

# Self-Organized Criticality In Genomes

*Physics Department Colloquium  
National Central University  
2006 May 23*

HC Lee  
*Computational Biology Lab  
Inst. Systems Biology & Bioinformatics  
Dept. Physics & Inst. Biophysics  
National Central University*

# Three questions we should ask

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

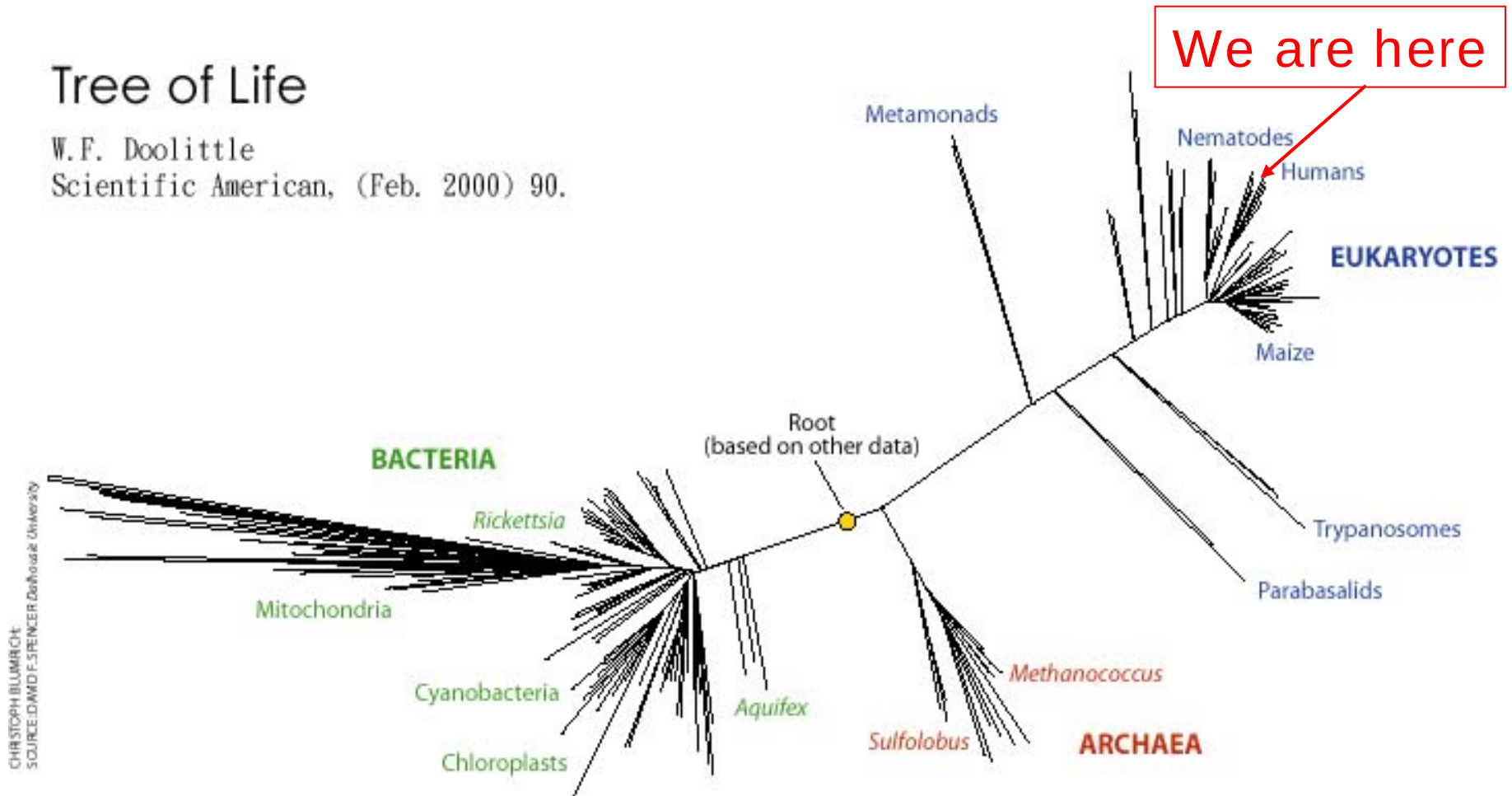
# Some concepts to be discussed

- Genome – book of life
- Genome as text
- Average and standard deviation
- The central limit theorem (in probability)
- Power-law in complete genomes
- Scaling and power-law
- Criticality and scale invariance
- Self-organized criticality
- Genome the blind critical self-copier

# Life is highly diverse and complex

## Tree of Life

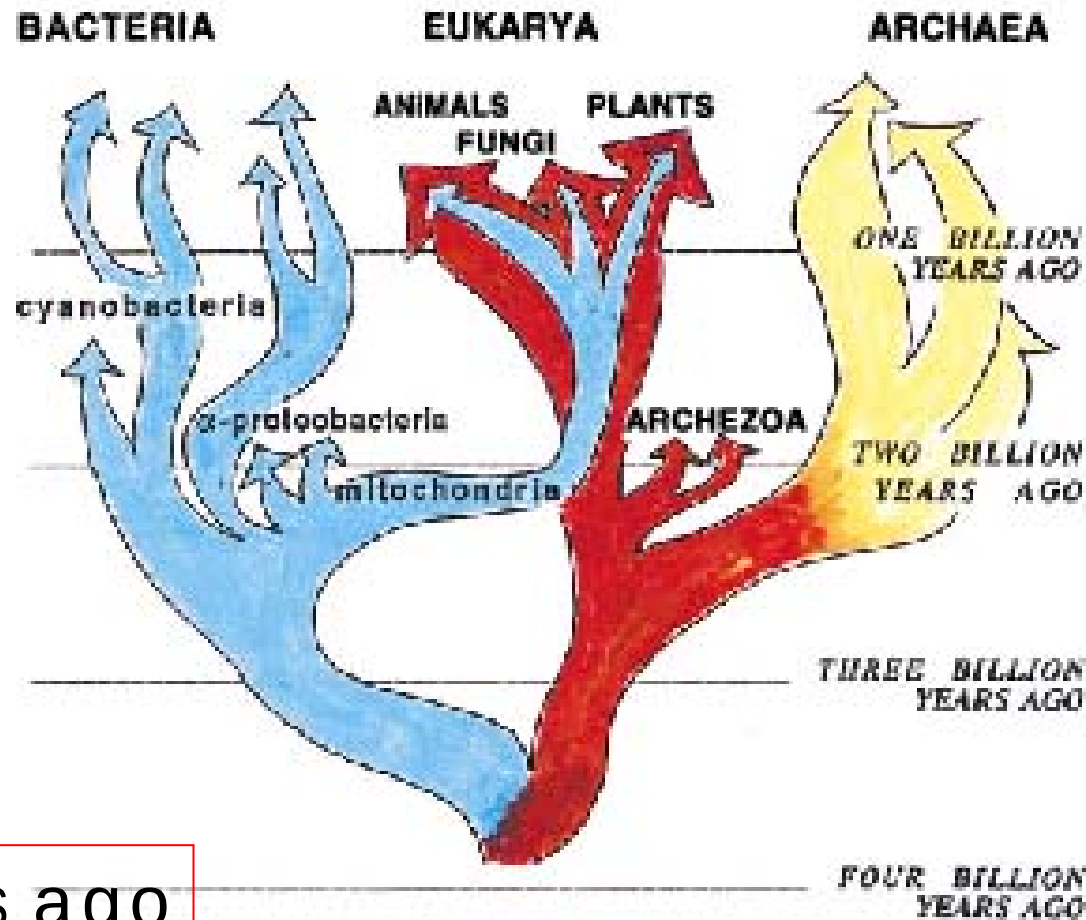
W.F. Doolittle  
Scientific American, (Feb. 2000) 90.



