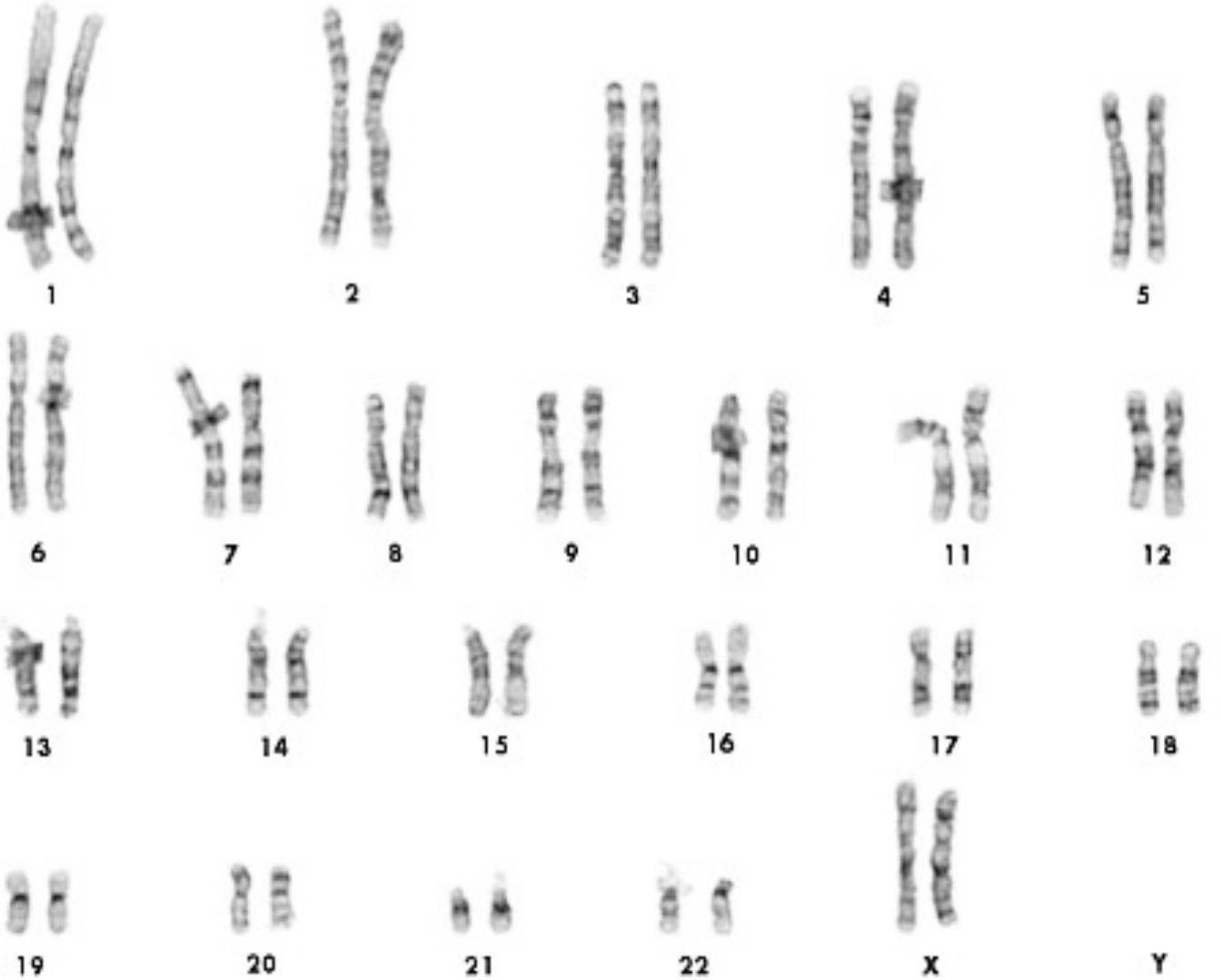
The Evolutionary Fate of Duplicated Neutral DNA
-
Breaking Sticks on Evolutionary Time Scales

Peter F Arndt and Florian Massip

The Evolutionary Fate and Consequences of Duplicate Genes

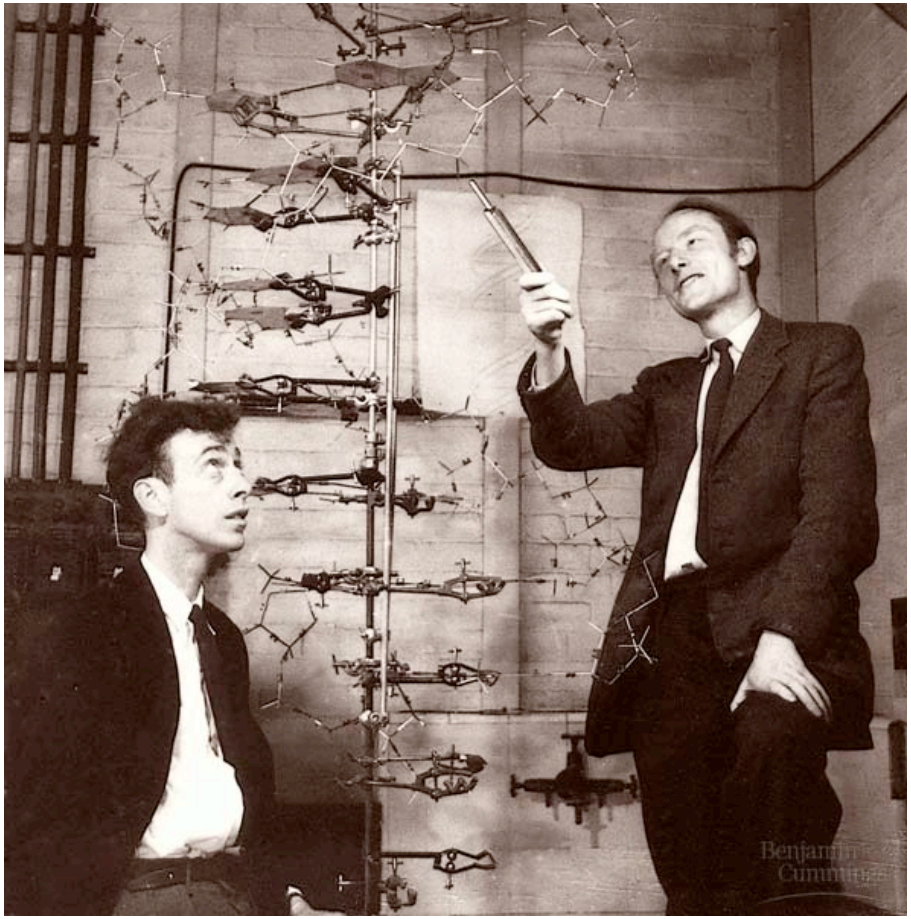
Michael Lynch[1]* and John S. Conery[2]

[Science 2000]

