The Blind Self-Copier — Universality and self-organized criticality in genomes

International Symposium on the Recent Progress in Quantitative and Systems Biology
2006 December 9-11

HC Lee

Computational Biology Lab
Inst. Systems Biology/Dept. Physics/Inst. Biophysics
National Central University

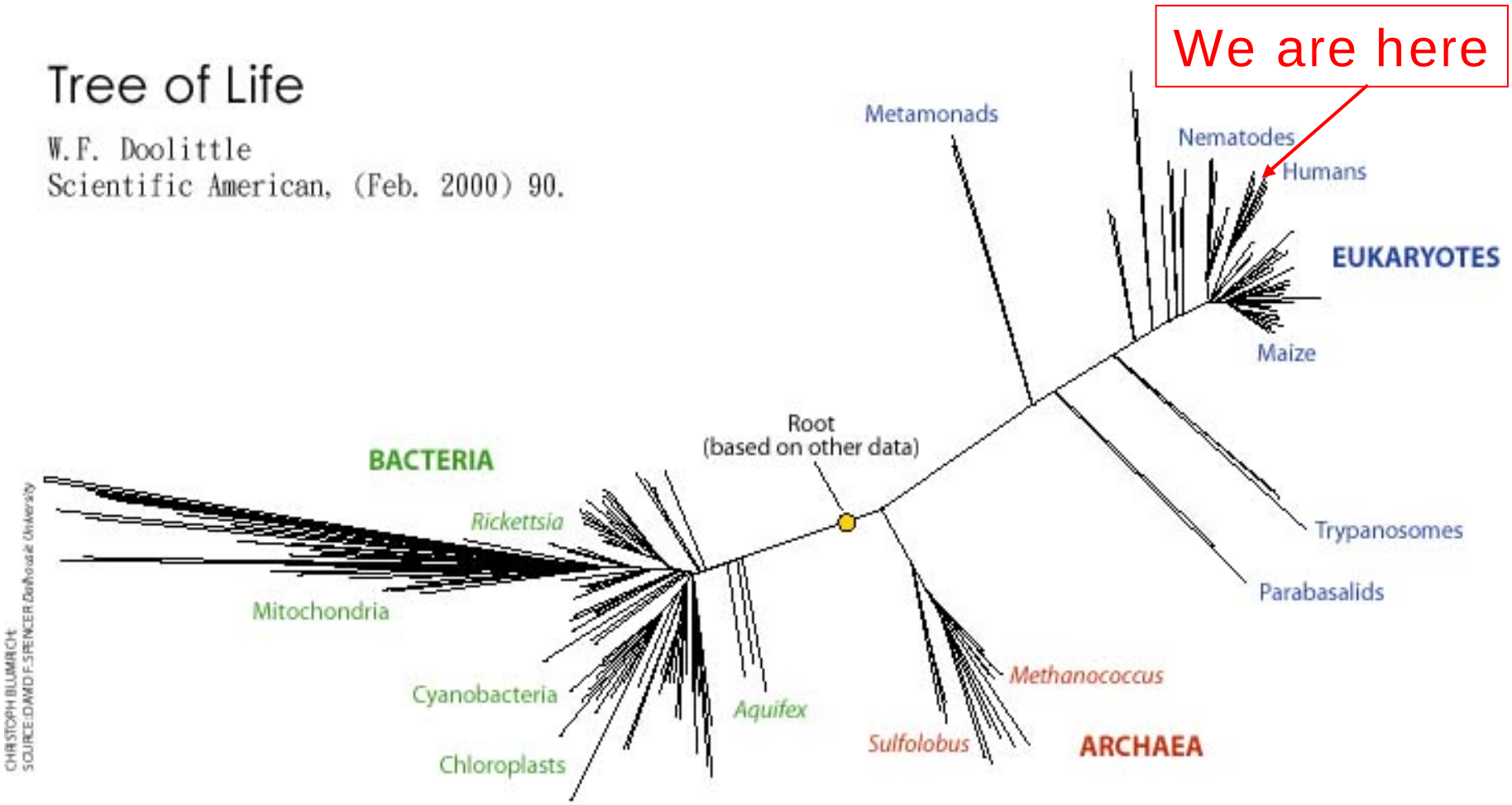
Three questions we should ask

- WHAT is the phenomenon?
  - What is strange/unusual/interesting?
    - Usual & tell-tail characteristics of genomic sequence statistics
- HOW did it happen?
  - (Physics)
    - Critical "blind self-copying"
- WHY did it happen?
  - (Biology)
    - Good for information growth
    - Natural selection

Some concepts to be discussed

- Randomness and order
- Second law of thermodynamics and genome growth
- Diversity and universality
- Blind self-copying
- Scaling, self-similarity and criticality
- The self-organized critical genome
- Biological manifests
- Role of natural selection

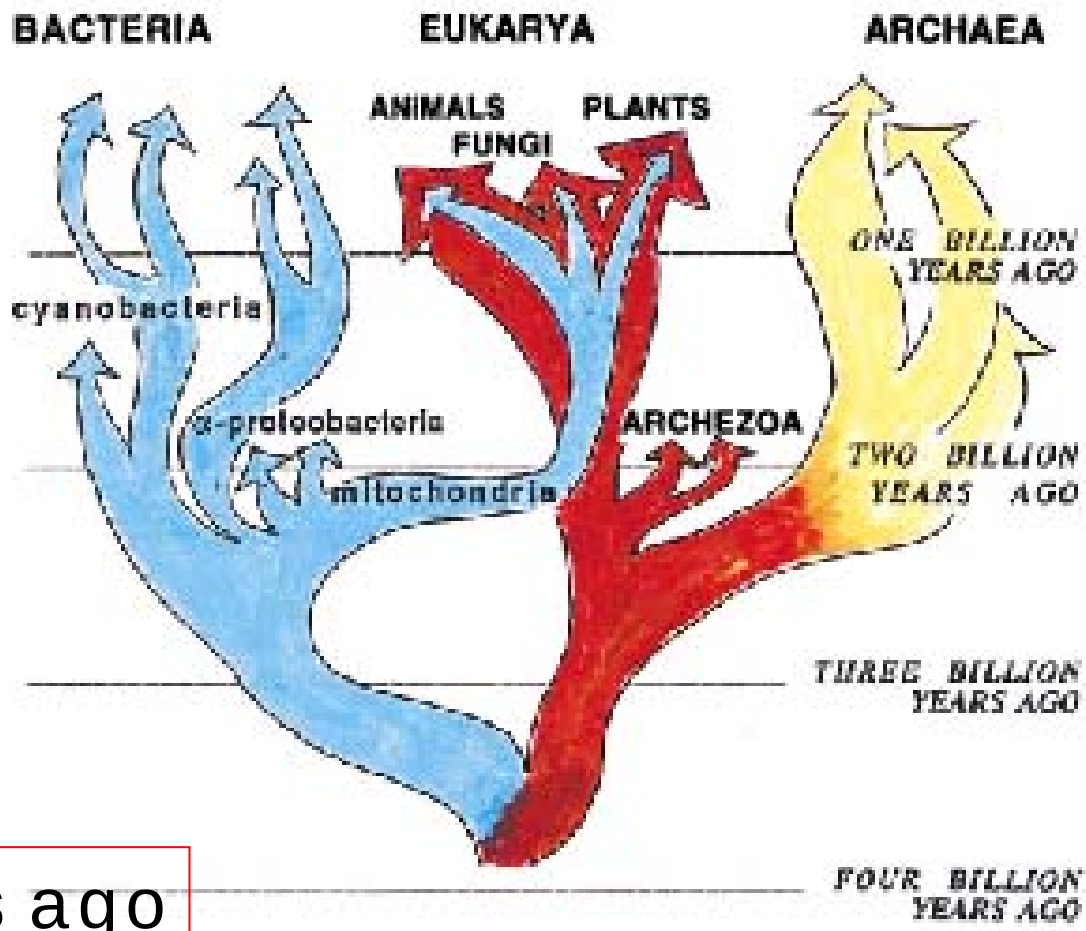
Life is highly diverse and complex



And it took a long time to get here

now

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

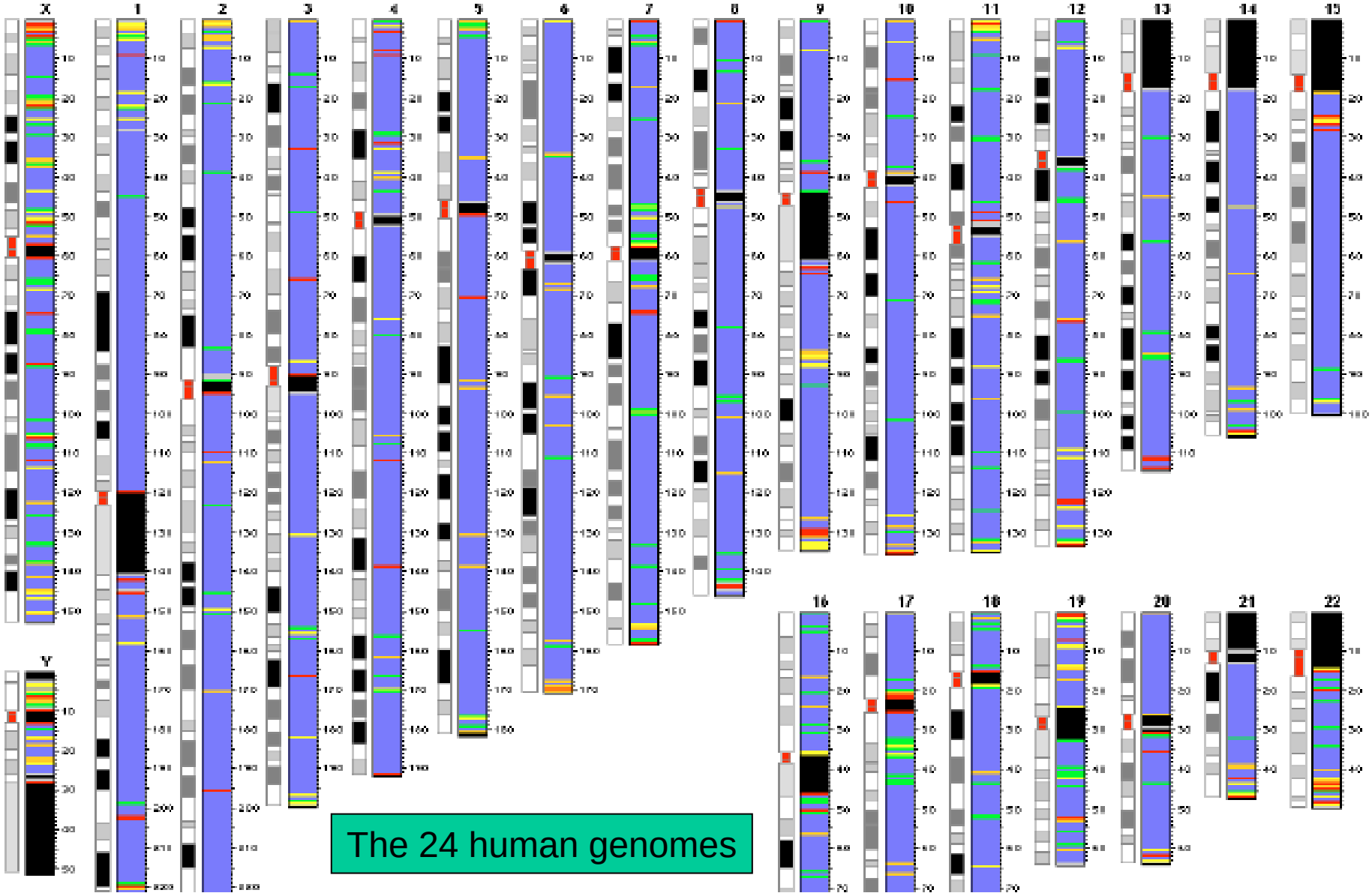


4 billion yrs ago

Evolution of life is recorded in genomes

- Genome is Book of Life
- A double helix - two strands of DNA
- DNA: String of four types of
- molecules – chemical letters - A, C, G, T
- Genome is a linear text written in four letters
- We believe all genomes have a common ancestor, or a small group of ancestors

Genome is an extremely complex



Genomes are BIG

A stretch of genome from the X chromo-some of Homo sapien

http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=2276452&db=Nucleotide&dopt=GenBank

The complete genome has 2,000,000 such pages

```
   1 tgctgagaaa acatcaagctg tgtttctcct tccccaaag acacttcgca gcccctcttg
  61 ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggcctg
 121 ctctcctcac gccggccccc acctgggcca cctgtggcac ctgccaggag gctgagctgc
 181 aaaccccaat gaggggcagg tgctcccgga gacctgcttc ccacacgccc atcgttctgc
 241 ccccggcttt gagttctccc aggccccctct gtgcacccct ccctagcagg aacatgccgt
 301 ctgccccctt gagctttgca aggtctcggt gataatagga aggtctttgc cttgcaggga
 361 gaatgagtca tccgtgctcc ctccgagggg gattctggag tccacagtaa ttgcagggct
 421 gacactctgc cctgcaccgg gcgccccagc tcctccccac ctccctcctc catccctgtc
 481 tccggctatt aagacggggc gctcaggggc ctgtaactgg ggaaggtata cccgccctgc
 541 agaggtggac cctgtctgtt ttgatttctg ttccatgtcc aaggcaggac atgaccctgt
 601 tttggaatgc tgatttatgg attttccagg ccactgtgcc ccagatacaa ttttctctga
 661 cattaagaat acgtagagaa ctaaatgcat tttcttctta aaaaaaaaaa aaaccaaaaa
 721 aaaaaaaaaa aaaccaaaaa actgtactta ataagatcca tgcctataag acaaaggaac
 781 acctcttgtc atatatgtgg gacctcgggc agcgtgtgaa agtttacttg cagtttgcag
 841 taaaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
 901 tatttttct aatctgggca acaatggtgc catctcggtt cactgcaacc tccgcttccc
 961 aggttcaagc gattctccgg cctcagcctc ccaagtagct gggaggacag gcacccgcca
1021 tgatgcccgg ttaatttttg tatttttagc agagatgggt tttcgccatg ttggccaggc
1081 tggtctcgaa ctcctgacct caggtgatcc gcctgccttg gcctcccaaa gtgctgggat
1141 gacaggcgtg agccaccgcg cccagccagg aatctatgca tttgcctttg aatattagcc
1201 tccactgccc catcagcaaa aggcaaaaca ggttaccagc ctcccgccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg atttttgctg aggttataag
1321 ccacttcctt catctgggtc tgagcttttt tgtattcggt cttaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgctgagtca aatacaaata
1441 gatttttaaa gtgtatttat tttaaacaaa aaataaaatc acacataaga taaaacaaaa
1501 cgaaactgac tttatacagt aaaataaacg atgcctgggc acagtggctc acgcctgtca
```

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

- How was genome able to simultaneously grow stochastically AND acquire information?

