 10k	> 5k
Sequence comparison relative word in order word content	 excellent poor	 poor moderate
Positional information local, relative large scale structure	 excellent poor	 poor good
Correlation short range long range	 good poor	 poor good
(Absolute) **Order of sequence**	---	excellent
Evolution time scale < 100 Mya (divergence time) >> 100 Mya (growth time)	 good ---	 --- good
Mode of genome growth	---	good

Methods

- Treat genome as a 4-letter pseudo-text
- Count frequency of occurrence of "words"
  - A *k*-letter word is called a "*k*-mer"
- Number of possible *k*-mers is 4^k
- Use a sliding window of width *k* and slide 1
- Analyze set of *k*-mer frequencies



After a single sweep have, for each k, a 4^k-component vector $\{f_i \mid i=4^k\}$ whose element f_i is the frequency of the *i*th *k*-mer.

Frequency set, *k*-spectrum & relative spectral width

Given freq. set $\{f_i\}$, define

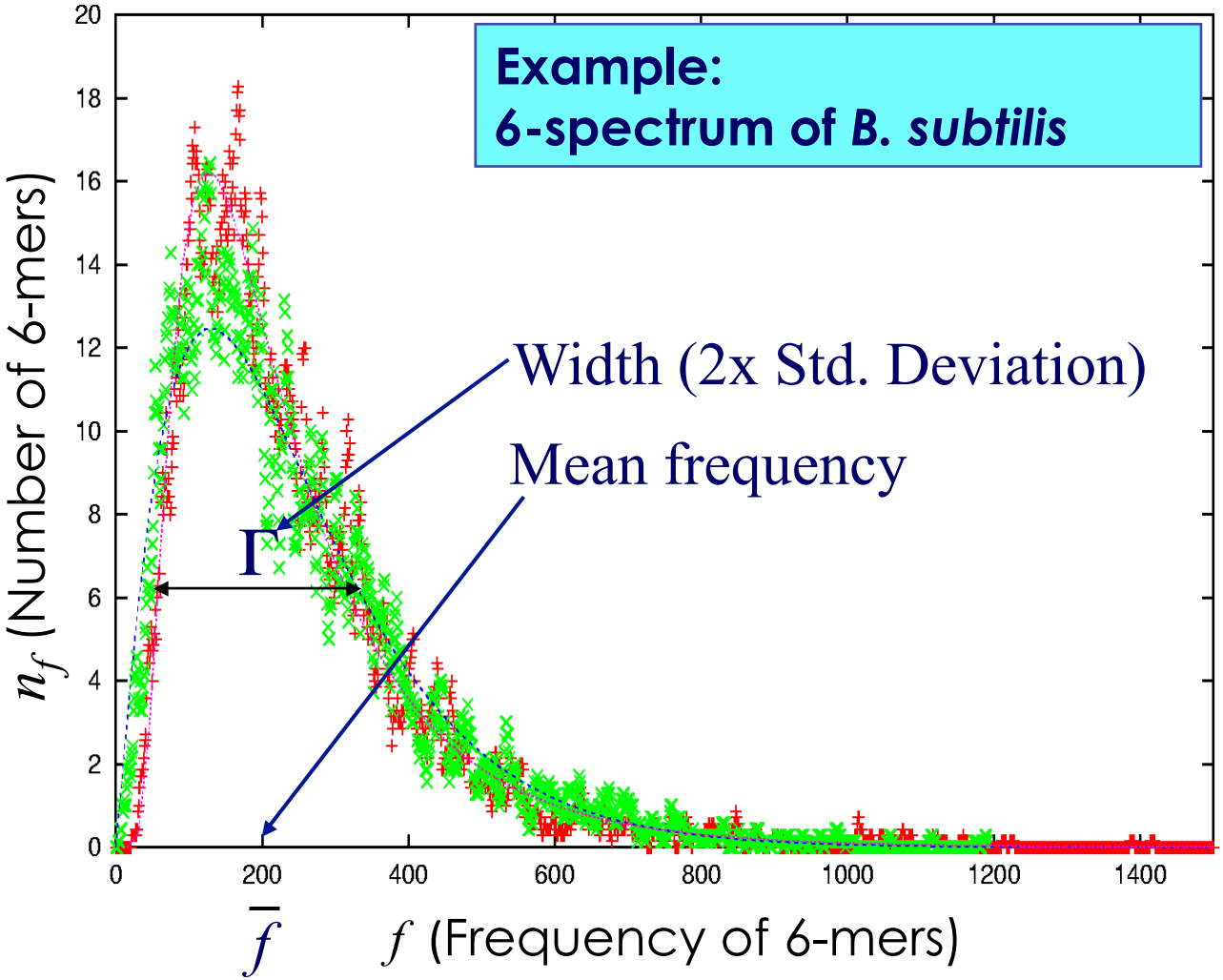
***k*-spectrum**

$\{n_f \mid f=1,2,\ldots\}$

$\Sigma_i f_i = \Sigma_n f\, n_f$

Relative spectral width

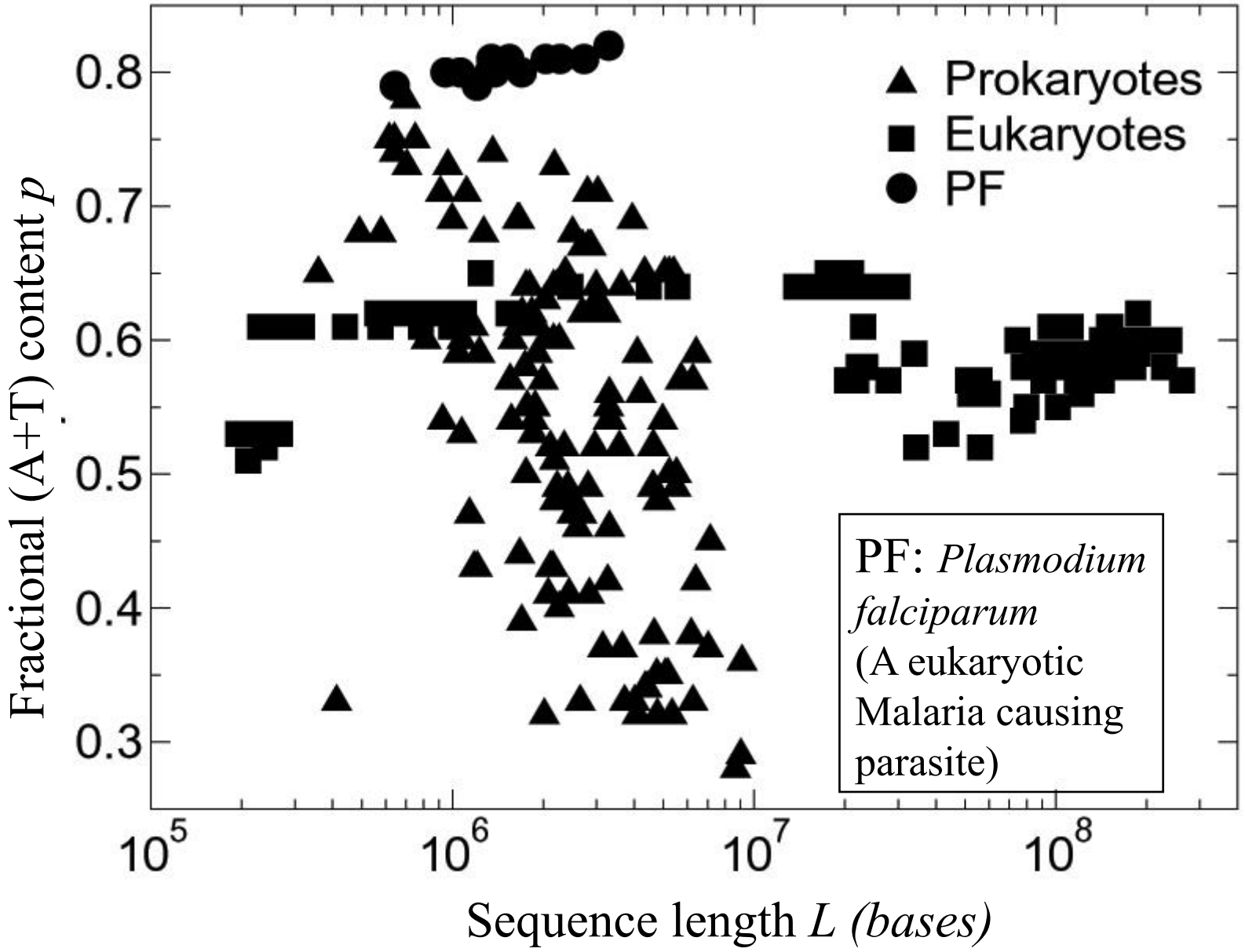
σ = std dev/<f>



Data set

- All sequences downloaded from *GenBank* (Nov 2006)
- 422 complete *prokaryotic* ("bacterial") genomes
  - Lengths 0.5 to 9 M base pairs (Mb)
- 19 complete *eukaryotic* genomes totaling 328 complete chromosomes
  - Lengths 0.2 to 250 Mb
- Total data set: 2.2×10^{10} base pairs

Complete Genomes are diverse



Four lectures

Lecture One. Genome: A Large System with Small System Statistics

Lecture Two. Genome: Information Carrier at the "Edge of Chaos"

Lecture Three. Symmetries in Genomes

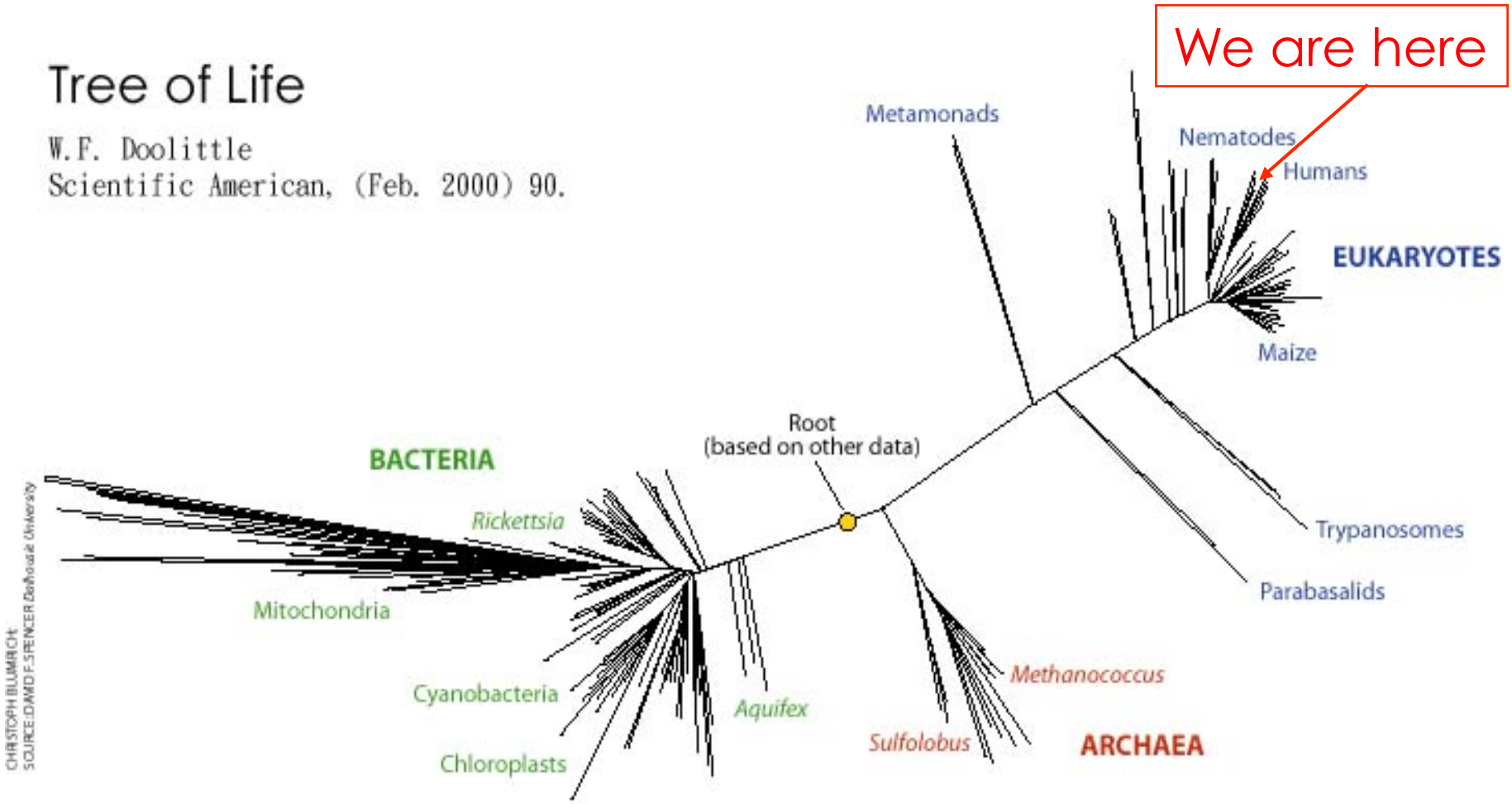
Lecture Four. Genome the Blind Self-Plagiarizer and the Mystery of the "Missing" Mutations