# And it took a long time to get here

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

now



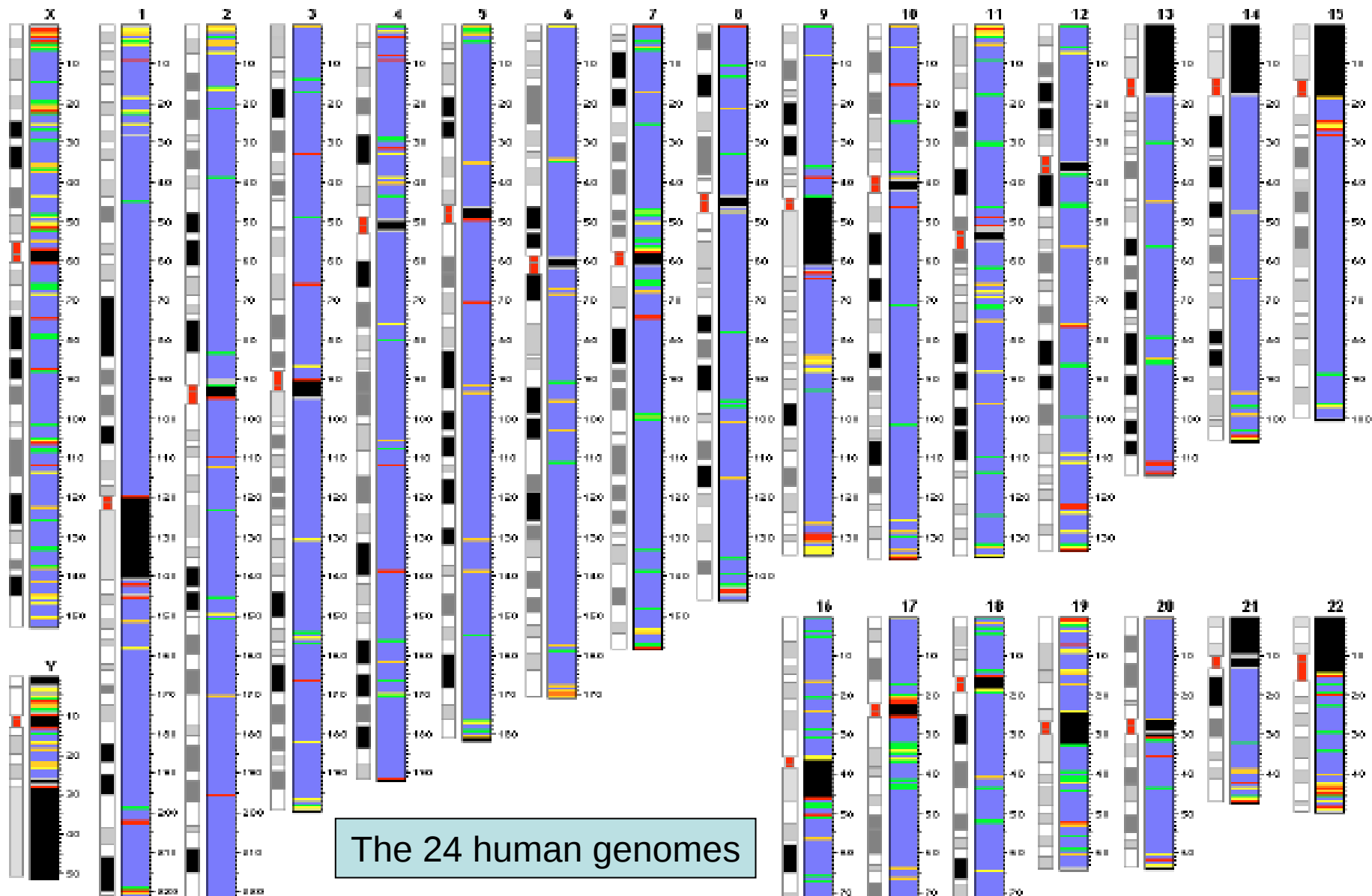
4 billion yrs ago

# 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](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 tcccaaag acacttcgca gcccctcttg
61 ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggctg
121 ctctctcac gccggcccc accctgggcca cctgtggcac ctgccaggag gctgagctgc
181 aaacccaat gaggggcagg tgctcccgga gacctgcttc ccacacgccc atcggttctg
241 cccgggcttt gagttctccc aggcccctct gtgcaccctt ccctagcagg aacatgccgt
301 ctgccccctt gagctttgca aggtctcggg gataatagga aggtctttgc ctgcaggga
361 gaatgagtc tccgtgctcc ctccgagggg gattctggag tccacagtaa ttgcagggt
421 gacactctgc cctgcaccgg gcgccccage tctccccac ctccctctc catccctgtc
481 tccggctatt aagacggggc gtcaggggc ctgtaactgg ggaaggata cccgccctgc
541 agaggtggac cctgtctgtt ttgatttctg ttcatgtcc aaggcaggac atgaccctgt
601 ttggaatgc tgatttatgg atttccagg cactgtgcc ccagatacaa tttctctga
661 cattaagaat acgtagagaa ctaaatgcat ttcttctta aaaaaaaaaa aaaccaaaaa
721 aaaaaaaaaa aaaccaaaaa actgtactta ataagatcca tgcctataag acaaaggaa
