

The Evolutionary Fate of Duplicated Neutral DNA

-

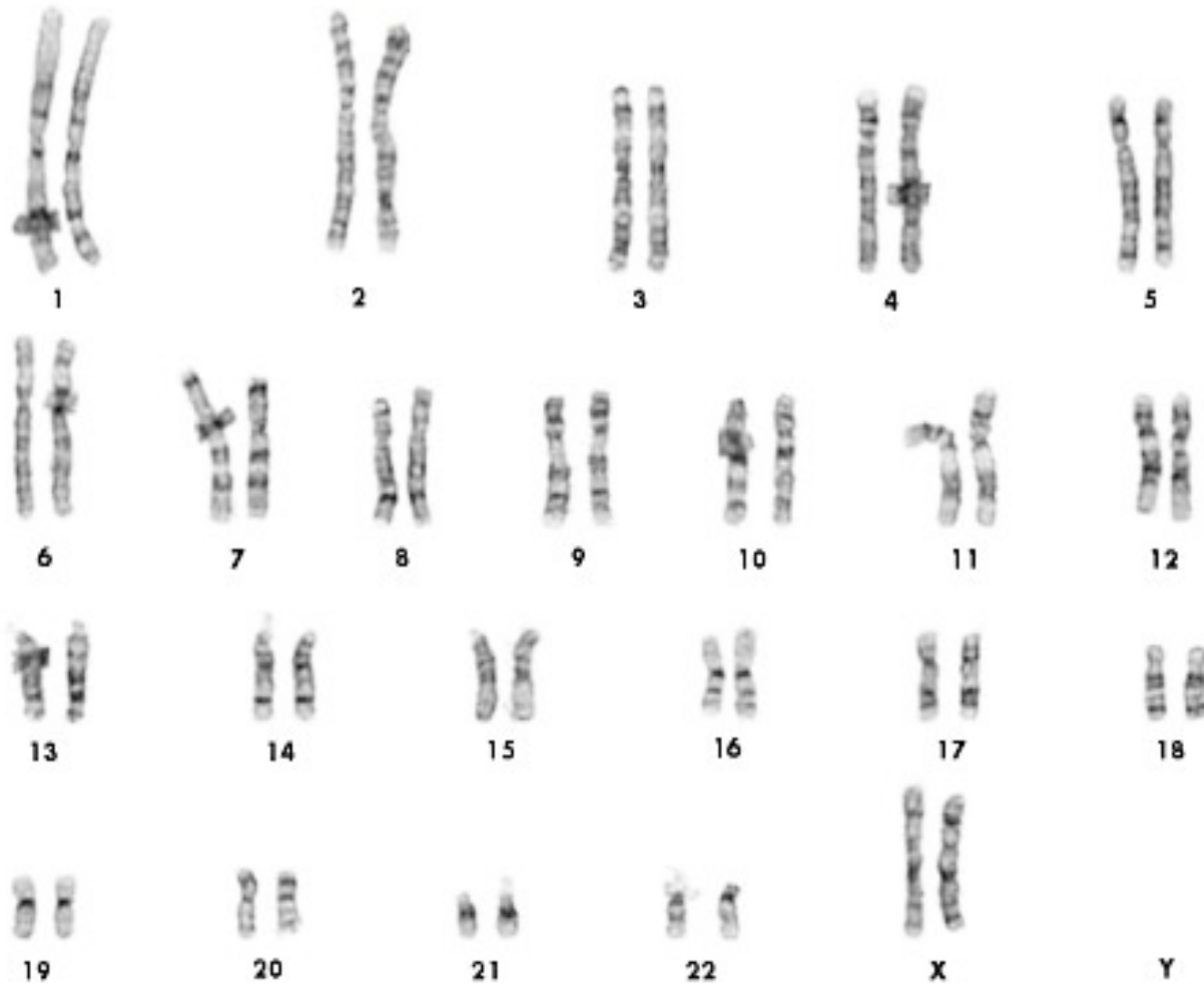
Breaking Sticks on Evolutionary Time Scales

Peter F Arndt and Florian Massip

The Evolutionary Fate and Consequences of Duplicate Genes

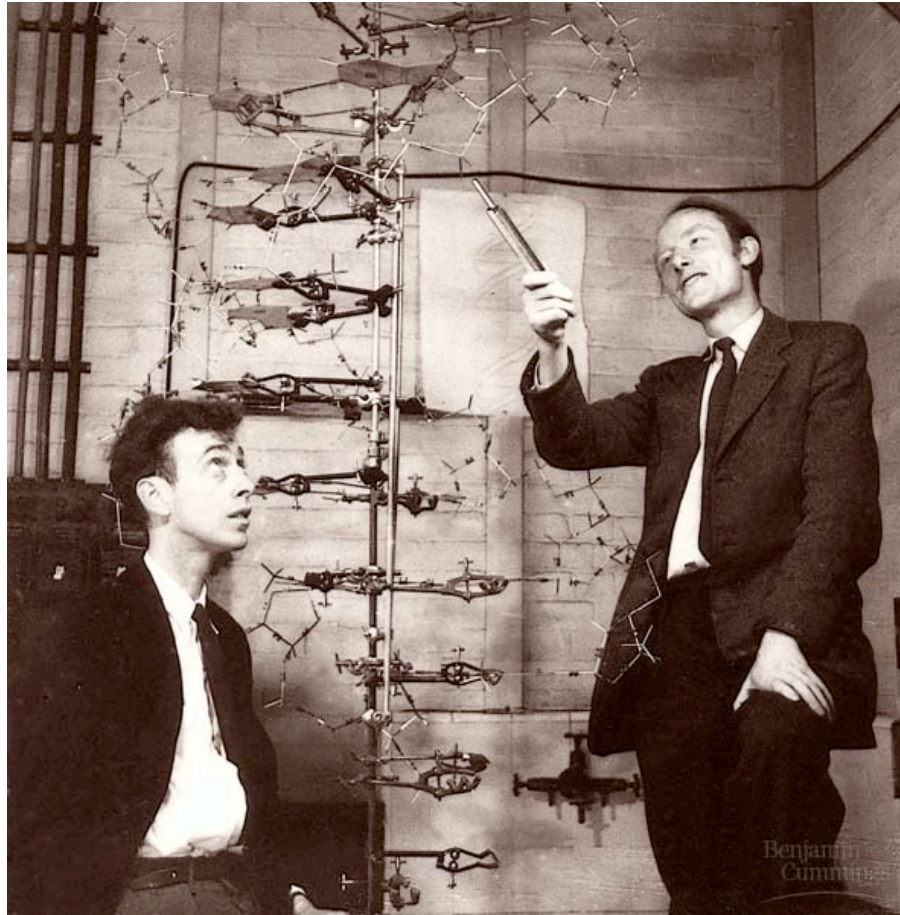
Michael Lynch^{1*} and John S. Conery²

[Science 2000]