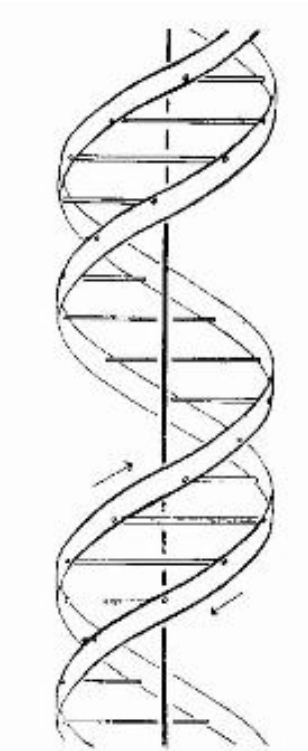


Giemsa-banding pattern of the human genome

Watson + Crick: 1953



James Watson + Francis Crick

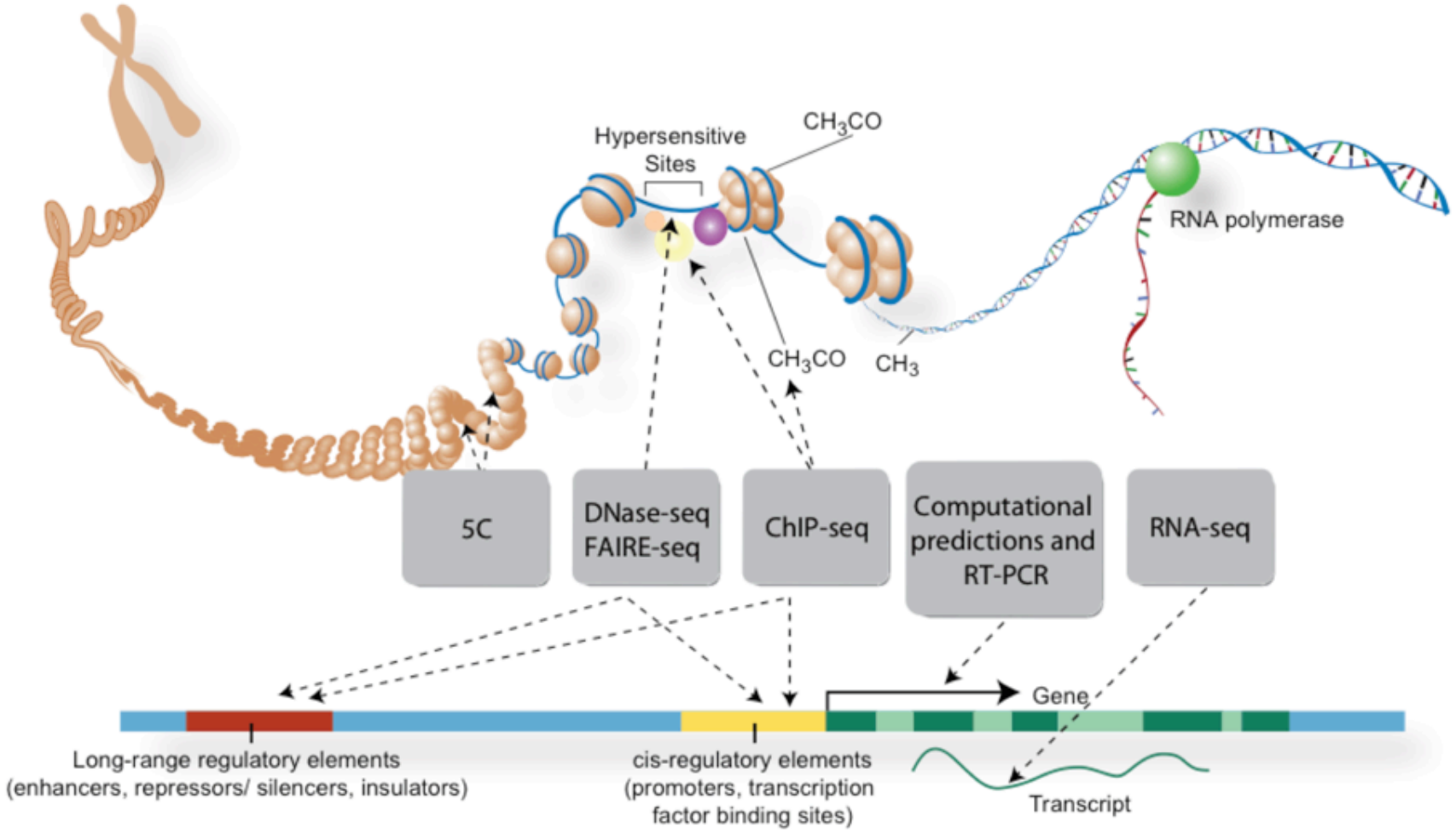


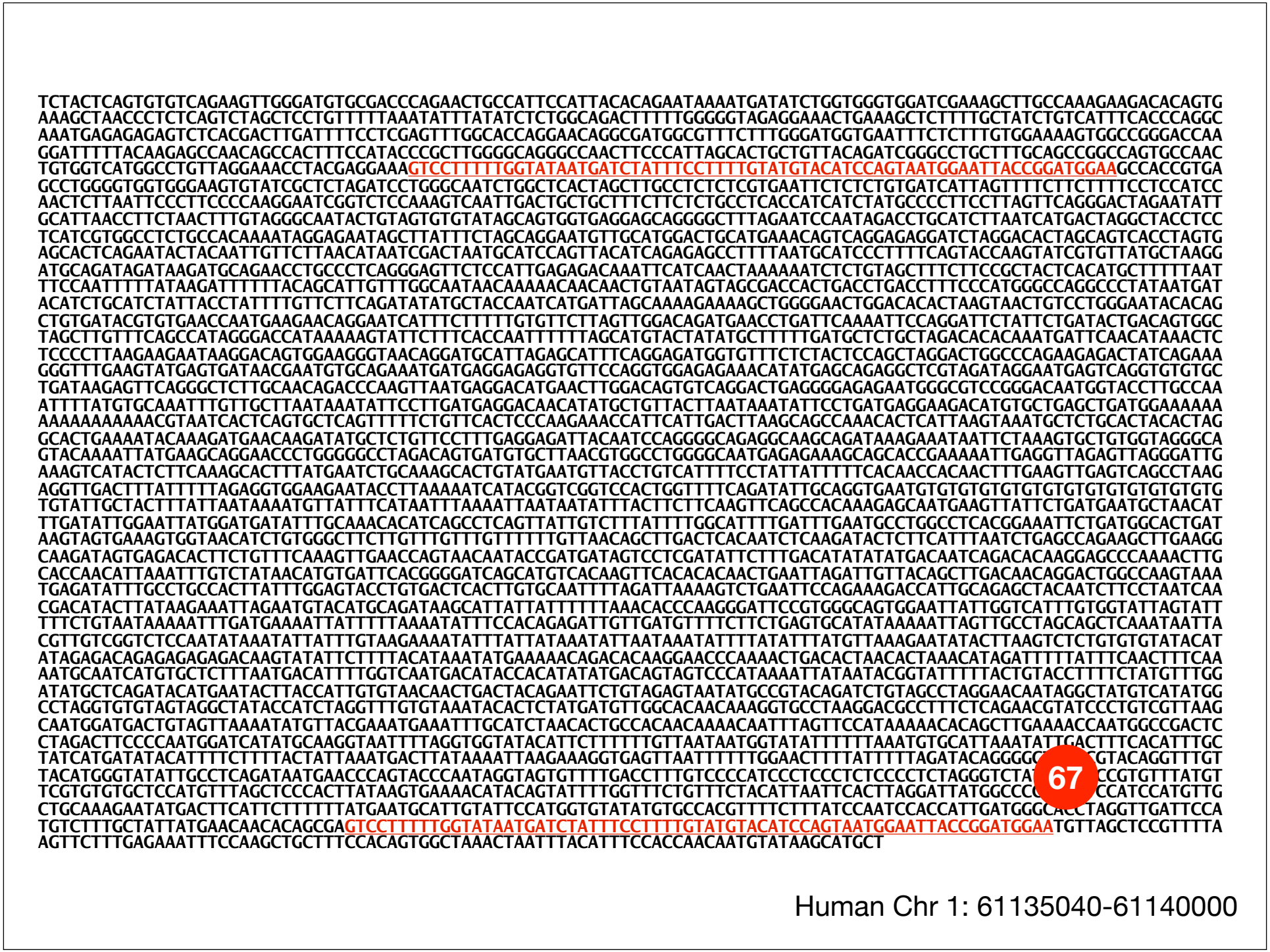
This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

MAX-PLANCK-GESELLSCHAFT

ENCODE: 2013

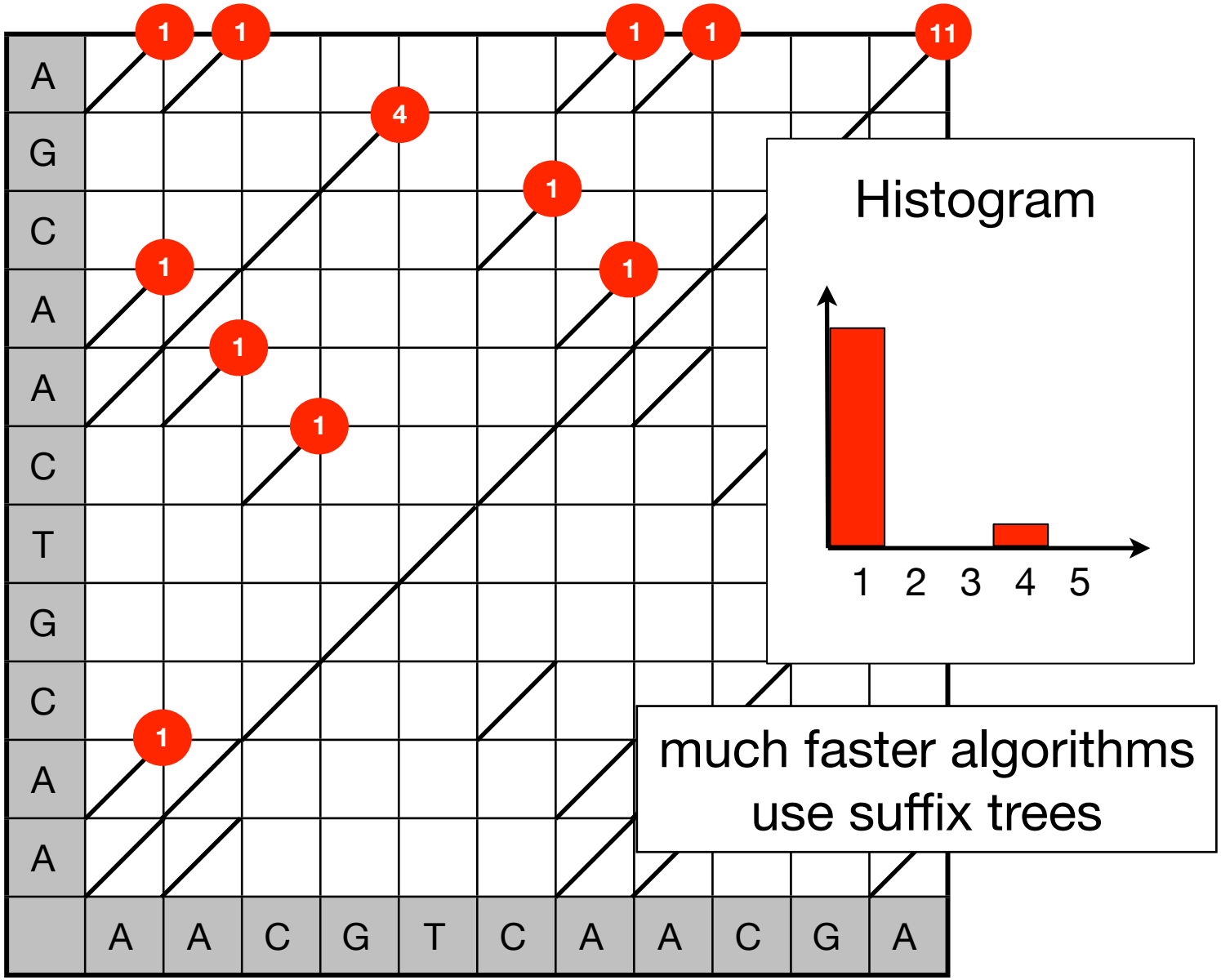
- comprehensive parts list of functional elements





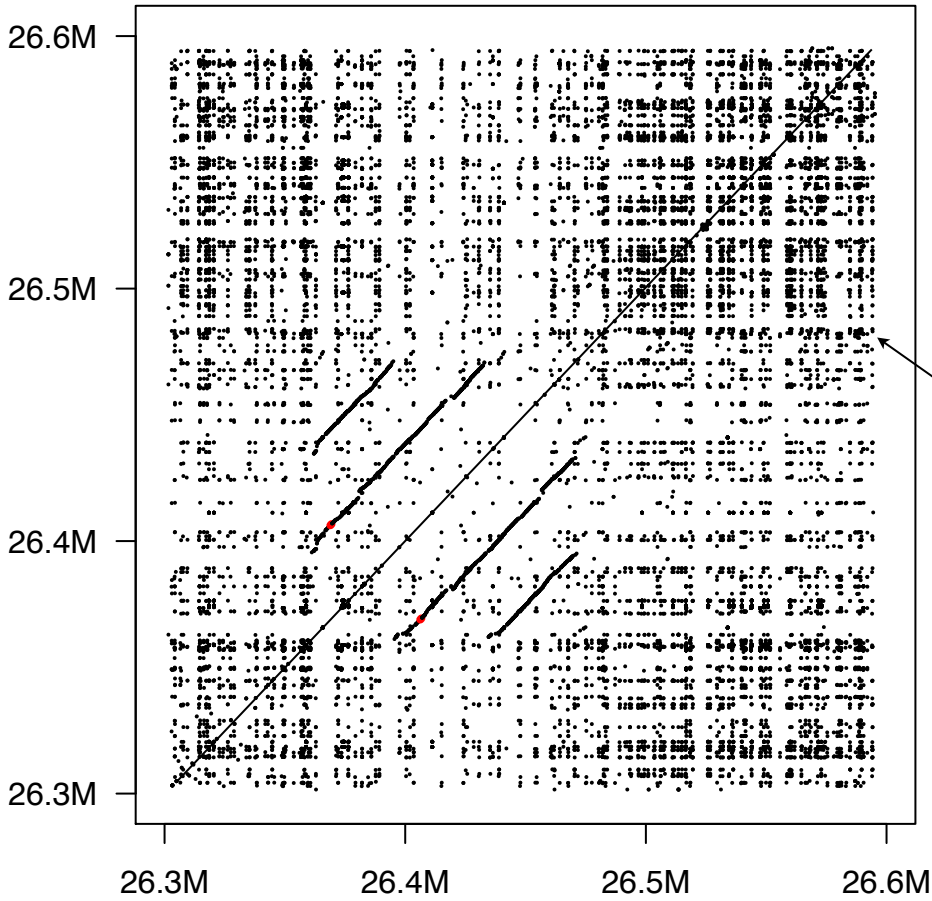
Human Chr 1: 61135040-61140000

Self-Alignment



Low Copy Number Repeats

- Dot-Plot of a Self-Alignment of Human DNA:



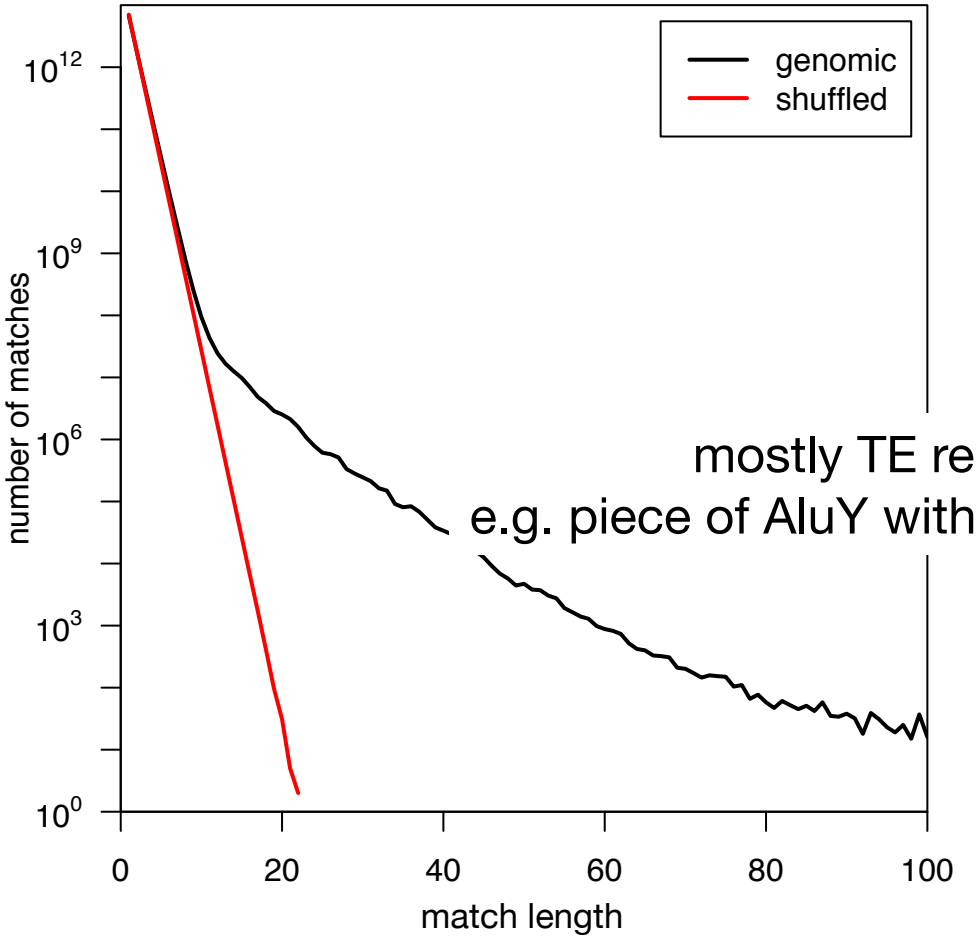
100% identical matches longer than 30 bp