781 acctctgtc atatatgtgg gacctcgggc agcgtgtgaa agtttacttg cagtttgca
841 taaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
901 tatttttct aatctgggca acaatgggtgc catctcgggt cactgcaacc tccgcttccc
961 aggttcaagc gattctccgg cctcagcctc ccaagtagct gggaggacag gcacccgcca
1021 tgatgcccg ttaattttg tatttttagc agagatgggt ttcgccatg ttggccaggc
1081 tggctcga ctcctgacct caggatgatcc gcctgccttg gcctcccaa gtgctgggat
1141 gacaggcgtg agccaccg cgccagccagg aatctatgca ttgccttg aatattagcc
1201 tccactgccc catcagcaaa aggcaaaaca ggttaccagc ctcccgccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg attttgctg aggttataag
1321 ccacttcctt catctgggtc tgagctttt tgtattcggg ctaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgctgagtc aatacaata
1441 gatttttaa gtgtatttat tttaacaaa aaataaaatc acacataaga taaaacaaa
1501 cgaaactgac ttatacagt aaaataaacg atgcctgggc acagtggctc acgcctgtca
```

# Genomes are Blind Self-Copiers - Summary of earlier work on genome sequence analysis

- Genomes form a **universality class**
  - universal effective lengths
  - maximal homogeneity in word-content