Characterization of Genomes

- Primary characterization of genomes
  - length in bp (base pair)
  - base composition p = A+T/(A+T+C+G)
  - word frequencies
- Secondary characterization
  - % coding region (microbials: ~85%; eukaryotes (2~50%)
  - number of genes (few hundred to 25K)
- Tertiary characterization
  - intron/exon (microbials, no; eukaryotes, yes)
  - other details

Short and universal genomic L_{eff}

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

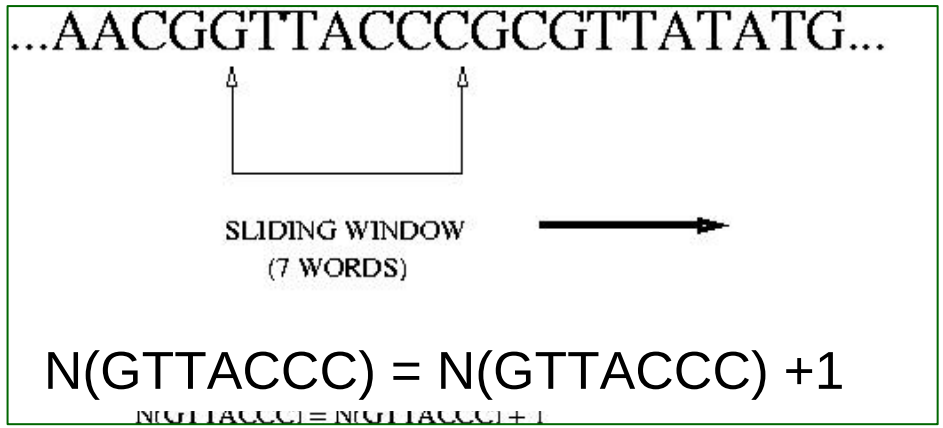
- Random events given by Poisson distribution
  - $\Delta^2 = f_{ave}$, or, $(CV)^2 = 1/f_{ave}$
  - That is, $(CV)^2 = \tau/L$

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

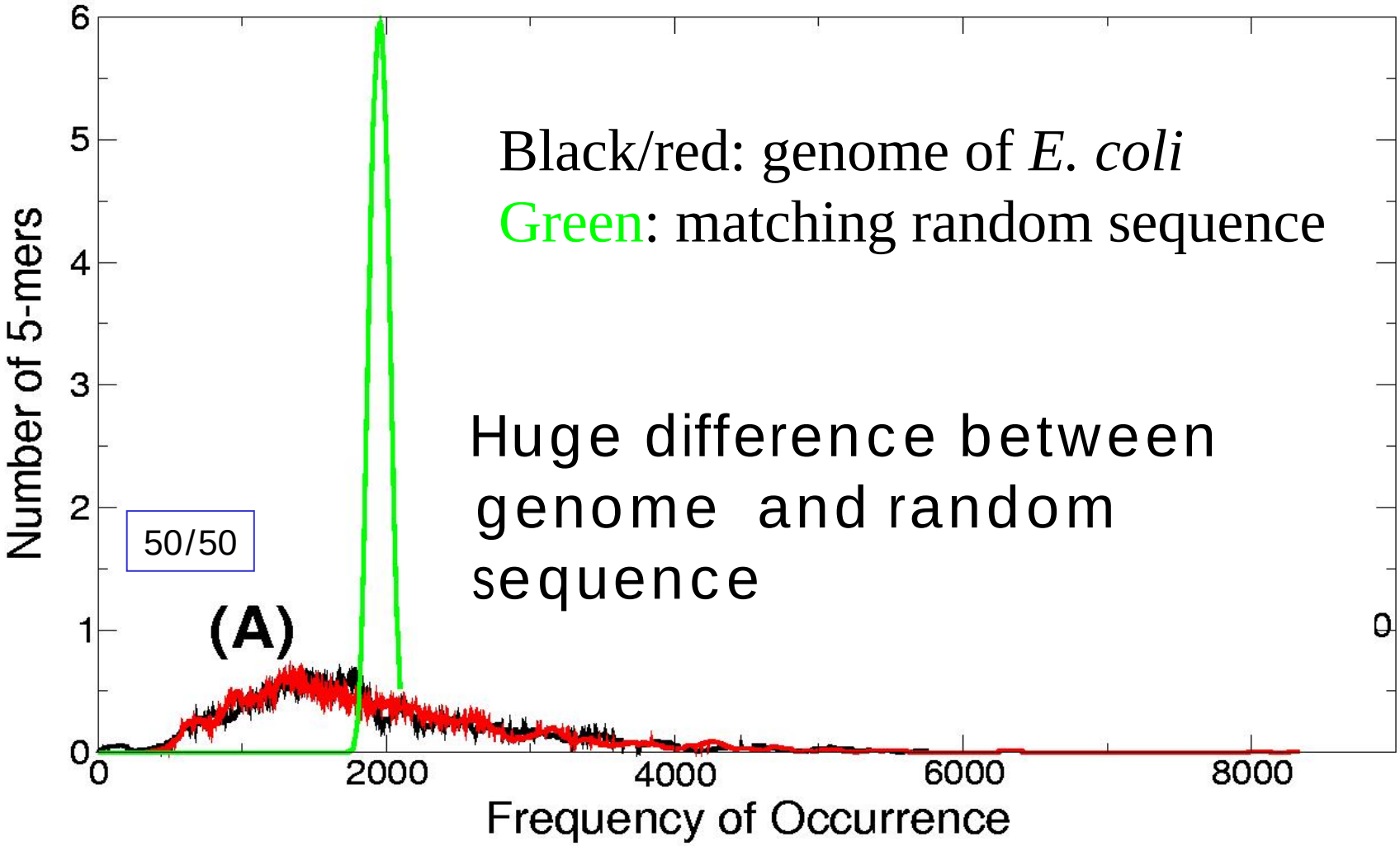
- For given τ, if CV is known, then
  - $L \sim \tau/(CV)^2$

Genome as text - Frequencies of *k*-mers

- Genome is a text of four letters – A,C,G,T
- Frequencies of k-mers characterize the whole genome
  - E.g. counting frequencies of 7-mers with a "sliding window"
  - Frequency set $\{f_i \mid i=1 \text{ to } 4^k\}$



For genomes: events=word occurrence; type of events τ=types of words = 4^k; distr.= distr. of frequency of occurrence



Two big surprises from complete genomes

Given τ and CV, define effective length

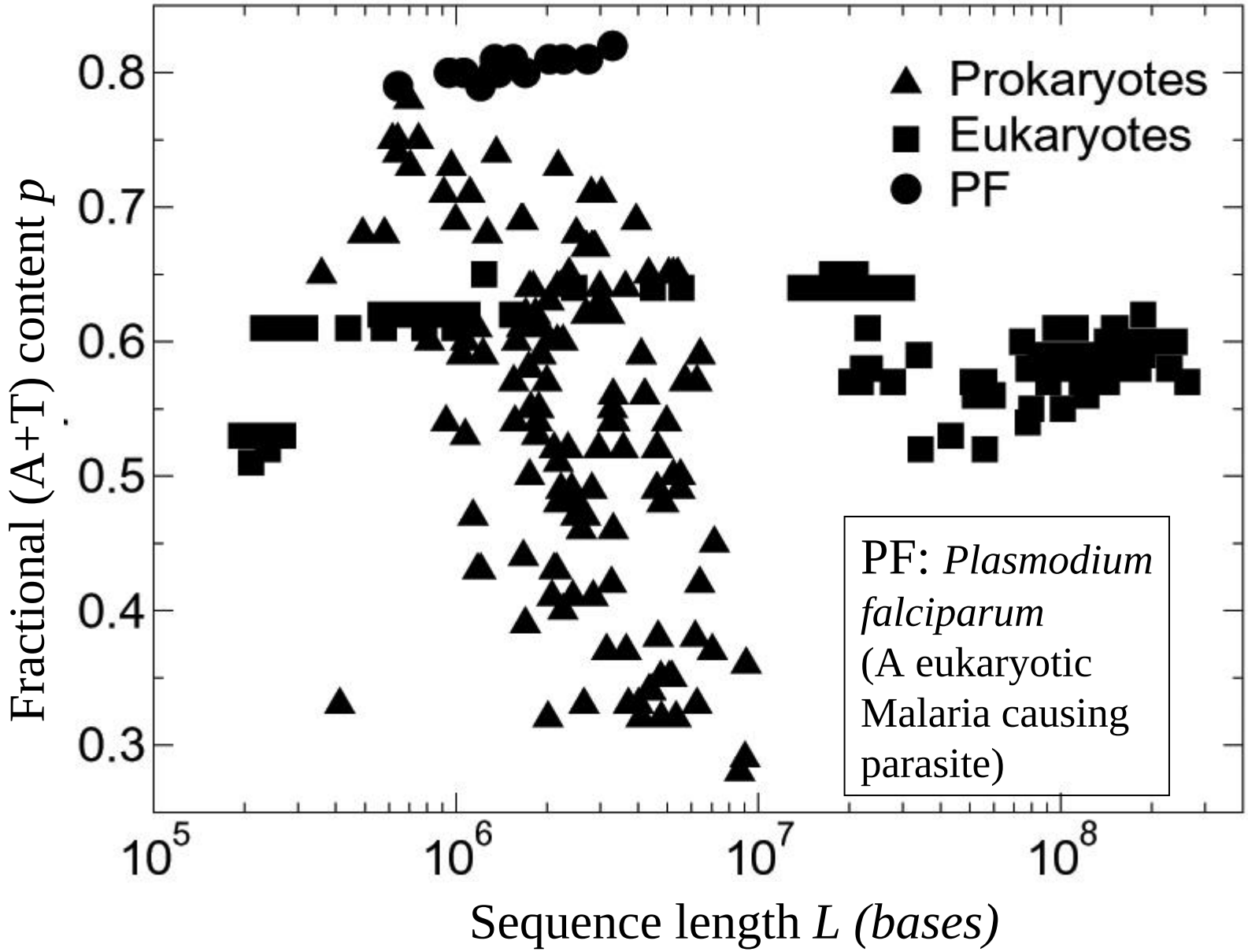
$$L_{eff} = \tau/(CV)^2$$

- The L_{eff} of complete genomes are far shorter than their actual lengths
- For a given type of event (word counts) L_{eff} is universal
  - Actual length varies by factor > 1000
  - "Information" in genomes growths as L

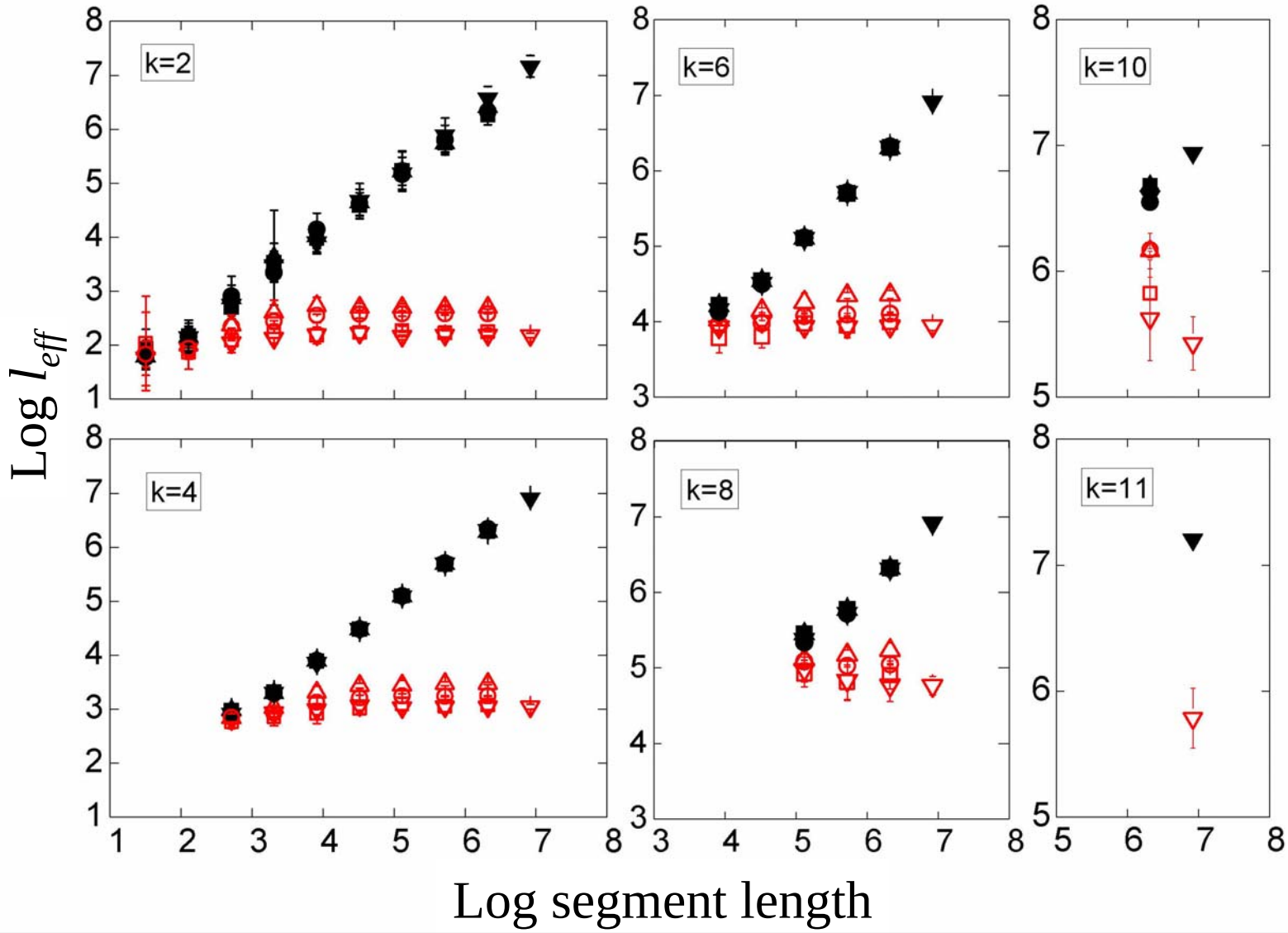
The genome data set includes all complete sequences

- All (~300) complete sequences from **GenBank** at the time of download (March 2006)
- Include 318 (293+25) complete prokaryotic (原核) genomes, 292 complete chromosomes from 15 eukaryotic (真核) genomes, 11 complete fungi chromosomes
- For each sequence compute $L_{eff}(k)$ for k= 2 to 10 (or to k_{max} such that sequence length > 4^{kmax})
  - Each k has about ~600 pieces of data
  - All told about 5400 pieces of data

Complete Genomes are diverse

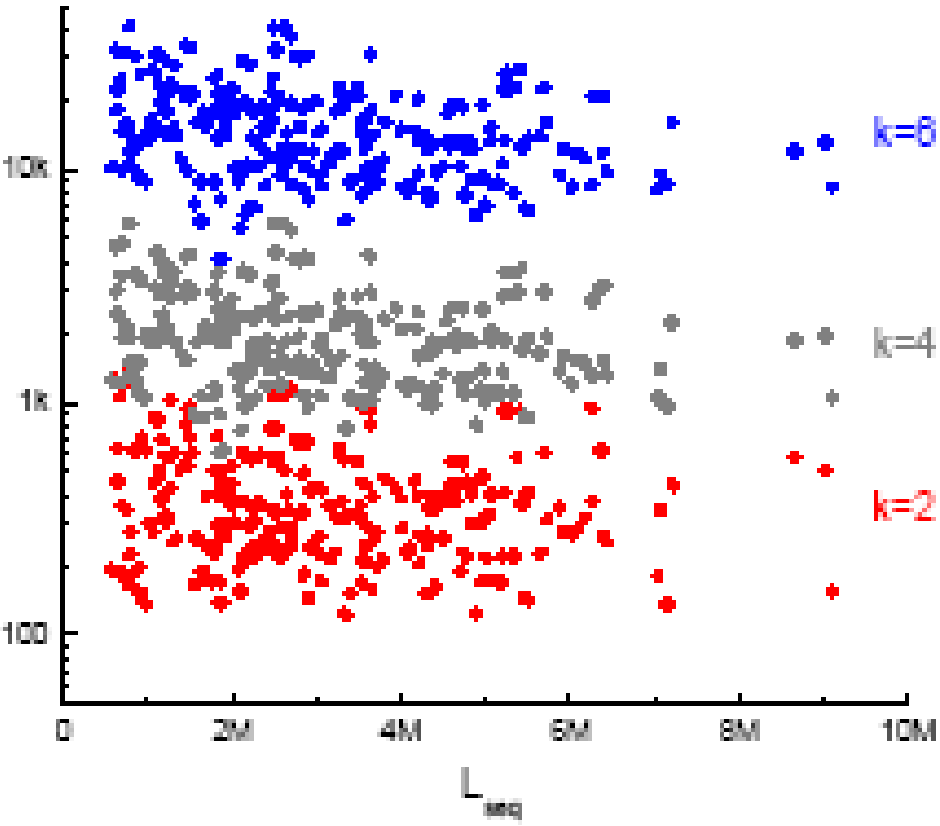


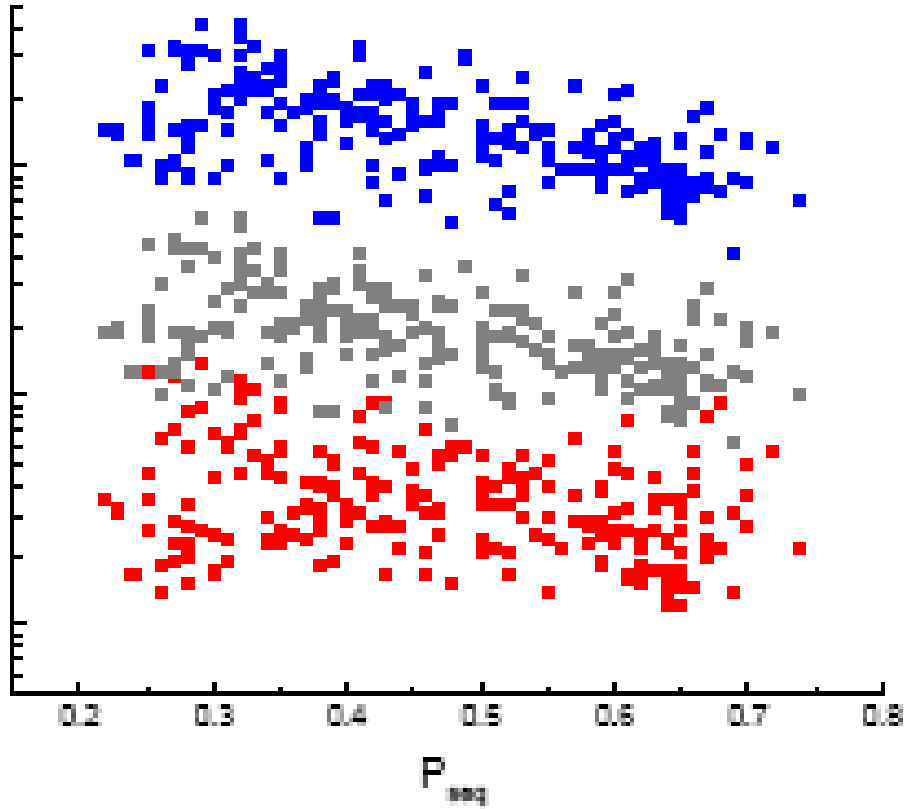
l_{eff} *vs* segmental length for E. coli, worm, mustard and human