Lecture One

Genome: A large System with Small-System Statistics

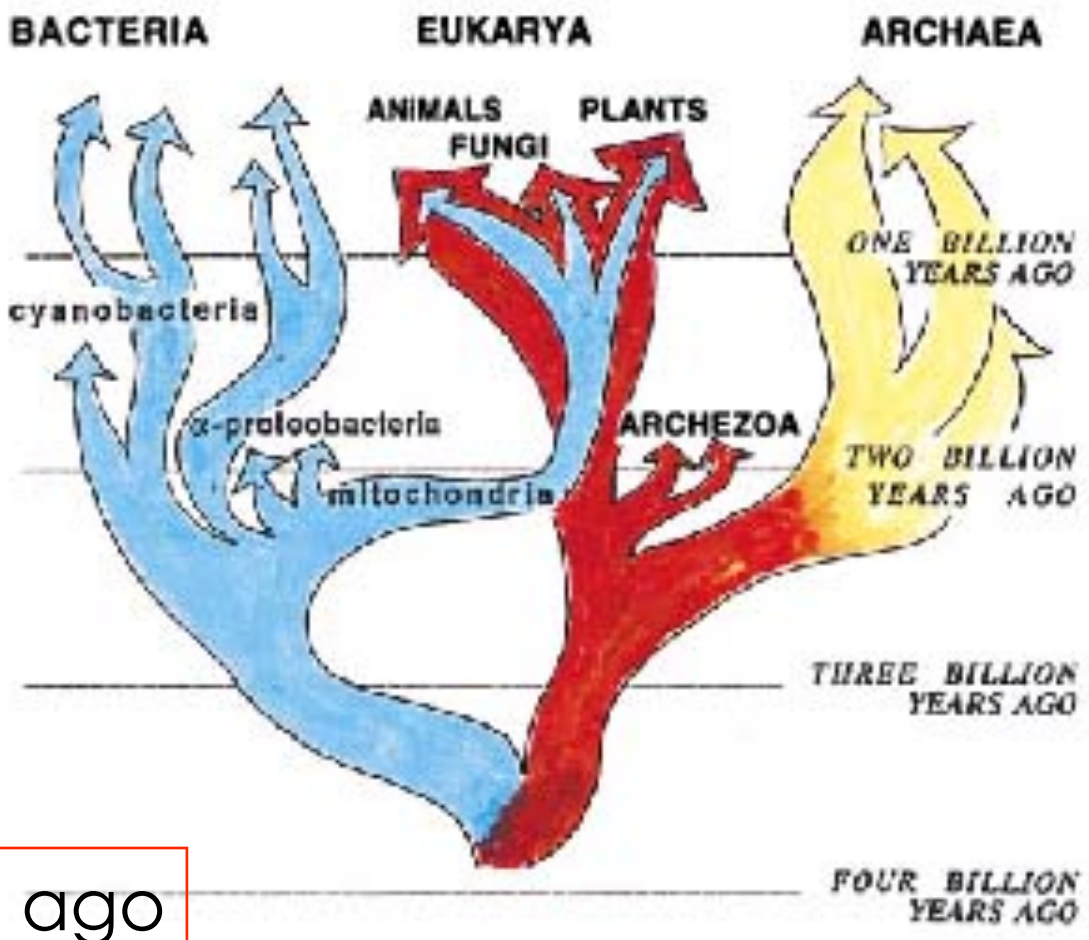
HC Lee
National Central University

Life is highly diverse and complex



And it took a long time to get here

Divergence of species
W.F. Doolittle, PNAS 94 (1997) 12751.



now

4 billion yrs ago

Life evolved super rapidly

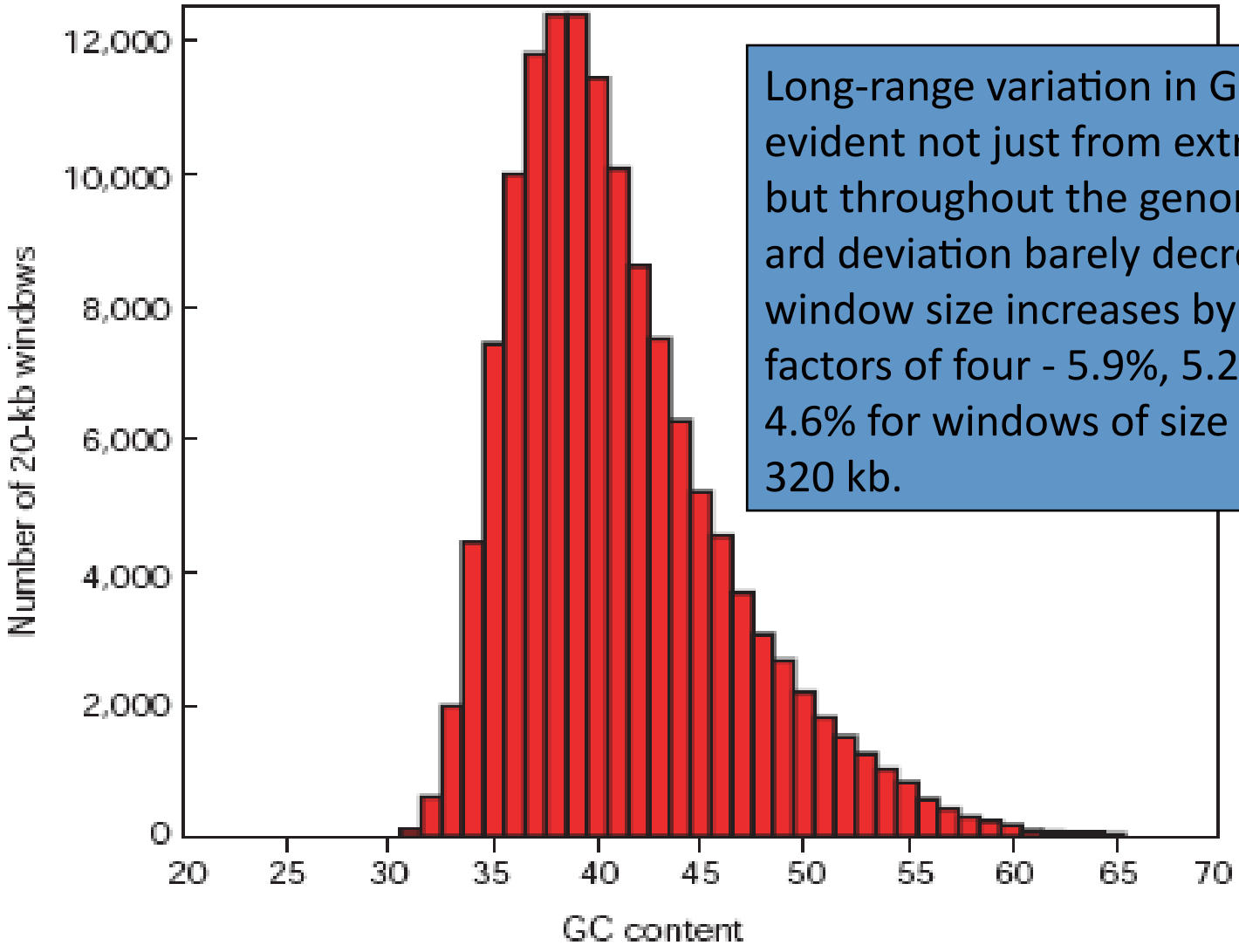
- Typical mutation rate is 1/site/Byr
- Life in the form of microbe existed less than 1 Byr after cooling of earth
- Suppose we were "given" a genome 1000 bp long at dawn of life
- After 1 Byr every site mutated once
- Size of sequence space is $4^{1000} \sim 10^{600}$
- Evolution could not have been driven by point mutation

Evolution of Genomes and the Second Law of Thermodynamics

- **Genomes grew & evolved stochastically**
  - modulated by natural selection
  - Bigger genomes carry more information than smaller ones
- **The second law of thermodynamics**:
  - the entropy of closed system can never decrease
  - a system that grows stochastically tends to acquire entropy
  - Increased randomness ⟶ more entropy
- **Shannon information**
  - Information decreases with increasing entropy

NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com

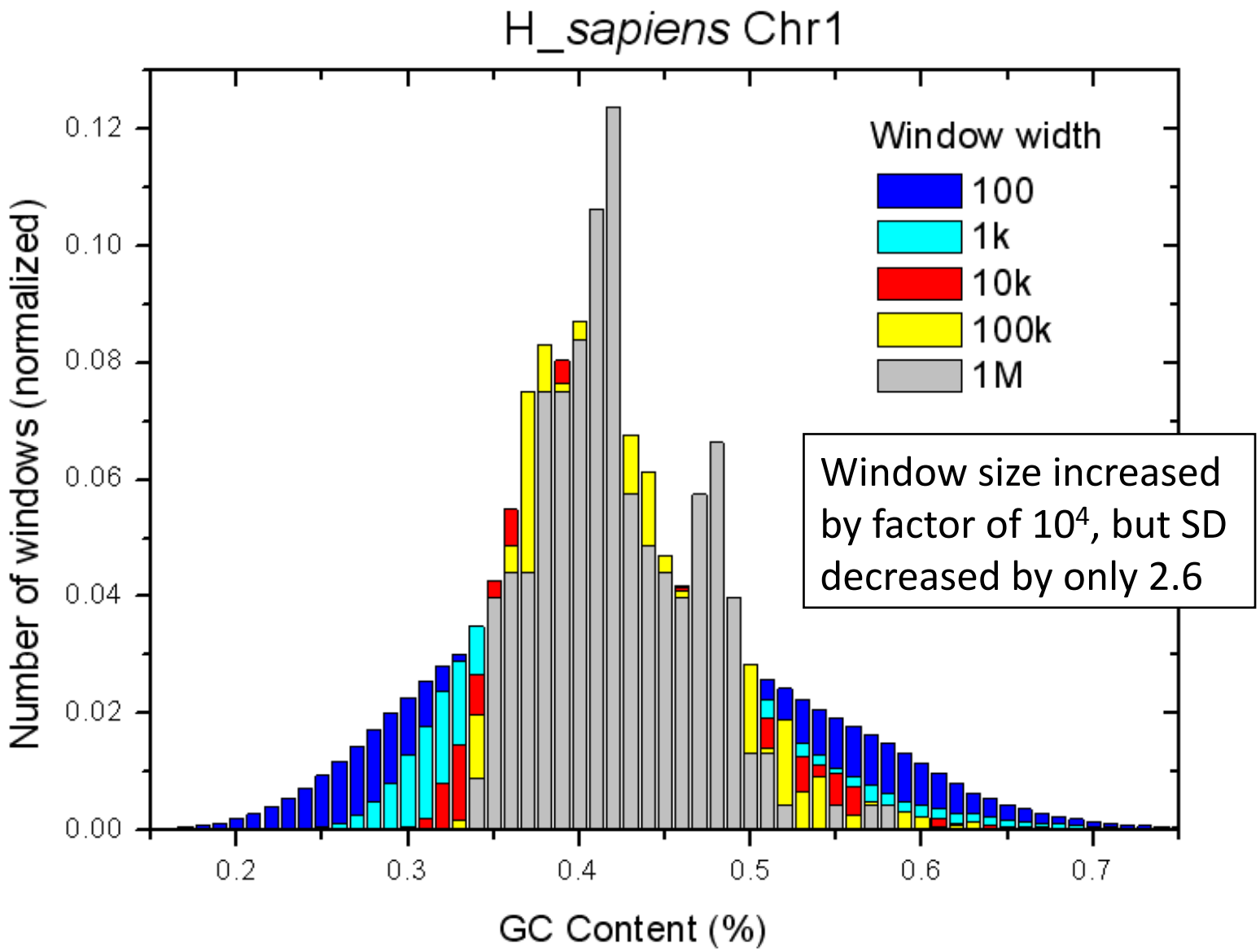
Long-range variation in GC content



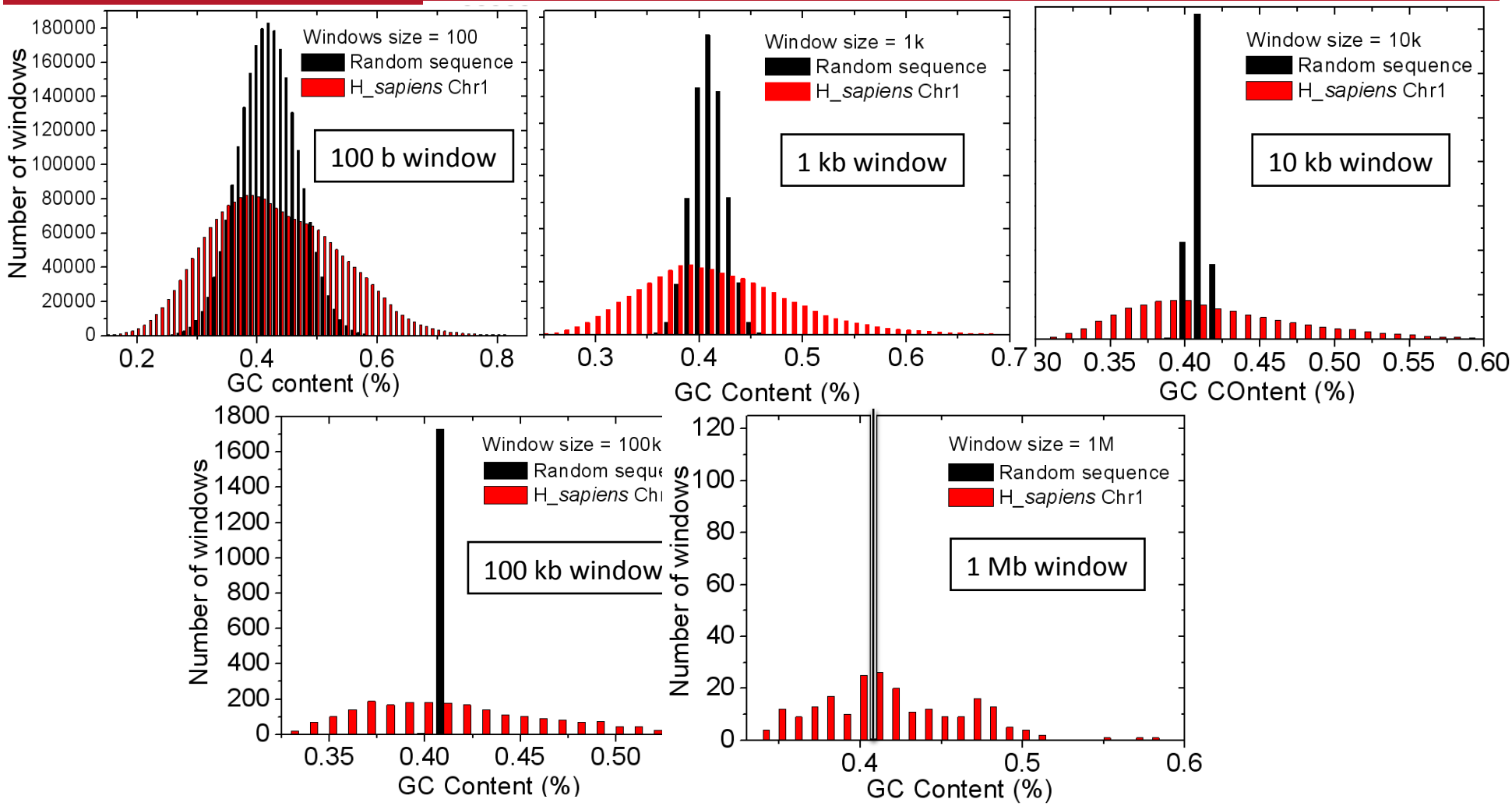
Long-range variation in GC content is evident not just from extreme outliers, but throughout the genome…the standard deviation barely decreases as window size increases by successive factors of four - 5.9%, 5.2%, 4.9% and 4.6% for windows of size 5, 20, 80 and 320 kb.

Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.

CG-content in Human genome has long-range variation

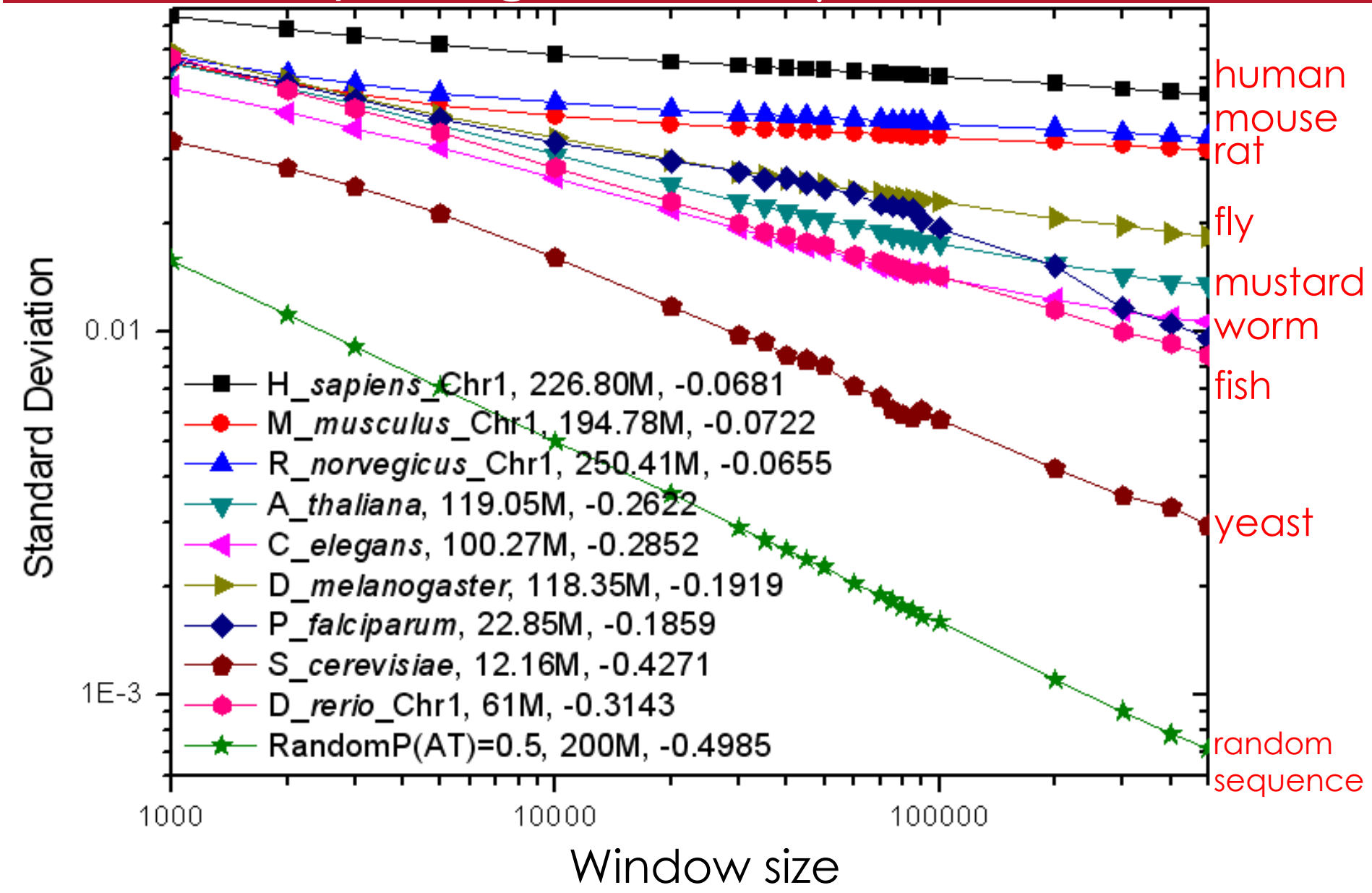


For GC-content histograms: sample size = window size



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

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



Distribution and Width

- Consider τ equally probable events occurring a total of L times.

- Distribution of occurrence frequency characterized by
  - mean frequency: $f_{ave} = L/\tau$
  - SD (standard deviation) Δ; or

    CV (coefficient of variation) = Δ/f_{ave}
  - Higher moments of distribution

Random events

- Random events given by Poisson distribution
  - $\Delta^2 = f_{ave}$, or, $(CV)^2 = 1/f_{ave}$
  - That is, $(CV)^2 = \tau/L$

- For fixed τ, $(CV)^2 \sim 1/L$
  - Large L limit (thermodynamic limit): $L \sim$ infinity, $CV \sim 0$