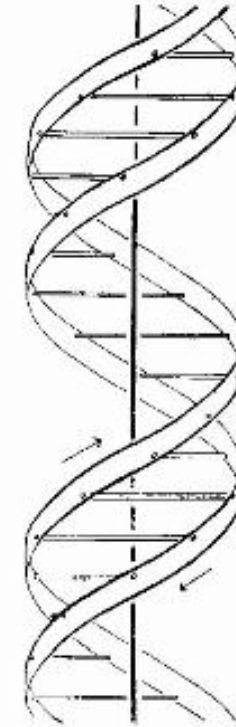


Giemsa-banding pattern of the human genome

Watson + Crick: 1953



James Watson + Francis Crick

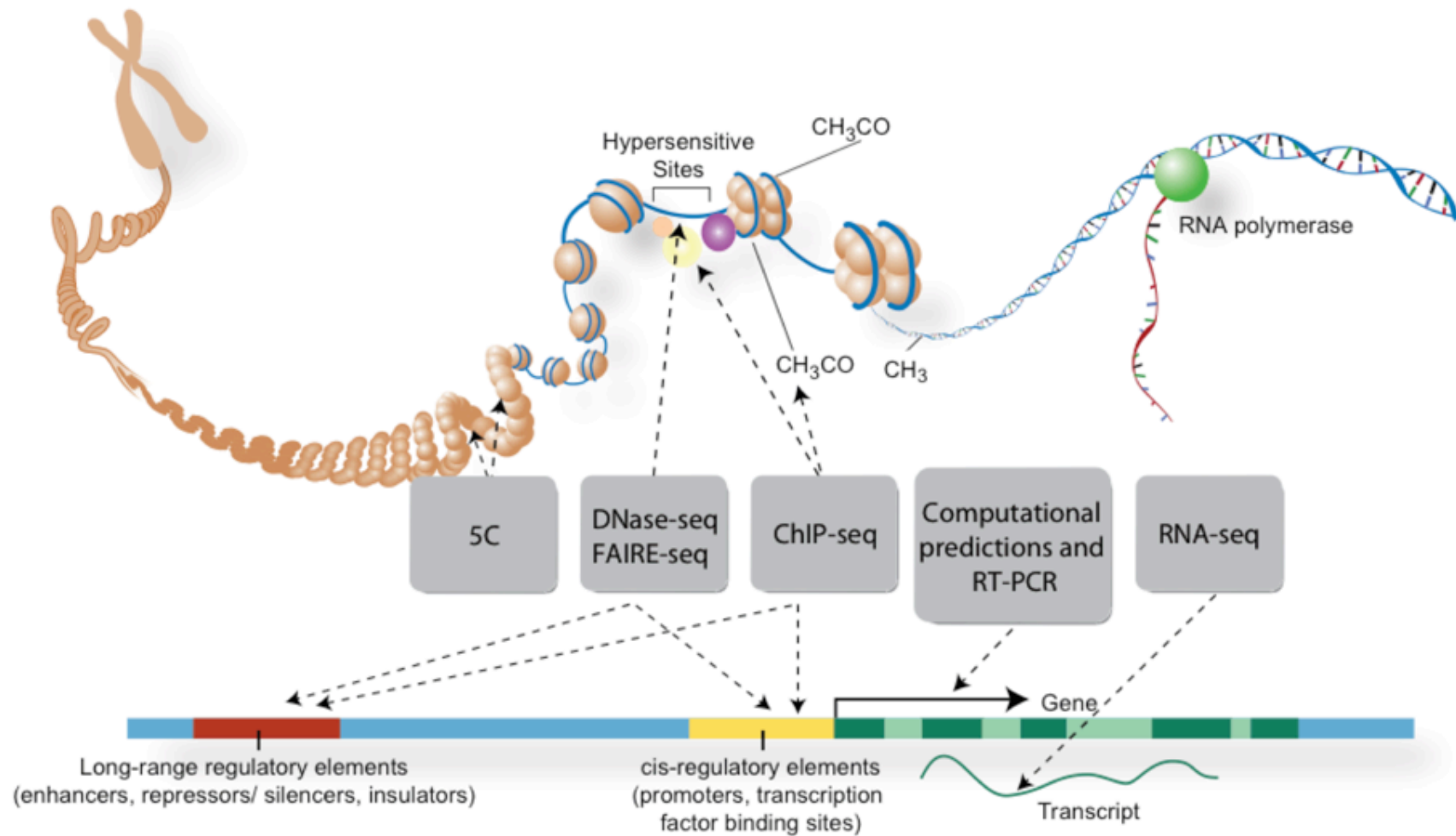


This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis



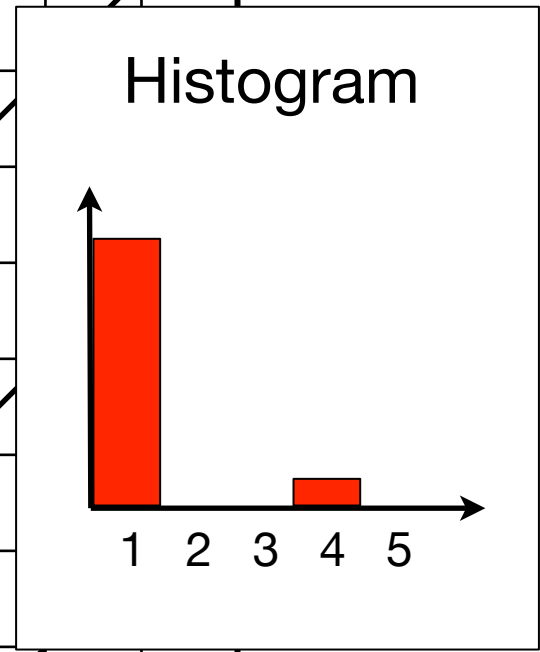
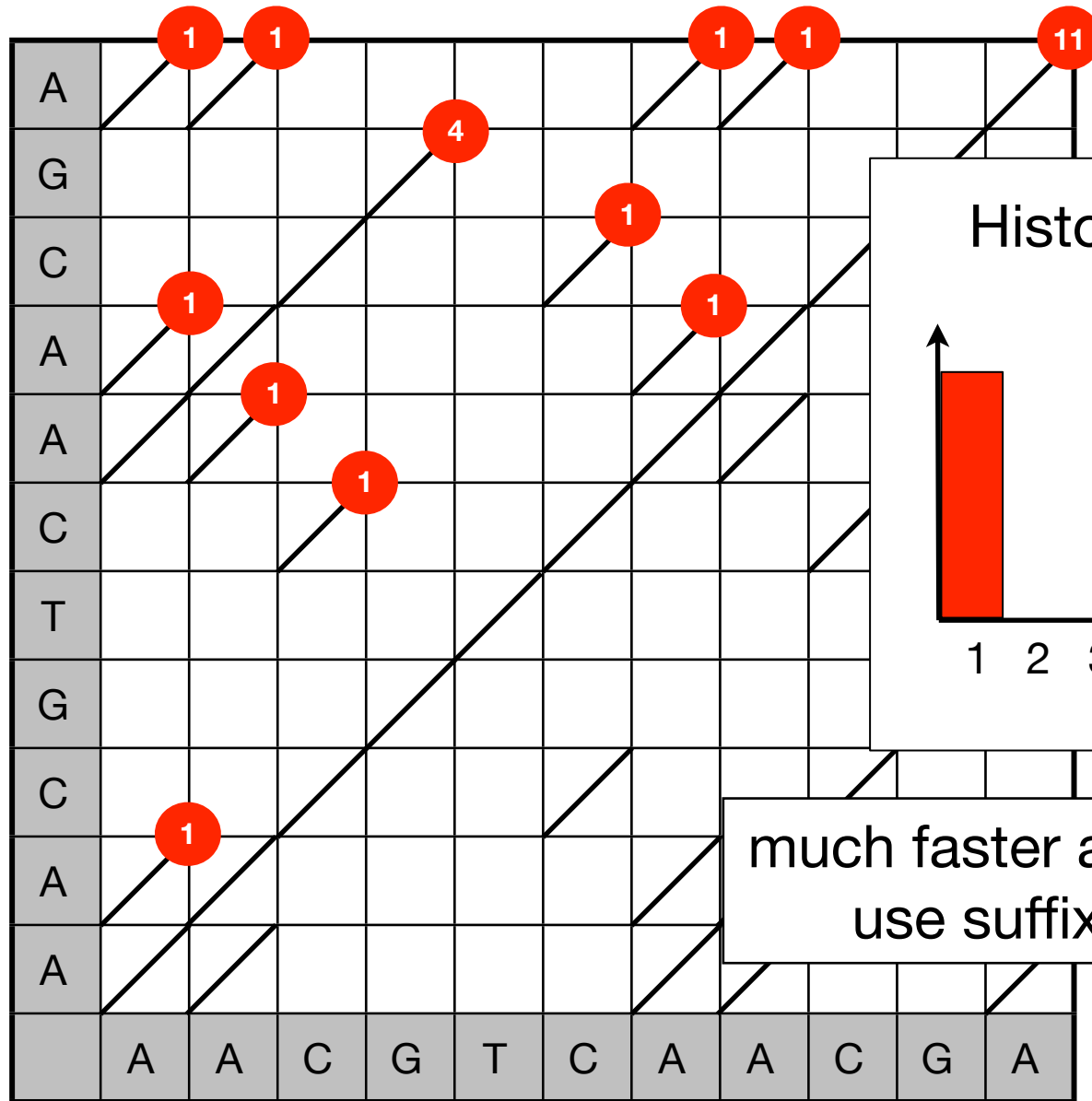
ENCODE: 2013