- Genomes are **Blind Self-Copiers**
  - Growth by **maximally random segmental duplication** (MRSD)
  - Very early onset of duplication process
- MRSD itself is chosen by natural selection

# Another clue from the human genome

NATURE | VOL 409 | 15 FEBRUARY 2001 | [www.nature.com](http://www.nature.com)

## articles

# Initial sequencing and analysis of the human genome (3.36 x 10<sup>9</sup> 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<sup>1-3</sup> 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

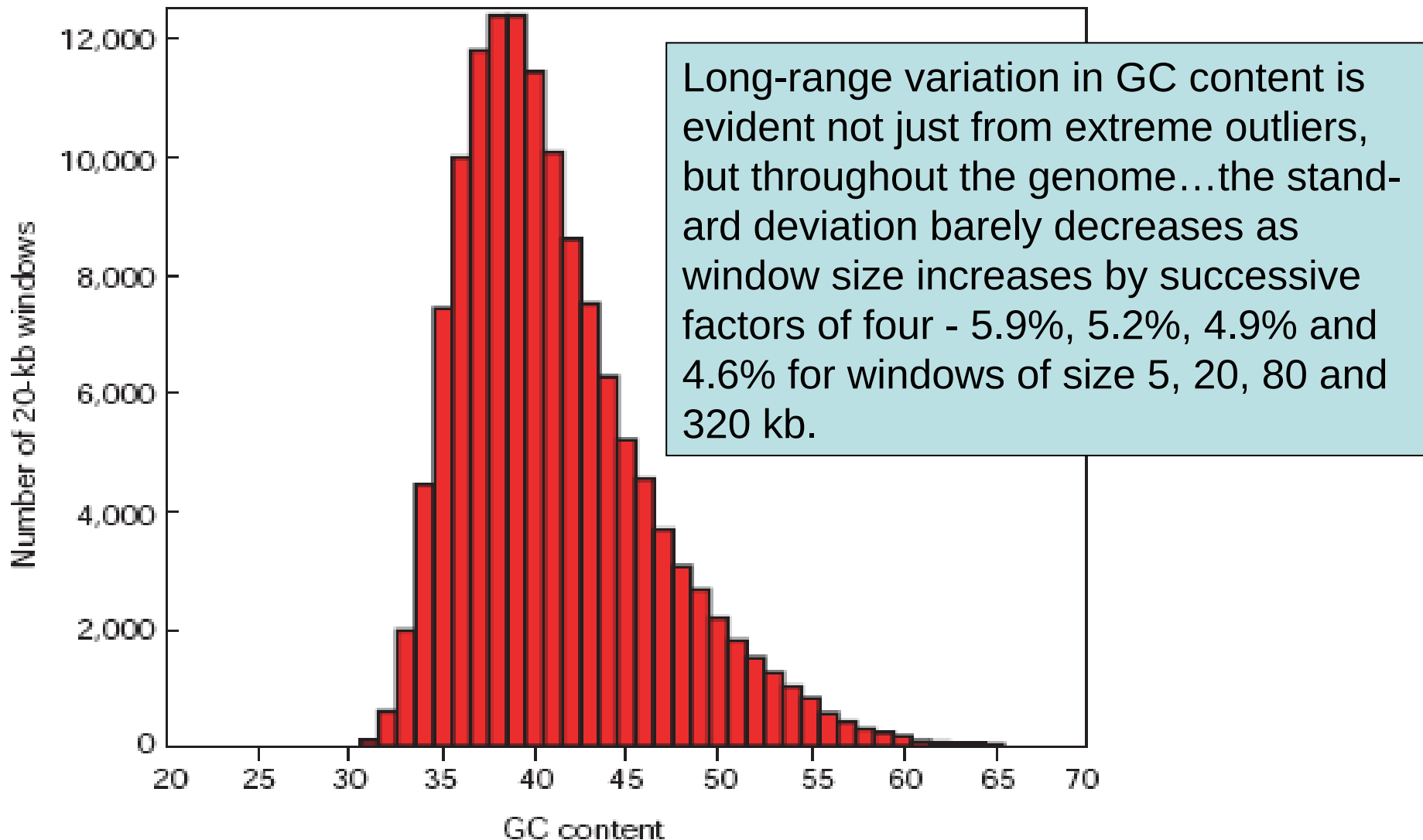
coordinate regulation of the genes in the clusters.

- 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-

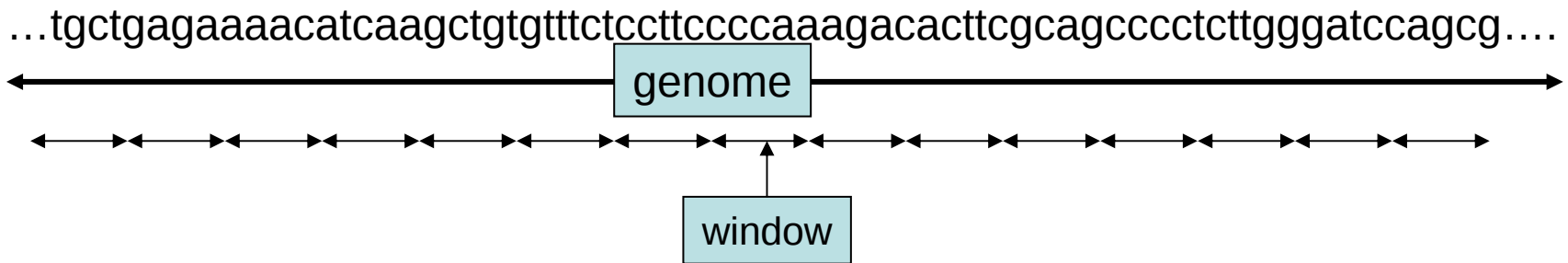
## Long-range variation in GC content



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

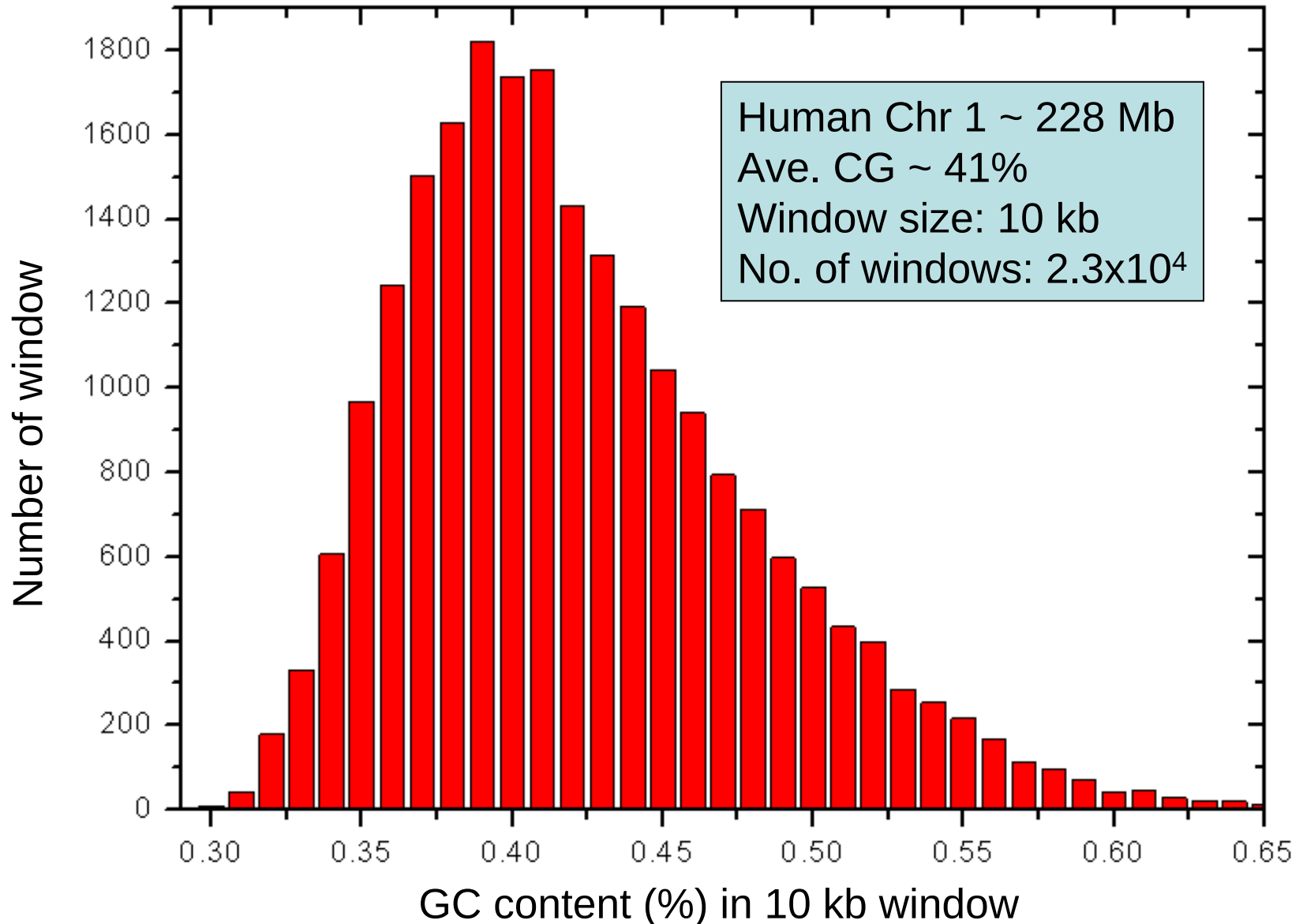
# What was measured? And why?

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

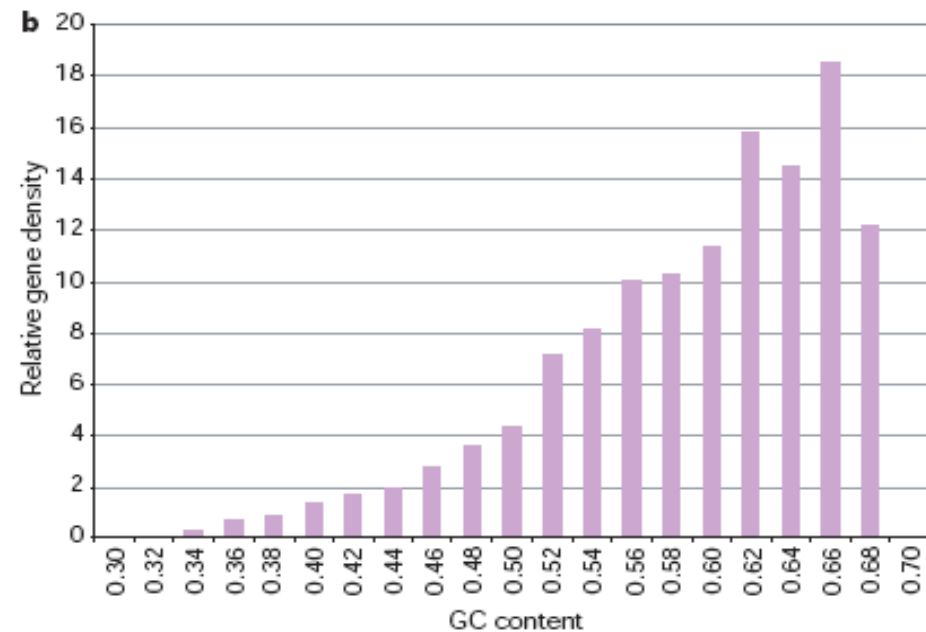
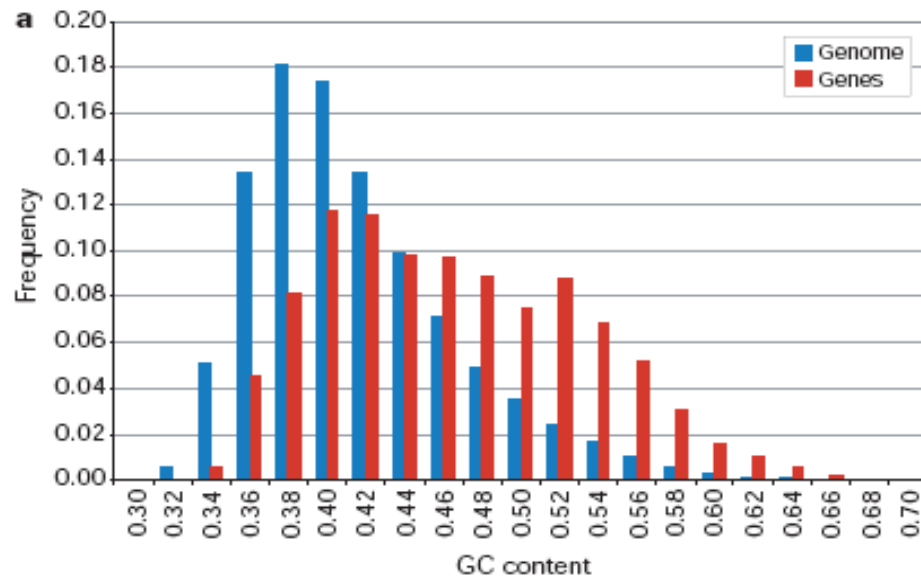


- Plot distribution: histogram of no. of windows vs. GC-content