Example: 300 MB on the human chromosome 7

MAX-PLANCK-GESELLSCHAFT

Self-Alignment of Genomes

- the length distribution of identical matches



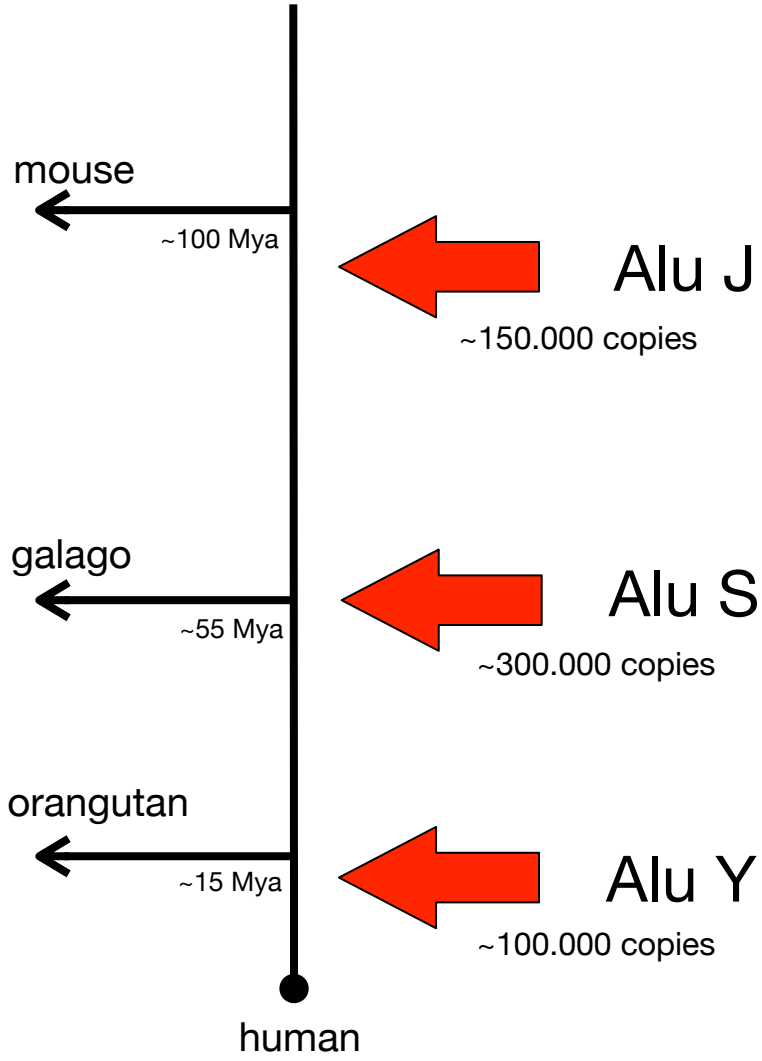
random matches

$$L^2(1-p)^2p^r$$

mostly TE related matches
e.g. piece of AluY with a piece of another AluY

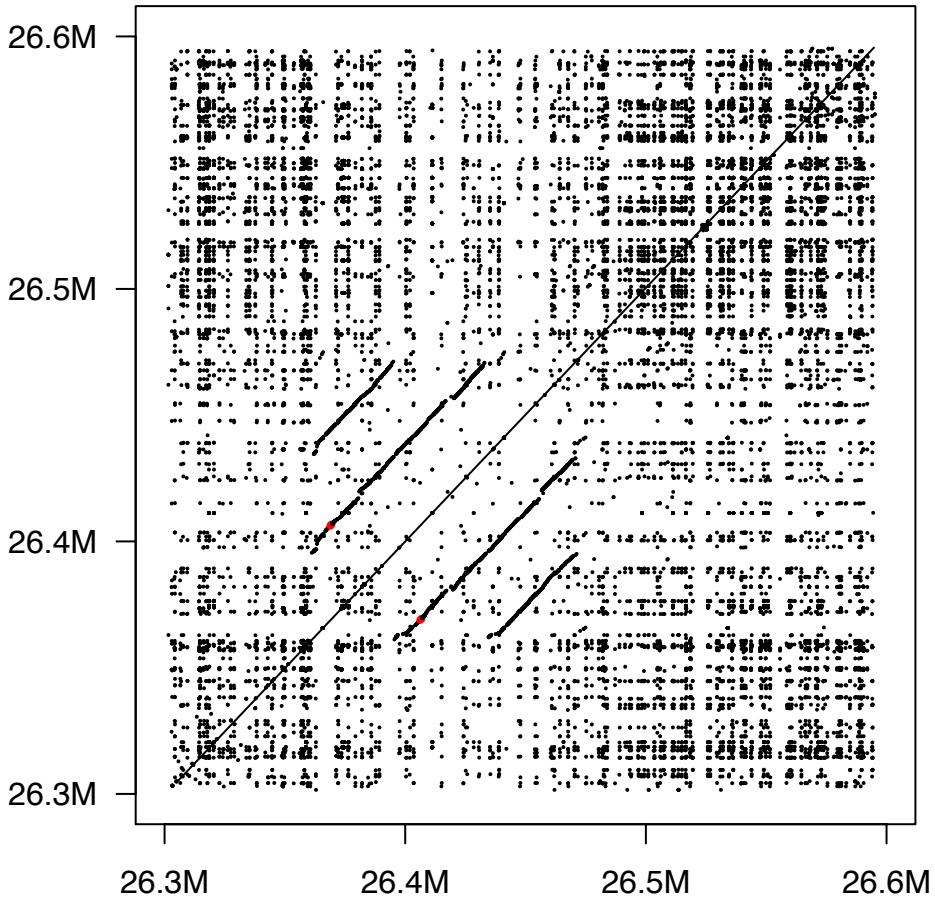
Transposable Elements

- retrotransposons entered the human genome in bursts



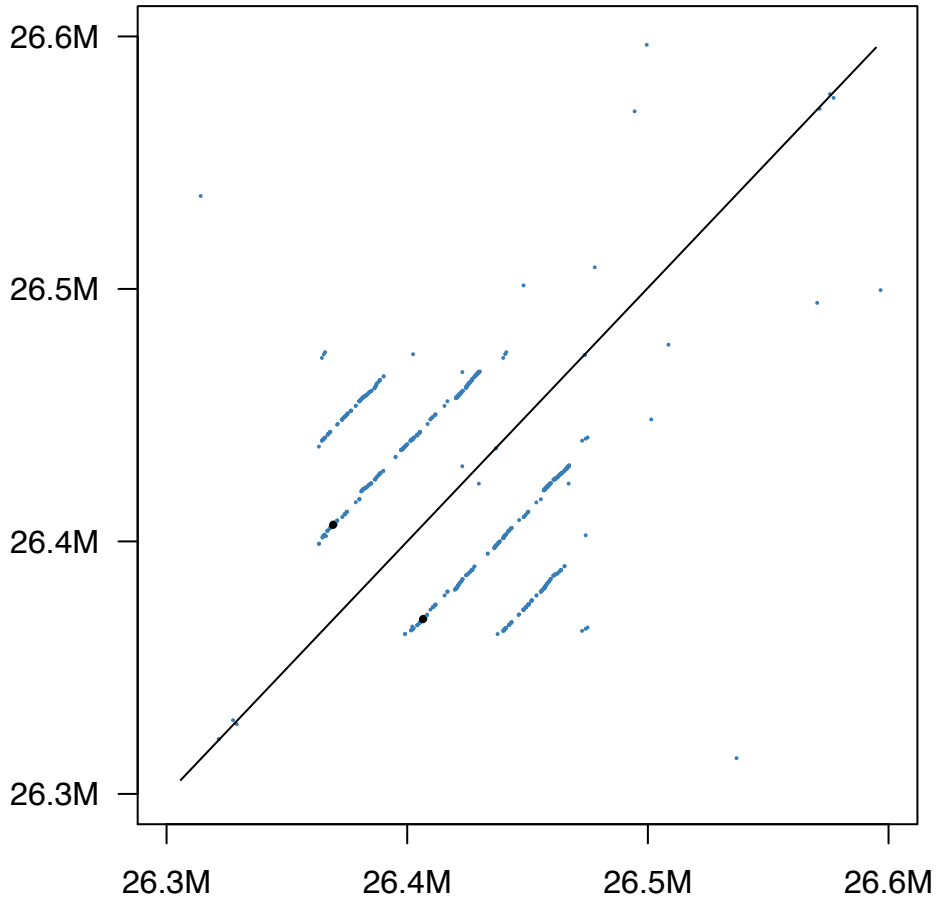
Self-Alignment

- Example: 300 MB on the human chromosome 7



Low Copy Number Repeats

- Example: 300 MB on the human chromosome 7



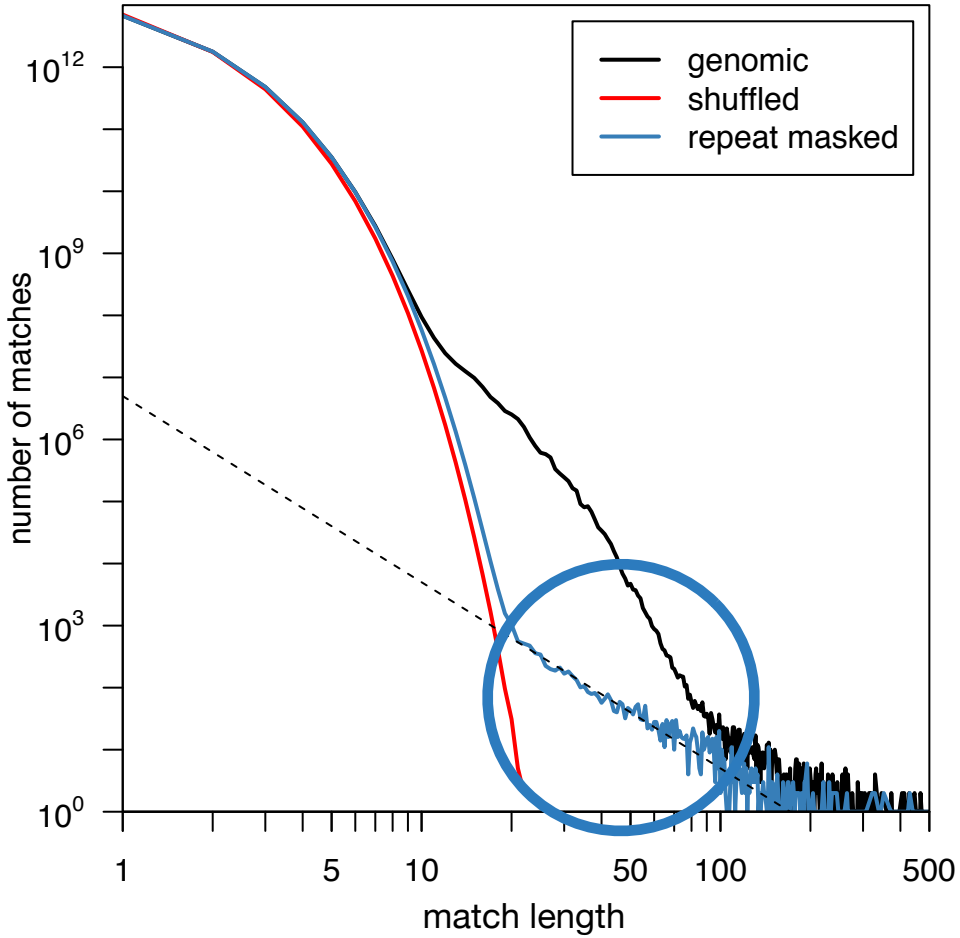
segmental duplication

here: repeat masked

[Bailey & Eichler, 2006]

Self-Alignment of Genomes

- the length distribution of identical matches



random matches

$L^2(1-p)^2p^r$

matches unrelated to repeats

α/r^3

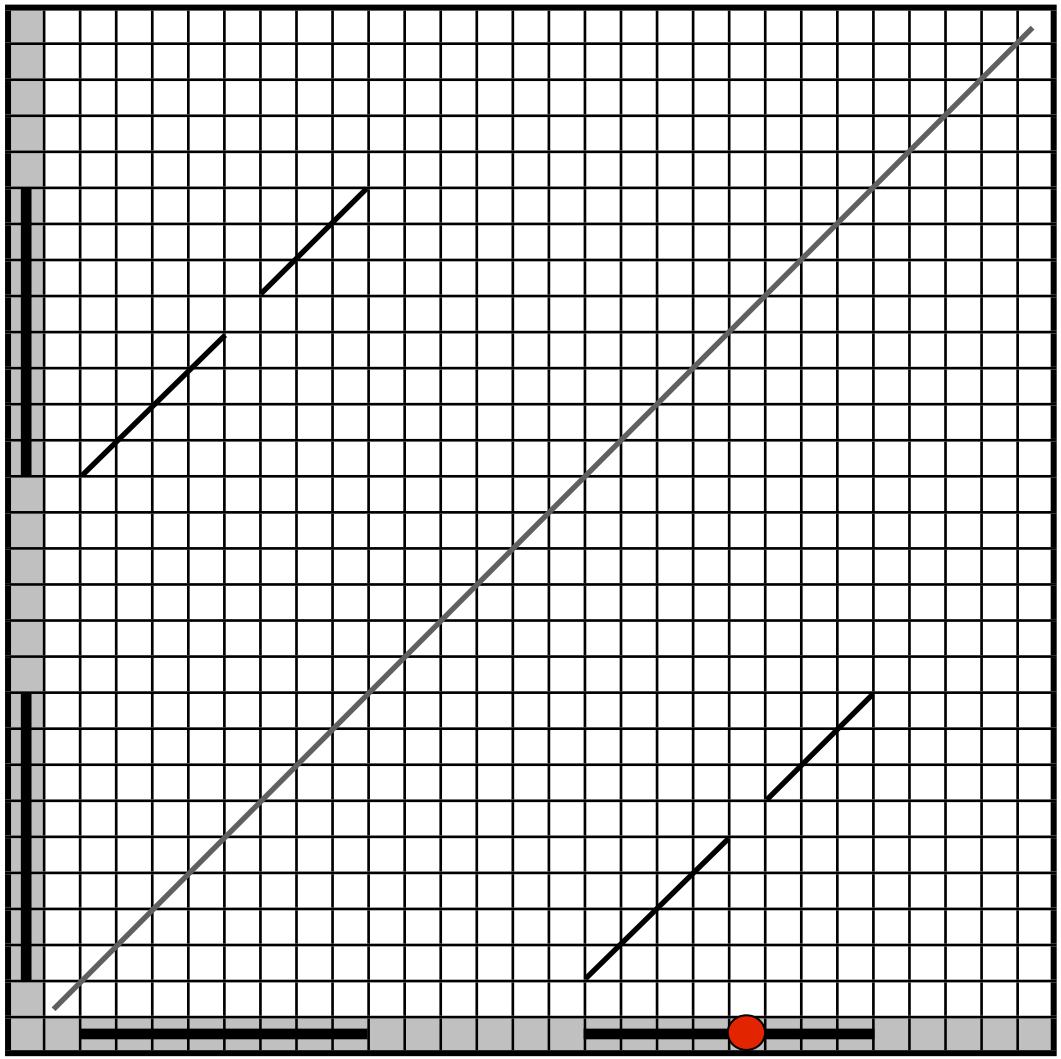
Selection?

[Gao & Miller, 2011]

MAX-PLANCK-GESELLSCHAFT

The Broken Stick Model

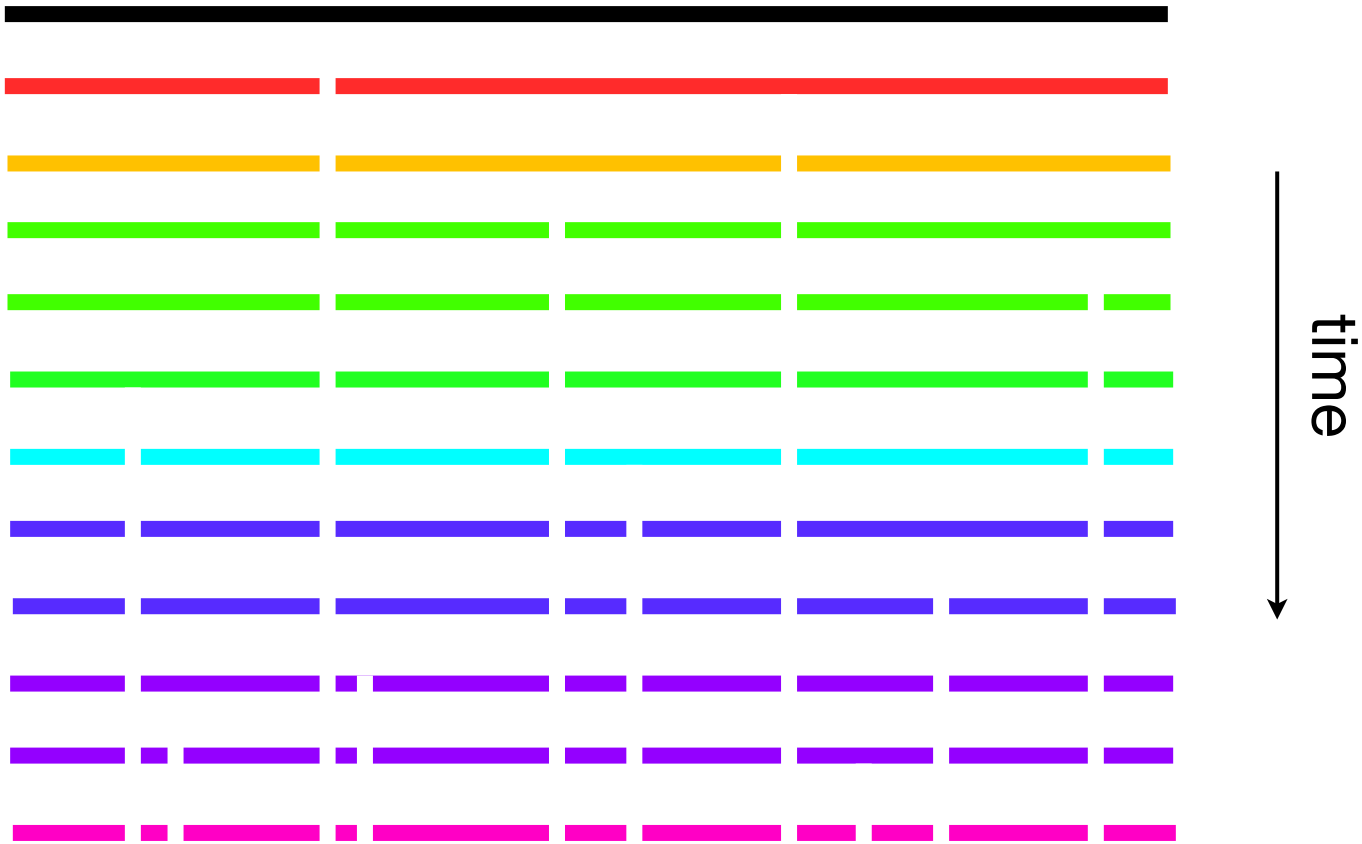
- a segmental duplication will dissolve by mutations