- comprehensive parts list of functional elements



TCTACTCAGTGTGTAGAAAGTTGGGATGTGGACCCAGAAGCTGCCATTCATTACACAGAATAAAATGATATCTGGTGGGTGGATCGAAAAGCTTGCCAAAAGAAGACACAGTGA
AAAGCTAACCCCTCAGTCTAGCTCCTGTTTTTAAATATTTATATCTCTGGCAGACTTTTTGGGGGAGAGGAAAGCTGAAAGCTCTTTTGCTATCTGTCAATTCACCCAGCC
AAATGAGAGAGAGTCTCAGACTGATTTTCTCGAGTTTGGCACCAGGAACCGCGTTCTTTGGGATGGTGAATTTCTTTGTGGAAGAGTGGCCGGGACAA
GGATTTTTTACAGAGCCAACAGCCACTTTCCATACCCGCTTGGGGCAGGGCCAAGCTTCCATTAGCACTGCTTTACAGATCGGGCCTGCTTTGCAGCCGGCCAGTGGCCAAC
TGTGGTCTGGCCTGTTAGGAAACCTACGAGGAAA**GTCCCTTTTTGGTATAATGATCTATTTCCCTTTTGTATGTACATCCAGTAATGGAAATACCCGGATGGAA**GCCACCCTGTA
GCCTGGGGTGGTGGGAAGTGTATCGCTCTAGATCCTGGGCAATCTGGCTCACTAGCTTGCCTCTCTCGTGAATTTCTCTGTGATCATTAGTTTTCTTCTTTTCTCCATCC
AACTCTTAATTCCTTCCCAAGGAATCGGTCTCCAAAGTCAATGACTGCTGCTTTCTTCTGCTCACCATCATCTATGCCCTTCCCTTAGTTCAGGGACTAGAATATT
GCATTAACCTTCTAACTTTGTAGGGCAATACTGTAGTGTGTATAGCAGTGGTGAAGGACAGGGCTTTAGAATCCAATAGACCTGCATCTTAATCATGACTAGGCTACCTCC
TCATCGTGGCCTCGCCACAAAATAGGAGAATAGCTTATTTCTAGCAGGAATGTTGCATGGACTGCATGAAAACAGTCAGGAGAGGATCTAGGACACTAGCAGTCACTAGTG
AGCACTCAGAATACTACAATTTGTTCTTAACATAATCGACTAATGCATCCAGTTACATCAGAGAGCCTTTTTAATGCATCCCTTTTTAGTACCAAGTATCGTGTATGCTAAGG
ATGCAGATAGATAAGATGCAGAACCTGCCCTCAGGGAGTCTCCATTGAGAGACAAAATCATCAACTAAAAATCTCTGTAGCTTTCTTCCGCTACTCACATGCTTTTTTAAAT
TTCCAATTTTTATAAGATTTTTTACAGCAATTTTTGGCAATAACAAAAACAACAACTGTAATAGTAGCGACCACTGACCTGACCTTTCCCATGGGGCAGGGCCTATAATGAT
ACATCTGCATCTATTACCTATTTTGTCTTCTAGATATATGCTACCAATCATGATTAGCAAAAGAAAAGCTGGGGAACCTGGACACTAAGTAACCTGCCTGGGAATAACACAG
CTGTGATACGTGTGAACCAATGAAGAACAGGAATCATTTCTTTTGTGTTCTTGTGTTCTTGTGTTGTTCTTGGACAGATGAACCTGATTCAAAATCCAGGATTCTATTCTGATACTGACAGTGGC
TAGCTTGTTCAGCCATAGGGACCATAAAAAAGTATTCTTCCCAATTTTTTGTAGCATGTACTATAGCTTTTTTGTGCTCTGTAGACACACAAAATGATTCACATAAACTC
TCCCCTTAAGAAGAAATAGGACAGTGAAGGGTAAACAGGATGCATTAGAGCATTTCAGGAGATGGTGTCTTCTACTCCAGCTAGGACTGGCCAGAGAGACTATCAGAAA
GGTGTGAAGTATGAGTGATAACGAATGTGCAGAAATGATGAGGAGAGGTTCCAGGTGGAGAGAAACATATGAGCAGAGGGCTCGTAGATAGGAATGAGTCAGGTGTGTGC
TGATAAGAGTTCAGGGCTCTTGAACAGACCCAAAGTTAATGAGGACATGAACCTTGGACAGTGTCAAGACTGAGGGAGAGAATGGGCGTCCGGGACAATGGTACTTGCCAA
ATTTTGTGCAAAATTTGTTGCTTAATAAATATTCCTTGATGAGGACAACATATGCTGTTACTTAAATAAATTTCCCTGATGAGGAAGACATGTGCTGAGCTGATGGAAAAA
AAAAAAAAAAACGTAATCACTCAGTGTCTCAGTTTTCTGTTCACTCCCAAGAAACCAATTCATTGACTTAAAGCAGCCAAACTCATTAAGTAAATGCTCTGCACTACACTAG
GCACTGAAAATACAAAGATGAACAAGATATGCTCTGTTCTTGGAGAGATTACAATCCAGGGCAGAGGCAAGCAGATAAAGAAATAATTCTAAAGTGCTGTGGTAGGGCA
GTACAAAATTTAAGCAGGAACCCCTGGGGGCTAGACAGTGTGCTTAACTGGCCTGGGGCAATGAGAGAAAGCAGCACCGAAAAAATTAGGGTTAGAGTTAGGGATTG
AAAGTCATACTCTTCAAAGCACTTTATGAATCTGCAAAGCACTGTATGAATGTTACCTGTCAATTTTCTAATTTTTTACAACCACAACCTTTGAAGTTGAGTCAGCCTAAG
AGTTGACTTTATTTTLAGAGTGGAAAGTAACCTTAAAATCATACGGTCTGCTGTTACTTTTTCAGATTTGCAAGTGAAATGCAGTGAATGTGTGTGTGTGTGTGTGTGTG
TGATTTGCTACTTTATAAATAAATGTTATTTTCAATAAATTAATAAATAAATTTACTTCTTCAAGTTACAGCAAAAGGCAATGAAGTTATTCTGATGAAATGCTAACAAT
TTGATATTGGAATTATGGATGATATTTGCAACACATCAGCCTCAGTTATTGTCTTTATTTTGGCATTGTGATTTGAATGCCTGGCCTCACGGAAATTTCTGATGGCACTGAT
AAGTAGTGAAGTGGAACATCTGTGGGCTCTTGTGTGTTTTGTTAAGCAGTTGACTCACAATCTCAAGATACTCTTCAATTAATCTGAGCCAGAAGCTTGAAGG
CAAGATAGTGAGACACTTCTGTTTCAAAGTGAACCAGTAACAATACCGATGATAGTCCCTCGATAATTCTTTGACATAATATGACAATCAGACACAAGGAGCCAAAAGCTTG
CACCAACATTAATTTGTCTATAACATGTGATTCACGGGGATCAGCATGTACAAAGTTCACACACAACCTGAATTAGATTGTTACAGCTTGACAACAGGACTGCCAAGTAAA
TGAGATATTTGCCCTGCCACTTATTTGGAGTACCTGTGACTCACTTGTGCAATTTTAGATTAATAAGTCTGAATTTCCAGAAAGACCAATTCAGAGGCTACAATCTTCTTAATCAA
CGACATACTTATAAGAAATTAGAATGTACATGCAGATAAAGCATTATTTTTTAAACACCAAGGGATCCGTTGGCAGTGGAAATTATTGGTCAATTTGGTATATAGTAT
TTCTGTAAATAAATTTGATGAAATTTATTTTTAAAATATTTCCACAGAGATTGTTGATGTTTTCTTCTGAGTGCATATAAAAAATTAGTTGCCTAGCAGCTCAATAAATTA
CGTTGTCCGCTCCAATATAAATATTTTGTAAAGAAAAATTTATTTAATAAATATTTAATAAATATTTTAAATAAATATTTTAAATAAAGAAATATACTTAAGTCTCTGTGTGATACAT
ATAGAGACAGAGAGAGACAAAGTATATCTTTTACATAAAATAGAAAACAGACACAAGGAACCCAAAAGTACACTAACTAAACATAGATTTTTTAACTTTTCAA
AATGCAATCATGTGCTCTTAAATGACATTTTGGTCAATGACATACCACATATATGACAGTAGTCCCAAAAAATTATAAATACGGTATTTTTACTGTACCTTTTCTATGTTTGG
ATATGCTCAGATACATGAATACTTACCATTGTGTAACAACTGACTACAGAAATTCTGTAGAGTAATATGCCGTACAGATCTGTAGCCTAGGAACAATAGGCTATGTCAATGG
CCTAGGTGTGATAGGCTATACCATCTAGTTTGTGTAATAACTCTATGATGTTGGCACAACAAAGGTGCCTAAGGACGCTTTCTCAGAACGATCCCTGTGCTTAAG
CAATGGATGACTGTAGTTAAAAATGTTACGAAATGAAATTTGCATCTAACACTGCCACAACAAAACAAATTTAGTTCCATAAAAAACAGCTTGAACCAATGGCCGACTC
CTAGACTTTCCCAATGGATCATATGCAAGGTAAATTTAGTGGTATACATTTCTTTTTGTAAATAATGGTATAATTTTTAAATGTGCATTAATAATGACTTTACATTTTGC
TATCATGATATACATTTTTCTTTTACTATTAATAAGTACTTATAAAATTAAGAAAGGTGAGTTAATTTTTTGGAACTTTTTATTTTTAGATACAGGGGCGTACAGGTTTGT
TACATGGGTATATTGCCCTCAGATAATGAACCCAGTACCCAATAGGTAGTGTGTTTACCTTTGTCCCCATCCCTCCCTCTCCCCCTAGGGTCTA**67**CCGTGTTTATGT
TCGTGTGCTCCATGTTTAGCTCCACTTATAAGTGAACACATACAGTATTTTGGTTTCTGTTTCTACATTAATTCAGTTAGGATTATGGCCCAATCCATCCATGTTG
CTGCAAGAAATATGACTTCATCTTTTTATGAATGCATTGTATTCCATGGTGTATATGTGCCAGTTTTCTTTATCCAACTCCACCATTGATGGGCACCTAGGTTGATTCCA
TGTCTTTGCTATTAATAACACAGCGA**GTCCCTTTTTGGTATAATGATCTATTTCCCTTTTGTATGTACATCCAGTAATGGAAATACCCGGATGGAA**TGTAGCTCCGTTTTA
AGTTCTTTGAGAAATTTCCAAGCTGCTTTCCACAGTGGCTAAAATAATTTACATTTCCACCAACAATGTATAAGCATGCT