  - Human genome is 41% GC
- Compute SD of distribution

# GC content variation in human chromosome 1



# Genes prefer higher GC regions

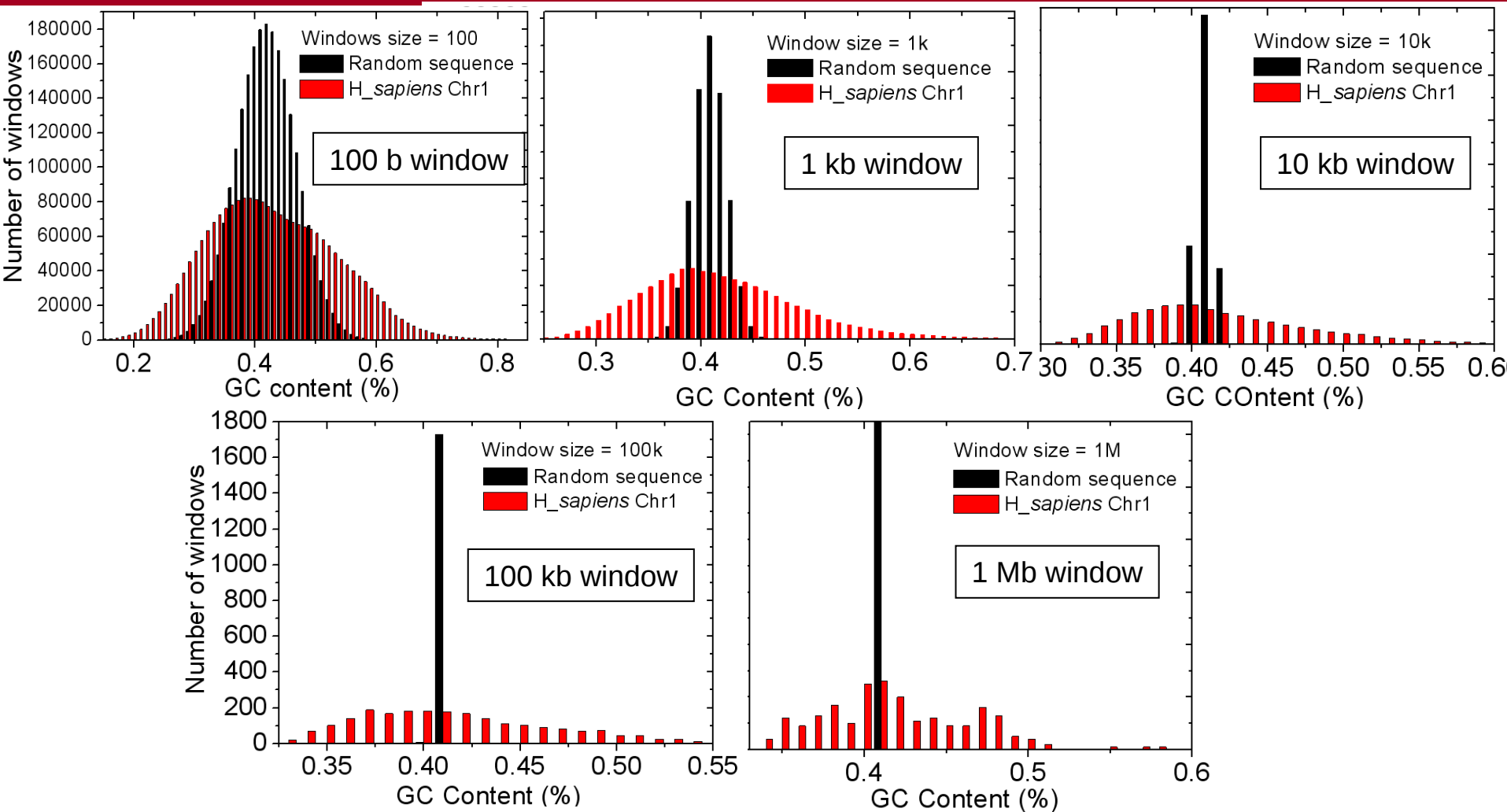


# Why was result strange?

## The Central Limit Theorem

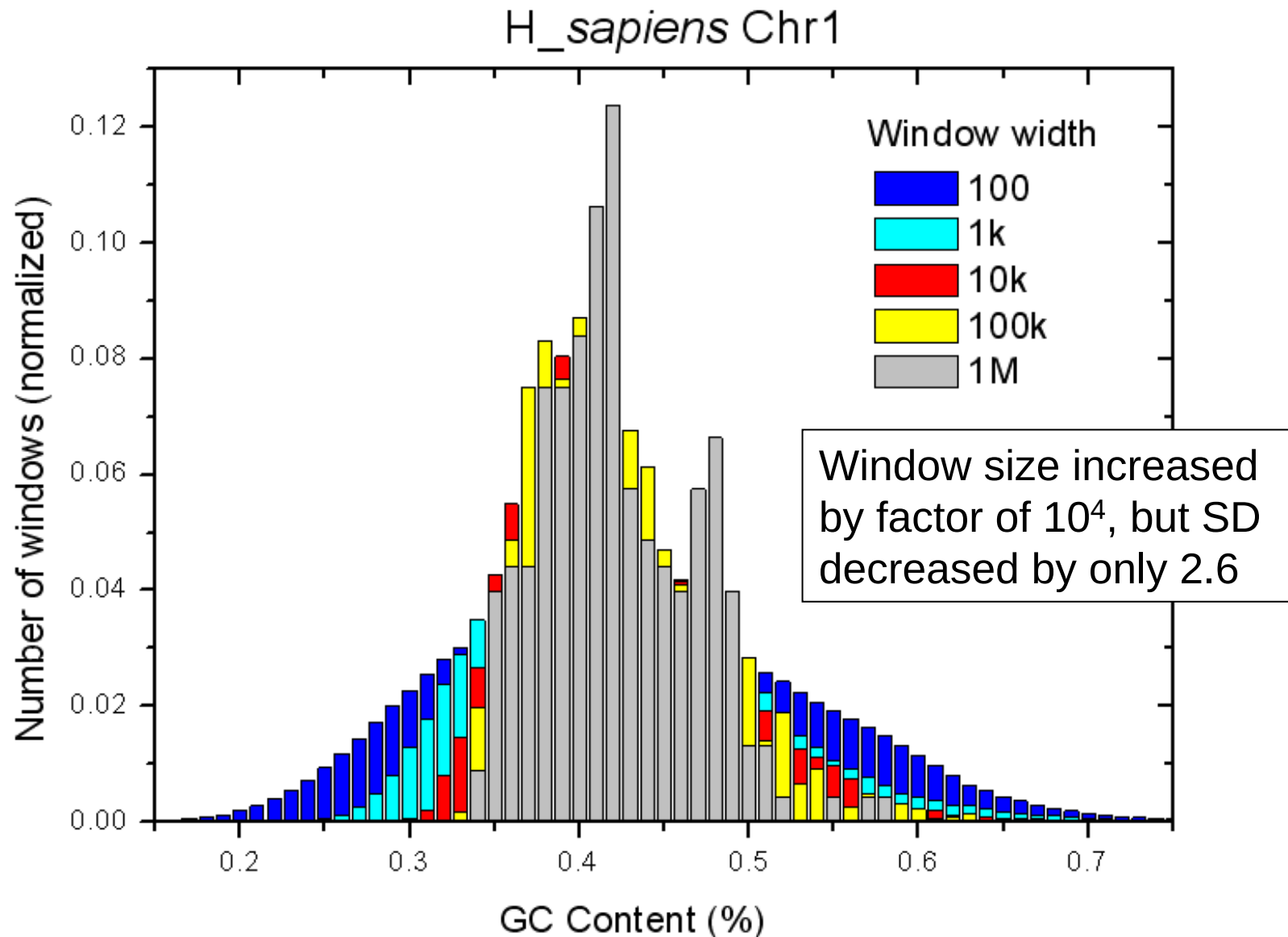
- A large number of independent observations from the same distribution has an approximate normal (Gaussian) distribution **whose variance is inversely proportional to sample size.**
  - PS Laplace 1810; A. Lyapunov 1899.
  - S. Bernstein, *Math. Ann.* 97:1-59 (1927) M. Rosenblatt, *PNAS* 42:43-47 (1955)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

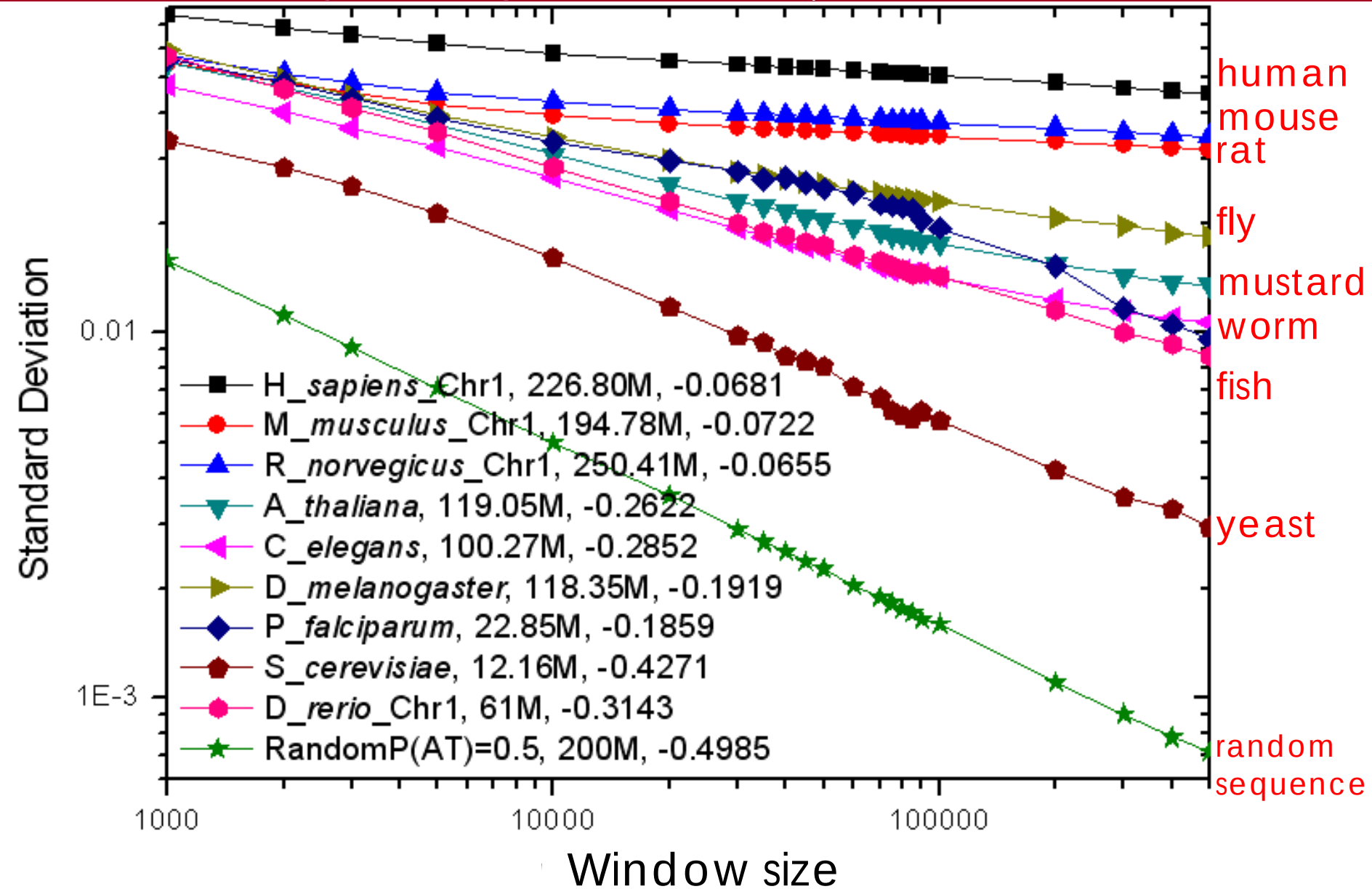
- Power law

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

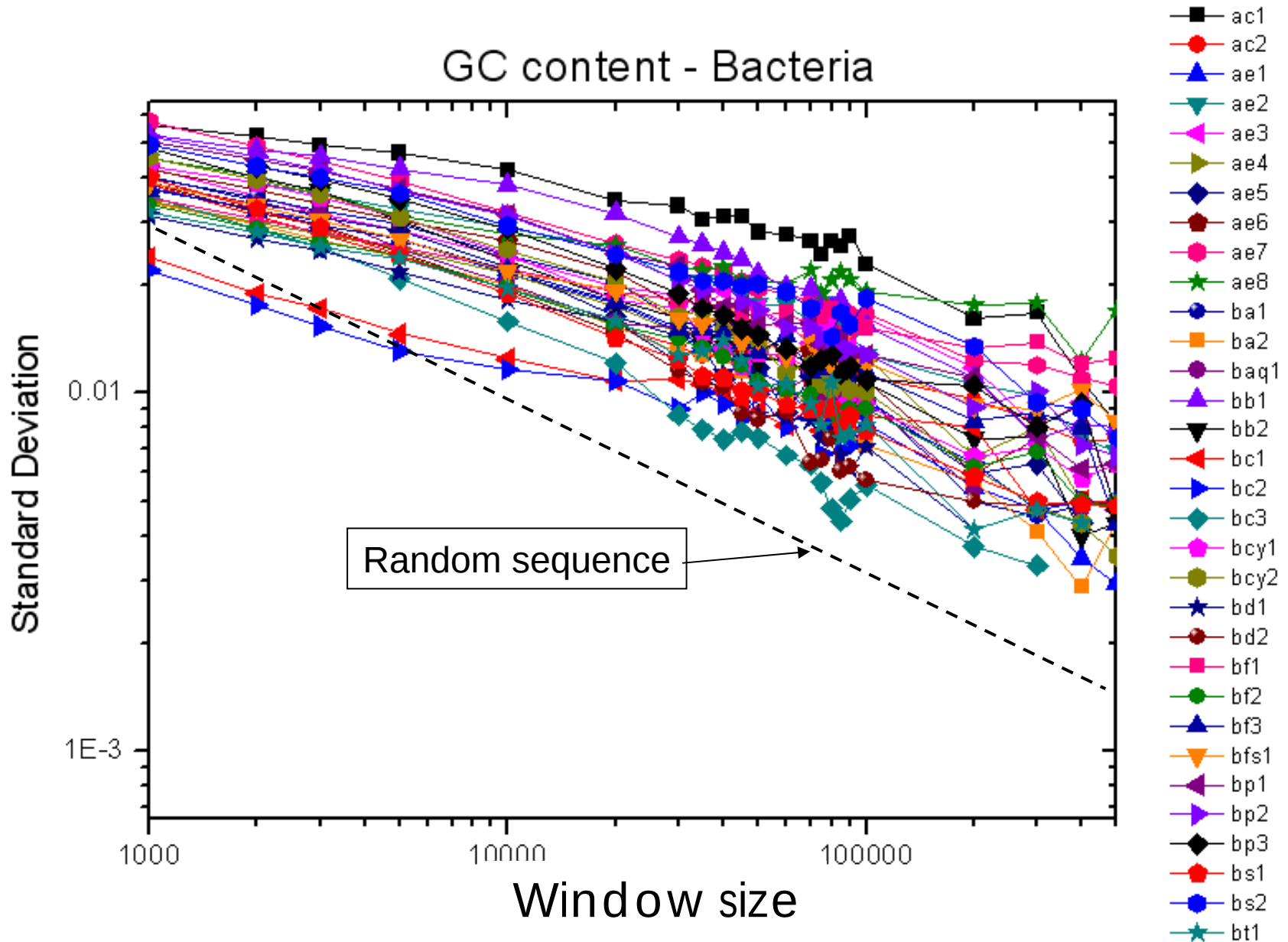
(Log-log plot is a straight line with slope  $\gamma$ )

- Central limit theorem:  $\gamma = -0.5$
- Human chromosomes:  $\gamma \sim -0.07$

# Power law is universal for complete eukaryotic genomes: $\gamma = -0.06$ to $-0.42$



And for bacteria:  $\gamma = -0.16$  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