- For given τ, if CV is known, then
  - $L \sim \tau/(CV)^2$

Frequency set, "*k*-spectrum" & relative spectral width

Given freq. set $\{f_i\}$, define

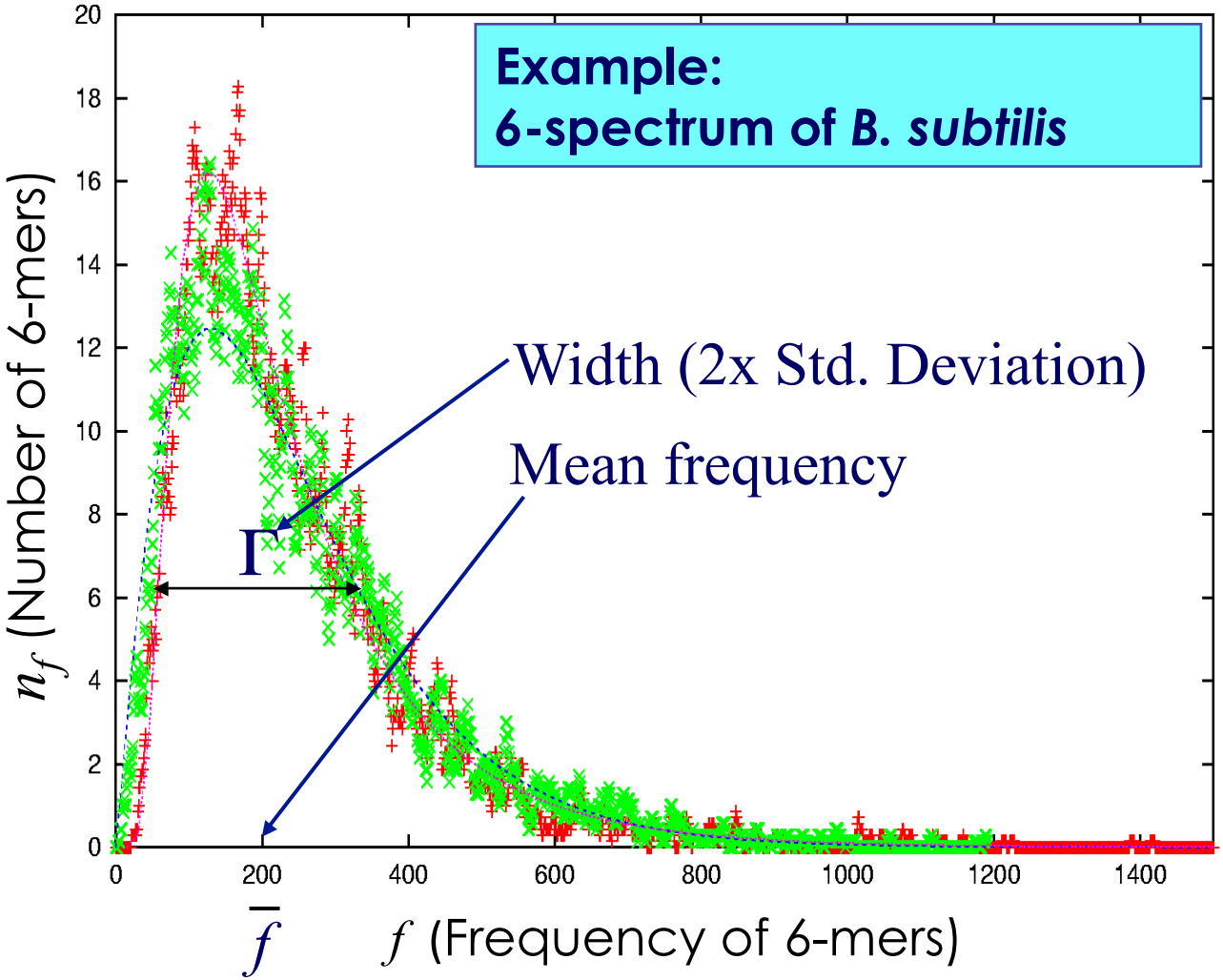
***k*-spectrum**

$\{n_f \mid f=1,2,\ldots\}$

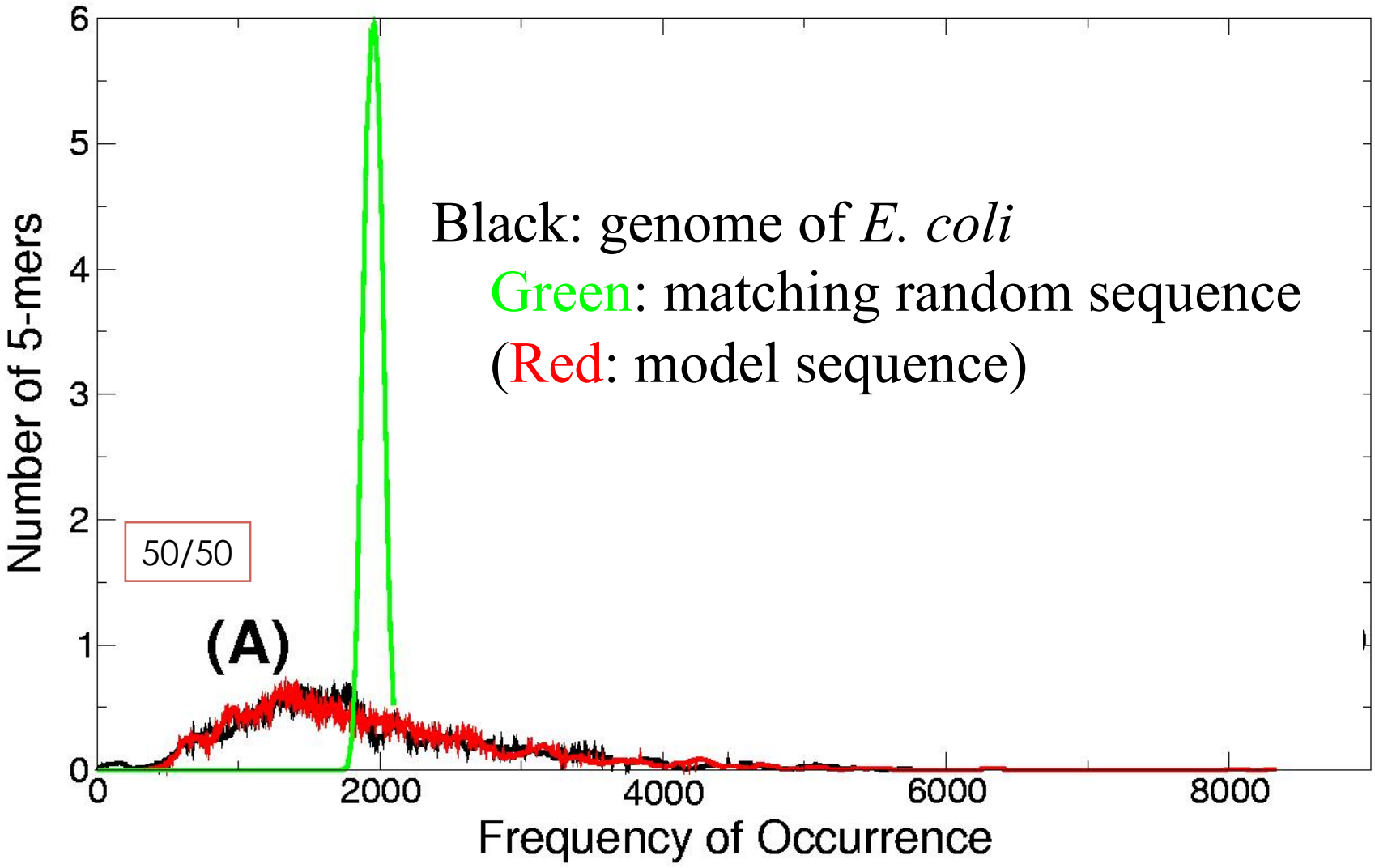
$\Sigma_i f_i = \Sigma_n f\, n_f$

Relative spectral width

σ = std dev/<f>



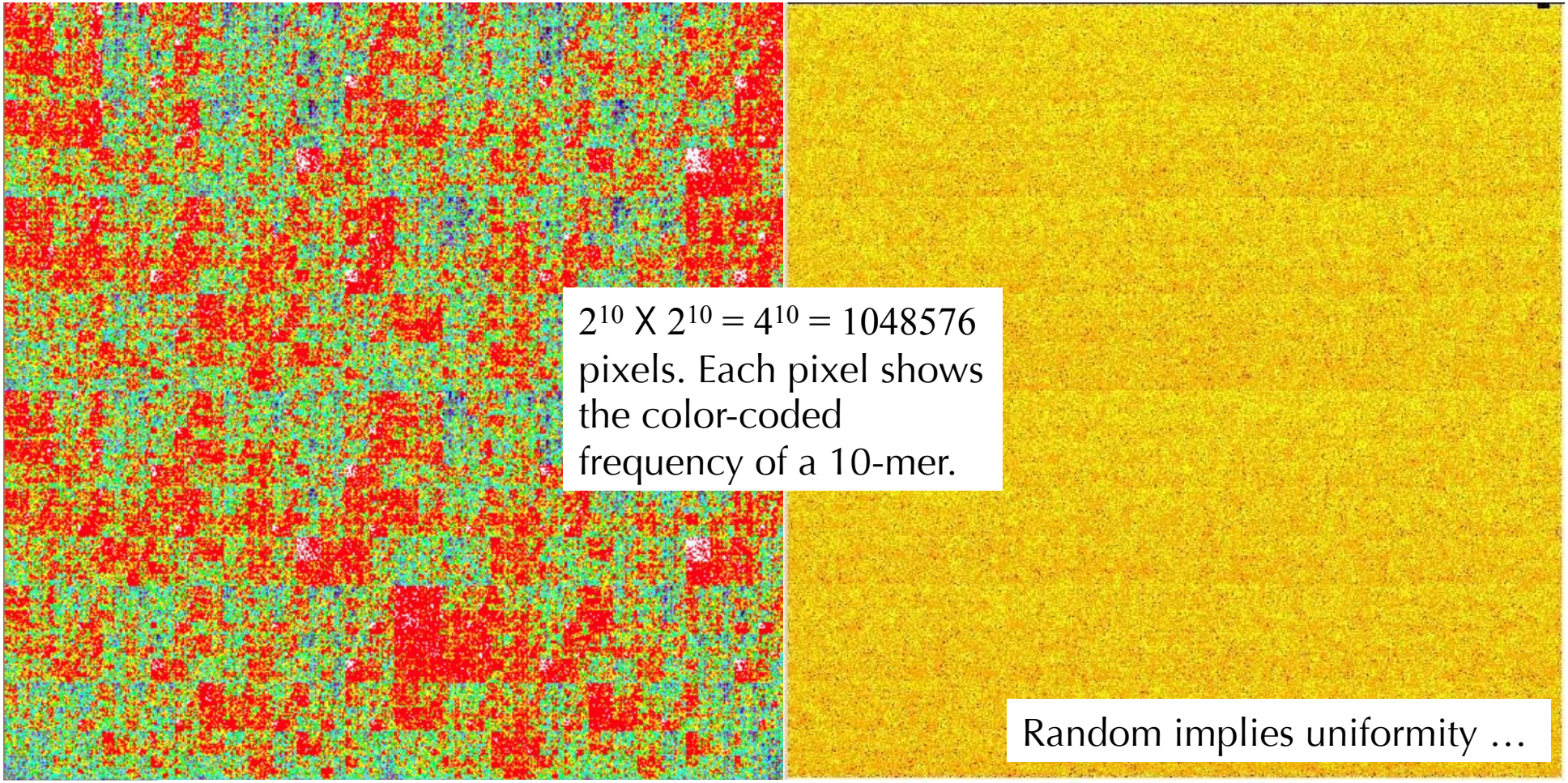
Huge difference between genomes and random sequences



2D "portrait" of *E. coli* genome

Frequencies of 10-mers *in E. coli*

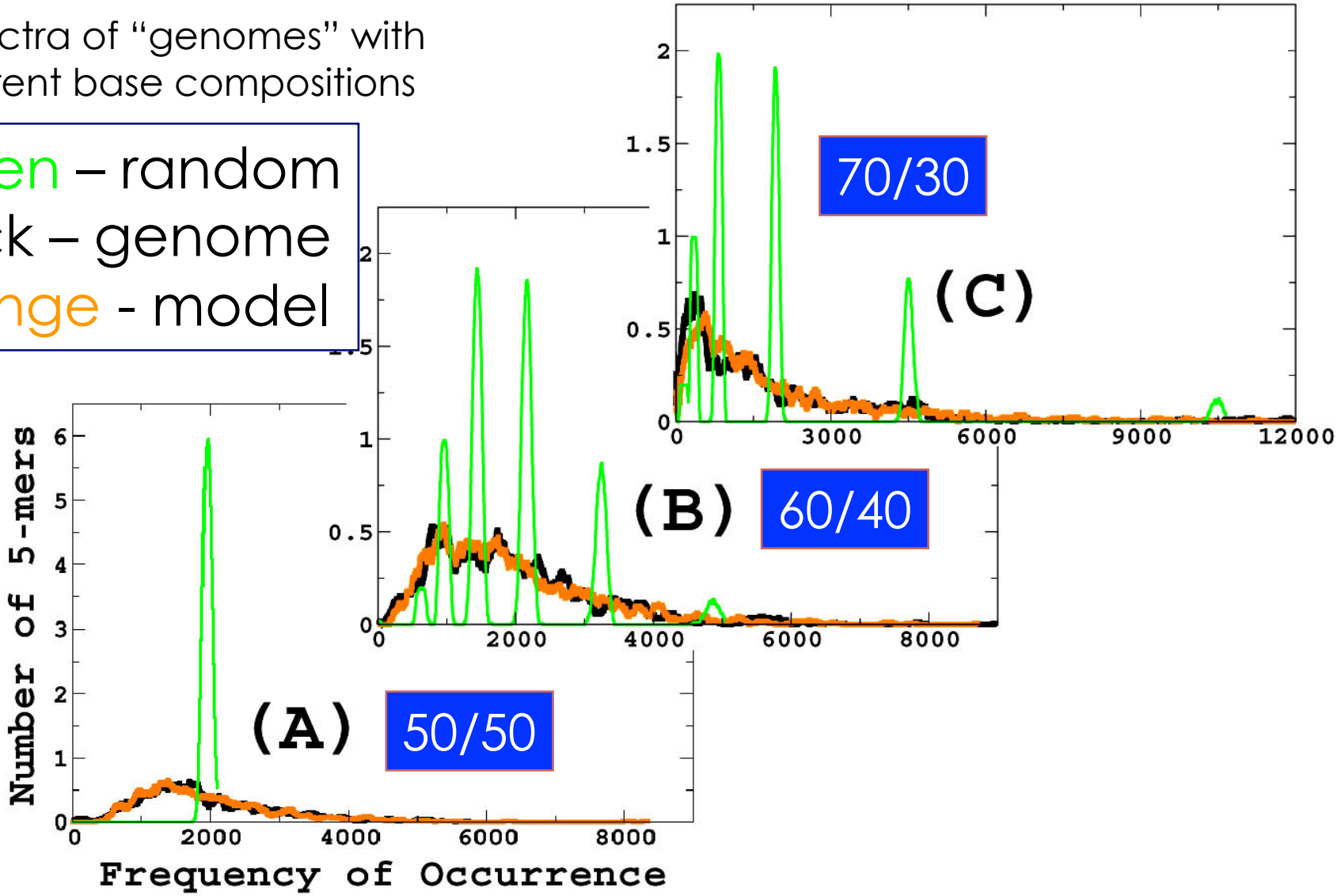
E. coli randomized



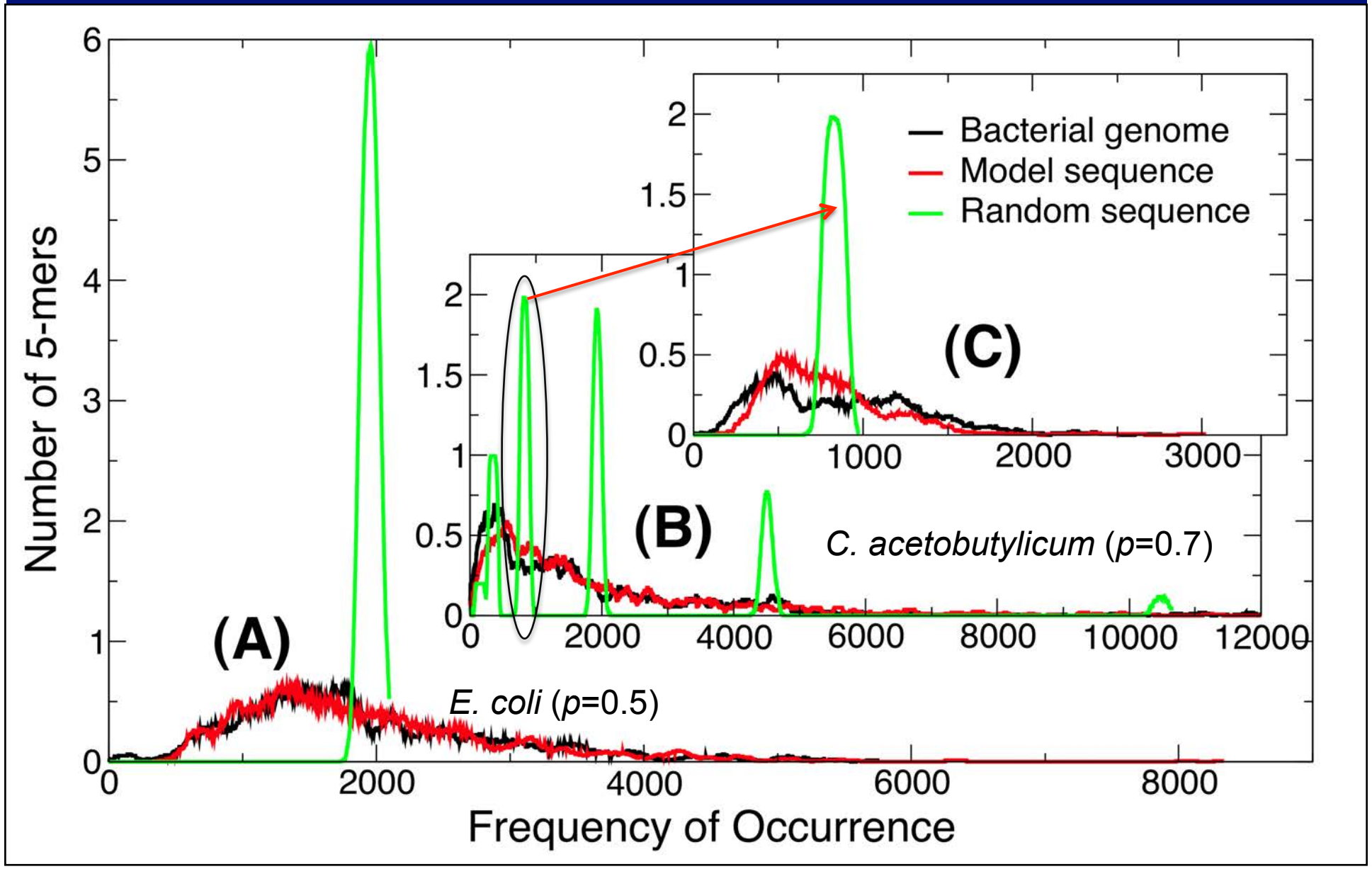
Random implies uniformity; but sometimes other factors intervene