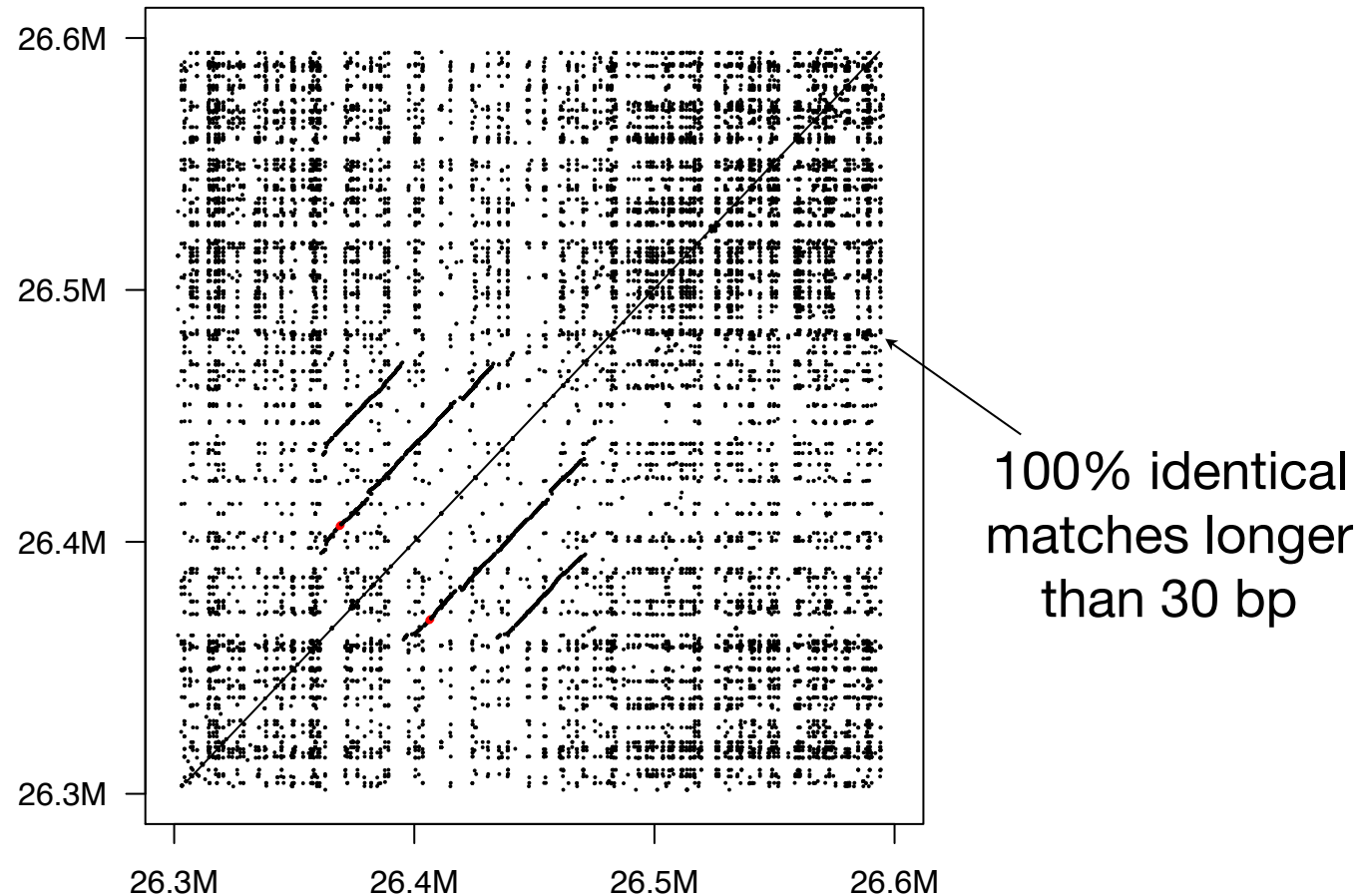
Self-Alignment



much faster algorithms
use suffix trees

Low Copy Number Repeats

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

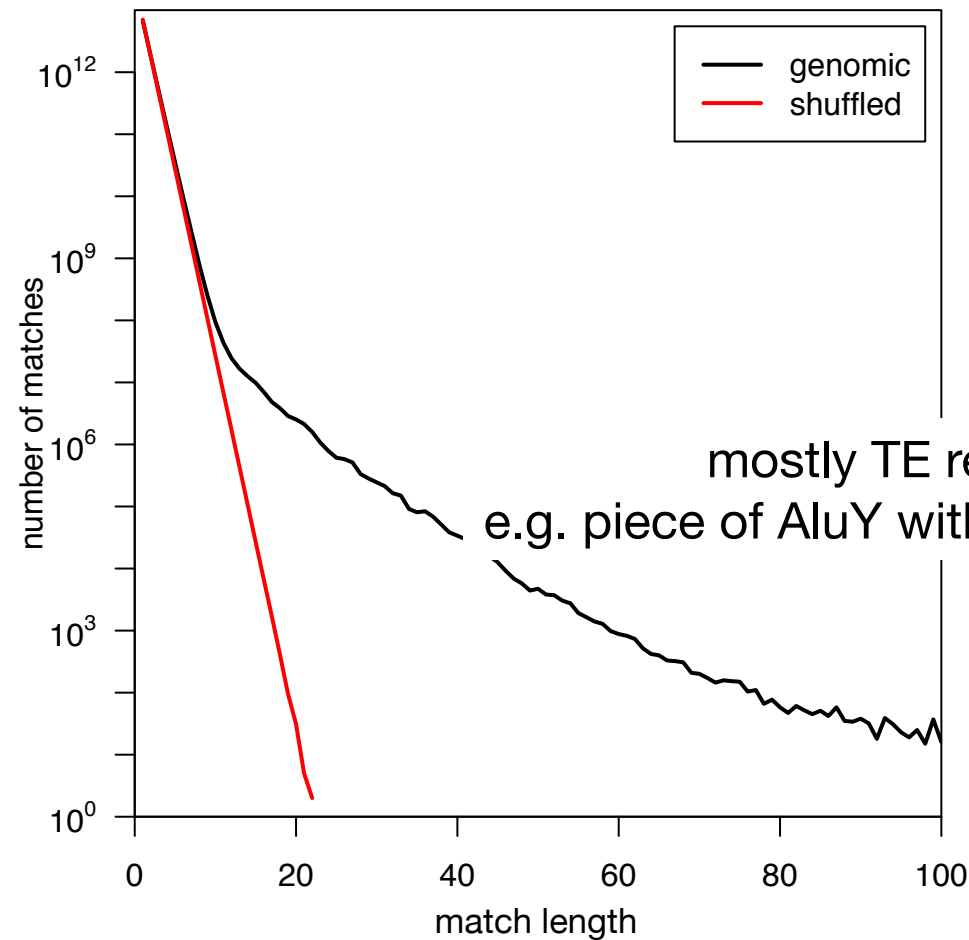


Example: 300 MB on the human chromosome 7



Self-Alignment of Genomes

- the length distribution of identical matches



random matches

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