MAX-PLANCK-GESELLSCHAFT

The Broken Stick Model

- consider a stick of initial length *K* which is broken with rate μ



- let *m(r,t)* be the number of sticks of length *r* at time *t*

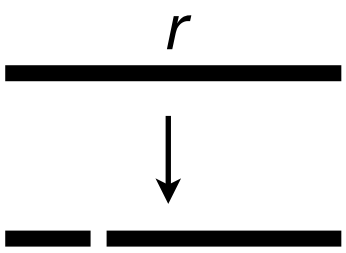
The Broken Stick Model

- the dynamics is given by:

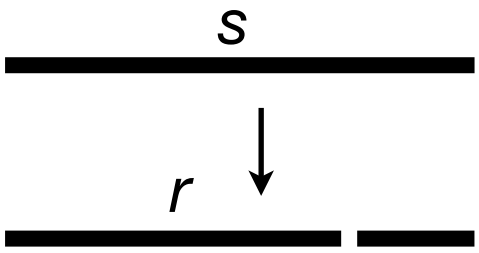
$$\frac{\partial m(r,t)}{\partial t} = \underbrace{-2\mu r\, m(r,t)}_{\text{loss}} + \underbrace{4\mu \int_r^\infty m(s,t)\, ds}_{\text{gain}}$$

μ mutation rate per bp

loss of matches by mutations in one copy



gain of matches through particular mutations in longer matches



MAX-PLANCK-GESELLSCHAFT

[Kuhn 1930, Saito, 1958]

The Broken Stick Model

- the dynamics is given by:

$$\frac{\partial m(r,t)}{\partial t} = -2\mu r\, m(r,t) + 4\mu \int_r^\infty m(s,t)\, ds$$

- the solution is

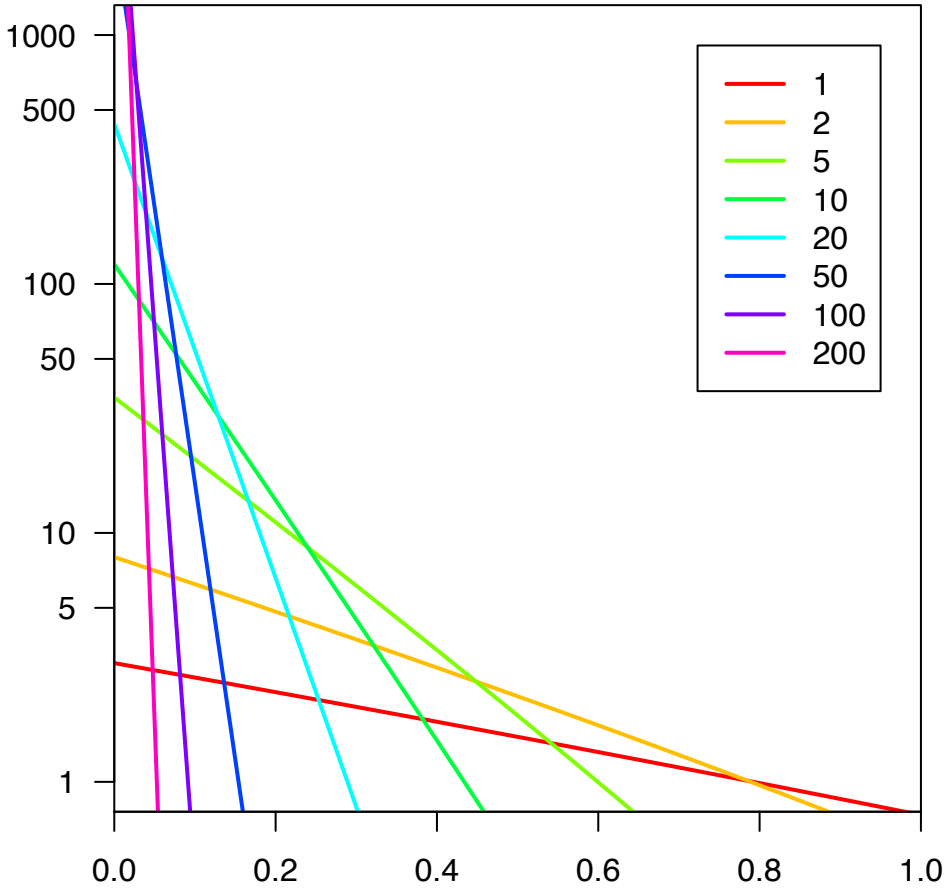
$$m(r,t) = [4\mu t + 4\mu^2 t^2 (K - r)]\; e^{-2\mu r t}$$

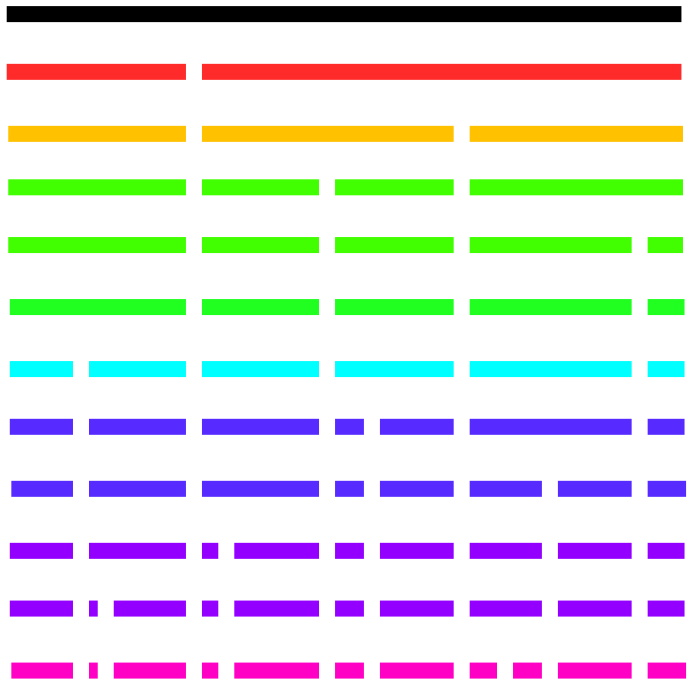
where K is the initial length of a duplication.

[Ziff & McGrady, 1985]

The Broken Stick Model

- breaking one stick: $m(r,t) = [4\mu t + 4\mu^2 t^2 (K - r)]\ e^{-2\mu r t}$





Where does the power law with exponent -3 comes from?

Integrated Broken Stick Model

- observing an ensemble of sticks of different age

$$m(r,t) = [4\mu t + 4\mu^2 t^2 (K - r)]\ e^{-2\mu r t}$$

- its time average:

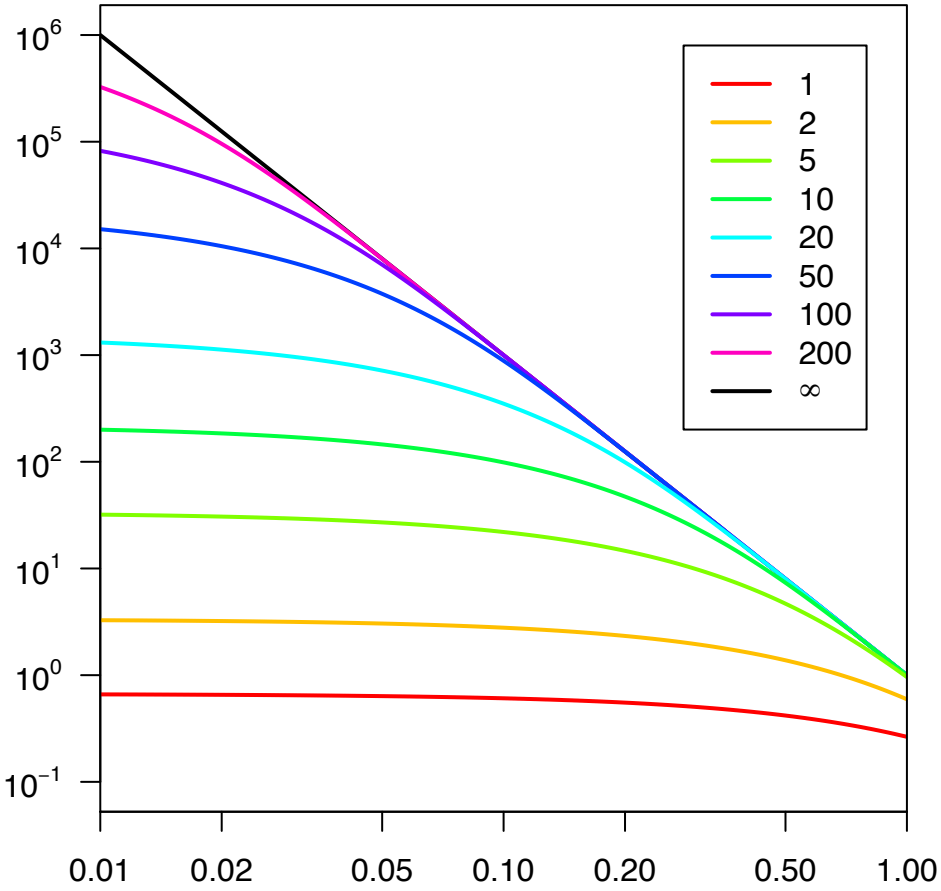
$$m(r) := \int_0^\infty m(r,t)\, dt = \boxed{\frac{K}{\mu r^3}}$$

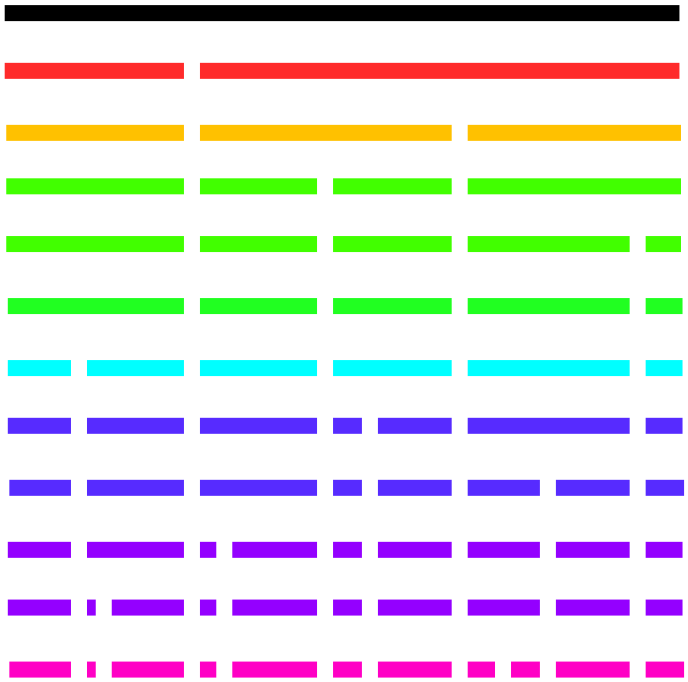
exponent is universal

[Massip & PA, PRL 2013]