5-spectra of "genomes" with different base compositions

Green – random
Black – genome
Orange - model



k-spectrum is composed of (*k*+1) sub-distribution of "*m*-sets"



Partition of *k*-mers into *m*-sets

- Consider genome with fractional AT-content *p* (then fractional GC-content *q=1-p*
- There will be more AT-rich words than GC-rich words
- Partition *k*-mers into sets, called *m*-sets, *m=0,1,…,k; k*-mers in an *m*-set all have *m* ATs.
- Total number of:
  - kinds of *k*-mers: $\tau = 4^k$
  - *k*-mers: *L* (sequence length); mean frequency is L/τ
  - kinds of *k*-mers in m-set: τ_m
  - *k*-mers in an m-set: L_m ; mean frequency is $f_m = L_m/\tau_m$

$$\tau_m = \binom{k}{m} 2^k, \quad L_m^{\{\infty\}} = L\binom{k}{m} p^m q^{k-m},$$

$$\bar{f} = L/\tau, \quad \tau = 4^k \qquad \bar{f}_m^{\{\infty\}} = L_m^{\{\infty\}}/\tau_m$$

f_m well approximated by its large-L limit

(k=5)	f_m					
Sequence	**(m=) 0**	**1**	**2**	**3**	**4**	**5**
p=0.492						
E. coli	2509	2245	1877	1760	1944	2656
Random	2101	2044	1987	1922	1857	1795
$\lim_{L\to\infty}$ Random*	2114	2048	1983	1920	1860	1801
p=0.691						
C. acetobutylicum	154	397	918	1951	4272	10300
Random	176	394	882	1970	4400	9832
$\lim_{L\to\infty}$ Random*	176	393	880	1968	4402	9845

All sequences normalized to a length of 2 Mb; $\bar{f} = 2\times 10^6/4^5 = 1953$. Random means matching random sequence, or sequence obtained by scrambling the genome. *Values of $\bar{f}_m^{(\infty)}$ given by Eq. (6).

doi:10.1371/journal.pone.0009844.t006

Decomposition of standard deviation into "non-fluctuating" and "fluctuating" parts

$$\sigma^2 = \tau^{-1} \sum_{u \in S} (f_u - \bar{f})^2 = \tau^{-1} \sum_{m=0}^{k} \sum_{u \in S_m} (f_u - \bar{f}_m + \bar{f}_m - \bar{f})^2$$

$$= \tau^{-1} \sum_{m=0}^{k} \left(\tau_m (\bar{f}_m - \bar{f})^2 + 2(\bar{f}_m - \bar{f}) \sum_{u \in S_m} (f_u - \bar{f}_m) + \sum_{u \in S_m} (f_u - \bar{f}_m)^2 \right). \quad (7)$$

$$\sigma^2 \equiv {\sigma_{nf}}^2 + {\sigma_{fl}}^2,$$

(generalization of parallel axis theorem)

non-fluctuating part

$${\sigma_{nf}}^2 = \sum_{m=0}^{k} \frac{\tau_m}{\tau} (\bar{f}_m - \bar{f})^2,$$

(depends only on averages)

fluctuating part

$${\sigma_{fl}}^2 = \sum_{m=0}^{k} \sum_{u \in S_m} \frac{(f_u - \bar{f}_m)^2}{\tau} \equiv \sum_{m=0}^{k} \frac{\tau_m}{\tau} \sigma_{m,fl}^2. \quad (10)$$

Large-*L* limit of *CV*

Large-*L* limit for non-fluctuating part is known (for all sequences)

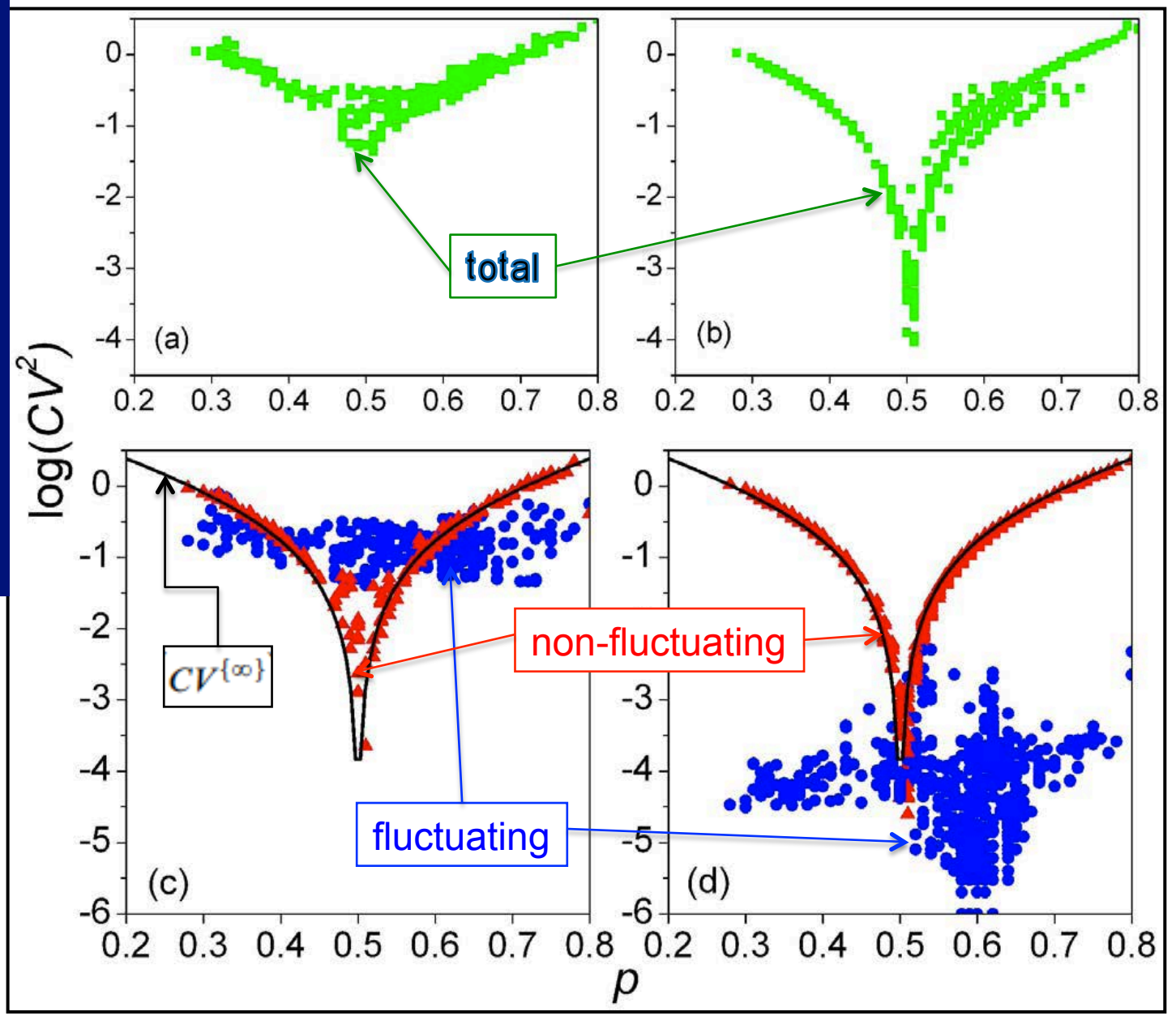
$$\left(CV^{\{\infty\}}\right)^2 \equiv \lim_{L\to\infty} CV_{nf}{}^2 = \sum_{m=0}^{k} 2^{-k}\binom{k}{m}\left(2^k p^m q^{k-m}-1\right)^2 \qquad (12)$$
$$= 2^k(p^2+q^2)^k - 1,$$

So is large-*L* limit for fluctuating part for random sequences

$$\lim_{L\to\infty} CV_{fl}{}^2 = \frac{1}{\bar{f}^2}\lim_{L\to\infty}\sigma_{fl}^2 = \frac{1}{\bar{f}^2}\sum_m \frac{\tau_m}{\tau}\bar{f}_m \qquad (13)$$
$$= \frac{1}{\bar{f}} = \frac{\tau}{L} \qquad \text{(random sequence)},$$

For genome-size random sequence, $CV_{nf} >> CV_{fl}$

Genome and random sequ's differ only in CV_{fl}



Equivalent length of a sequence

Because $CV_{fl} \sim \tau/L$, we define for any sequence an equivalent length l_e

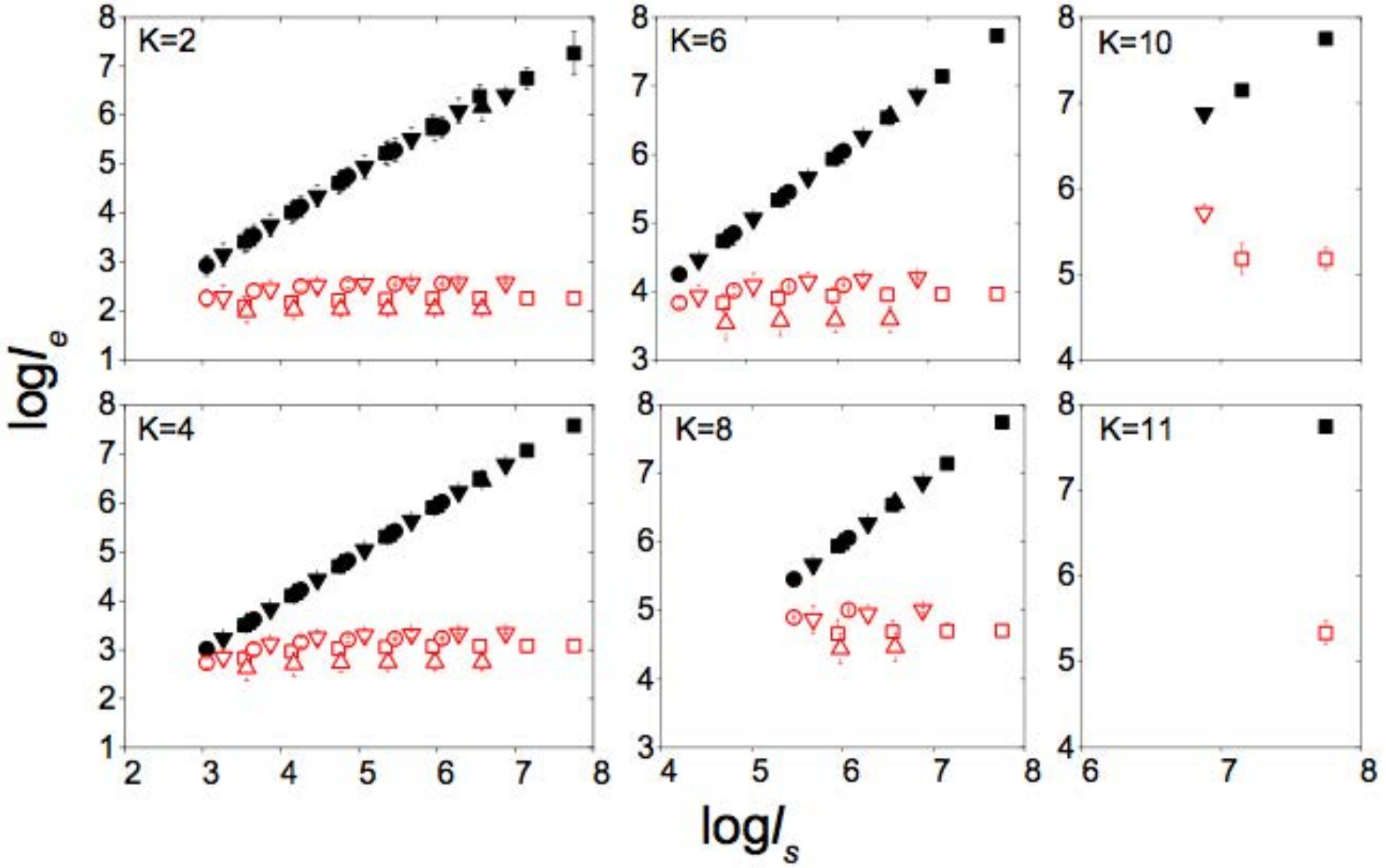
$$l_e = b_k \tau / CV_{fl}^2 \qquad (\text{for } k \geq 2) \qquad (14)$$

[$b_k = 1 - 1/2^{k-1}$; =1/2 for k=2 and approaches 1 quickly for larger k's]

(note: for random sequence $L = l_e$). The l_e of a sequence is the length of the random sequence whose CV_{fl} is the same as that of the sequence.