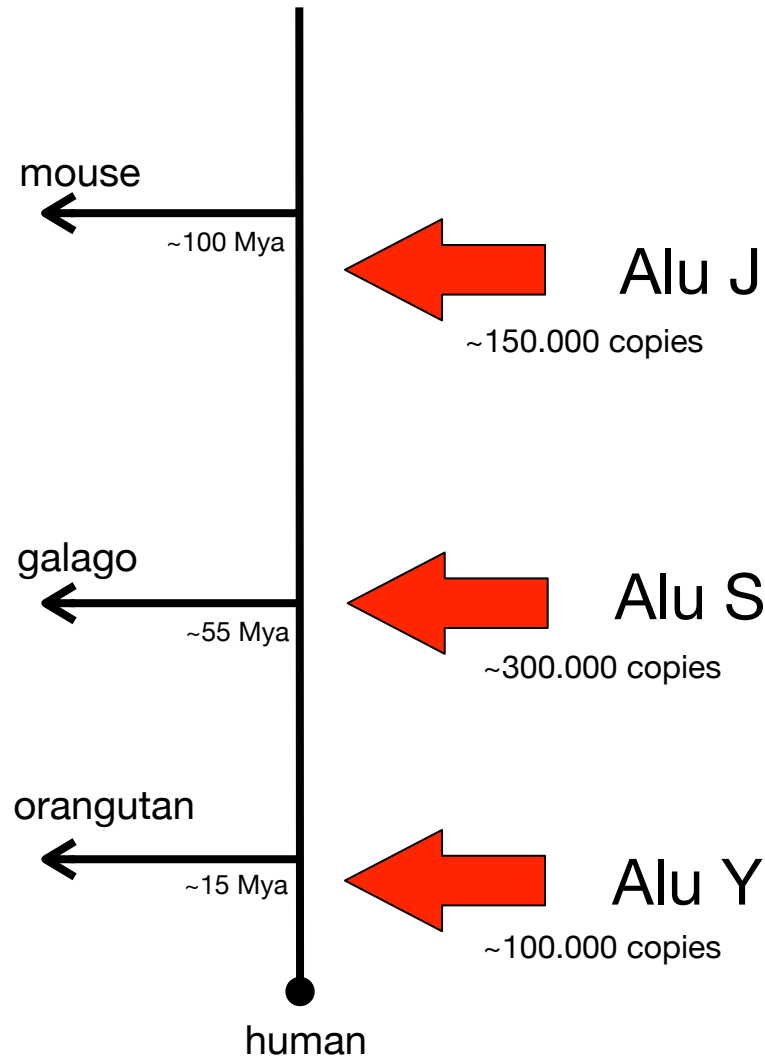
mostly TE related matches

e.g. piece of AluY with a piece of another AluY



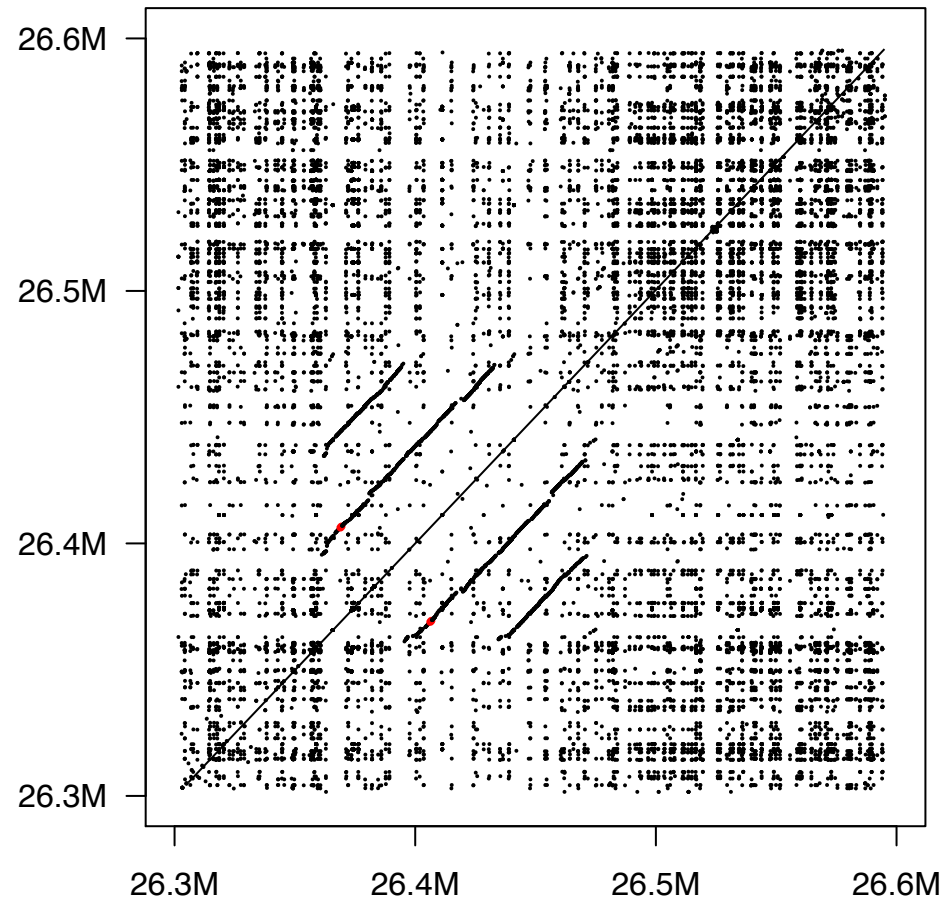
Transposable Elements

- retrotransposons entered the human genome in bursts



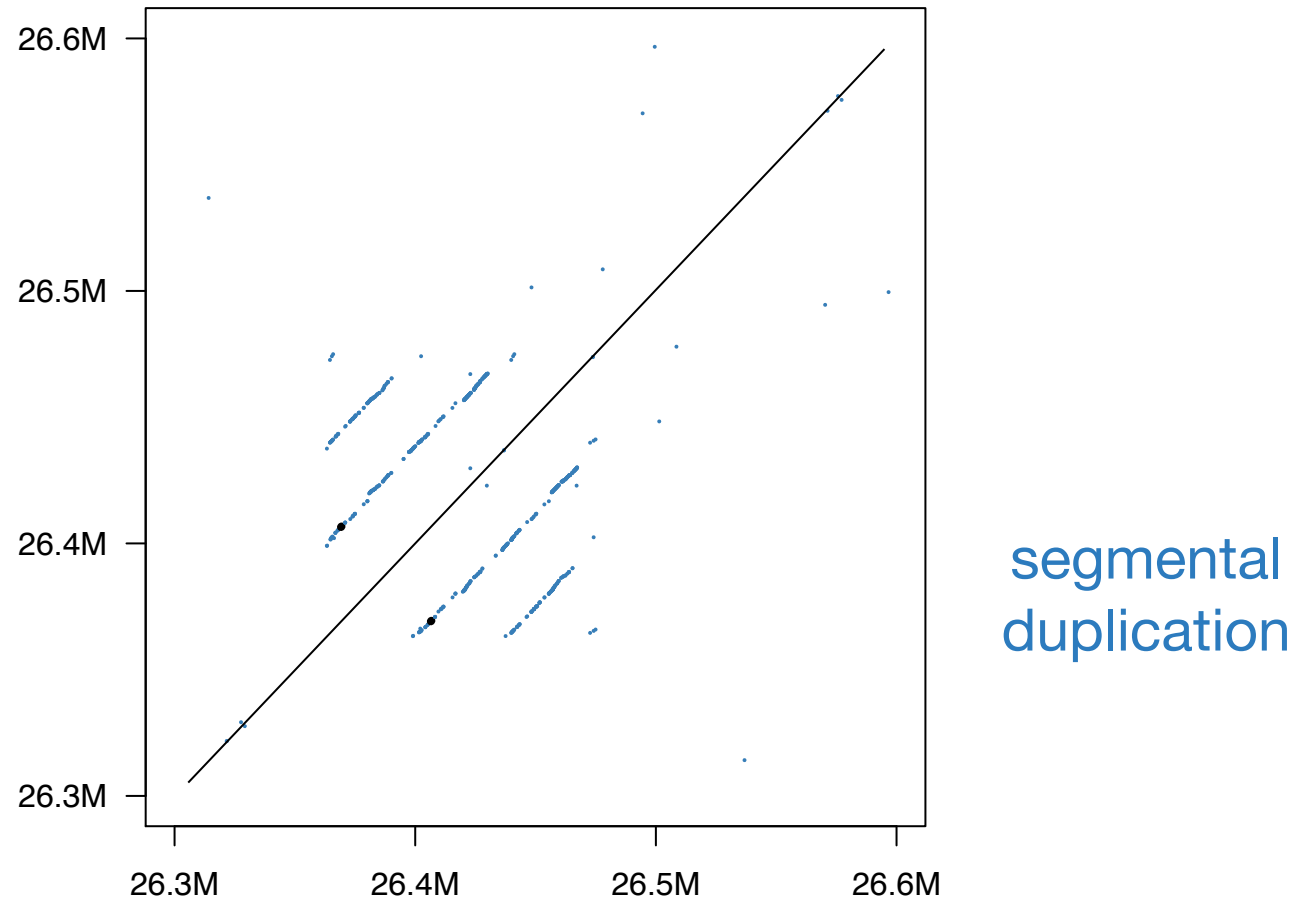
Self-Alignment

- Example: 300 MB on the human chromosome 7



Low Copy Number Repeats