Many Broken Sticks

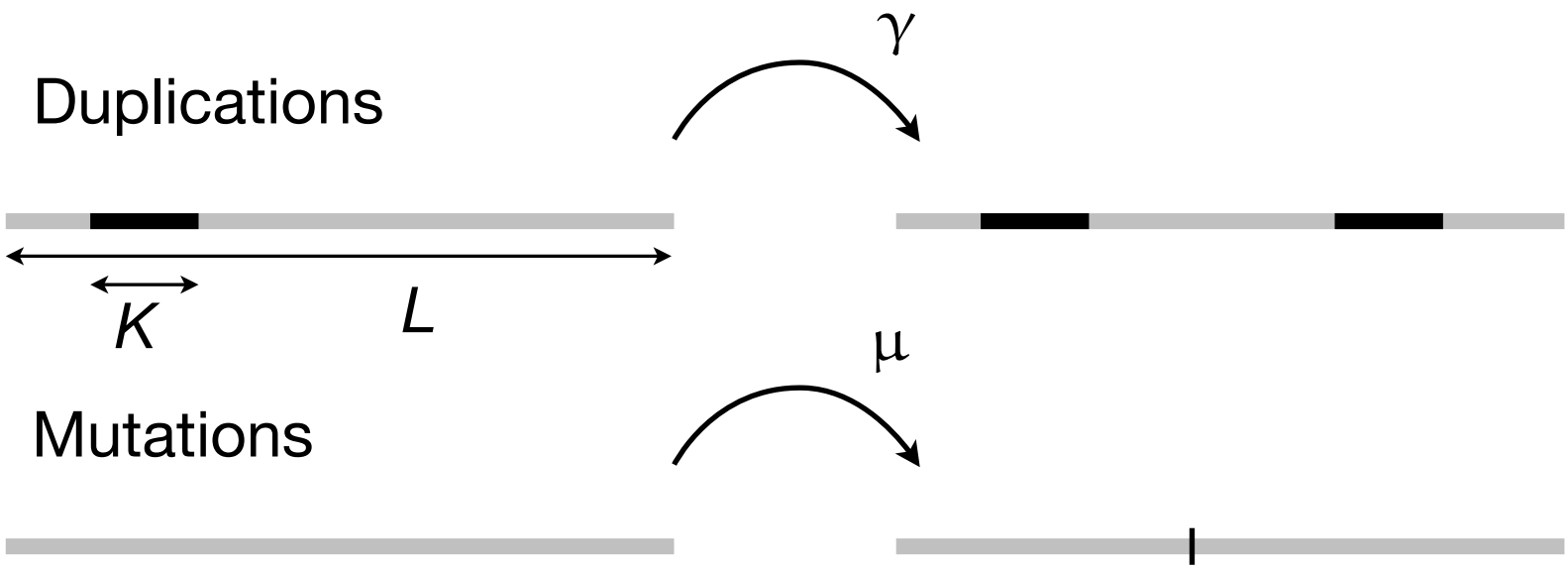
- adding up sticks of different age





Duplication-Mutation Model

- for simulations consider the following processes:



- start with a random iid sequence
- the stationary state is reached for $t > \max(1/\mu, 1/K\gamma)$

[Massip & PA, PRL 2013]

Duplication-Mutation Model

- the dynamics is now given by:

$$\frac{\partial m(r,t)}{\partial t} = -2\mu r\, m(r,t) + 4\mu \int_r^{\infty} m(s,t)\, ds \boxed{+ \gamma L\, \delta(r-K)}$$

- the **stationary solution** is defined by:

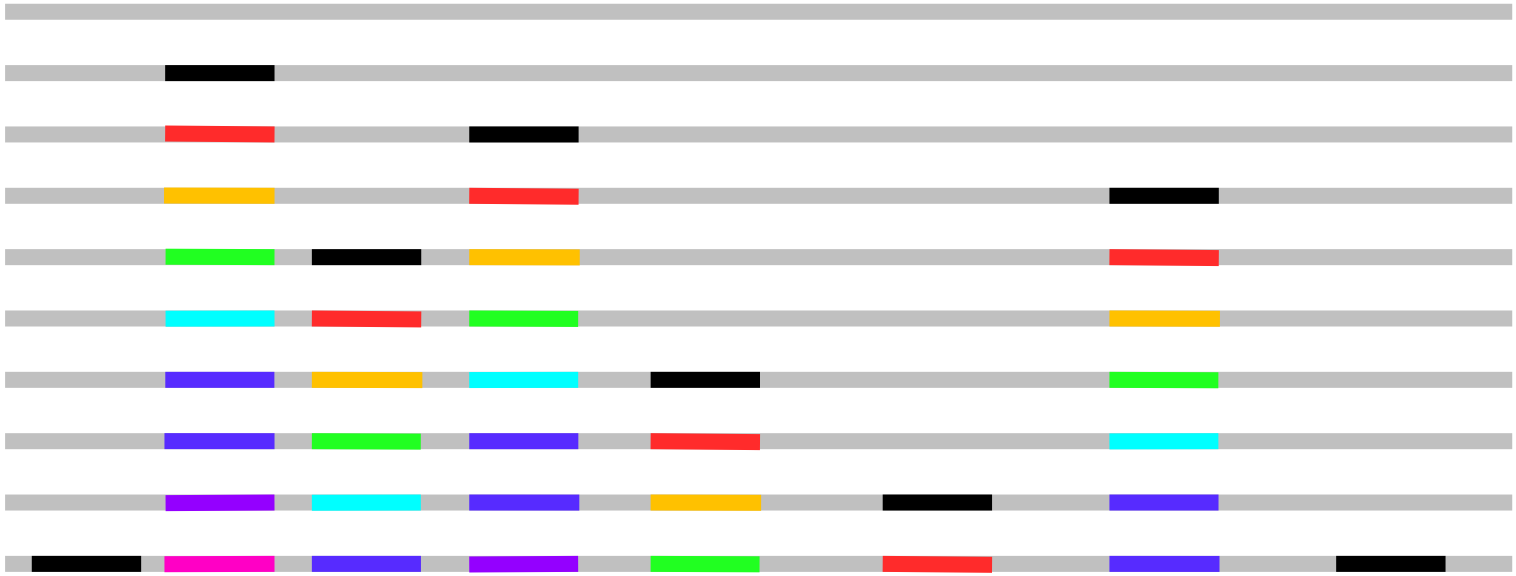
$$\frac{\partial}{\partial t} m(r,t) = 0$$

and given by

$$m_{\text{stationary}}(r) = \frac{\gamma}{\mu} K L \frac{1}{r^3}$$

Duplication-Mutation Model

- starting with a random sequence:

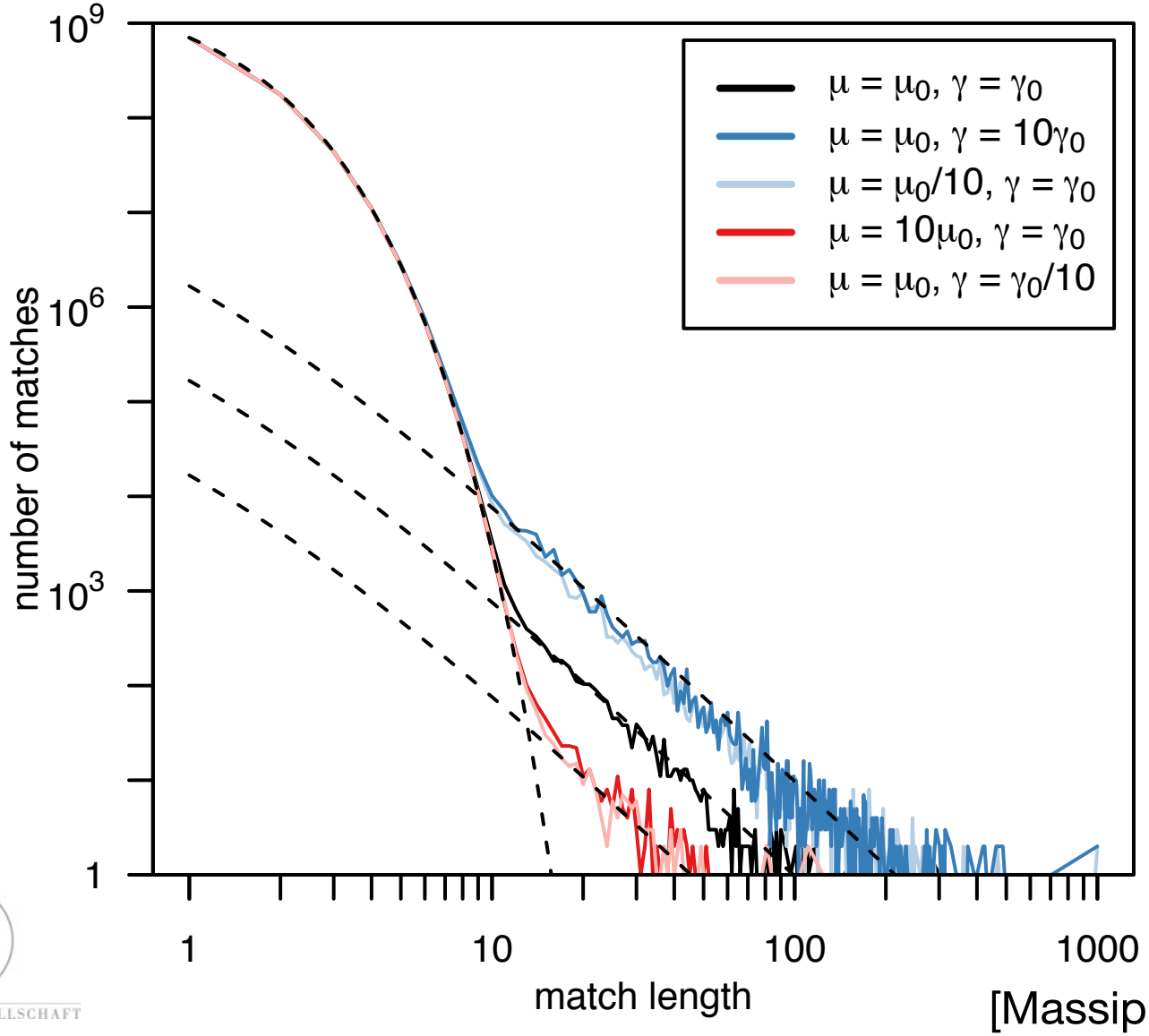


- one reaches a stationary state:

$$m_{\text{stationary}}(r) = \frac{\gamma}{\mu} K L \frac{1}{r^3}$$

[Massip & PA, PRL 2013]