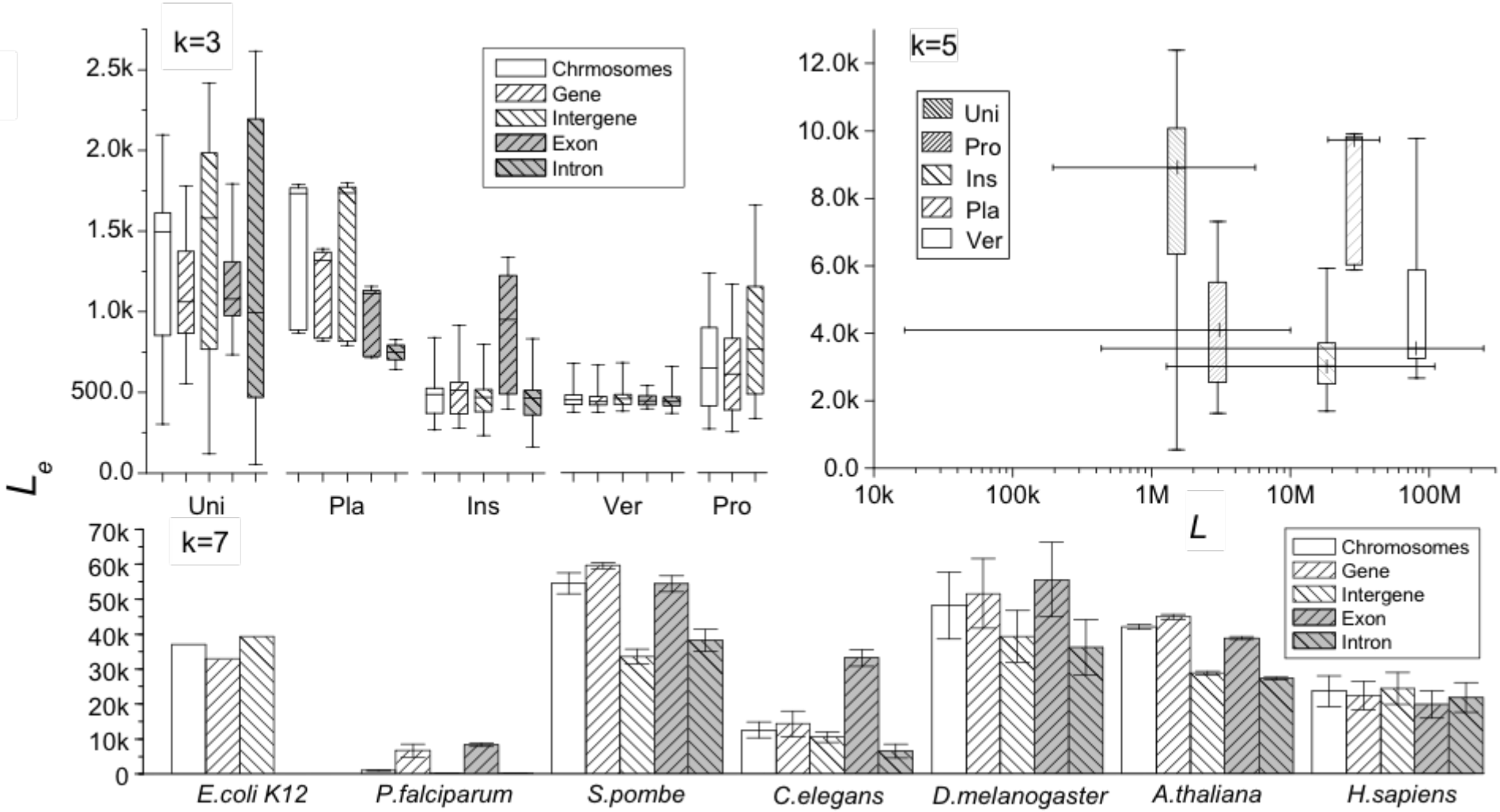
Notation: use l_e for a segment, and L_e for a whole genome/chromosome

Genomic (*E. coli*, worm, mustard and human) l_e does not grow with sequence length



Red: segments from genomes. Black: segments from random sequences.

L_e of coding and non-coding parts not much different (I)



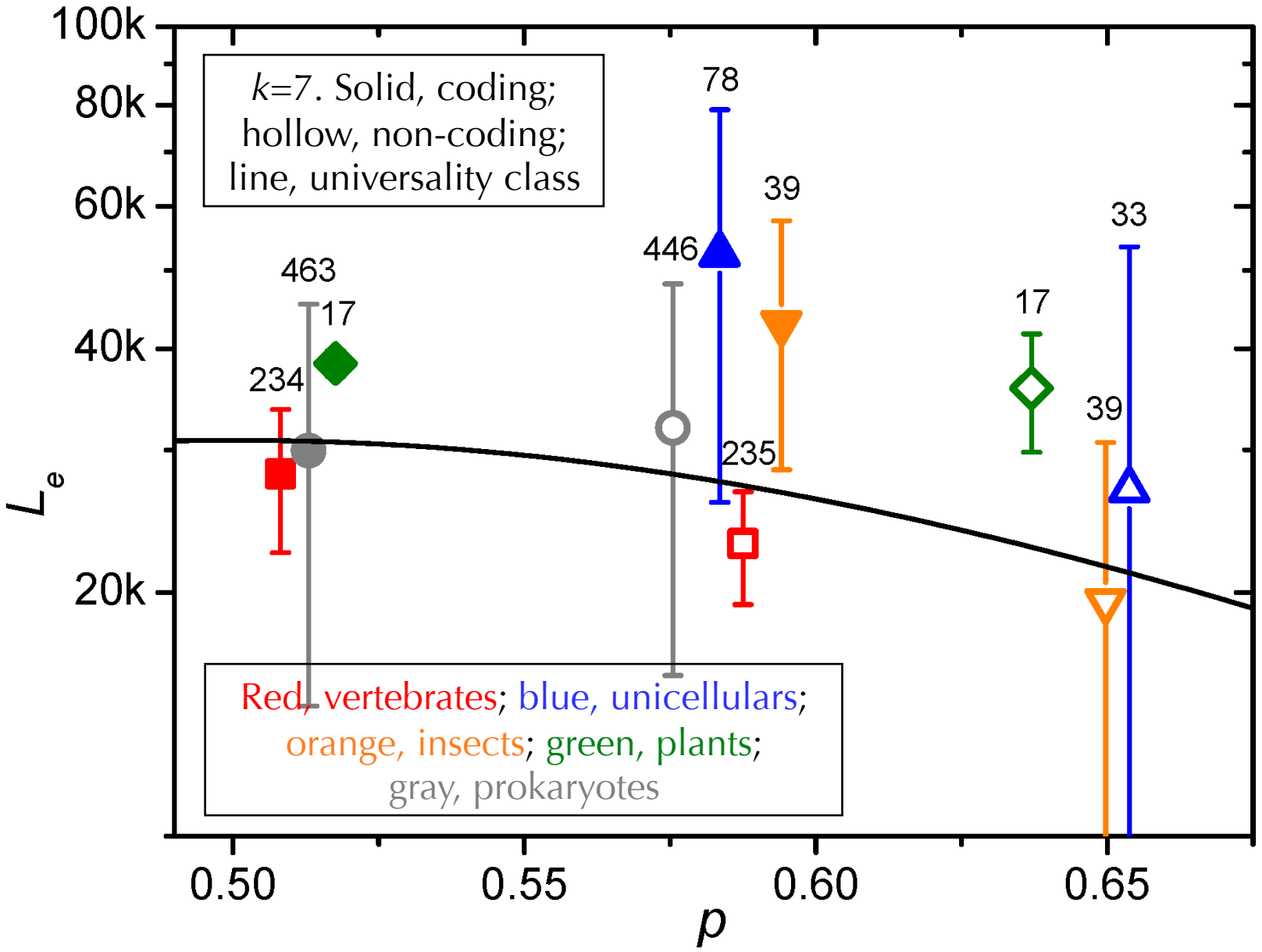
L_e of coding and non-coding parts not much different (II)

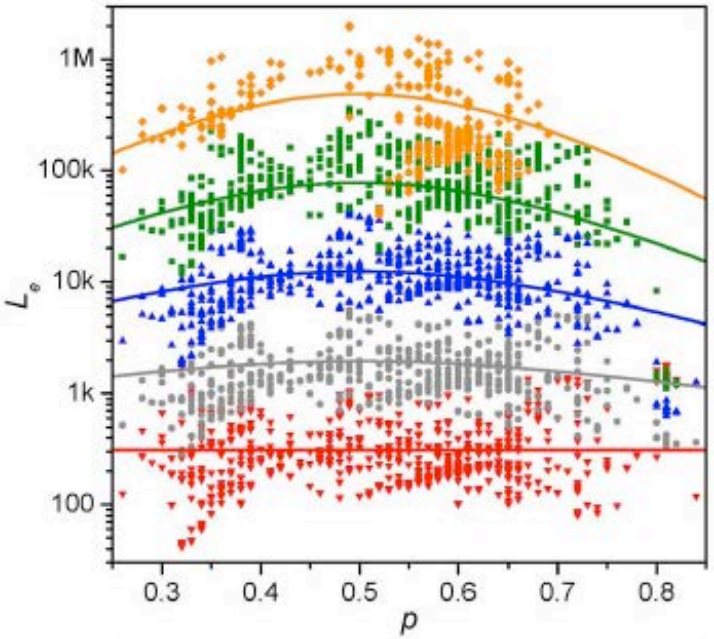
	L_e (kb)			
Category	**(k =) 2**	**5**	**7**	**10**
All	$.359^{+.333}_{-.172}$	$4.56^{+3.60}_{-2.01}$	$33.7^{+30.0}_{-15.9}$	388^{+524}_{-223}
gn (41.8%)	$.317^{+.253}_{-.141}$	$4.21^{+2.82}_{-1.67}$	$31.2^{+23.7}_{-13.4}$	337^{+396}_{-186}
ig (59.6%)	$.462^{+.879}_{-.302}$	$4.99^{+4.49}_{-2.36}$	$31.6^{+26.9}_{-14.5}$	213^{+170}_{-95}
ex (3.3%)	$.292^{+.215}_{-.122}$	$4.40^{+2.55}_{-1.62}$	$35.3^{+20.8}_{-13.1}$	620^{+298}_{-201}
in (31.8%)	$.348^{+.679}_{-.230}$	$3.65^{+2.55}_{-1.50}$	$23.5^{+13.9}_{-8.7}$	213^{+206}_{-105}
$L_e^{\{uc\}}$ (p = 0.5)	$.310^{+.290}_{-.150}$	$4.90^{+4.58}_{-2.24}$	$30.1^{+28.1}_{-13.8}$	487^{+455}_{-235}
RSD model	$.597^{+.756}_{-.351}$	$4.79^{+0.82}_{-0.70}$	$32.0^{+7.0}_{-5.8}$	510^{+211}_{-149}

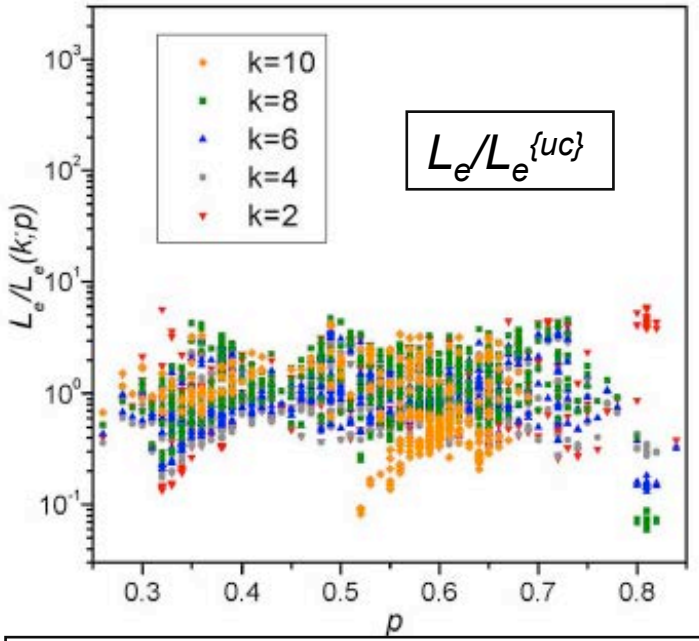
$L_e(k)$, k = 2, 5, 7 and 10, averaged over 865 chromosomes. Total sequences length is about 2.2×10^{10} bases. Abbreviations: All, complete chromosome; *gn*, genes; *ig*, intergenic; *ex*, exons; *in*, introns. Percentage given indicates portion of complete sequence. $L_e^{\{uc\}}$ is defined in Eq. (1) and RSD results are averaged over 200 model sequences. See Table S4 for $L_e(k)$ of other k values.

doi:10.1371/journal.pone.0009844.t002

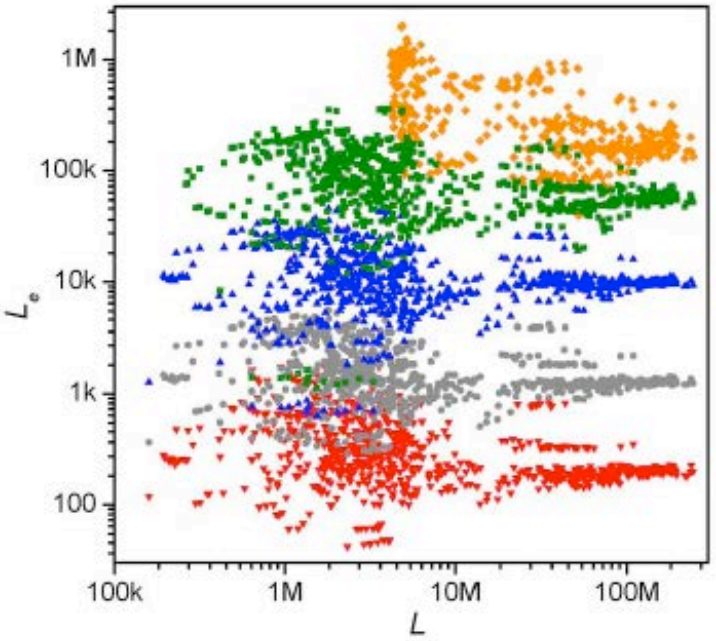
Difference in L_e of coding and non-coding parts mostly, but not all, caused by difference in p







Genomes have "universal" equivalent lengths



Equivalent (L_e) has only weak dependence on base composition (p), sequence length (L) or species.