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

□  $\gamma$  is the **scaling exponent**

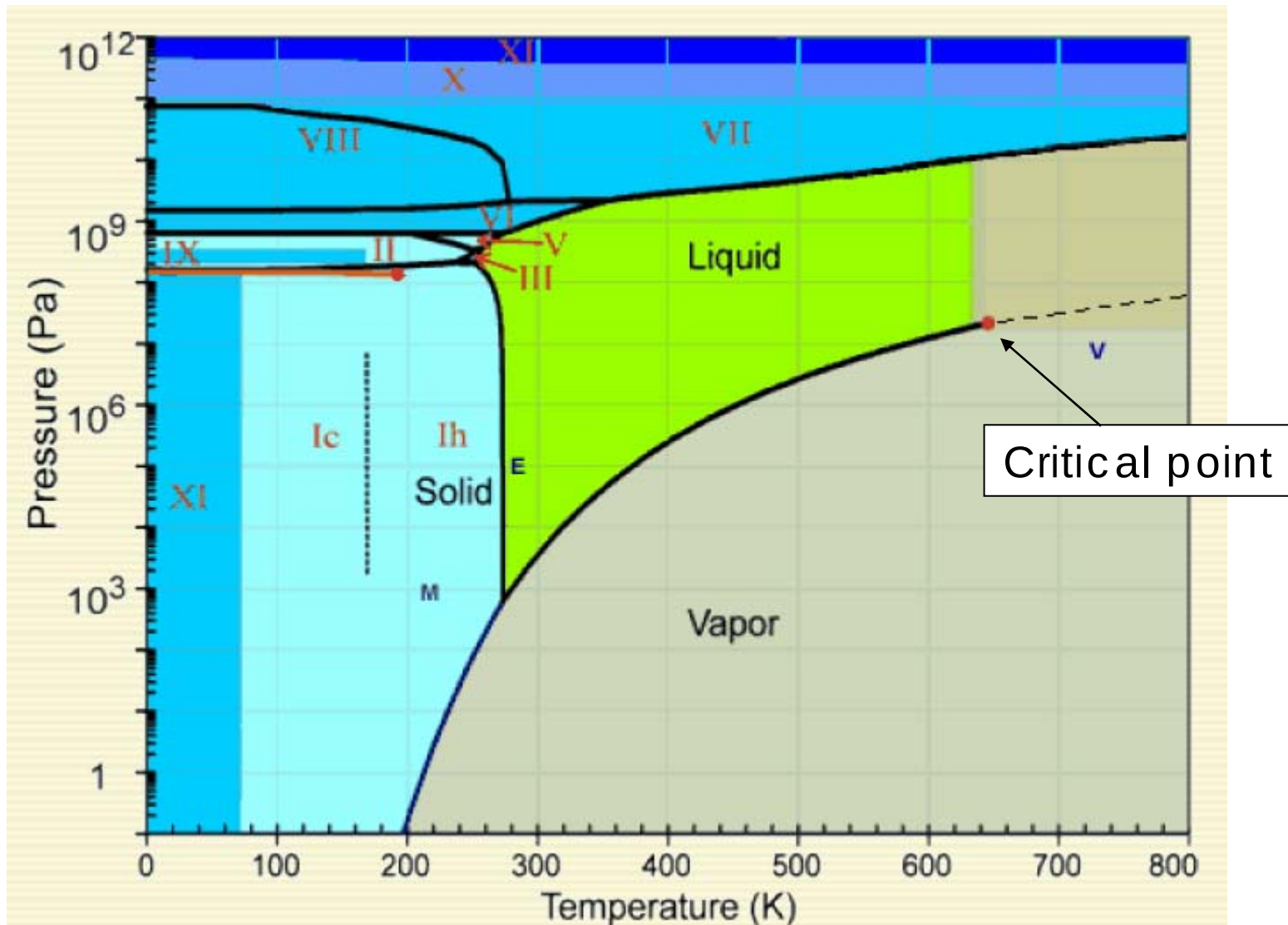
- **Exercise:** Show that  $f$  obeys **power-law**:

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

# 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 &  $2 \times 10^7$  Pa

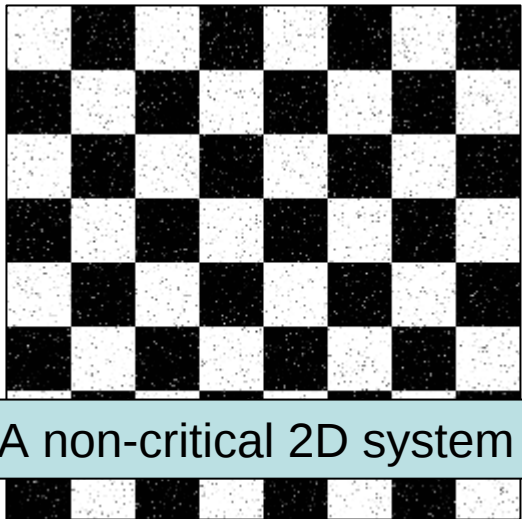


# Criticality = Scale invariance + self-similarity

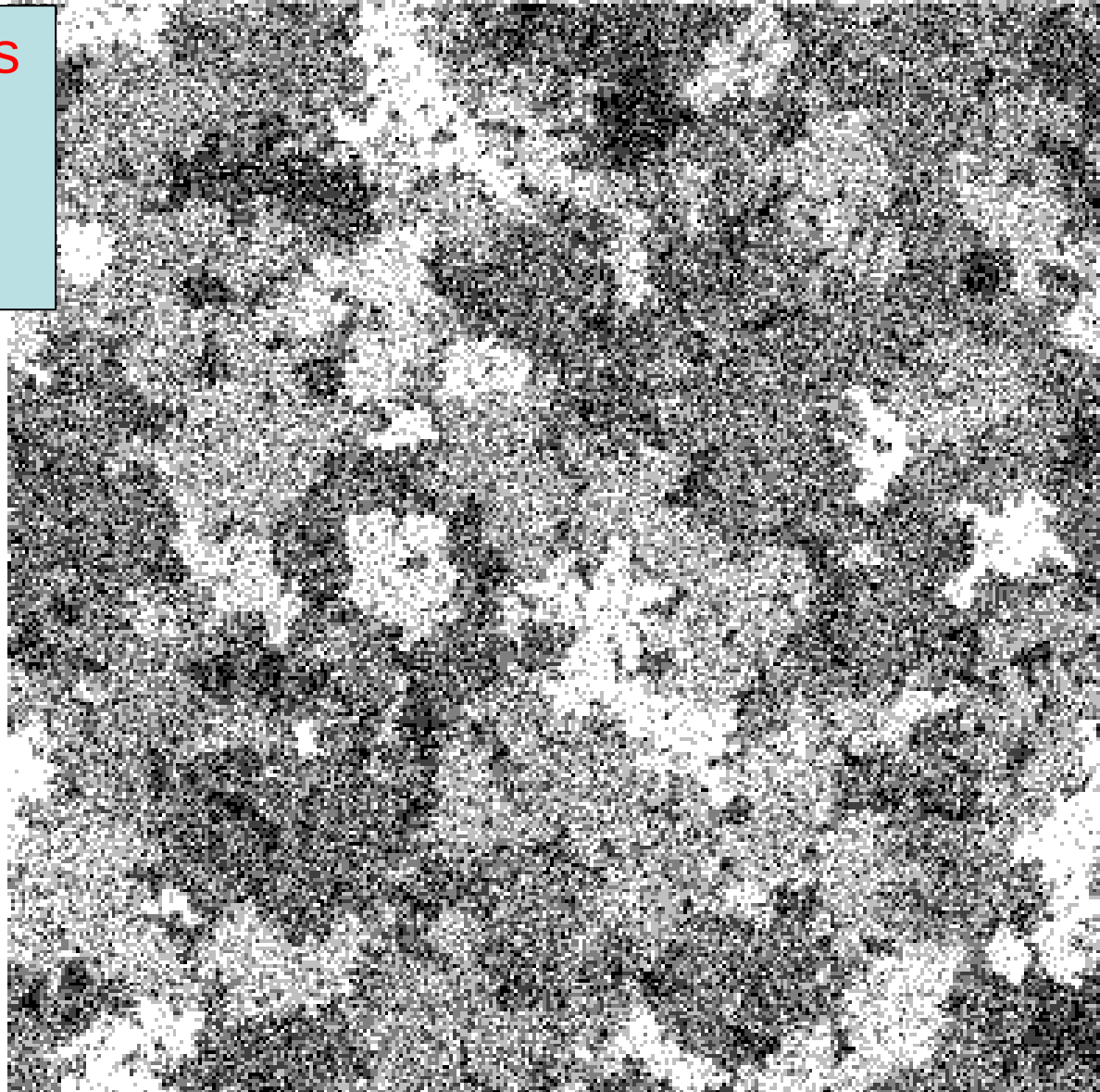
- Scale invariance: there are domains of all sizes
- Self-similarity: there are domain (of all sizes) within domains

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

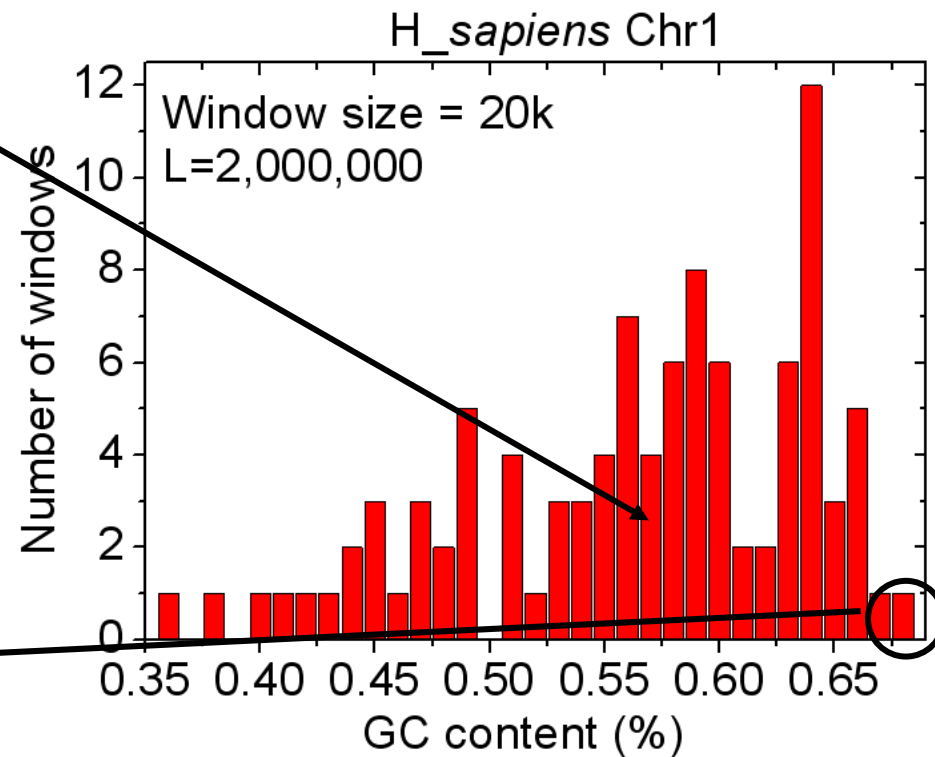
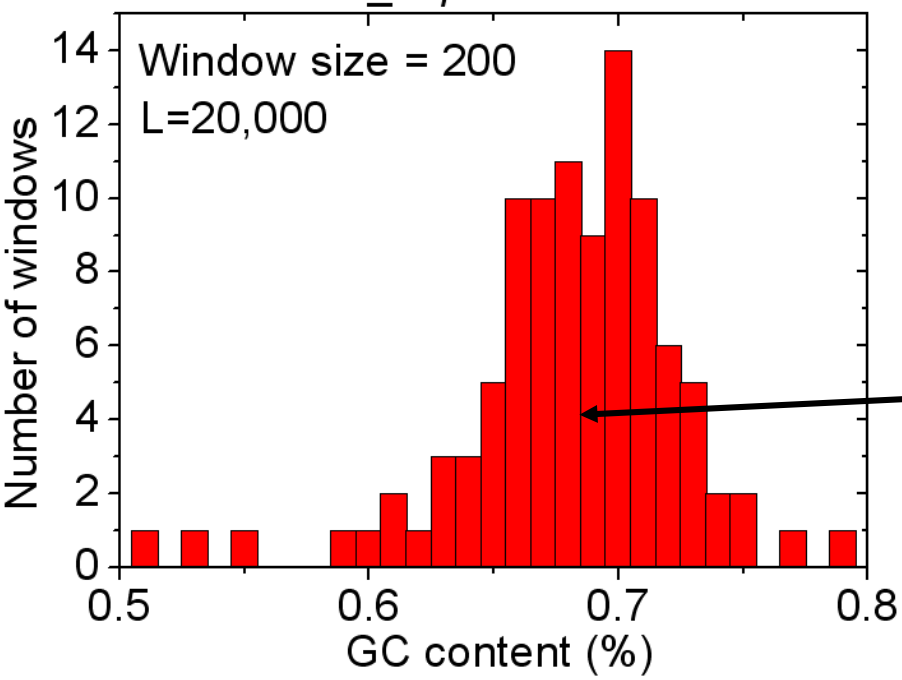
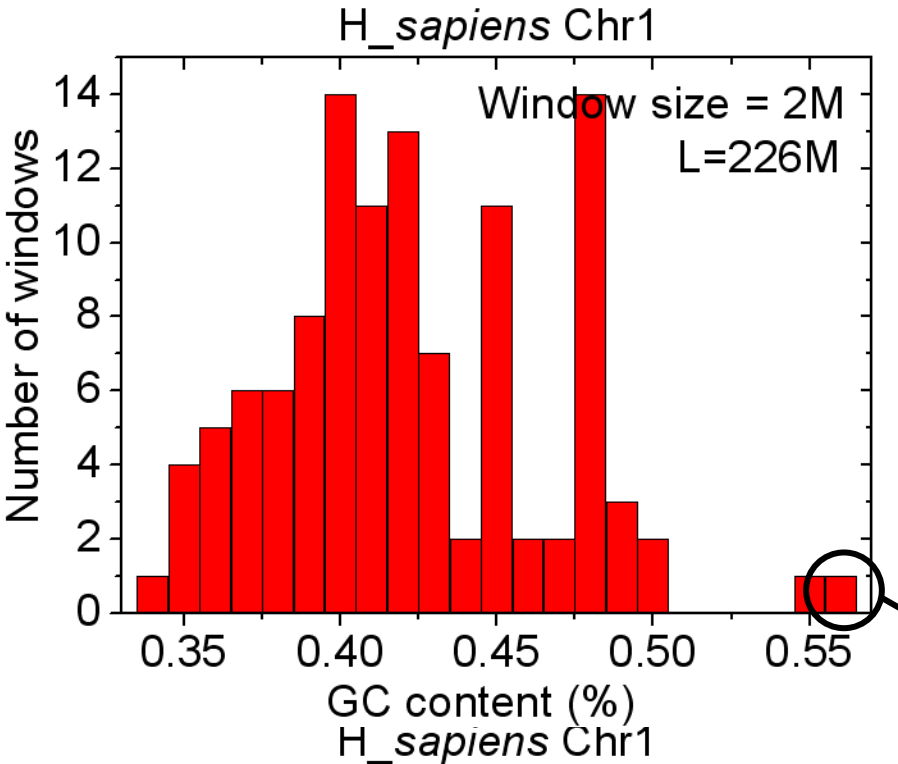
There are **domains of all sizes**, and there are domain within domains



A non-critical 2D system

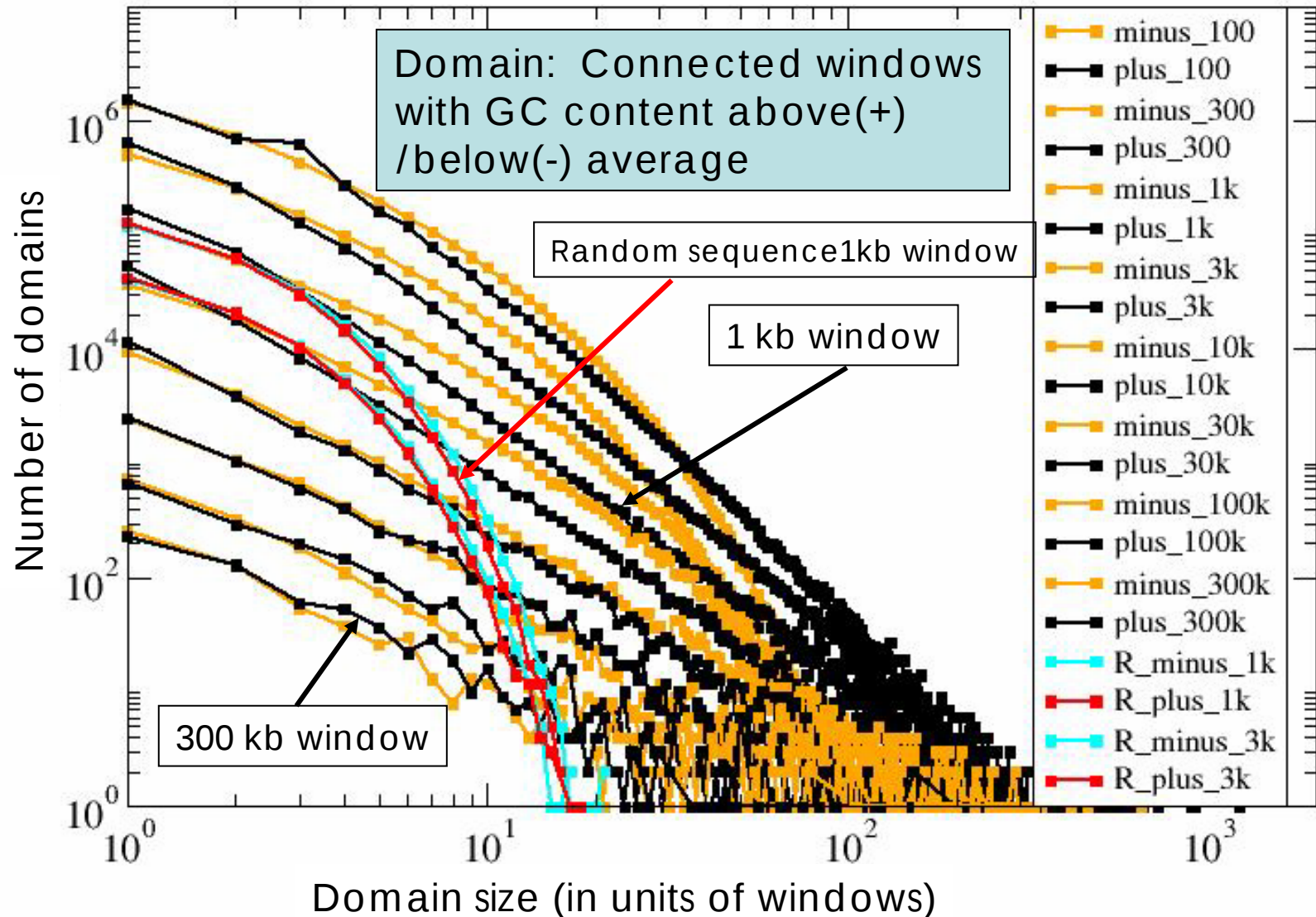


# Self-similarity in human chromosome 1



# Another look at HS genome: scaling in domain size

## Human Chromosome (3 Bb)



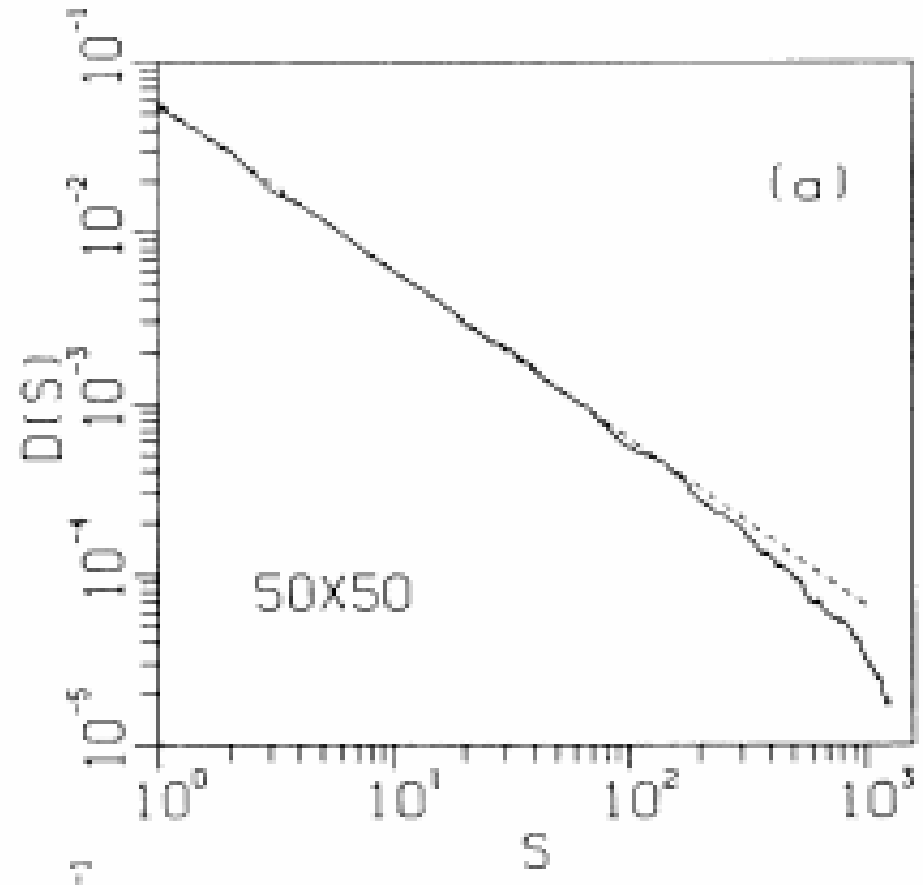