Simulated Sequences

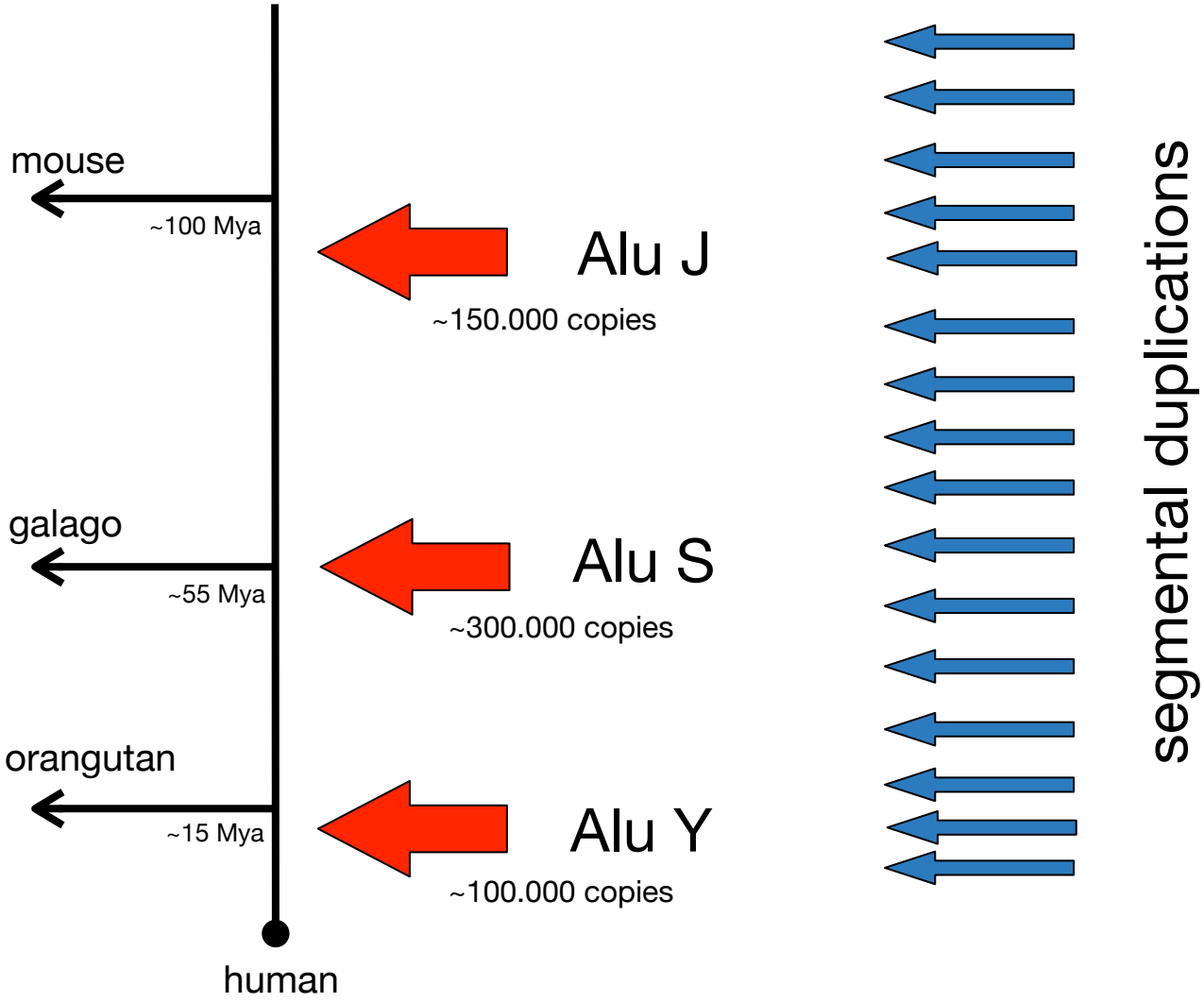


[Massip & PA, PRL 2013]

MAX-PLANCK-GESELLSCHAFT

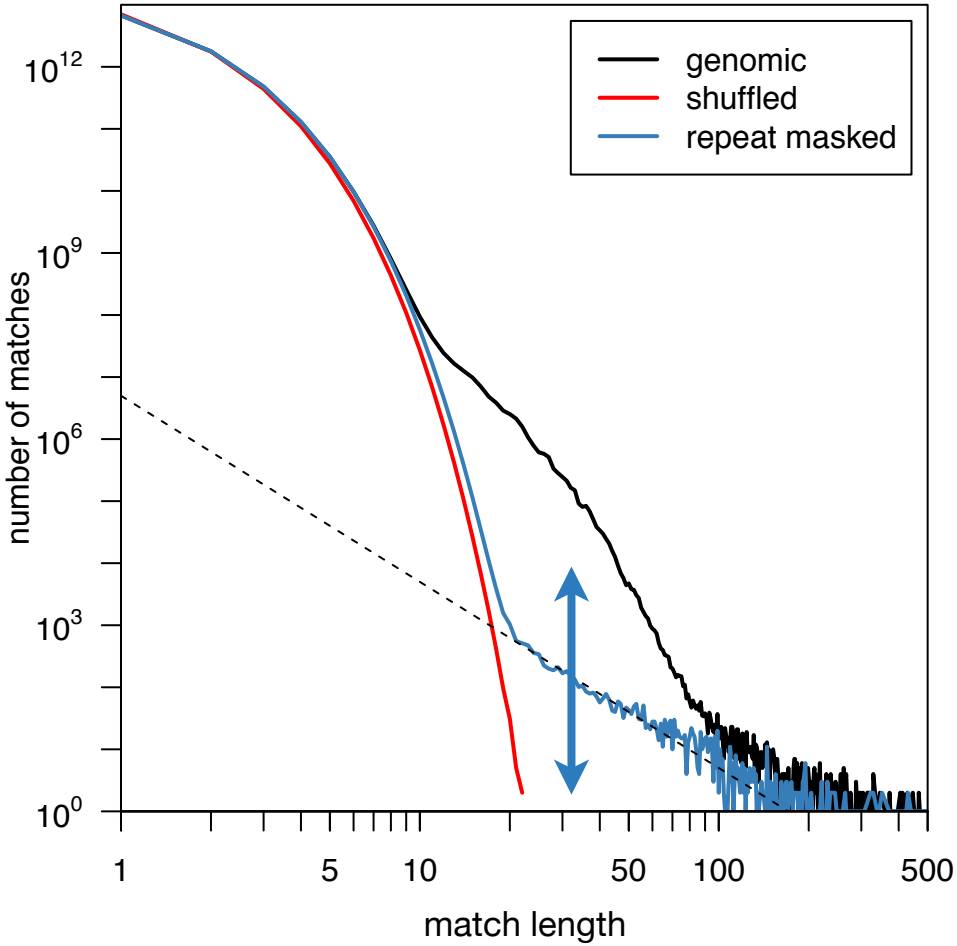
Segmental Duplications

- continuous generation of new stick due to SDs



[Bailey & Eichler, 2006]