Genomes form a k-dependent "universal class" in L_e

$$L_e^{\{uc\}}(k;p) = L_{e2}\exp((k-2)a(p)); (2 \le k \le 10) \qquad (1)$$

$$a(p) = \frac{a_0}{1+\epsilon\tan\left[\left(p^2+(1-p)^2-0.5\right)\pi\right]} \qquad (2)$$

where $a_0 = 0.92$, $L_{e2} = 310^{+290}_{-150}$ b, and $\epsilon = 0.50 \pm 0.05$.

Universality class strengthens with increasing genome length (or statistics)

Fraction of k-mers whose P-value is less than P_3, P_6, or P_8

	$k=2$ ($L_e=310$ b)			$k=9$ ($L_e=194$ kb)		
Length (Mb)	**$P<P_3$**	**$P<P_6$**	**$P<P_8$**	**$P<P_3$**	**$P<P_6$**	**$P<P_8$**
0.8	0.953	0.906	0.875	0.139	0.0031	0.0001
4.6	0.980	0.960	0.955	0.538	0.418	0.100
30	0.992	0.985	0.979	0.809	0.628	0.519
226	0.997	0.994	0.992	0.930	0.860	0.815

P-values for k-mer distribution given by Eq. (1) (at $p=0.5$). Null theory assumes genomes are random sequences. The P-values $P_3=2.7\times10^{-3}$, $P_6=2.0\times10^{-9}$, and $P_8=1.3\times10^{-15}$ correspond to z-values of three, six and eight, respectively.

doi:10.1371/journal.pone.0009844.t004

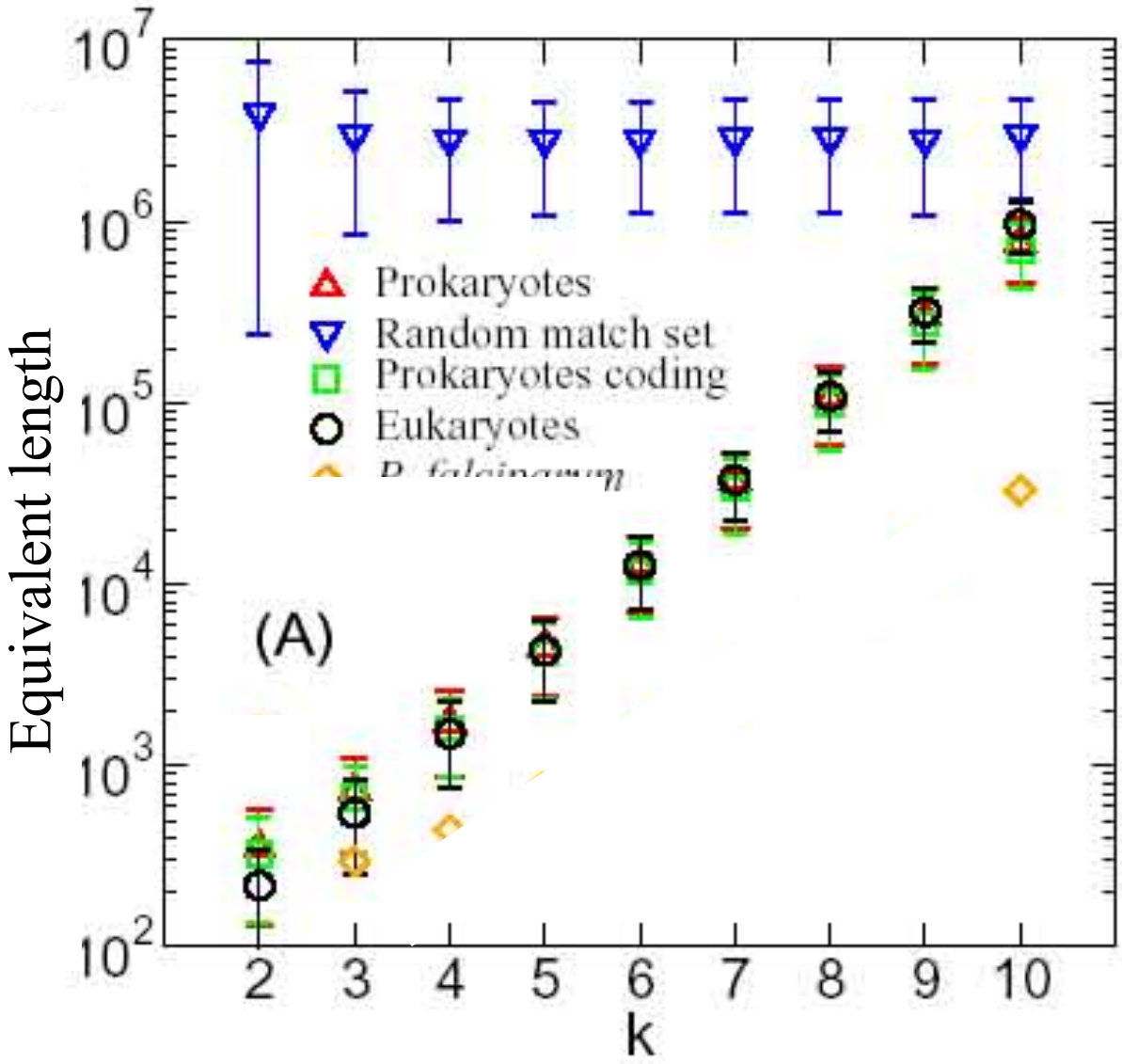
Universality Class of genomic equivalent length

$\log L_e(k) \sim a\,k + B$

$a = 0.40$

$B = 1.69 +- 0.28$

Mild exception: Plasmodium



Is universality class a consequence of similarity in genomes?

Similarity index and similarity matrix

Given a pair of equal-length sequences α and β, the similarity index $\eta_{sim}(\alpha,\beta)$ for the pair is defined as

$$\eta^2_{sim}(\alpha,\beta) = \frac{1}{k+1}\sum_m \frac{1}{2\tau_m}\sum_{u\in S_m} \frac{(f_u^{\{\alpha\}} - f_u^{\{\beta\}})^2}{\sigma_m^{\{\alpha\}}\sigma_m^{\{\beta\}}} \qquad (15)$$

where S_m is an m-set and σ_m^2 is the variance of the frequency of the k-mers in S_m. The pair are similar (in k-mer-content) when $\eta_{sim} \ll 1$, are (considered to be) identical when $\eta_{sim} = 0$, and are highly dissimilar when $\eta_{sim} \gtrsim 1$.

Homogeneity of Chromosome on a scale of 10 kb

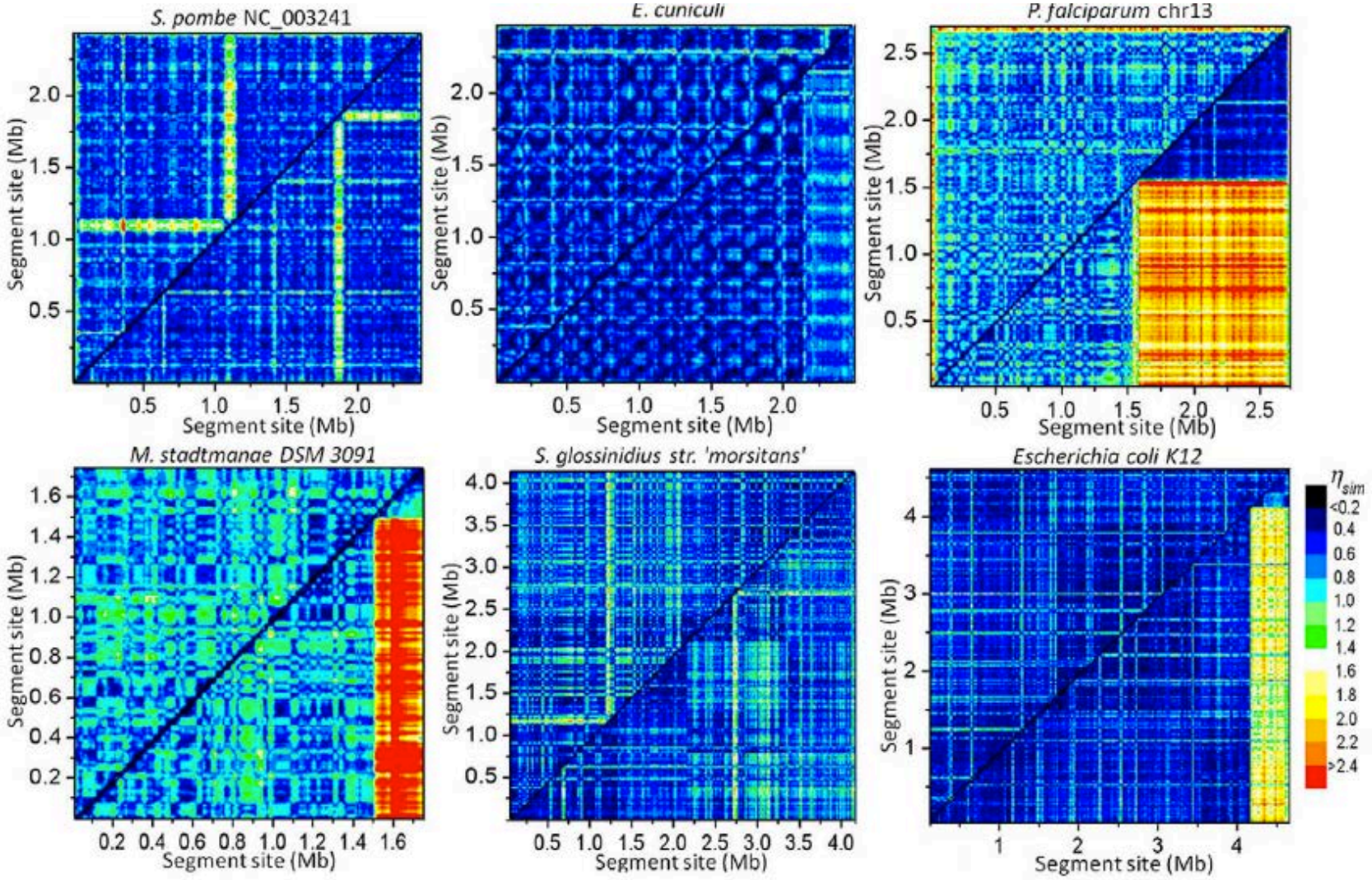


Figure 5. Intra-chromosomes similarity plots. Plots are for $k = 2$ (Methods). Sliding window has width 25 kb and slide 10 kb; pixel size is 10 kb by 10 kb. In each plot, the coordinates for the upper-left triangle are sites along the chromosome (*chr*), and those for the lower-right triangle are along a concatenate composed of gene (*gn*, left side) and intergene (*ig*, right side) parts. In effect, the upper-left triangle shows *chr-chr* similarity, and the lower-right triangle shows *gn-gn* (lower-left sub-triangle), *ig-ig* (upper-right sub-triangle), and *gn-ig* (rectangular) similarities in three separate regions. The lengths of the *gn* and *ig* parts are given in Table 3.