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

Genomic l_{eff} *vs* genome length for ~300 complete bacterial genomes



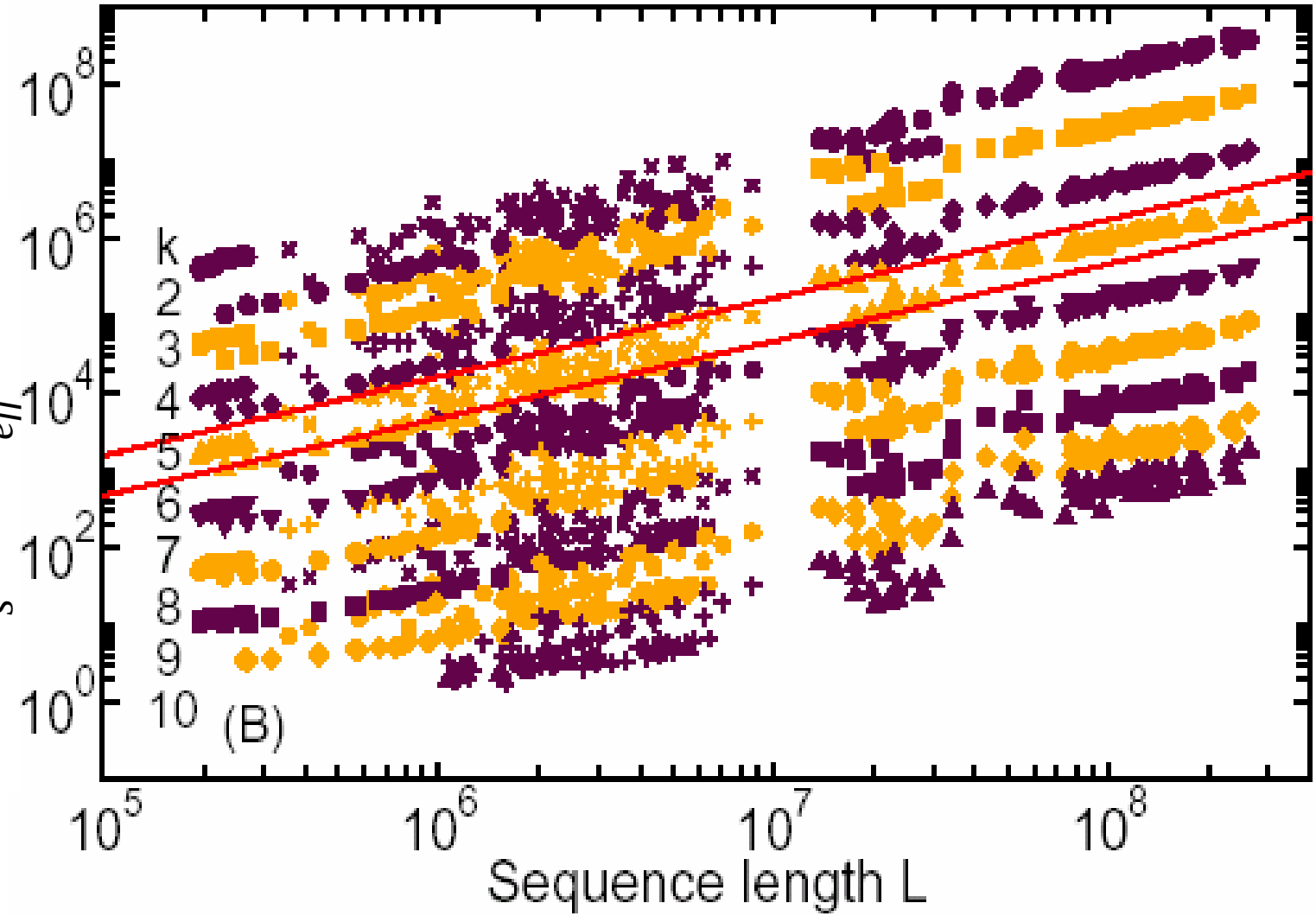


Genomic l_{eff} *classes of* genomes
Coding and non-coding regions

k	2	3	4	5	6	7	8	9	10
Archean coding	336±181	712±278	1695±634	4561±1721	13023±5113	37603±15114	108731±43415	292206±106916	649087±200105
Archean noncoding	494±306	1058±534	2597±1273	6782±3341	17515±8396	40807±17843	78727±32763	122749±66460	160732±121444
Archean(25)	354±192	757±309	1819±709	4918±1939	14104±5771	40917±17087	118581±48975	318381±119729	705566±224601
Bacteria coding	378±208	805±398	1923±856	5142±2238	14566±6281	41746±17624	118737±47706	313342±115410	699514±246337
Bacteria noncoding	541±402	1148±681	2705±1346	6979±3197	18196±7997	44573±19433	93476±42861	156376±81148	210188±124110
Bacteria(293)	408±233	880±466	2125±1014	5677±2649	16148±7445	46457±20834	132241±56128	348265±135153	774623±285942
Human coding	158±13	409±35	1161±108	3305±352	9373±1202	25311±4077	60848±12293	120804±29041	195897±51724
Human noncoding	158±11	412±32	1175±100	3369±345	9672±1264	26754±4612	66936±15053	139805±38540	236562±73315
Human(24)	158±12	412±32	1170±103	3346±348	9571±1243	26264±4423	64825±14081	133020±35161	221791±65522
A_thaliana coding	391±4	984±11	2460±29	6750±85	19750±256	57266±767	169938±2193	493778±6880	1274412±44219
A_thaliana noncoding	617±10	1386±21	3581±59	9894±182	28033±649	78161±2685	211429±11934	506382±46480	959651±131698
A_thaliana	493±6	1200±14	3068±39	8497±112	24715±341	71267±1097	208863±3829	599845±13923	1600349±54980
c_elegans	188±20	475±53	1317±160	3817±521	11413±1770	34657±6107	103900±21272	295278±71927	758801±215515
Drosophila	477±209	1186±417	3217±1085	8917±2978	25215±7807	69465±17234	183727±30109	474035±86329	1205065±270008
M_musculus(21)	160±5	397±13	1125±37	3188±120	9107±395	24167±1206	56882±3255	105771±6743	160584±10980
S_cerevisiae	598±22	1466±47	3733±107	10447±353	30253±1571	83271±8268	211921±28285	605830±24735	2762769±525662
Plasmodium(14)	919±90	763±37	1619±54	3661±119	8893±400	21232±1503	50167±5321	119733±14754	297077±32645
D_rerio	313±64	732±147	2068±37	5540±1113	15355±3092	39584±8042	88900±18468	159160±34431	238671±51647
Apis(16)	236±17	539±38	1351±98	3621±279	9862±833	26339±2508	65831±7463	147198±20417	306468±47480
Bos(31)	154±7	397±17	1132±48	3277±133	9550±327	26985±618	68781±2195	146678±11808	253960±33183
Canis(40)	159±11	406±28	1153±79	3287±229	9436±654	25767±1759	63025±4587	130456±11885	220071±25863
Gallus(40)	138±17	351±42	940±133	2696±420	7976±1374	24036±4596	73138±15764	219612±55803	623849±200480
Pan(25)	158±12	412±32	1174±101	3372±343	9704±1229	26908±4430	67550±14440	141902±37354	242363±72473
R_norvegicus(22)	169±6	414±20	1165±60	3261±204	9135±699	23400±2255	52130±5928	90116±11161	129126±15898
S_pombe(3)	534±14	1374±37	3751±100	10964±328	33418±1165	101986±4351	305556±19514	826566±95294	1759142±349874
Fungi(11)	629±257	1462±551	3518±1275	9601±3683	27676±11575	78441±35638	214429±104704	512002±269981	955007±575721

Average	300+250 -150	750+500 -300	1.8k+1.2k -0.6k	4.5k+3k -1.5k	12k+8k -3k	30k+20k -10k	120k+100k -50k	300k+250k -150k	700k+700k -500k

L/L_{eff} plot (one data per seq. per k)



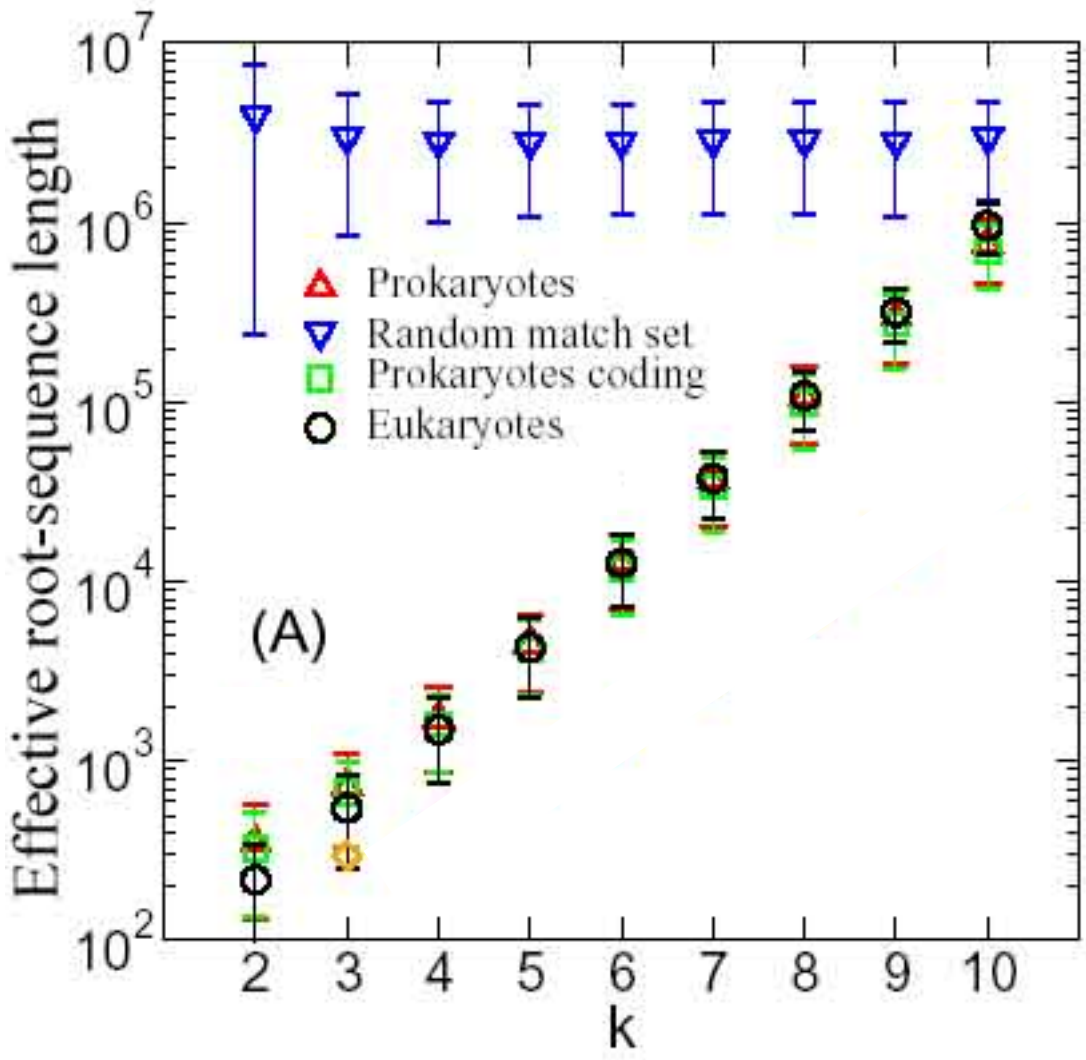
Straight line implies $L_{eff} \sim$ constant

Genomes are in Universality Classes

- Each *k-band* defines a **universal constant** $L/M = L_{eff} \sim$ constant (Effective root-sequence length)
- Obeys

  log L_r(k) = *a k + B*

  1989 pieces of data given be two parameters.
  a = 0.398+-0.038
  B = 1.61+- 0.11
- Defines a **universal class**
- Mild exception: Plasmodium



Black: genome data; green: artificial

Simple model for genome growth- The Blind Self-Copier

Order, Randomness, L_{eff} and duplications

- If we take random sequence of length L_0 and replicate it n time, then total sequence length (L) is nL_0 but L_{eff} of sequence remains L_0
- Smaller L_{eff} implies higher degree of ORDER
- Larger L_{eff} implies higher degree of RANDOMNESS
- Small L_{eff} of genomes suggests many DUPLICATIONS

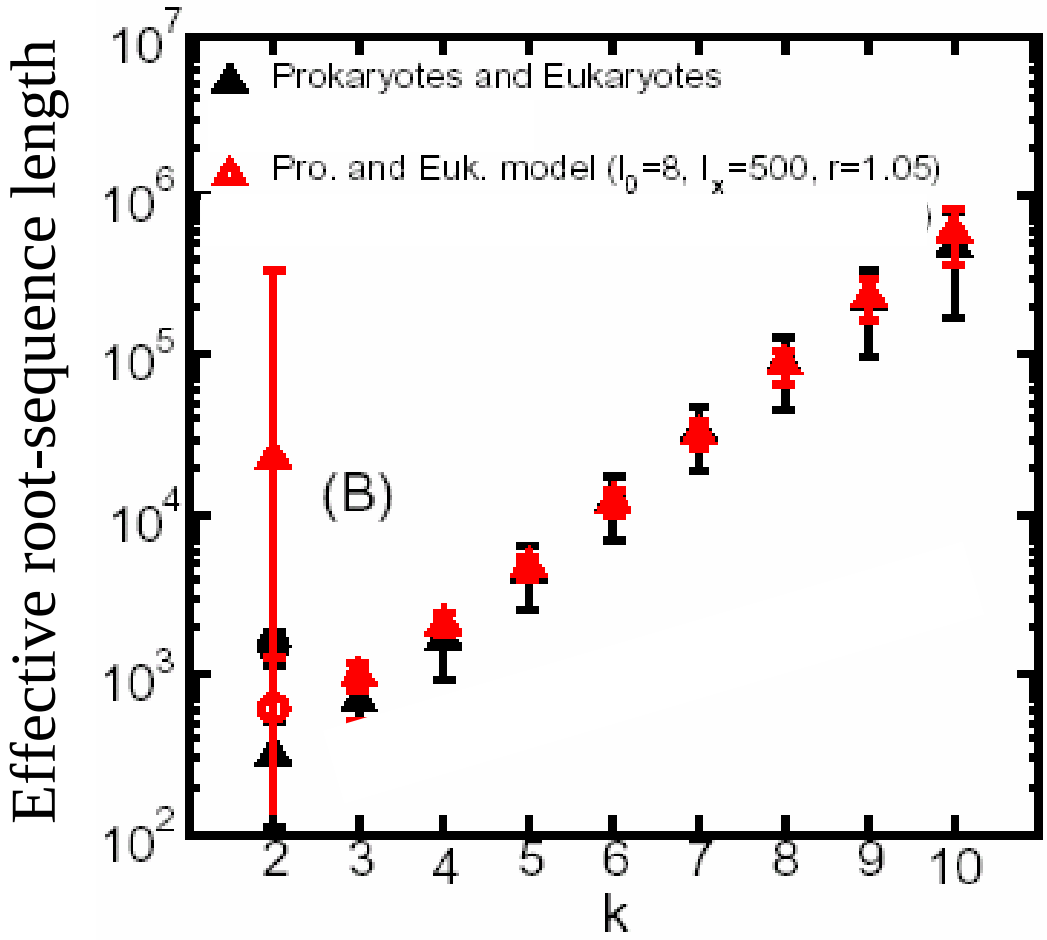
A Universal Model for Genome Growth

A model: at a universal initial length, genomes grew (and diverged) by maximally stochastic segmental duplication

1. Universal initial length - Common ancestor(?), universal L_{eff}.
2. Segmental duplication – *L*-independent *CV*
3. Maximum stochasticity – self-similarity, random word interval

Self copying – strategy for retaining and multiple usage of hard-to-come-by coded sequences (i.e. genes)

Model with three universal parameters – successfully generates universal L_{eff}



Red symbols are from 278 genome matching model sequences

Long-range variation & criticality in genome

Another clue from the human genome

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

articles

Initial sequencing and analysis of the human genome (3.36 x 10^9 base pairs)