- Example: 300 MB on the human chromosome 7

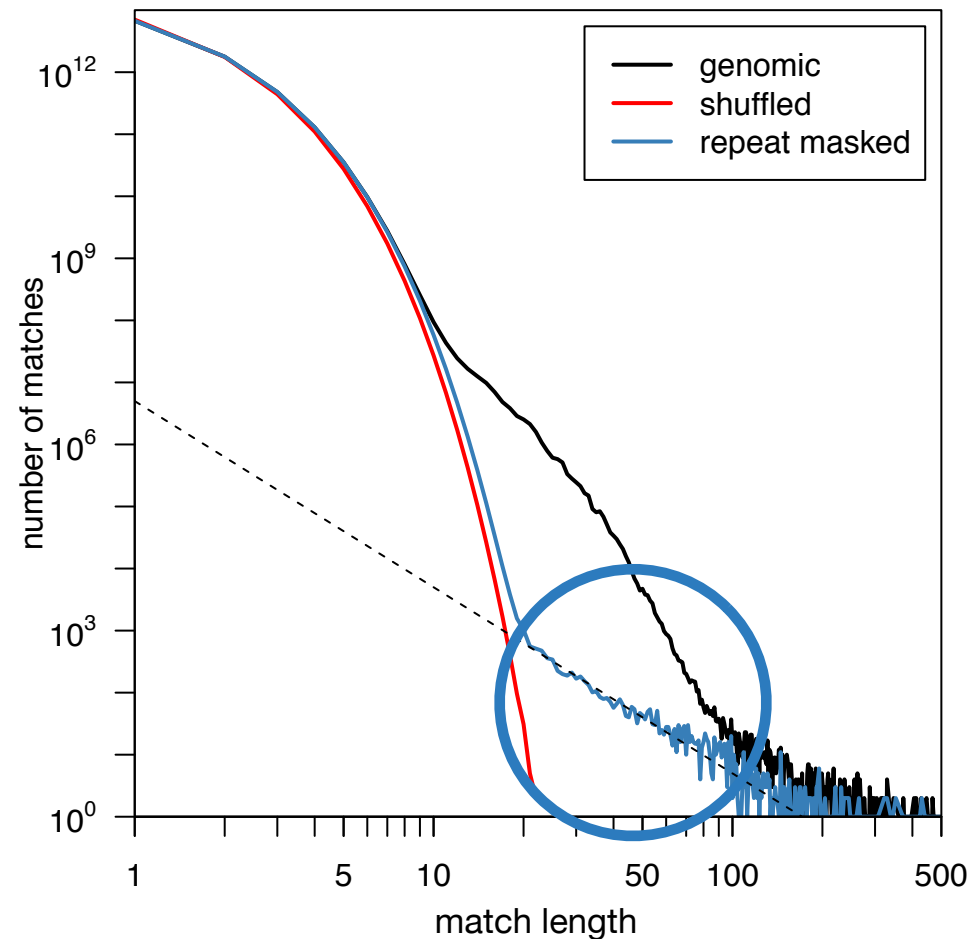


here: repeat masked

[Bailey & Eichler, 2006]

Self-Alignment of Genomes

- the length distribution of identical matches



random matches

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

matches unrelated
to repeats

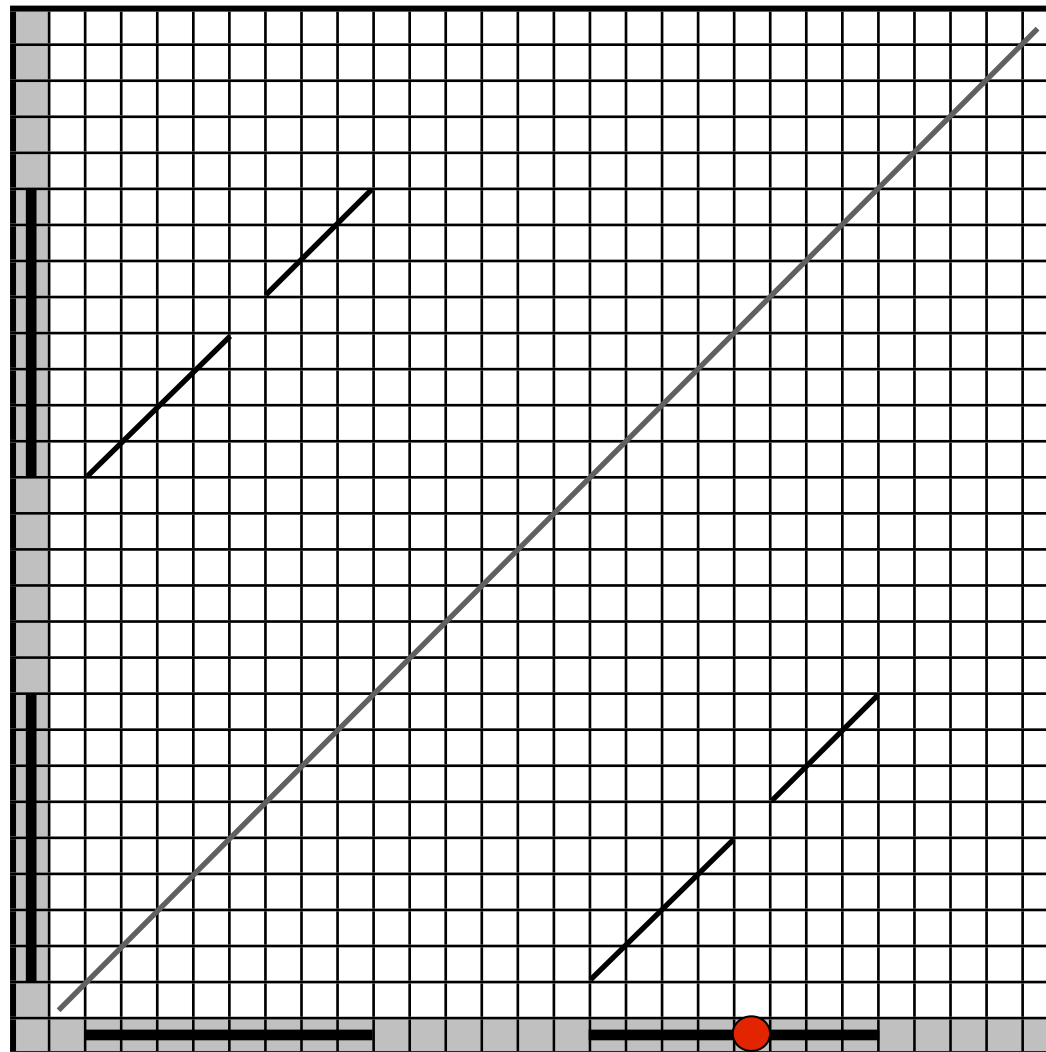
$$\alpha / r^3$$

Selection?