doi:10.1371/journal.pone.0009844.g005

Universality class not due to similarity in genomes

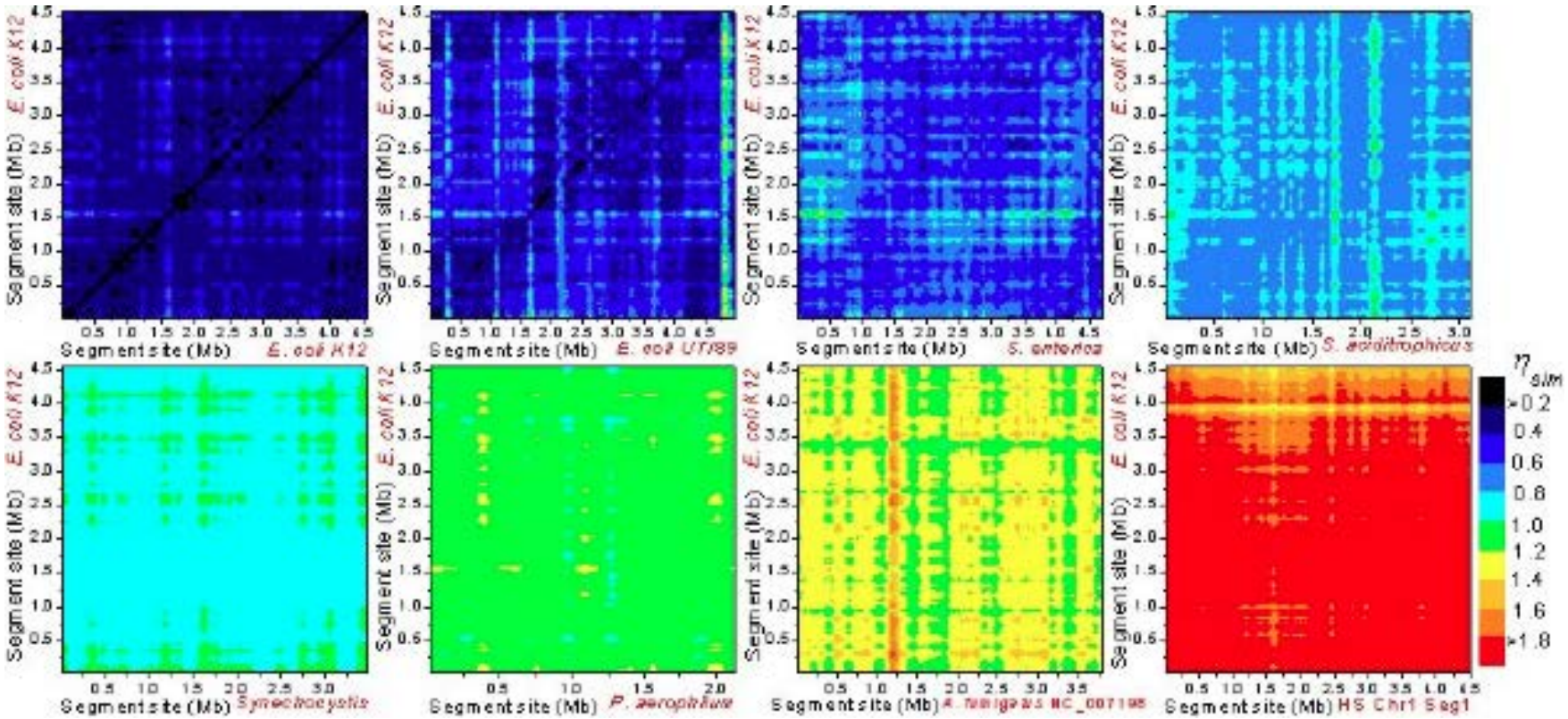


Figure 6. Intra- ***E. coli*** **and inter-chromosome similarity plots.** The plots are those of *E. coli* chromosome *vs.* the chromosomes of, left to right and top to bottom, *E. coli*, *E. coli UT189*, *Salmonella*, the delta-proteobacteria *S. aciditrophicus*, the cyanobacteria *Synechocystis*, the archaea *P. aerophilum*, chromosome 5 of the fungus *A. fumigatus*, and the first 4.5 Mb segment from chromosome 1 of *H. sapiens*. Coordinates are sites along the sequence. Sliding window width is 100 kb and slide is 25 kb, pixel size is 25 kb by 25 kb.

doi:10.1371/journal.pone.0009844.g006

Summary on genomic L_e

- Universality class – for fixed word length k, L_e is (approximately) the same for all genomes
  - Log $L_e(k) = ak + B$; a, B universal constants (for ~900 complete sequence)
- For $k<8$, L_e is much shorter than true genome length; L_e (k=2) ~ 300 b
- L_e for coding and non-coding parts about the same
- Universality not due sequence similarity

Small L_e is a signature of segmental duplication

- Recall: L_e is the length of random sequence having genomic *CV*
- Take a random sequence of length L_e and replicate it *n* times, then sequence length is nL_e but equivalent length is still L_e (for all $k << L_e$)
- Hint: small genomic L_e caused by segmental duplication

Fun with l_e and concatenates

$$l_e = \tau/(CV)^2 = \tau/(\sigma/f)^2 = L^2\,\tau^3/\sigma^2 \quad (\tau = 4k)$$

Let X be genomic, and R random, sequence of same length;
RX be concatenate of R and X.
Then $\sigma(X) >> \sigma(R)$, $l_e(X) << l_e(R) = L(R) = L(X)$.

$$l_e(RX) = L(RX)^2\,\tau^3/\sigma(RX)^2 \sim (L(R) + L(X))^2\,\tau^3/\sigma(X)^2$$
$$= 4L(X))^2\,\tau^3/\sigma(X)^2 = 4\;\; l_e(X)$$

Similarly,

$$l_e(RXX) \sim (3/2)^2\, l_e(X); \qquad l_e(RRX) \sim 9\, l_e(X)$$

Similar (notation ~): similar in word content. Suppose X and X′ are same length sequences from the same genome, Y from a different genome, and R is a random sequence. XY denote concatenate of X and Y.

- $l_e(X) \sim l_e(X') \sim l_e(XX')$
- If X~Y, then $l_e(X) \sim l_e(Y)$, $l_e(XY) \sim l_e(X)$,
- If X not~ Y, then $l_e(XY) > \min(l_e(x), l_e(Y))$

- $l_e(RX)$ approx 4 $l_e(X)$
- $l_e(RXY)$ approx 2.3 $l_e(XY)$
- $l_e(RR'X)$ approx 9 $l_e(X)$

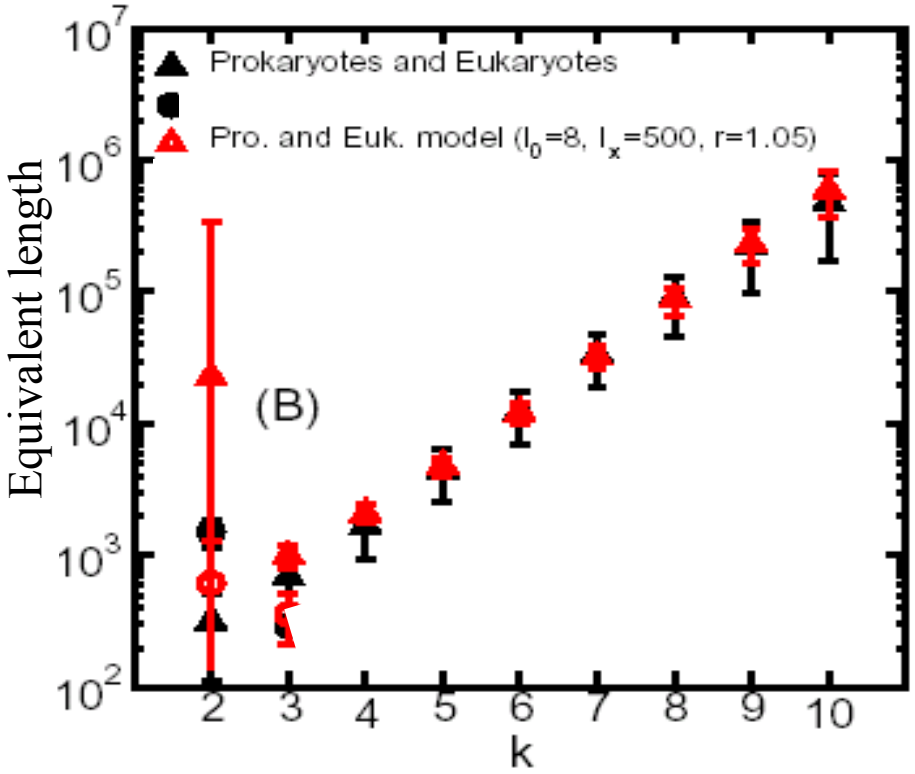
	l_e			
	$k=2$		$k=6$	
Sequence	$l=50$	$l=200$	$l=50$	$l=200$
R	47.5±28.2	154±126	48.6±1.5	192±5
RR′	37.0±16.2	124±46	48.2±1.2	197±5
A	.348±.037	.360±.033	9.55±.69	11.7±.7
AA′	.357±.046	.352±.023	9.88±1.07	11.1±.7
AC_1	.351±.061	.361±.021	9.37±1.01	11.5±.6
AC_2	.354±.043	.384±.045	9.18±.83	11.6±.9
AC_3	.359±.051	.371±.034	11.0±.9	14.2±1.5
AD_1	.411±.044	.423±.024	11.8±.9	14.3±.6
AD_2	.942±.275	1.05±.09	14.9±1.4	20.4±1.1
AD_3	.598±.104	.613±.052	17.9±1.6	24.0±1.6
AD_4	.324±.052	.383±.055	11.2±1.9	16.9±1.9
B	.124±.029	.166±.099	5.17±.68	6.54±2.00
BB′	.232±.155	.258±.183	6.16±1.94	7.54±2.30
AB	.463±.241	.502±.263	11.2±1.9	15.2±3.5
RA	1.19±.09	1.34±.20	22.6±1.2	38.5±3.0
RB	.575±.321	.754±.637	15.6±4.2	23.3±8.5
RAB	.873±.424	1.10±.49	18.4±3.2	31.3±6.0
RR′A	2.63±.66	3.16±.30	31.5±2.1	72.2±6.8
RR′B	1.03±.62	1.37±.70	22.9±4.5	44.7±14.3

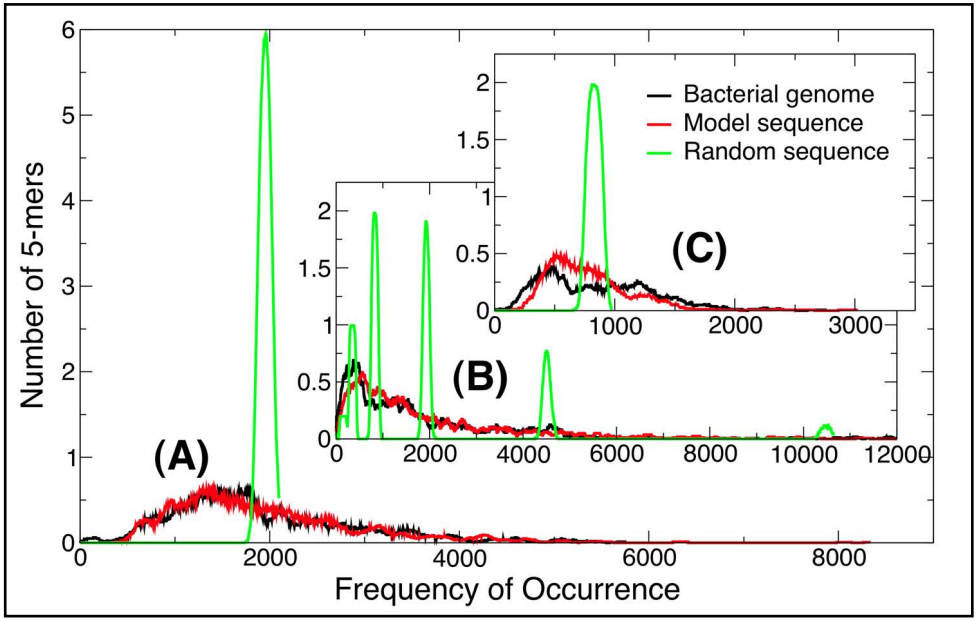
l_e of a sequence after n-fold growth

Equivalent length (l_e) of a sequence after an n-fold increase in length via three basic modes of growth. Initial sequence length is $l_0 >> 1$, final sequence length is $L = nl_0$.

Sequence type	Initial l_e	Mode of growth	Final l_e
Random	l_0	Random base-by-base growth	$L(=nl_0)$
		Whole-sequence replication (n-1 times)	$\approx l_0$
		Segmental duplication	$l_0 < l_e << nl_0$
Non-random	l_{e0} ($<< l_0$)	Random base-by-base growth	$\approx \min(n^2 l_{e0}, L)$
		Whole-sequence replication (n-1 times)	$\approx l_{e0}$
		Segmental duplication	$l_{e0} < l_e << \min(n^2 l_{e0}, L)$

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





Note: Deep consequences in understanding speed of evolution (Lecture 4)

HD Chen, et al. Universal Global Imprints of Genome Growth and Evolution - Equivalent Length and Cumulative Mutation Density. PLoS ONE (2010) 5(4): e9844.

End of Lecture One