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

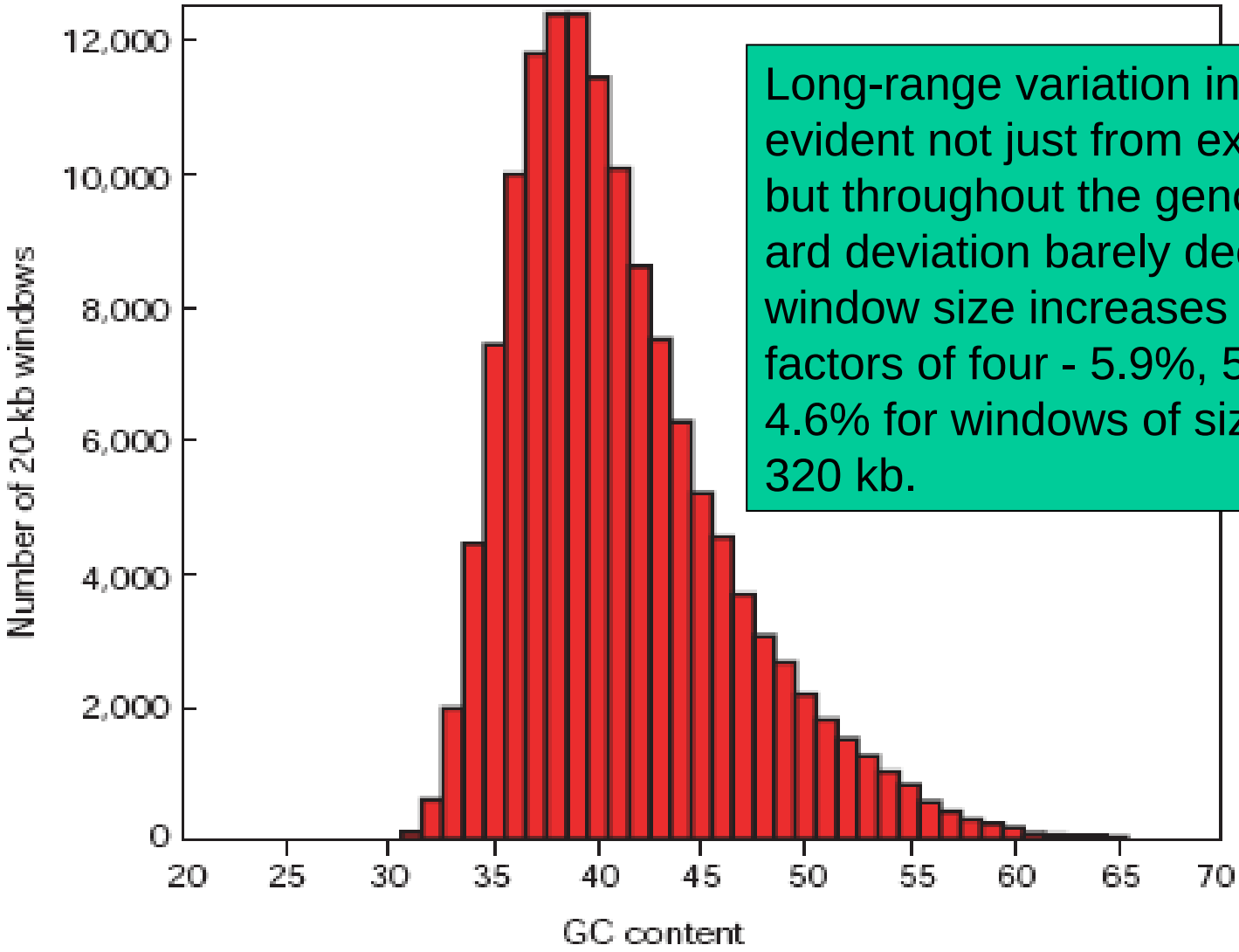
The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field

coordinate regulation of the genes in the clusters.

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from trans-

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

Long-range variation in GC content

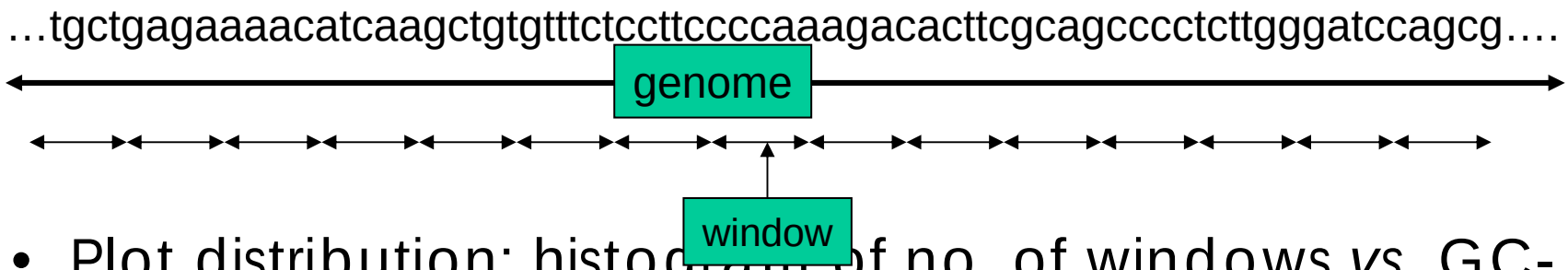


Long-range variation in GC content is evident not just from extreme outliers, but throughout the genome…the standard deviation barely decreases as window size increases by successive factors of four - 5.9%, 5.2%, 4.9% and 4.6% for windows of size 5, 20, 80 and 320 kb.

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

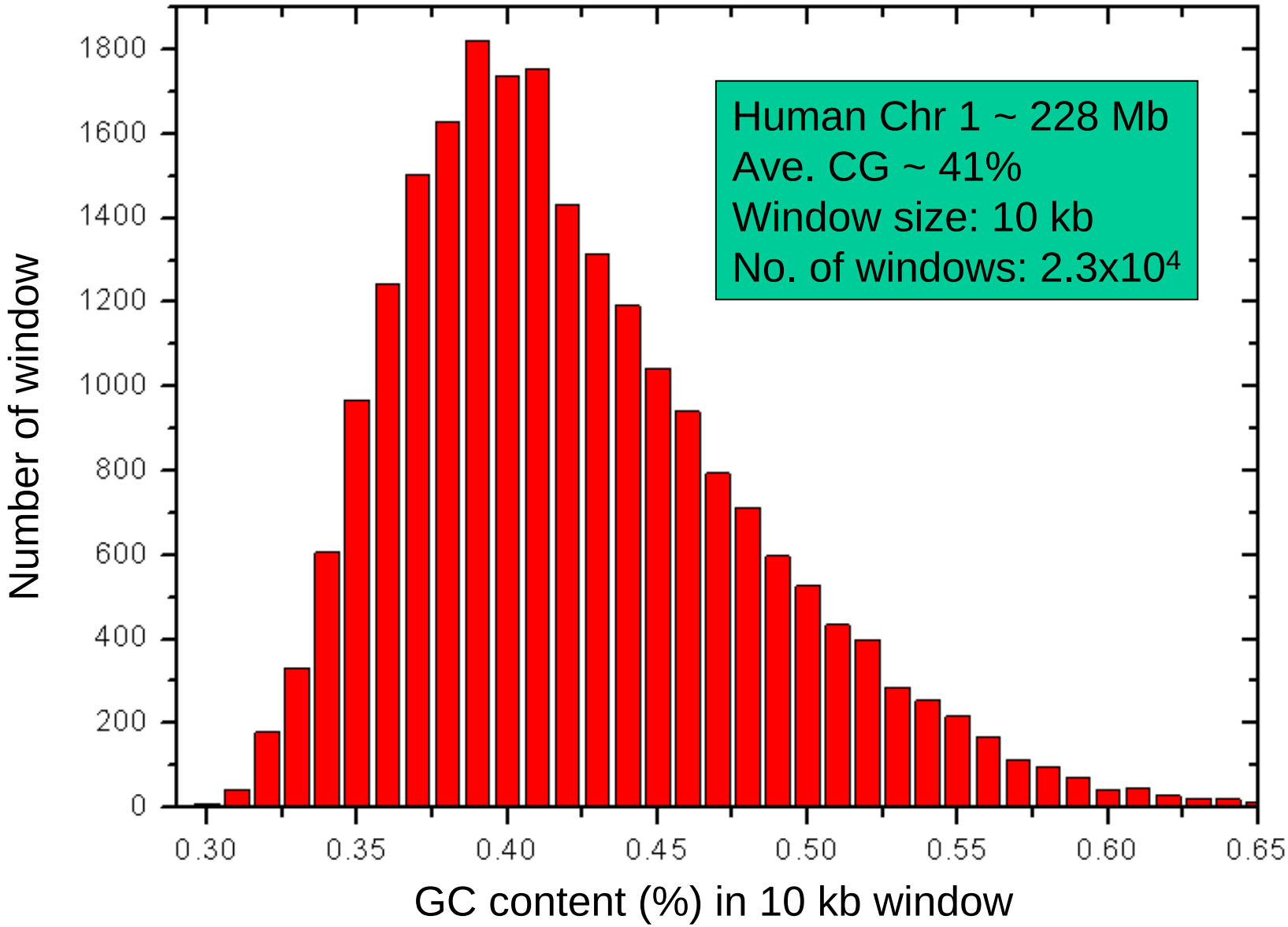
What was measured? And why?

- Cut genome into fixed-sized windows and computer the GC-content (percentage words that are G or C) in each window

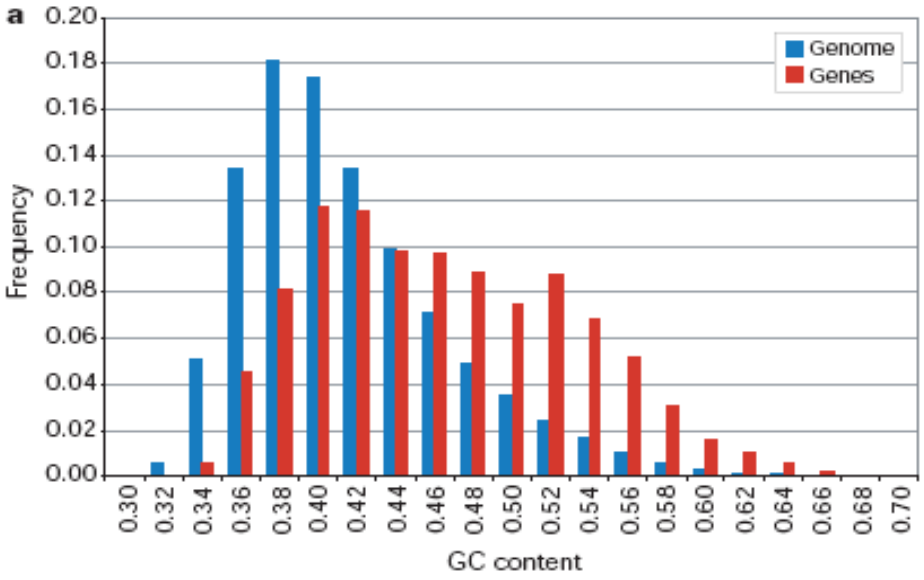


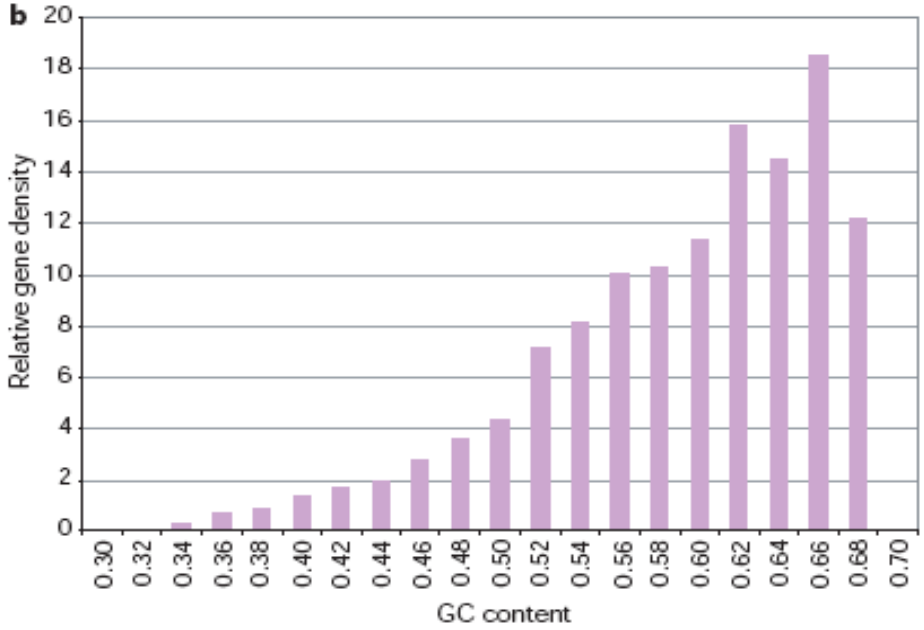
- Plot distribution: histogram of no. of windows *vs.* GC-content
  - Human genome is 41% GC
- Compute SD of distribution

GC content variation in human chromosome 1



Genes prefer higher GC regions





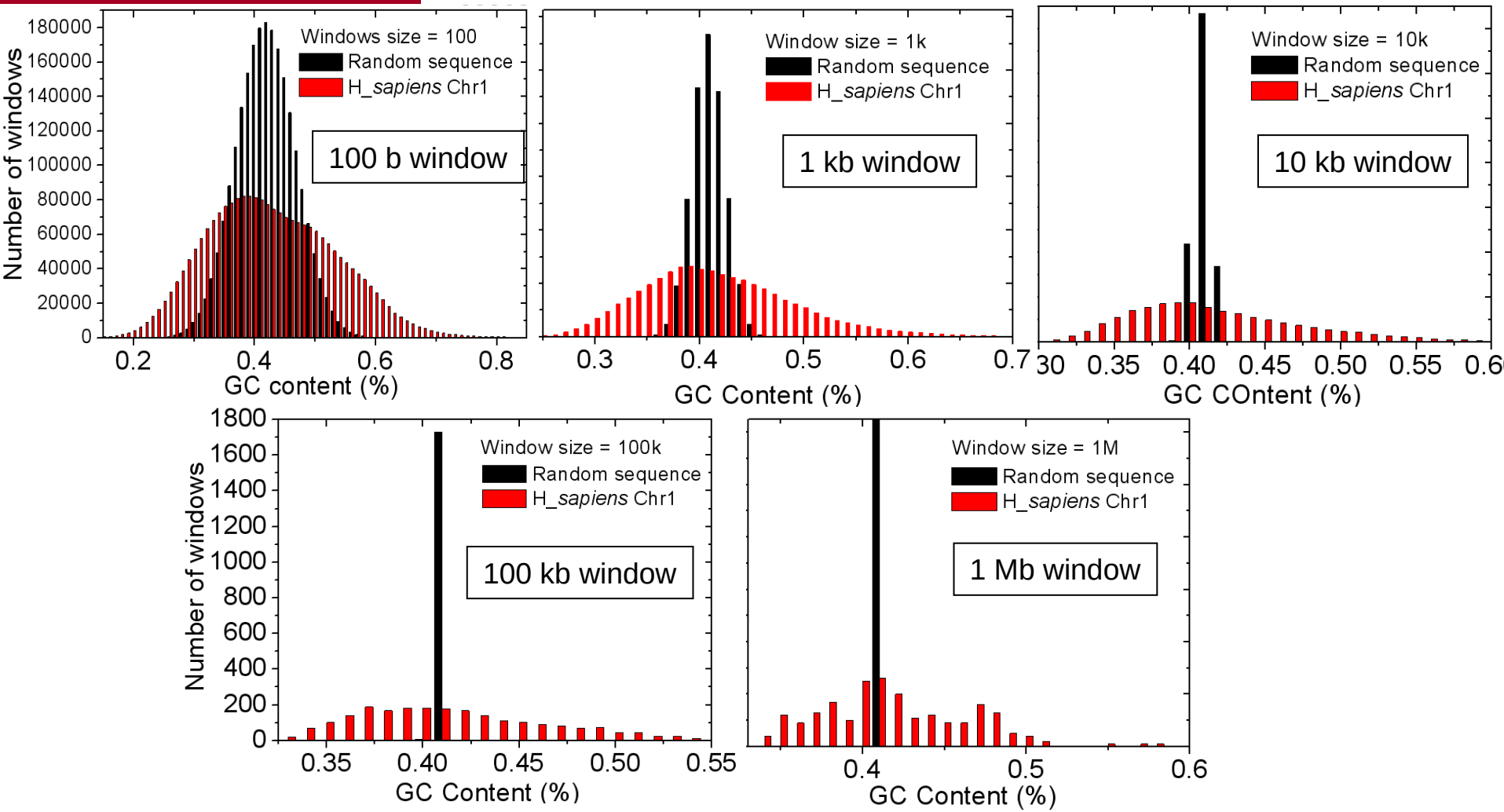
Why was result strange? The Central Limit Theorem