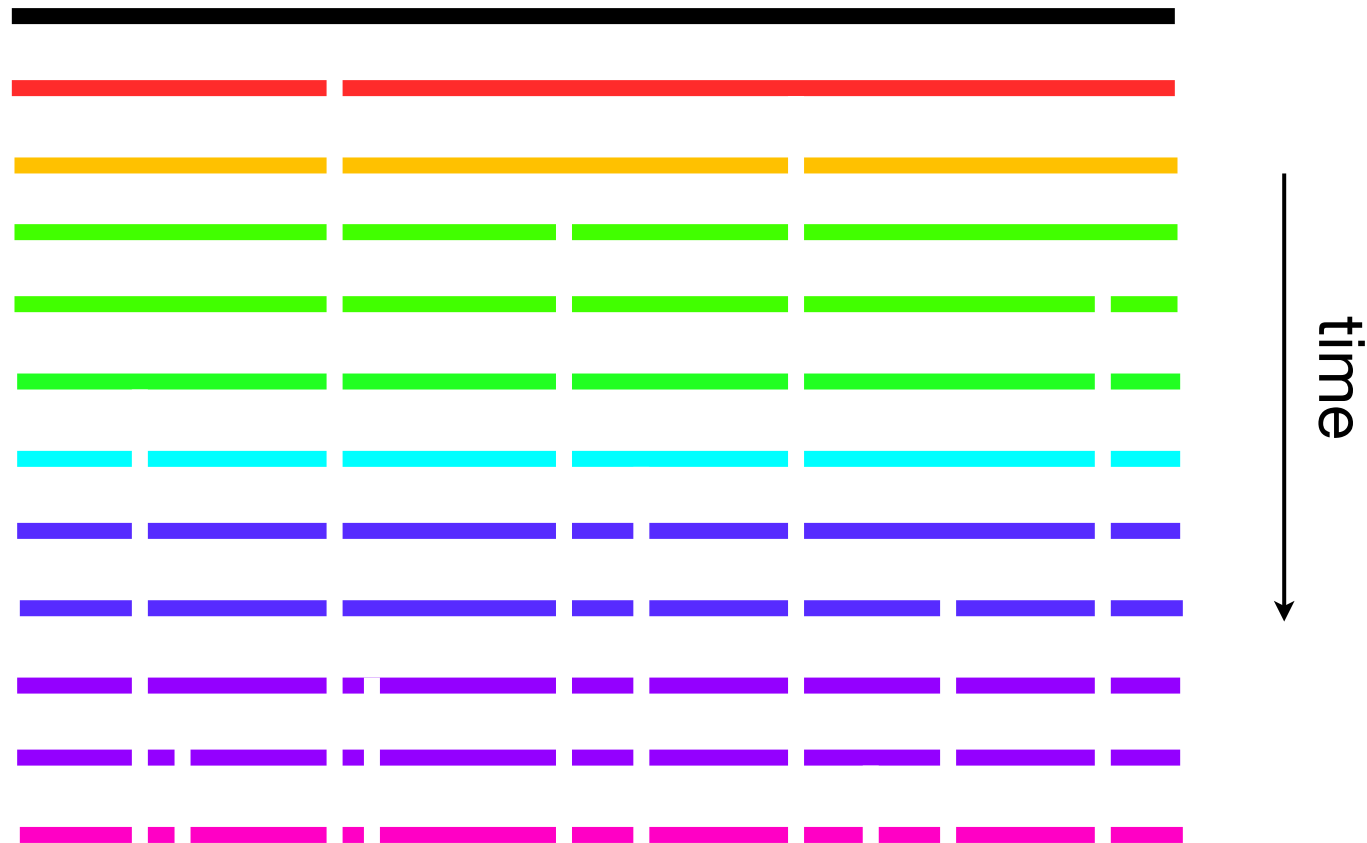
The Broken Stick Model

- a segmental duplication will dissolve by mutations



The Broken Stick Model

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



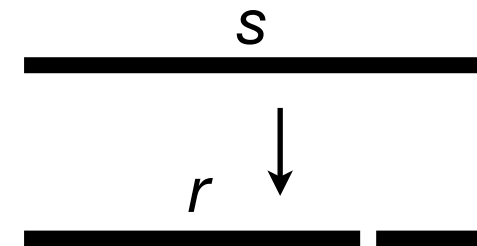
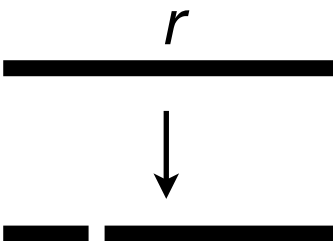
- let $m(r,t)$ be the number of sticks of length r at time t

The Broken Stick Model

- the dynamics is given by:

$$\frac{\partial m(r, t)}{\partial t} = \underbrace{-2\mu r m(r, t)}_{\text{loss of matches by mutations in one copy}} + \underbrace{4\mu \int_r^\infty m(s, t) ds}_{\text{gain of matches through particular mutations in longer matches}}$$

μ mutation rate
per bp



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

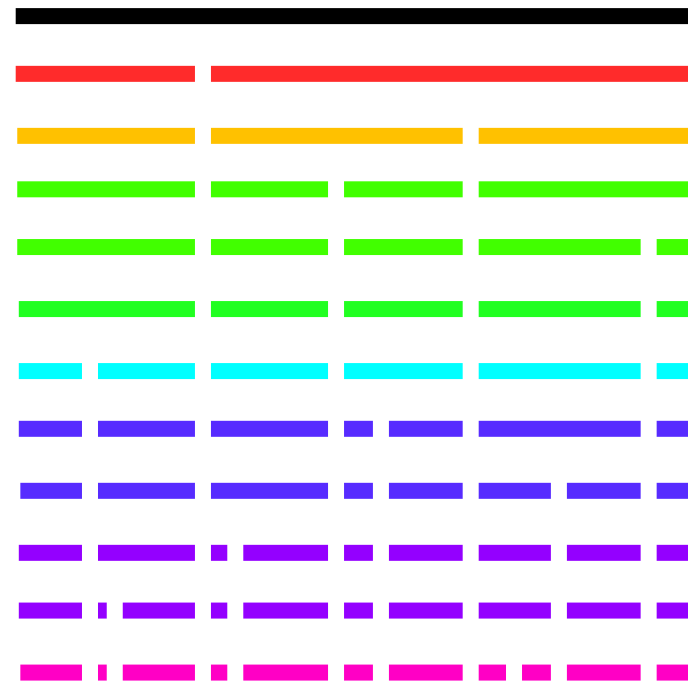
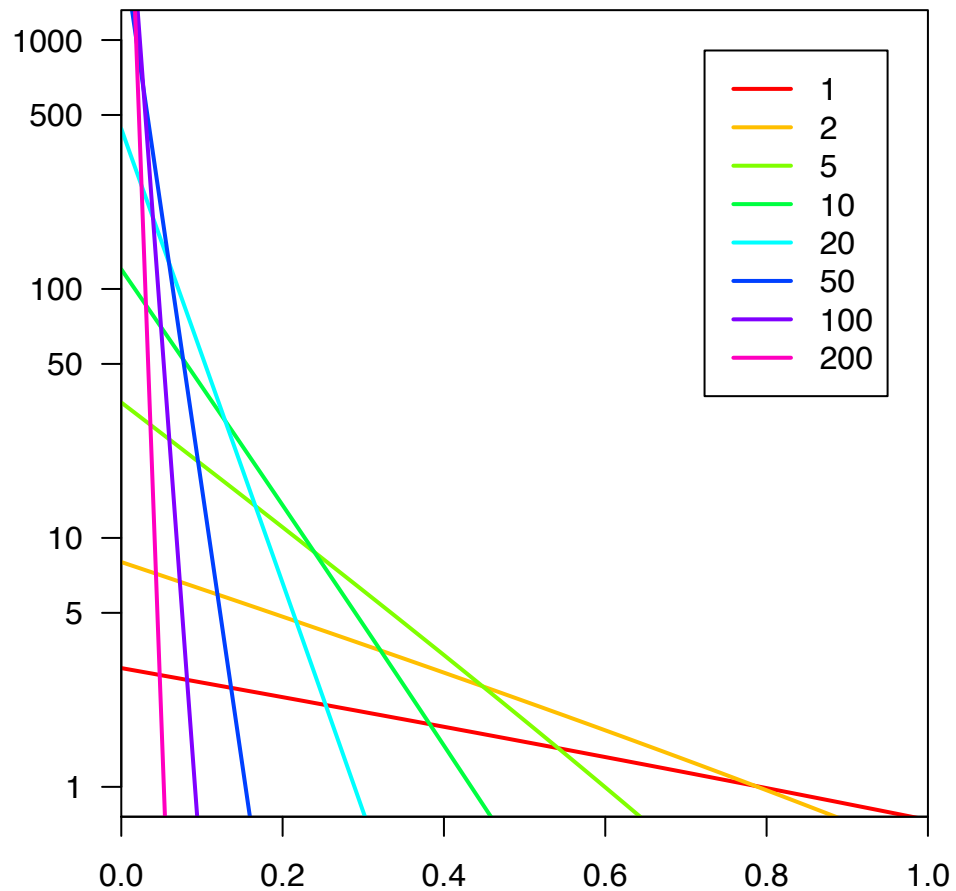
$$m(r, t) = [4\mu t + 4\mu^2 t^2 (K - r)] \underline{e^{-2\mu r t}}$$