# Genomes are critical

- 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
- How did genomes become critical

# 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-Wiesenfeld **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



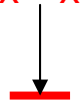
Bak-Tang-Wiesenfeld PRL (1987)

# Genome the critical blind self-copier

- Previous growth model: Genome the blind self-copier
  - Maximally random segmental duplication
  - Duplicated segments: randomly selected over a fixed length (~2000 bases)
  - No long range variation > 2000 b; no scaling
- New model: Genome the **critical** blind self-copier
  - Still MRSD
  - Duplicated segment
    - Any length up to current genome length
    - Repeated a random number of times before insertion

# Five steps of critical maximally random segmental duplication

1. Original genome



2. Duplication - copy any segment of any length

3. Replication - repeat segment any number of times



4. Insertion – at any site

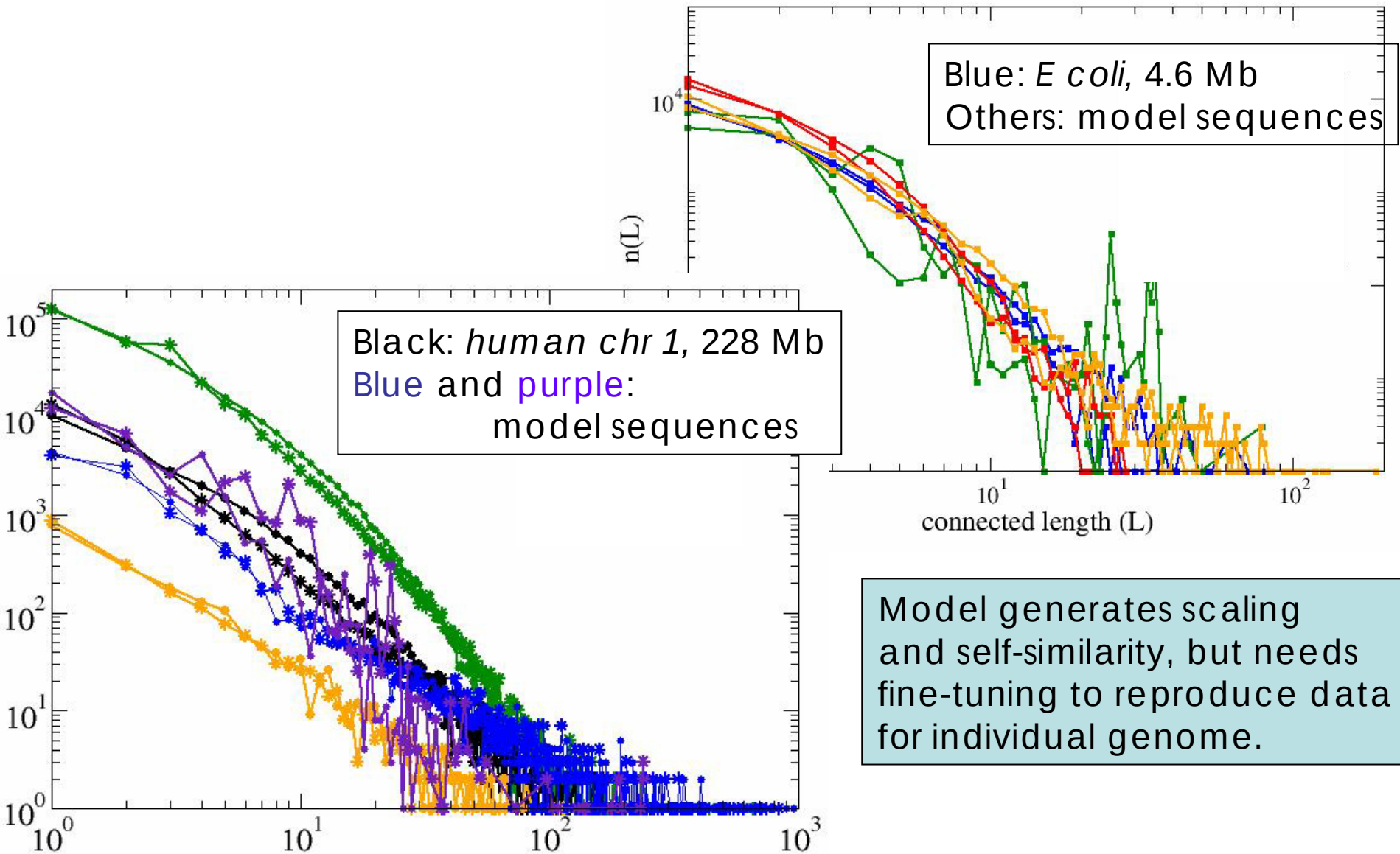


5. Longer new genome has repeated copied segment