- A large number of independent observations from the same distribution has an approximate normal (Gaussian) distribution whose variance is inversely proportional to sample size.
  - PS Laplace 1810; A. Lyapunov 1899.
  - S. Bernstein, *Math. Ann.* 97:1-59 (1927) M. Rosenblatt, *PNAS* 42:43-47 (1955)

- Roughly:

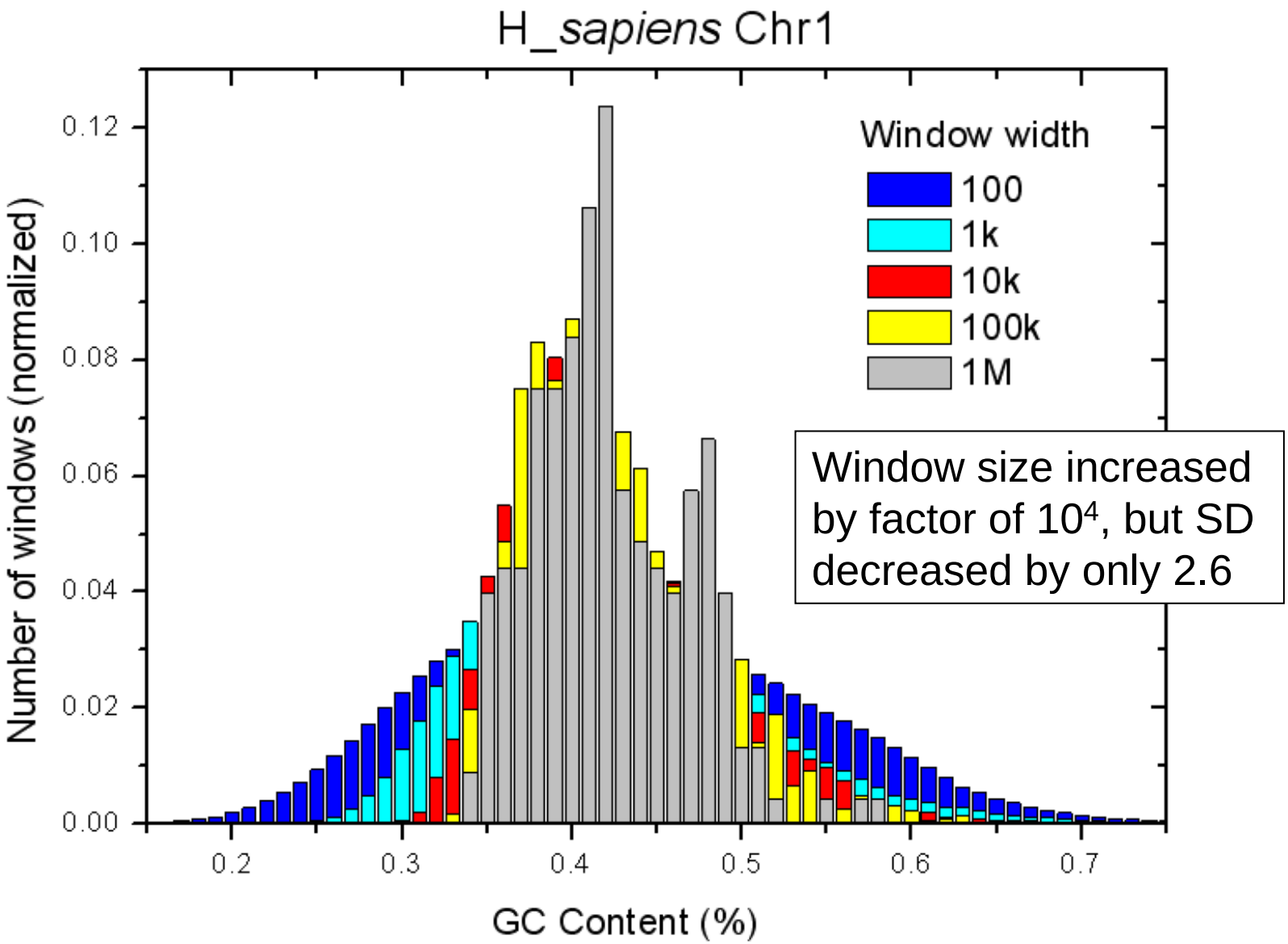
$$(SD)^2 \sim 1/(\text{sample size})$$

For GC-content histograms: sample size = window size



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

CG-content in Human genome has long-range variation



But it does obey a power law

- Power law

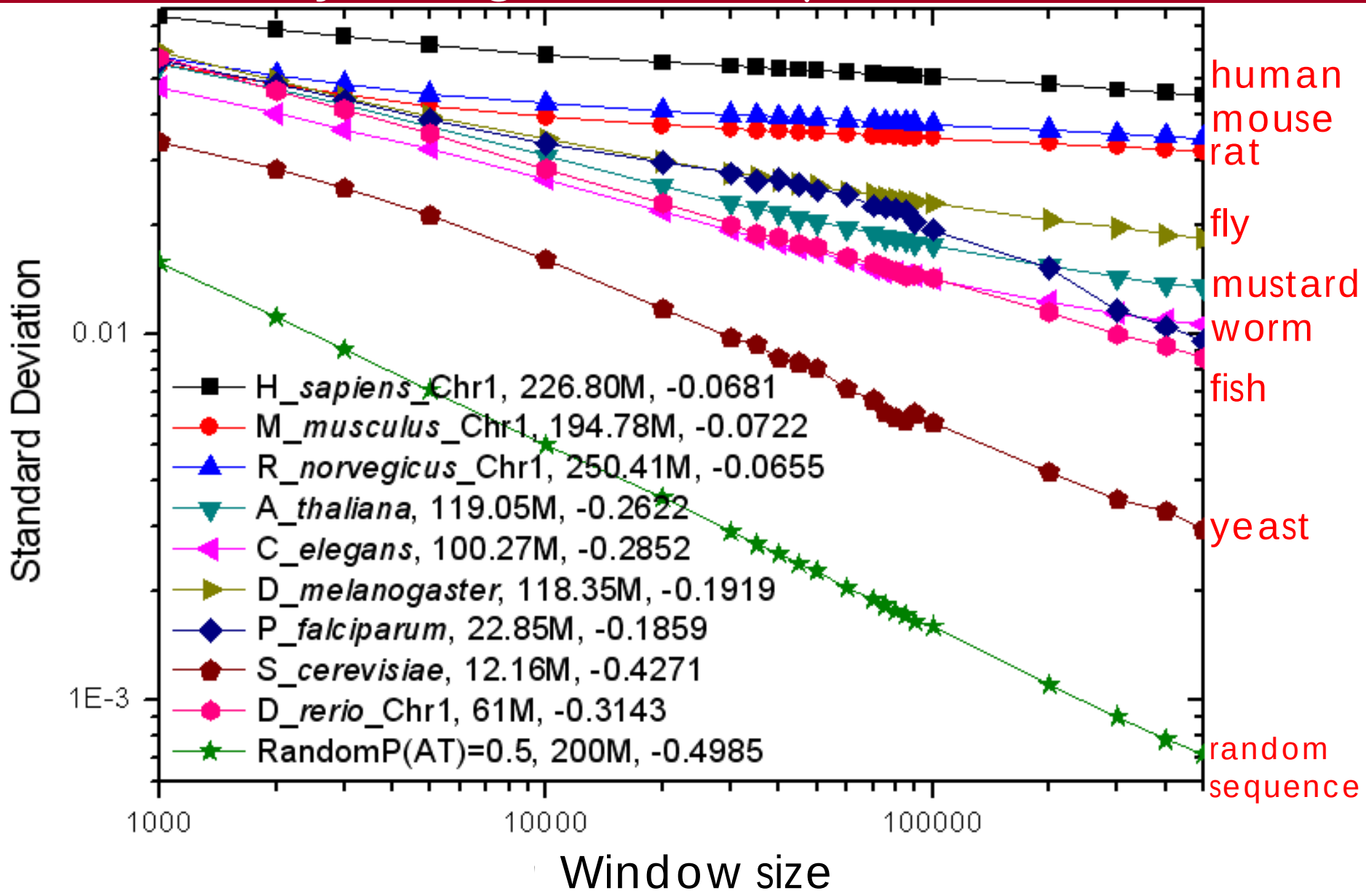
$$SD \sim (\text{window size})^{\gamma}$$

(Log-log plot is a straight line with slope γ)

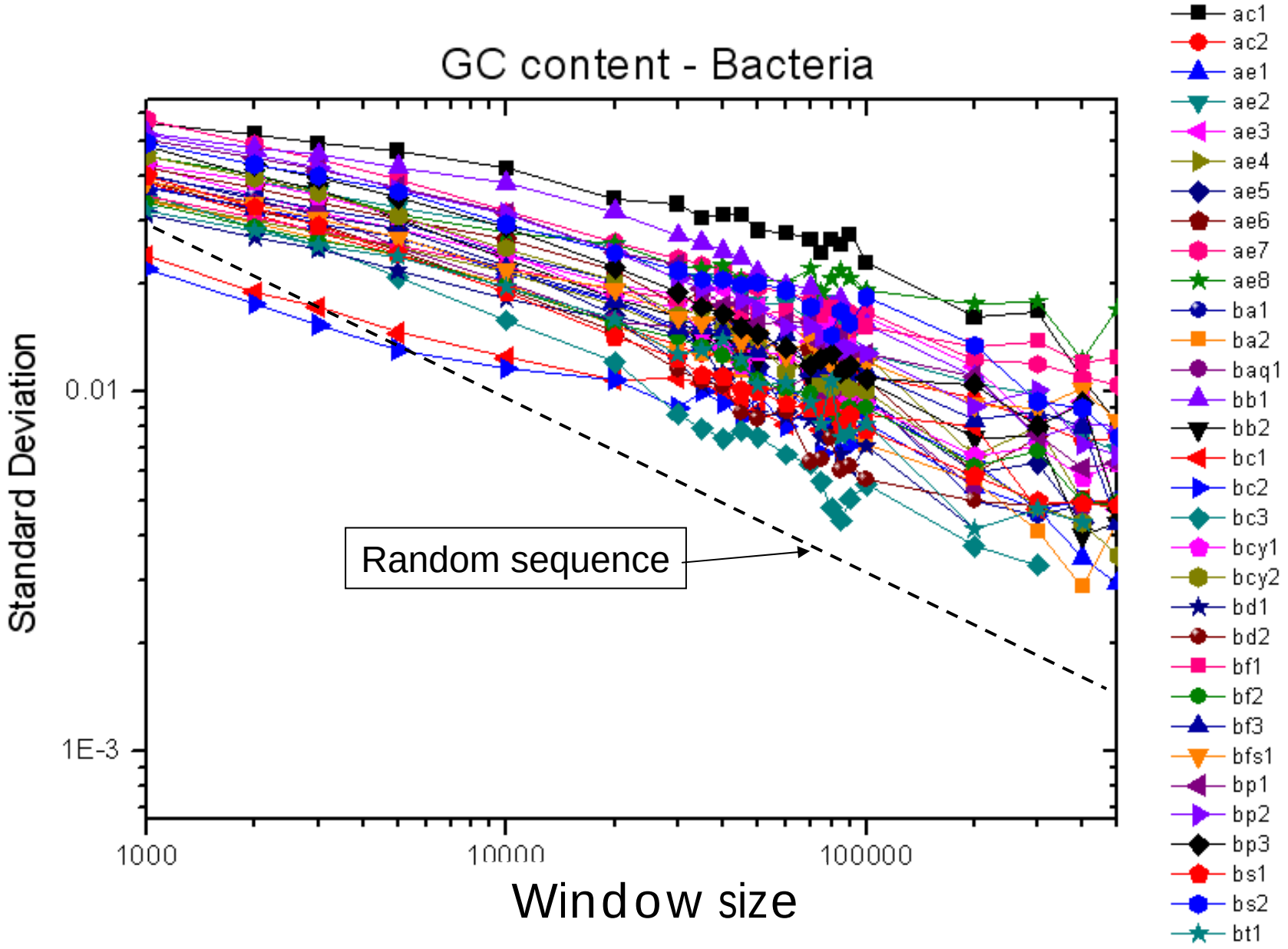
- Central limit theorem: $\gamma = -0.5$

- Human chromosomes: $\gamma \sim -0.07$

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



And for bacteria: γ=-0.16 to -0.45



Power law results from scale invariance

- Let $f(x)$ be a function of a scale (i.e., a length) variable
- Consider the property of f when x is changed by a scale factor: $x \rightarrow \lambda x$
- The function f is scale invariant if

$$f(\lambda x) = \lambda^{\gamma} f(x)$$

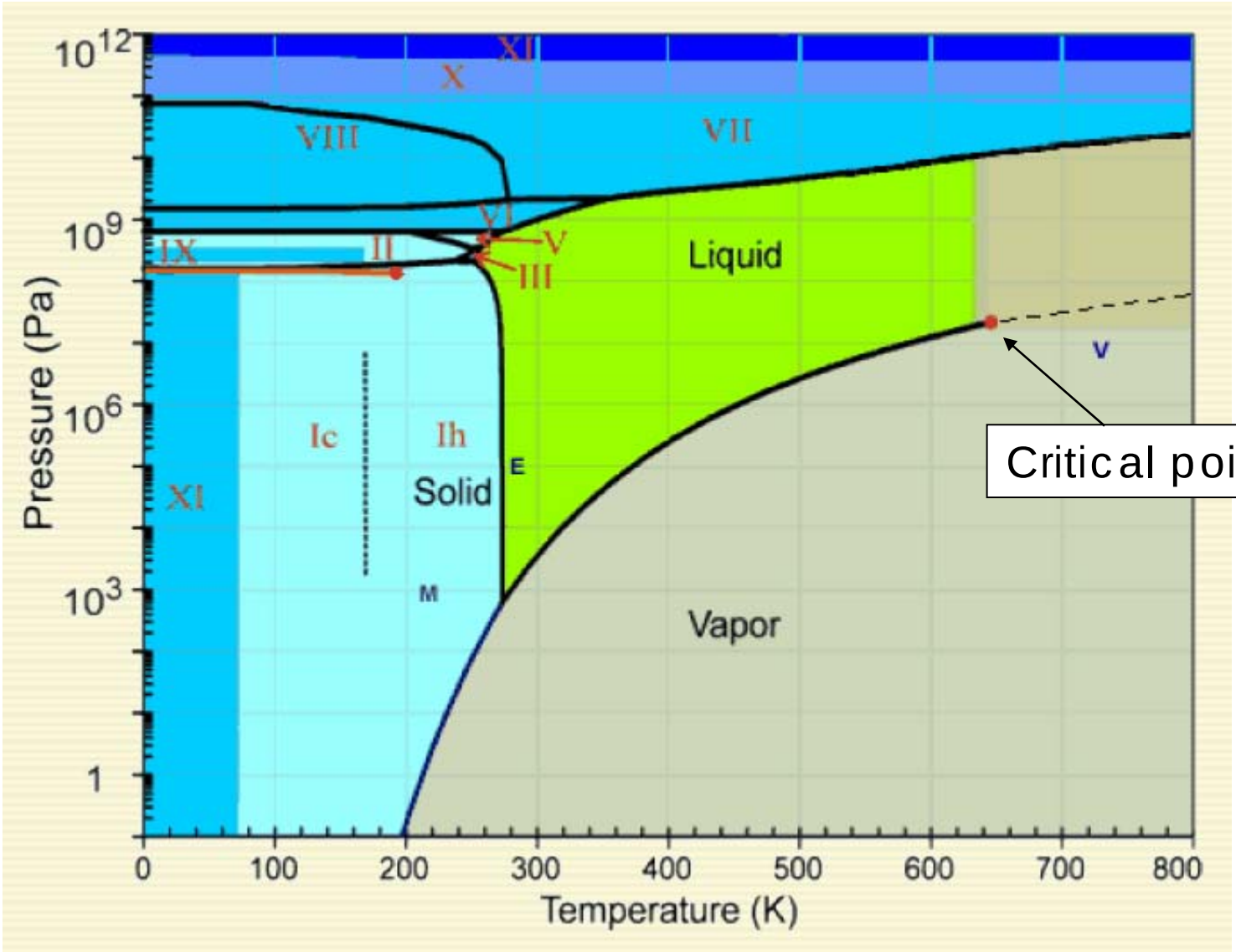
  - γ is the scaling exponent
- Exercise: Show that f obeys power-law:

$$f(x) \sim x^{\gamma}$$

Criticality & scale invariance

- Criticality refers to the behaviour of extended systems at a phase transition where scale invariance and self-similarity prevails.
  - Criticality in material (often) requires fine-tuning in external conditions such as temperature, pressure, etc.
  - Challenging field of study in theoretical physics (water, spin-systems, condensed matter)