Self-Alignment of the Human Genome



$$m_{\text{stationary}}(r) = \frac{\gamma K}{\mu} \frac{L}{r^3}$$

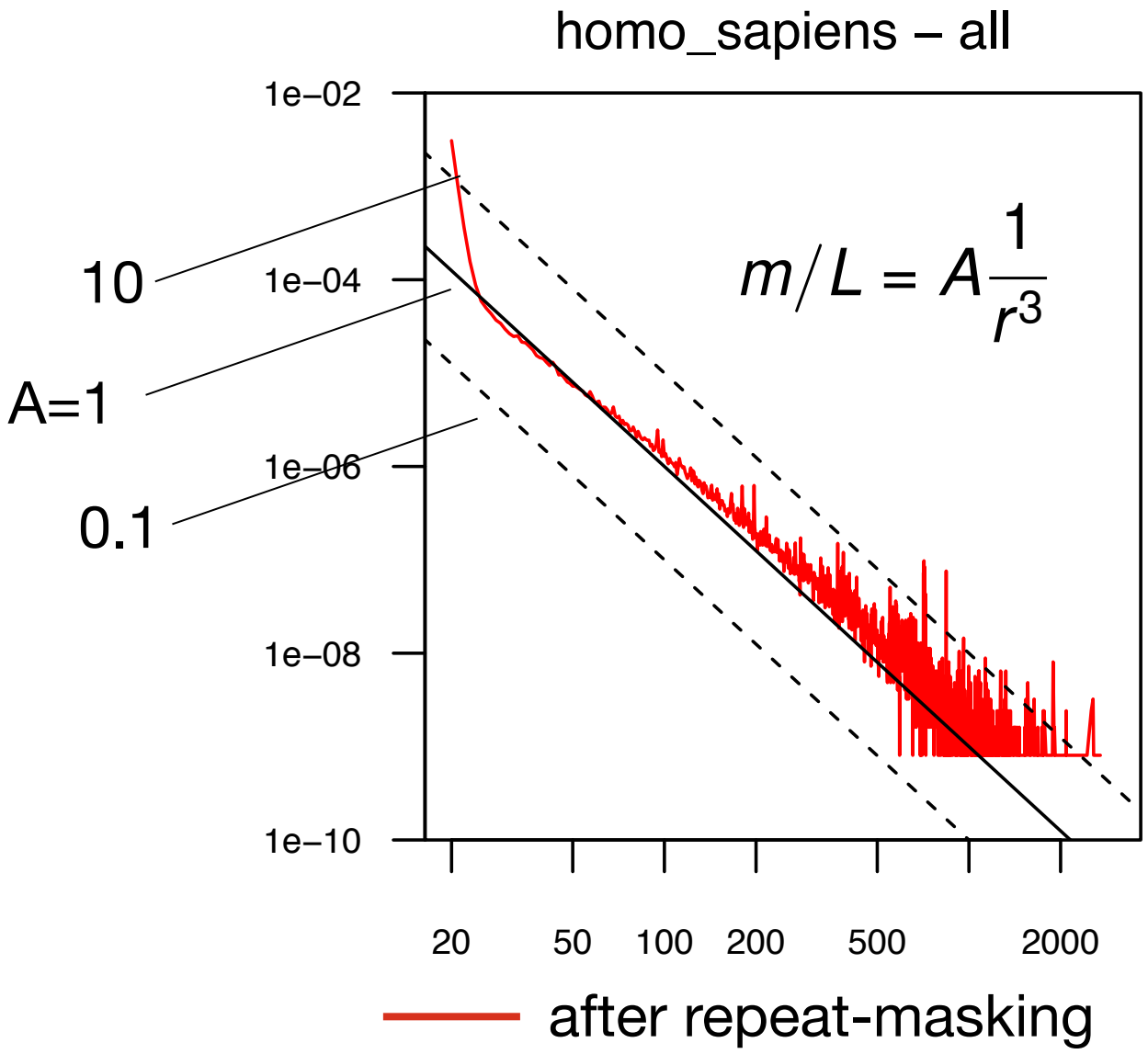
$$A = \frac{\gamma K}{\mu}$$

on average: as many back-up copies as mutations

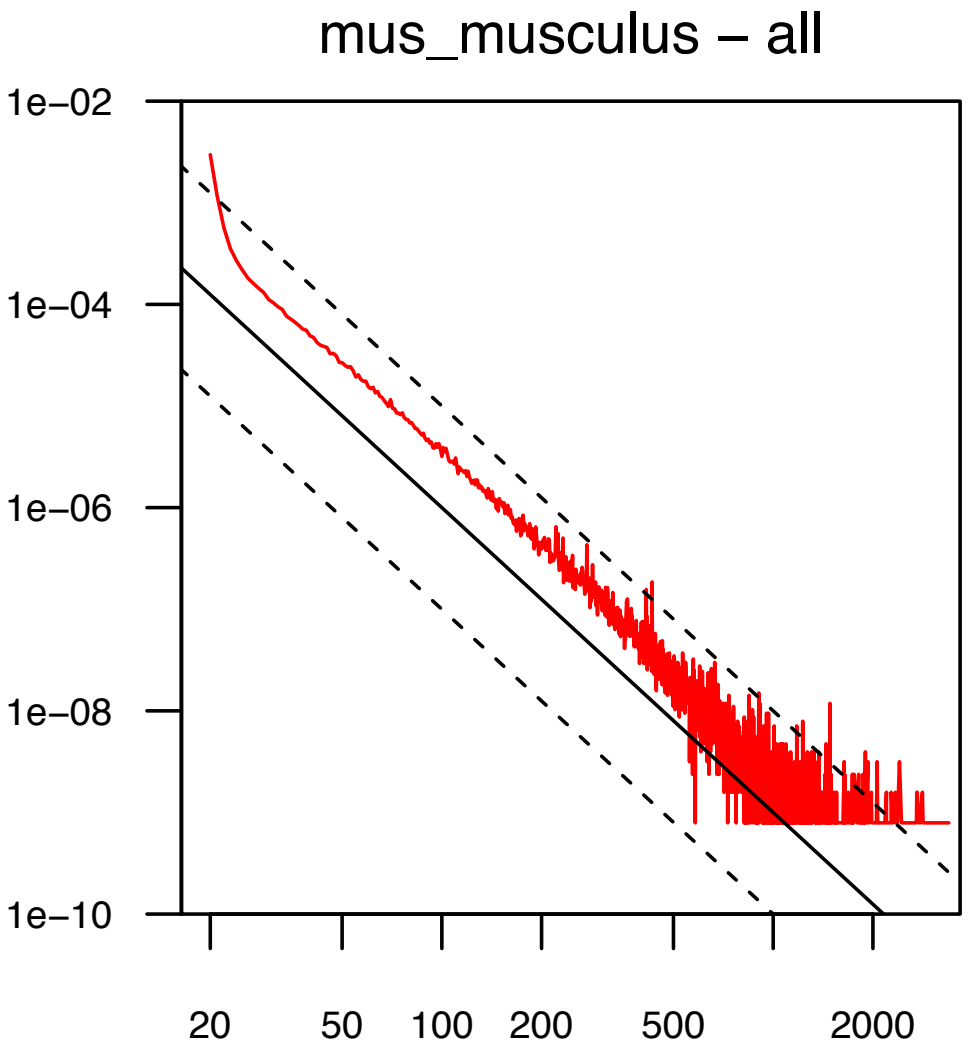
[Massip & PA, PRL 2013]

MAX-PLANCK-GESELLSCHAFT

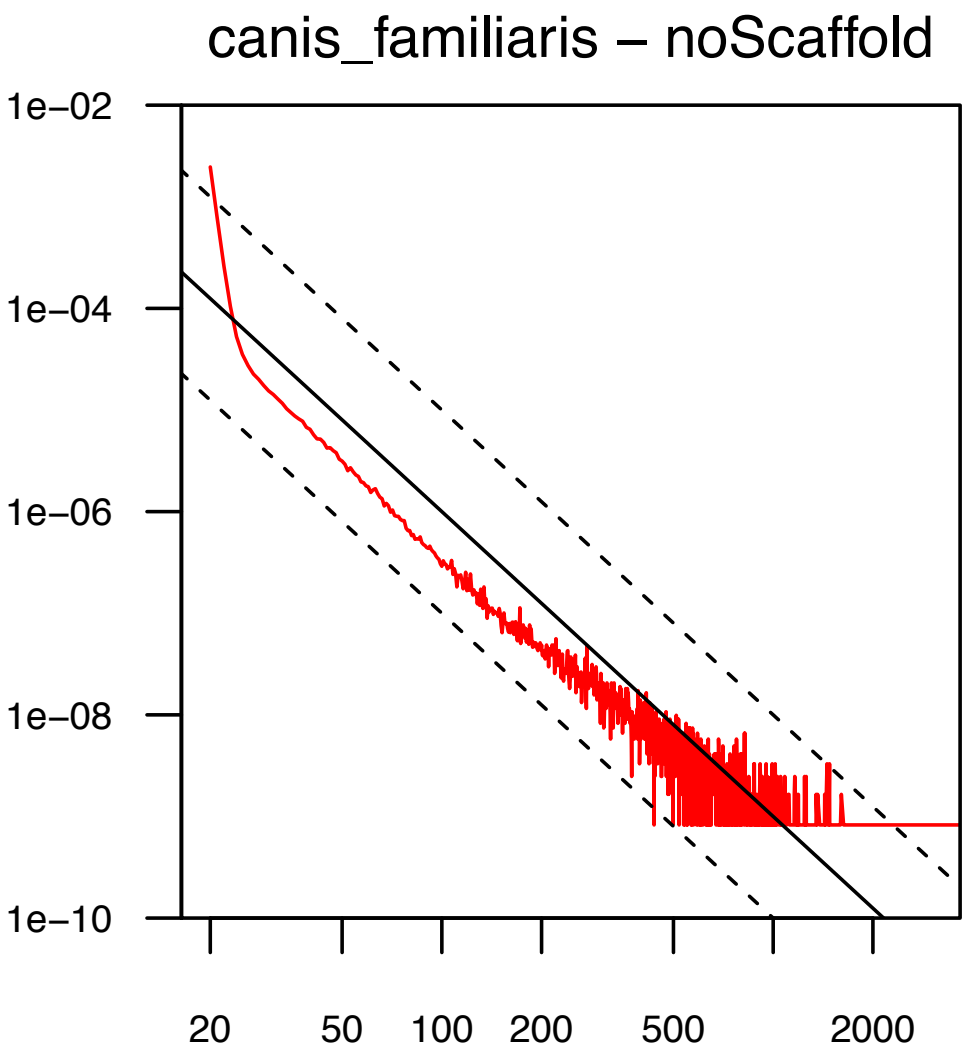
Human


homo_sapiens – all
1e−02
1e−04
1e−06
1e−08
1e−10
10
A=1
0.1
$m/L = A\frac{1}{r^3}$
20
50
100
200
500
2000
after repeat-masking

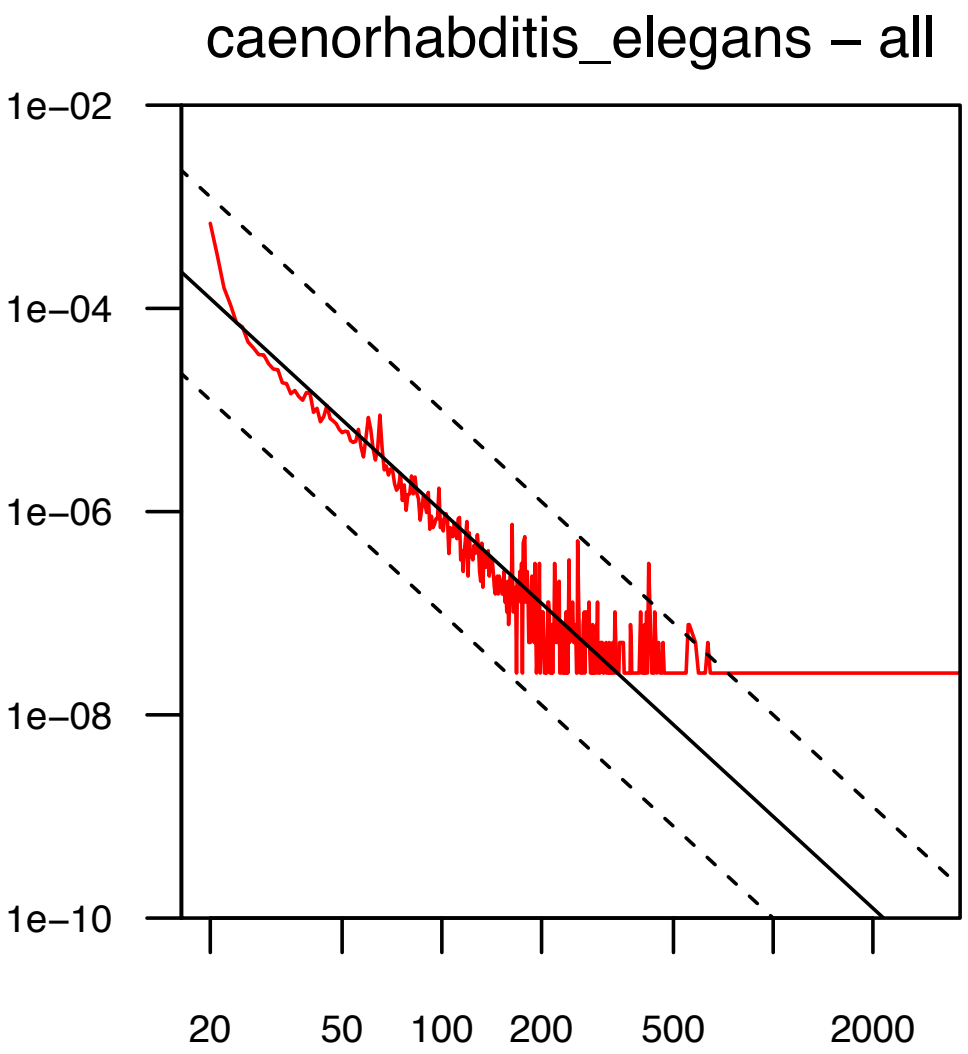
Mouse


mus_musculus – all
1e−02
1e−04
1e−06
1e−08
1e−10
20
50
100
200
500
2000

Dog


canis_familiaris – noScaffold
1e−02
1e−04
1e−06
1e−08
1e−10
20
50
100
200
500
2000

Worm


caenorhabditis_elegans – all
1e−02
1e−04
1e−06
1e−08
1e−10
20
50
100
200
500
2000

Acknowledgements

- **Florian Massip (INRA Paris)**
- **Irina Czogiel**
- Barbara Wilhelm
- Yves Clement (ISEM Montpellier)
- Mike Love
- Martin Vingron
- Navodit Misra
- Dmitri Petrov (Stanford)

Reference:
Massip and PA, Phys.Rev.Lett. 2013