Maximally random: All selections are random

# Preliminary modeling results promising



# Why would genome grow by Critical MRSD?

- **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 random self-copying is a result of **natural selection**

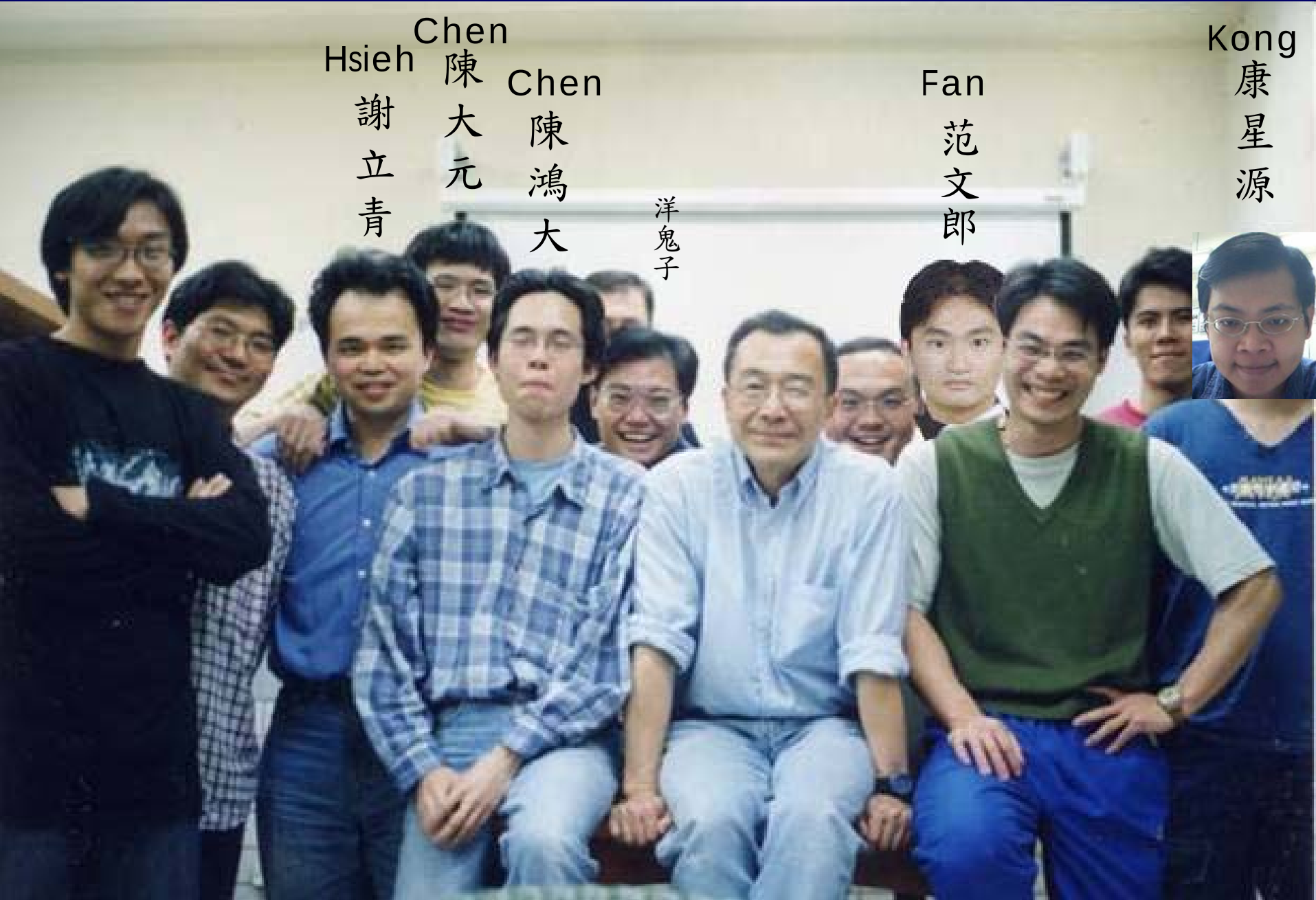
# Many biological phenomena explained by model

- Preponderance of **homologous genes** in all genomes
  - All genes belong to homologous families
- Genome is full of **non-coding repeats**
- Large-scale genome “**rearrangements**”
- Huge species **diversity**
- Apparent “**missing link**”
  - Coexistence of gradualism and punctuated equilibrium
- Many more ...

# People at CBL who work on project

- Dr. Hsieh Li-Ching – now with Genome Research Center, Academic Sinica
- Dr. Chen Da-Yuan – now at He-Hsin Cancer Research Center
- Chen Hong-Da - PhD student
- Kong Sing-Kuan - PhD student
- Fan Wen-Lang - PhD student

# Computation Biology Laboratory (2003)



Hsieh  
謝立青

Chen  
陳大元

Chen  
陳鴻大

洋鬼子

Fan  
范文郎

Kong  
康星源

Our papers and PDF version of  
talk are found at  
Google: HC Lee

Thank you!