Opalescence (fusing of liquid & vapor phases) in water occurs at 650 K & 2×10^7 Pa



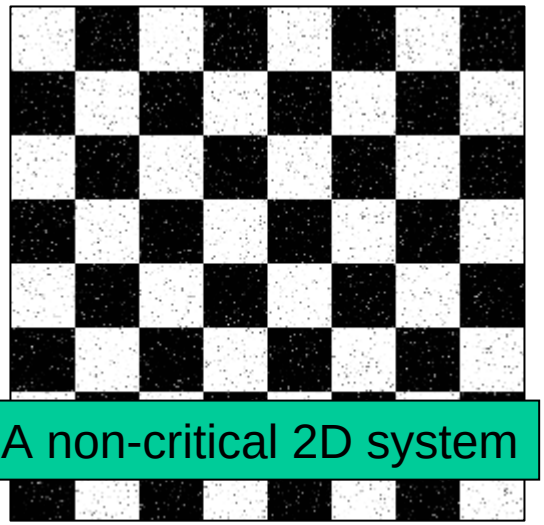
Criticality = Scale invariance + self-similarity

- Scale invariance: there are domains of all sizes

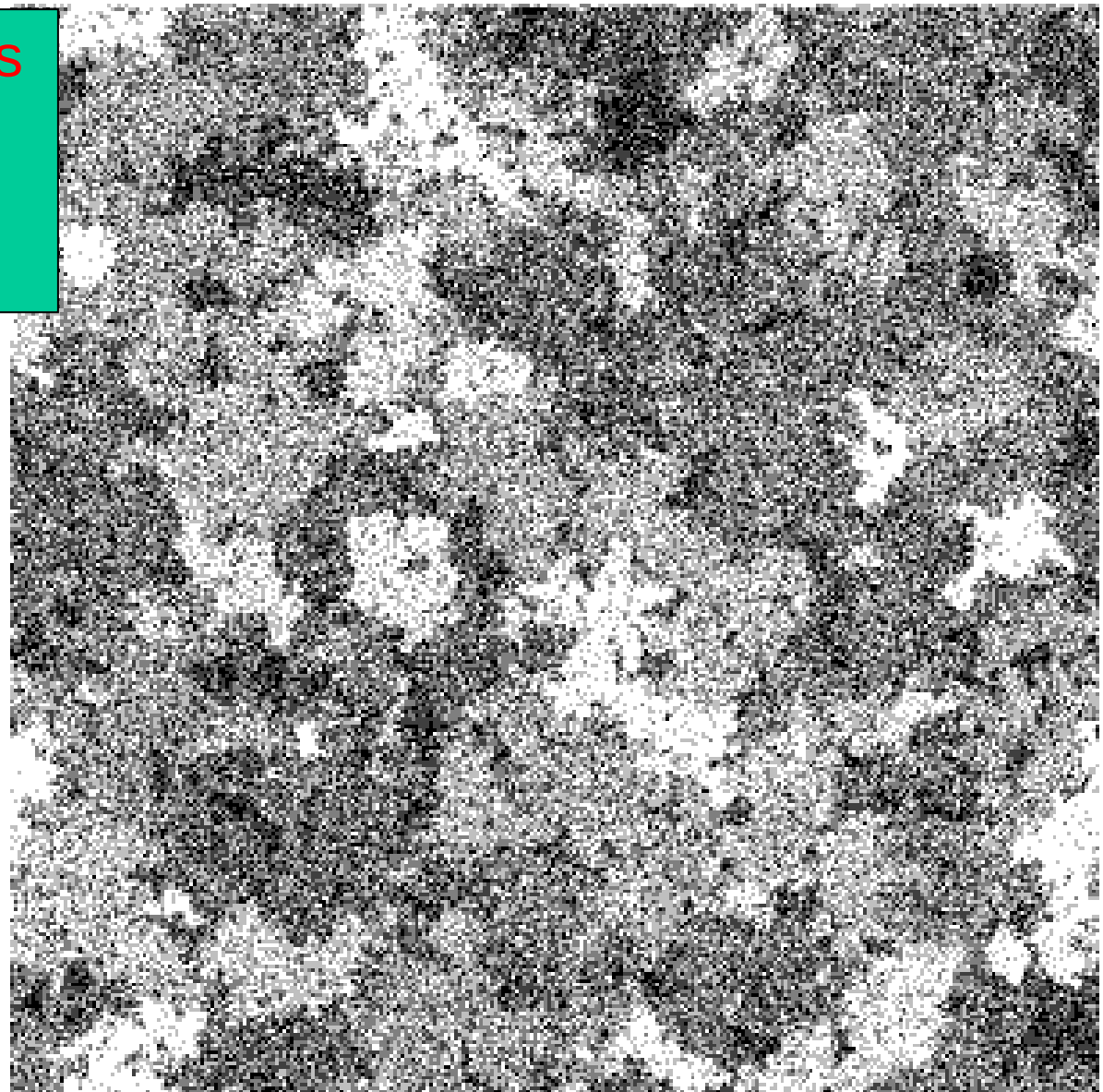
- Self-similarity: there are domain (of all sizes) within domains

A 2D critical spin-system
(black: spin-up; white: spin-down)

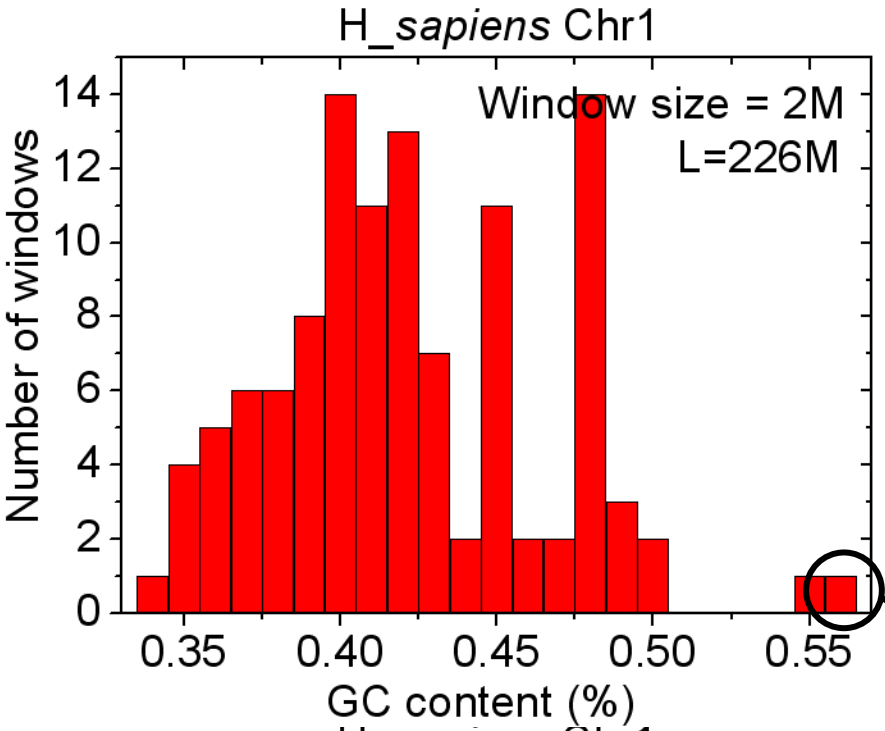
There are domains of all sizes, and there are domain within domains

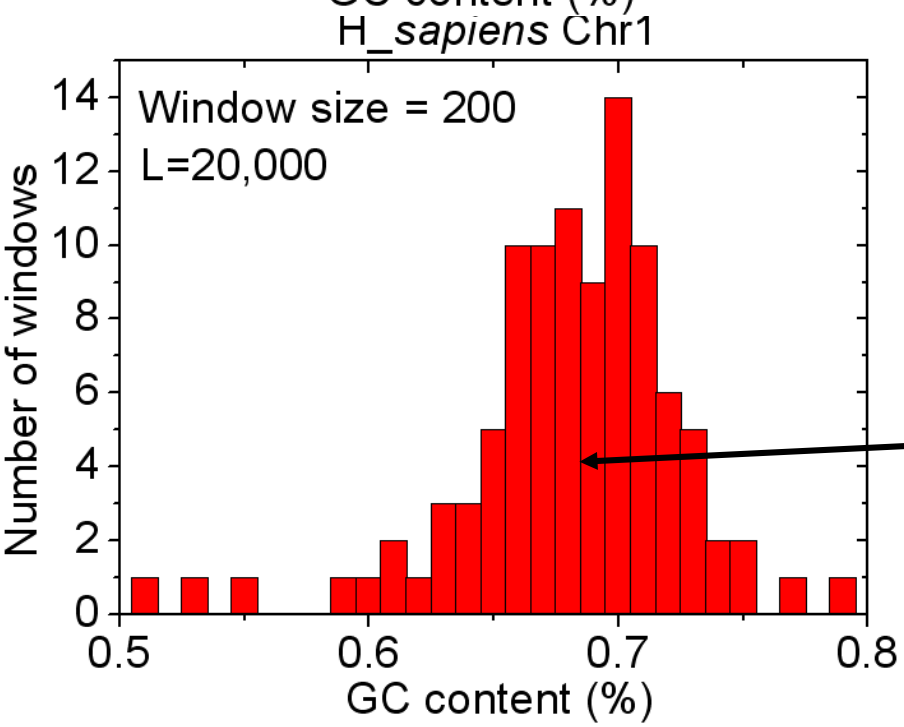


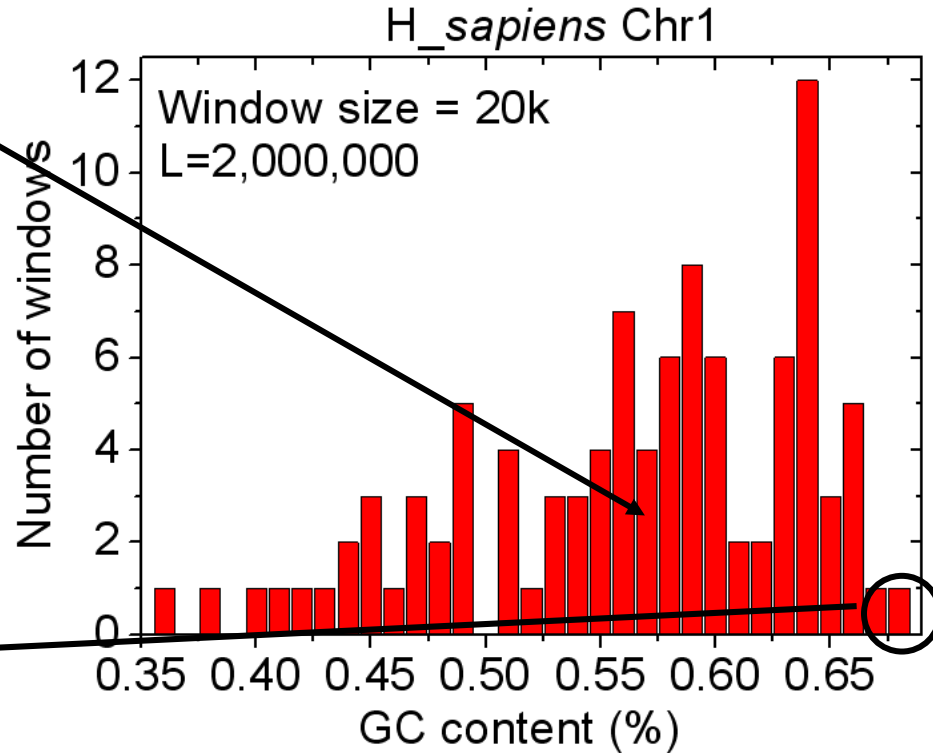
A non-critical 2D system



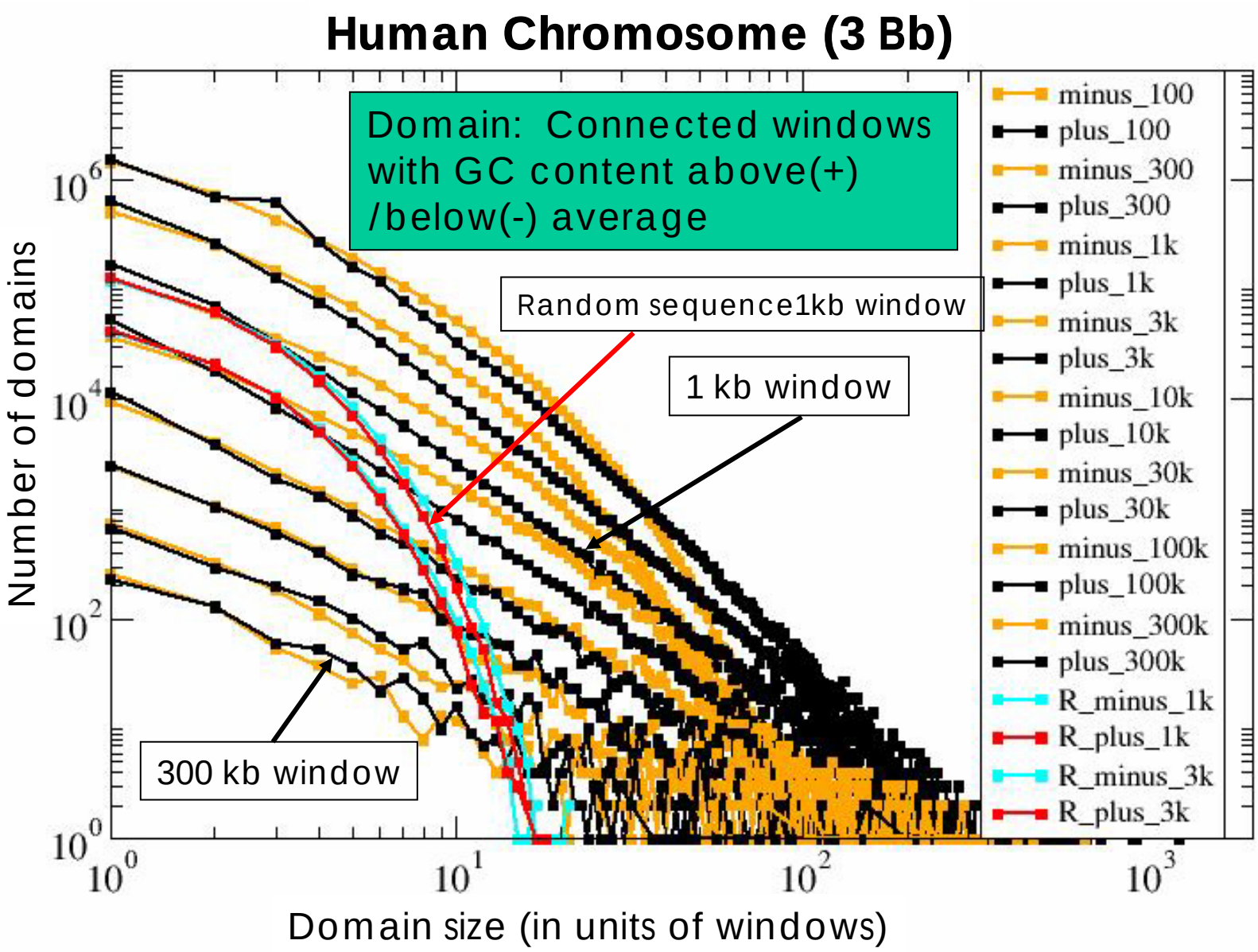
Self-similarity in human chromosome 1







Another look at HS genome: scaling in domain size



Genomes are critical

- Genomes exhibit non-trivial power-law behavior
  - Long-range variation in GC content
- Genomes exhibit self-similarity
  - Within each GC-specific domain there is another level of long-range variation
  - Genomes do not have isochores

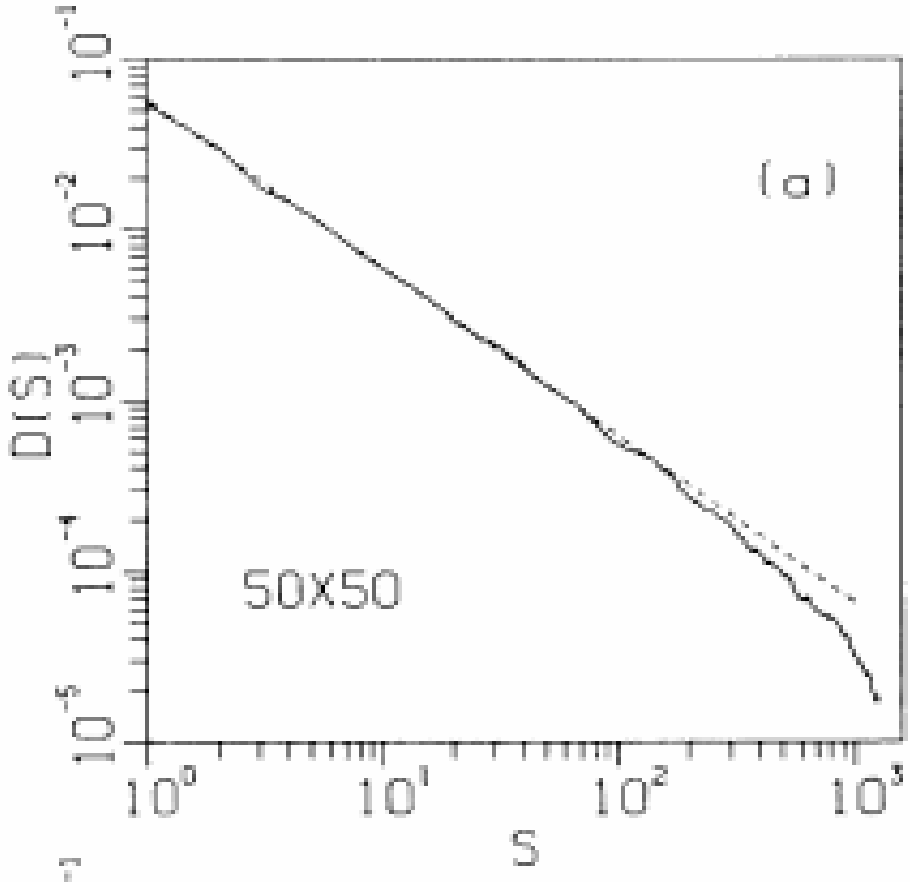
Two highly non-trivial properties of genome

- Universality in L_{eff}:
  - Log $L_{eff}(k) = ak + B$
    a, B are universal constants
- Criticality in GC content
  - Similarly in other "contents"
- What is the cause of these properties?

Self-Organized Criticality

- Many critical systems in Nature are self-organized: earthquakes in seismic systems, avalanches in granular media and rainfall in the atmosphere.
- Bak-Tang-Wiesenfield sandpile model
  - Phys. Rev. Lett. 59, 381–384 (1987)
  - extended dynamical systems governed by simple rules
  - robust critical fixed point
  - dissipative to stay at criticality

Sandpile model: size of avalanche has power-law distribution



Bak-Tang-Wiesenfield PRL (1987)

The └Critical Blind Self-Copier

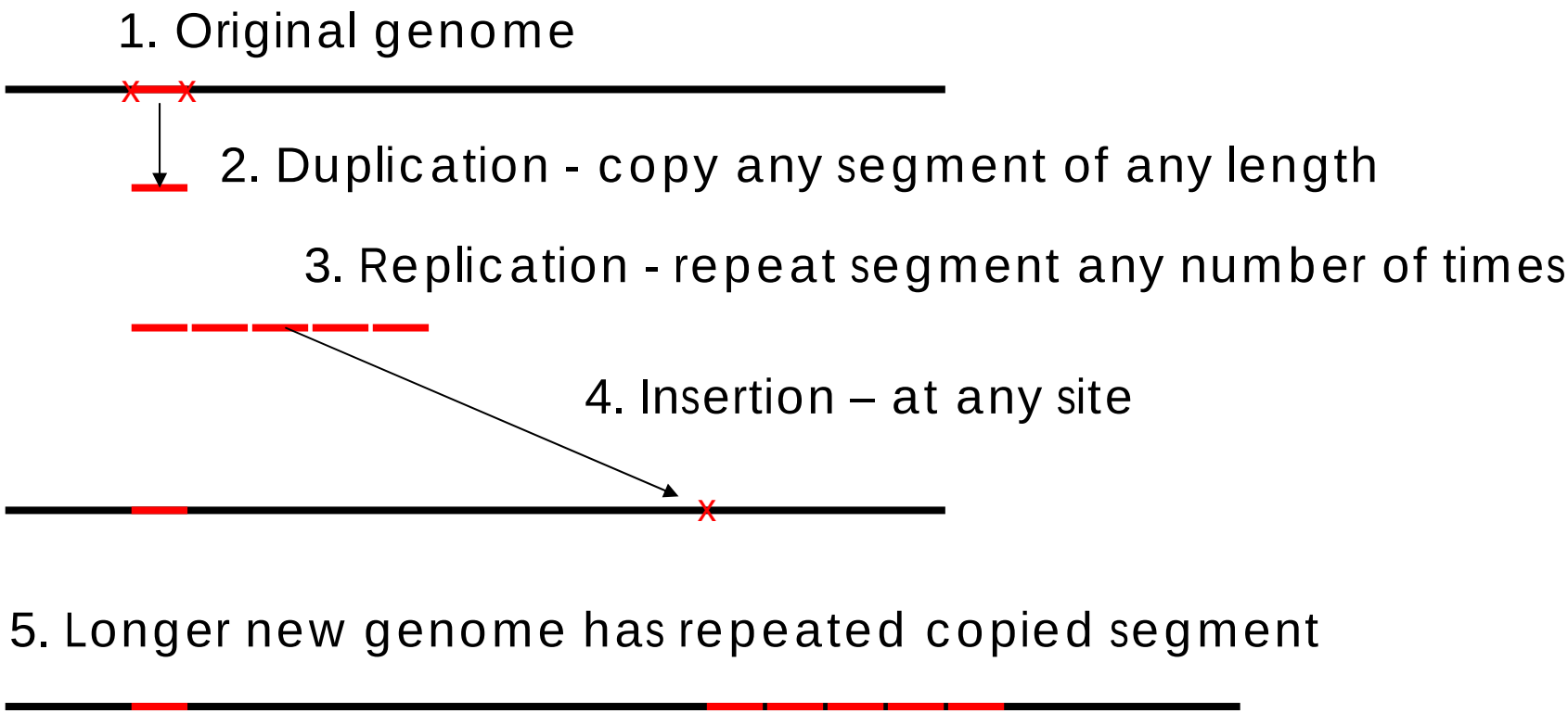
Simple growth model does not have long range variation

- Model has two scales: initial length and maximum segment length
  - HIDDEN parameter: distribution of lengths of copied segments
  - Used square distribution with maximum length of several kb (result insensitive to maximum length if >~ 1kb)
- Hence model sequence cannot be scale invariant

Growing genome must outgrow its own scale

- Critical segmental duplication
- Duplicated segment
  - Any length up to current genome length (power law)
  - Repeated a random number of times before insertion

Five steps of critical segmental duplication

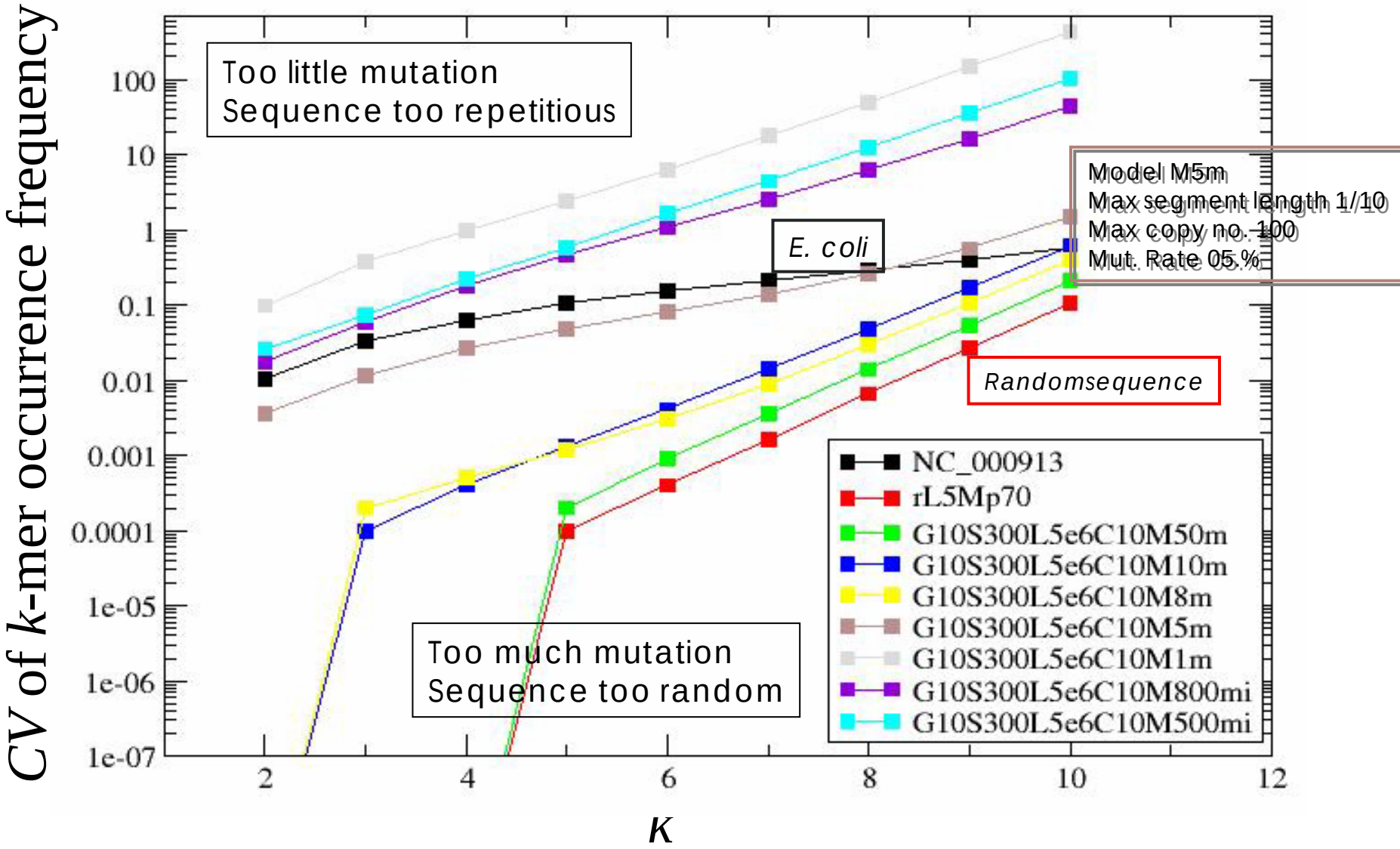


Maximally random: All selections are random

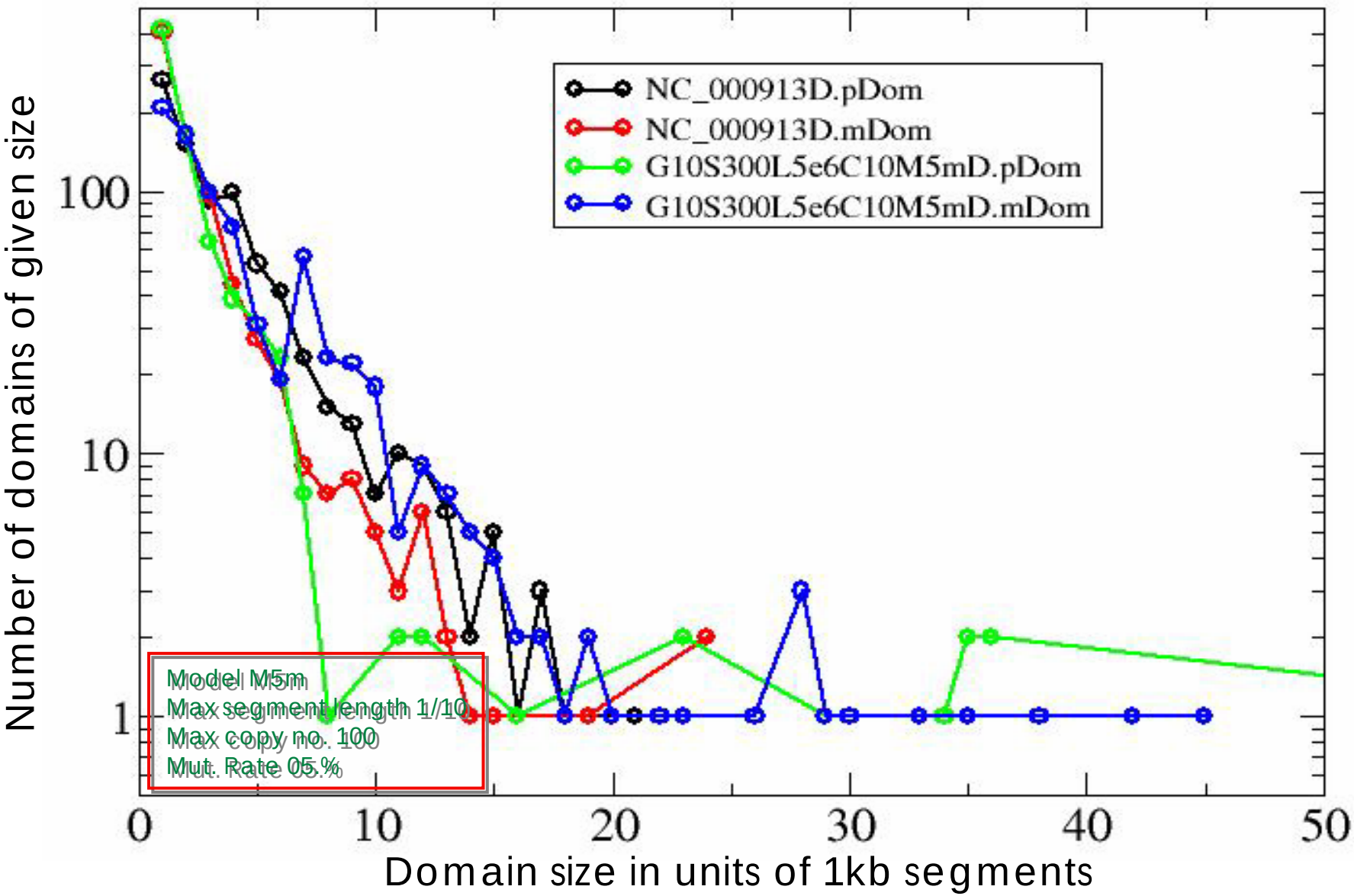
The right stuff is in the detail

- Distribution of duplicated segment length: power law $P(l) \sim l^{-1}$
- Maximum duplicated segment length: $l_{max} = L_{current}/10$
- Maximum times segment repeated: $n = 100$
  - $n >> 1$ is key to criticality
- Point mutation per duplication event per sequence length $R \sim 0.001$ to 0.01
- Generated sequence length: $L \sim 5$ Mb

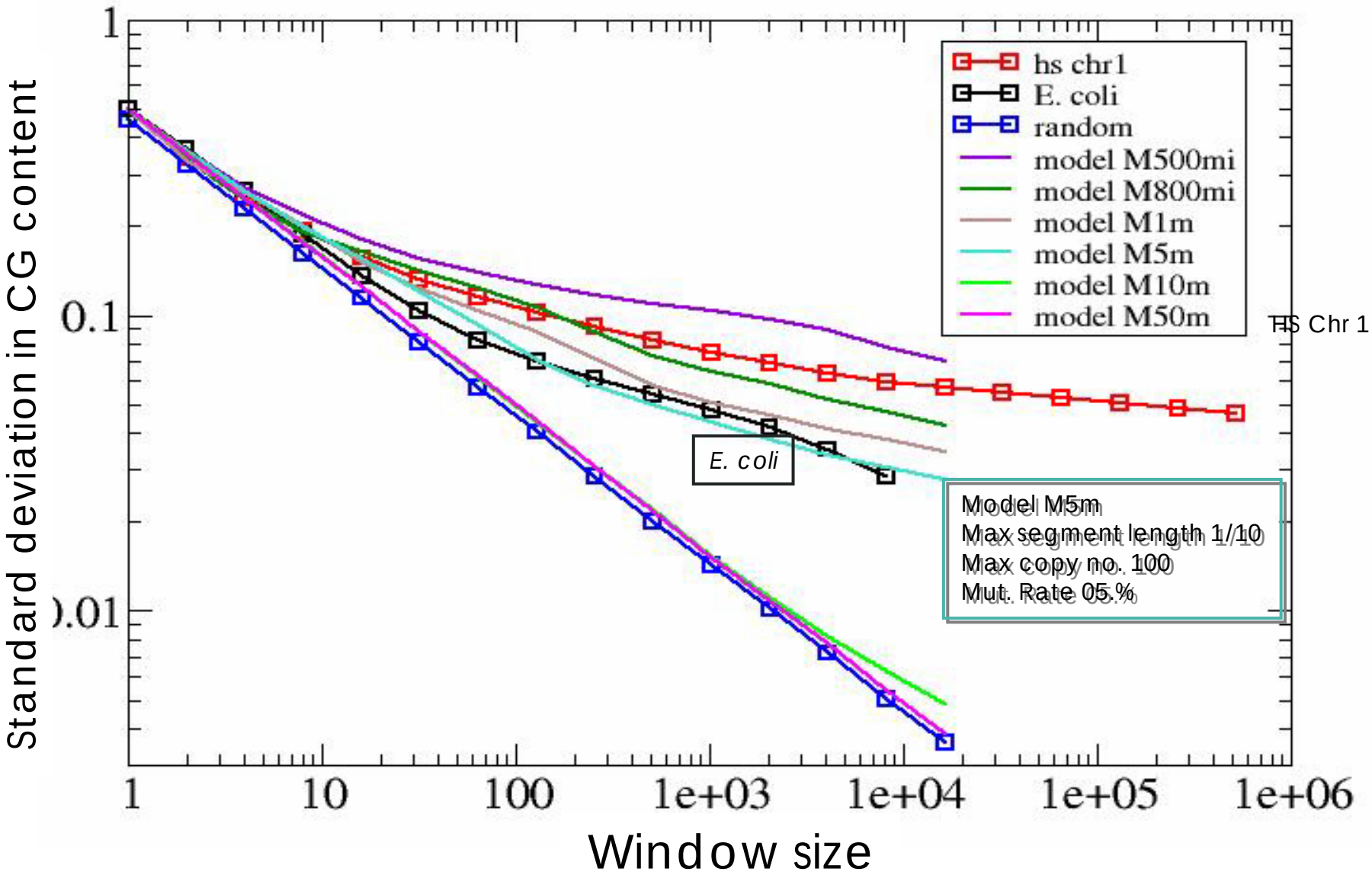
Correct *CV* value requires delicate balance between duplication and mutation



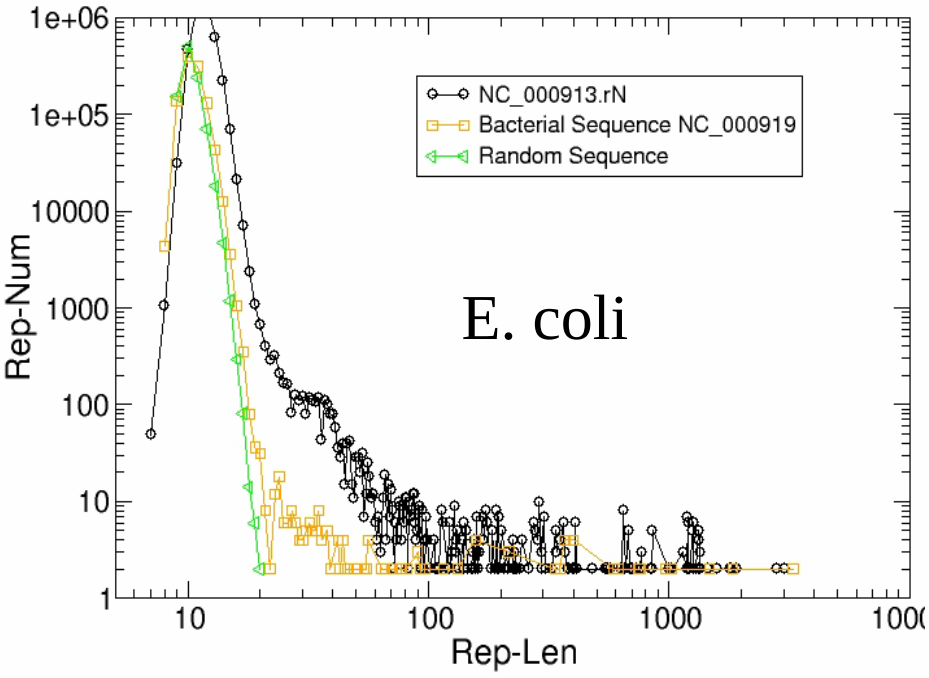
Size distribution of GC+ domains

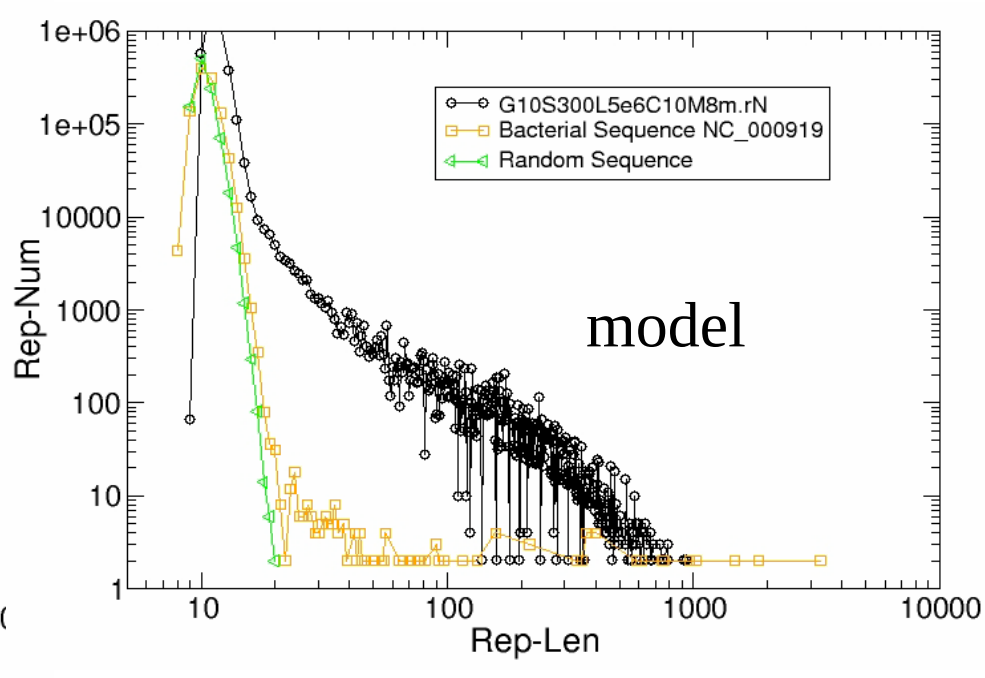


Size distribution of GC+ domains



Work in progress: model sequence has too many mid-length repeats but too few very long ones





More fine-tuning needed

Why would genome grow by Critical segmental duplication?

- Rapid **rate of evolution** - random self-copying is an extremely efficient way for information accumulation; it is genome's way to "beat" the **2nd law of thermo-dynamics**

- Growth by critical self-copying is a result of **natural selection**

Many biological phenomena explained by model

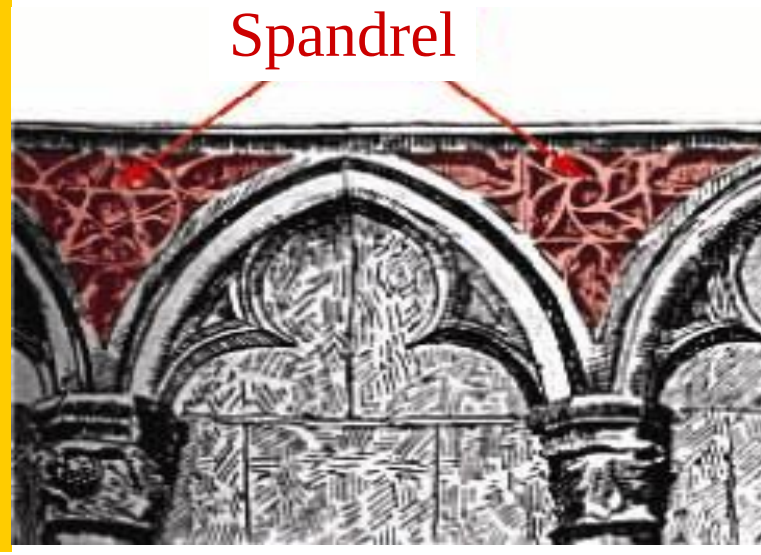
- Preponderance of **homologous genes** in all genomes
  - All genes belong to homologous families
- Genome is full of **non-coding repeats**
- **Transposons** and "**operons**"
- Large-scale genome "**rearrangments**"
- Huge species **diversity**
- Apparent "**missing links**" are natural
- Coexistence of gradualism (Dawkins) and punctuated equilibrium (Gould)
- Many more ...

Successful model should be able to explain the statistics of:

- Intra-genomic and Inter-genomic distributions of frequency of occurrence of homologous genes (orthologs and paralogs)
  - Same for non-coding genes
- All kinds of repeats
- Size and frequency of operons
- Correlations among control sequences and signals and genes (and operons)
- Others …

Are genes “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)*

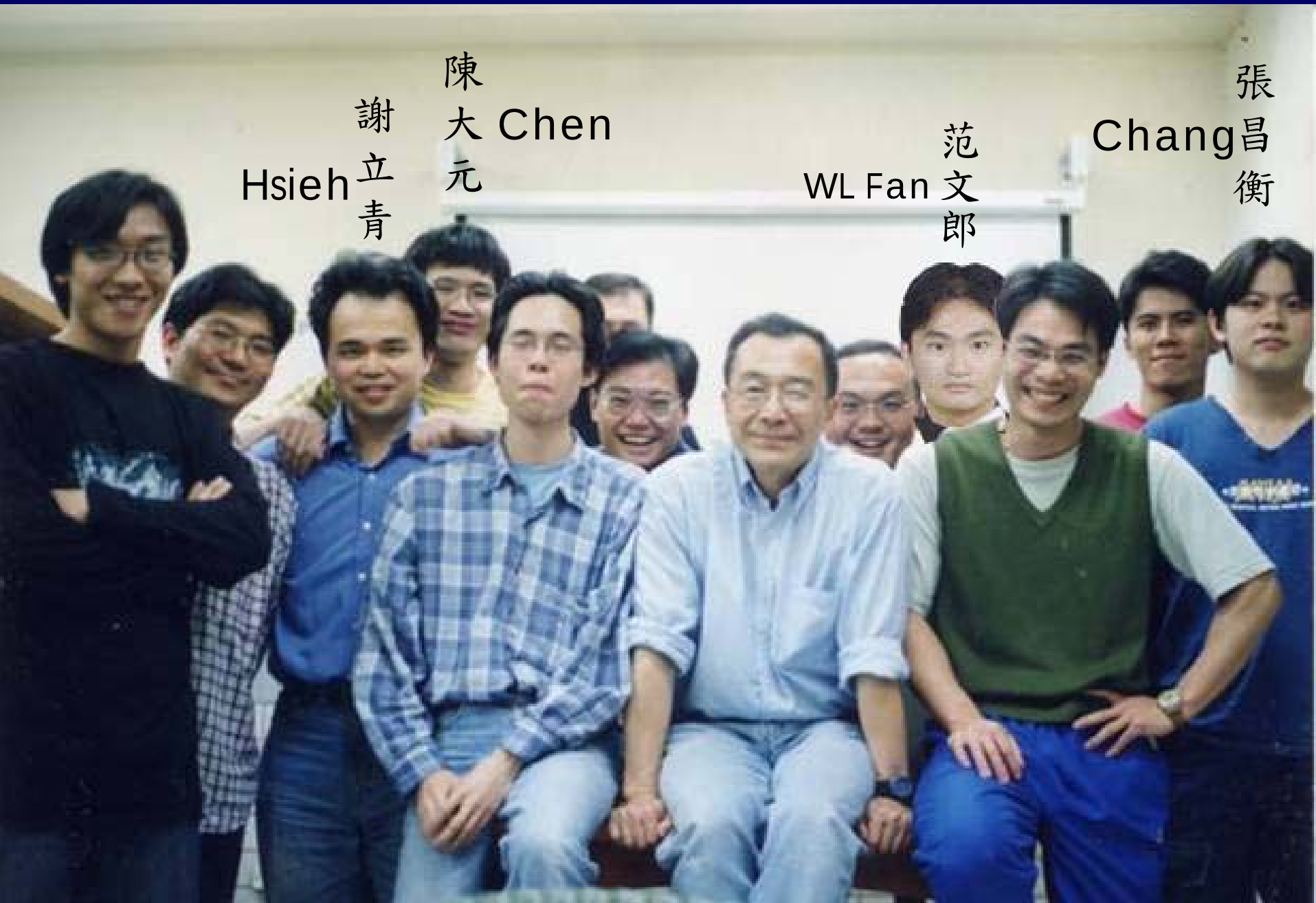


- Duplications to a genome are what the construction of arches, walls and ceilings are to a cathedral
- The product of duplication, the textual structure manifested in short L_{eff} are the spandrels and genes are décorations in the spandrels

People

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

Computation Biology Laboratory (2003)



Our papers are found at Google: HC Lee

Thank you!

Large CV, or small L_{eff}, implies more "information"

Compare L_{eff} with true length L for all complete genomes for 2-10 letter words

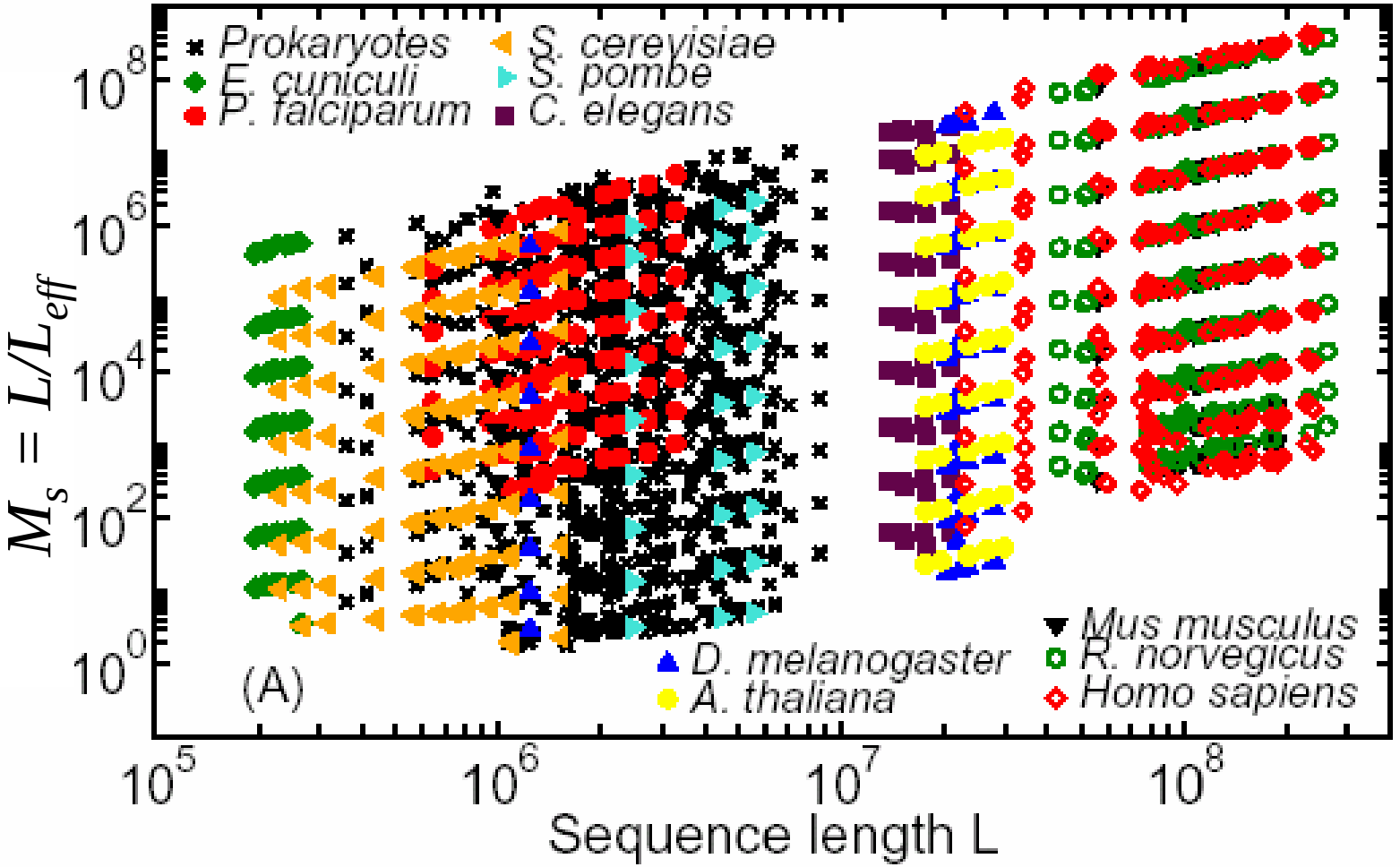
$$(CV_{genome})^2 = \tau / L_{eff}$$

$$\overset{def}{} \quad (CV_{random})^2 = \tau / L$$

$$M_s = (CV_{genome})^2 / (CV_{random})^2 = L / L_{eff}$$

Note: technical details when p not equal to 0.5

Results: color coded by organisms



Each point from one *k*-spectrum of one sequence; ~2000 data points. Black crosses are microbials. Data shifted by factor 2^{10-k}.