where K is the initial length of a duplication.



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

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

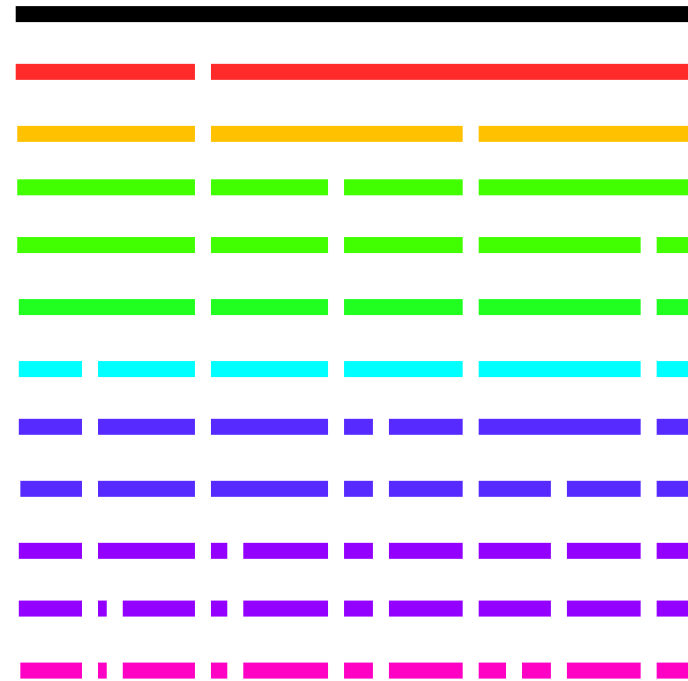
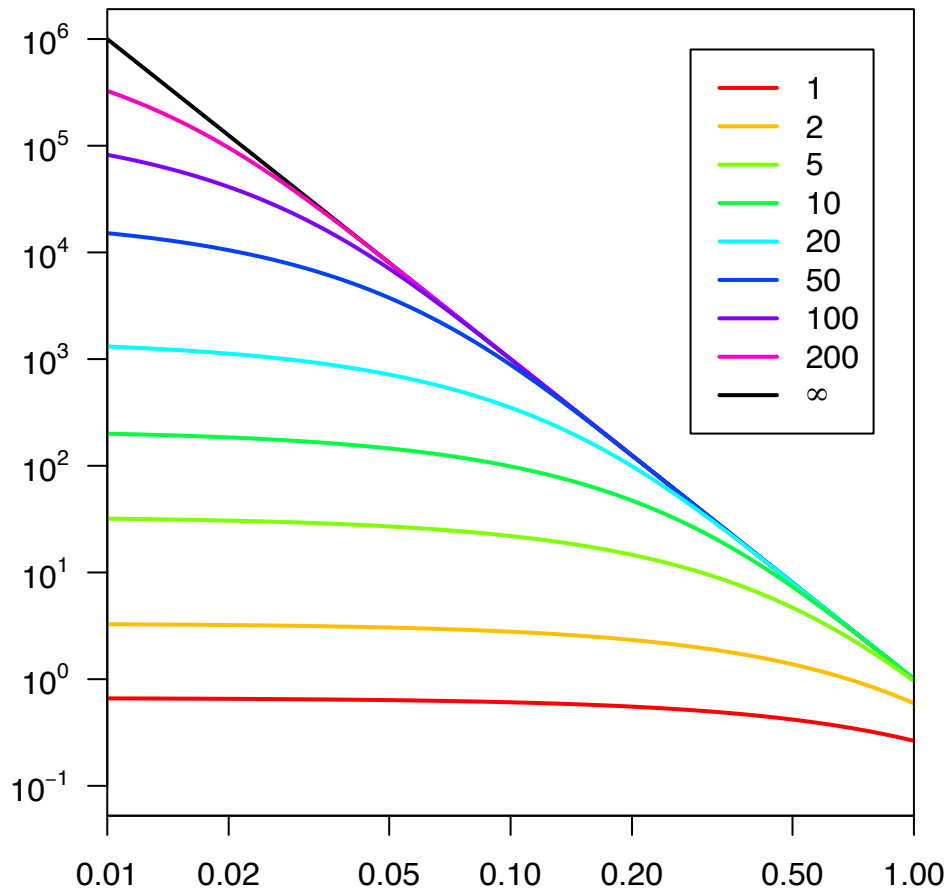
- its time average:

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



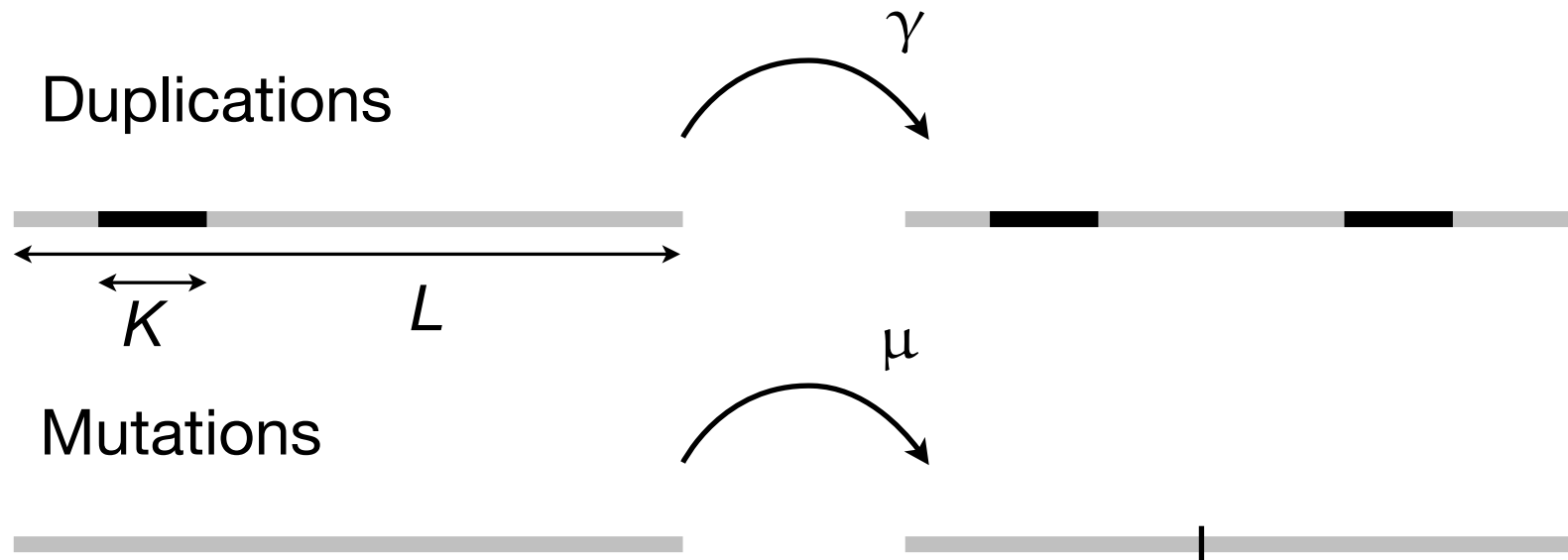
Many Broken Sticks

- 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:

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

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

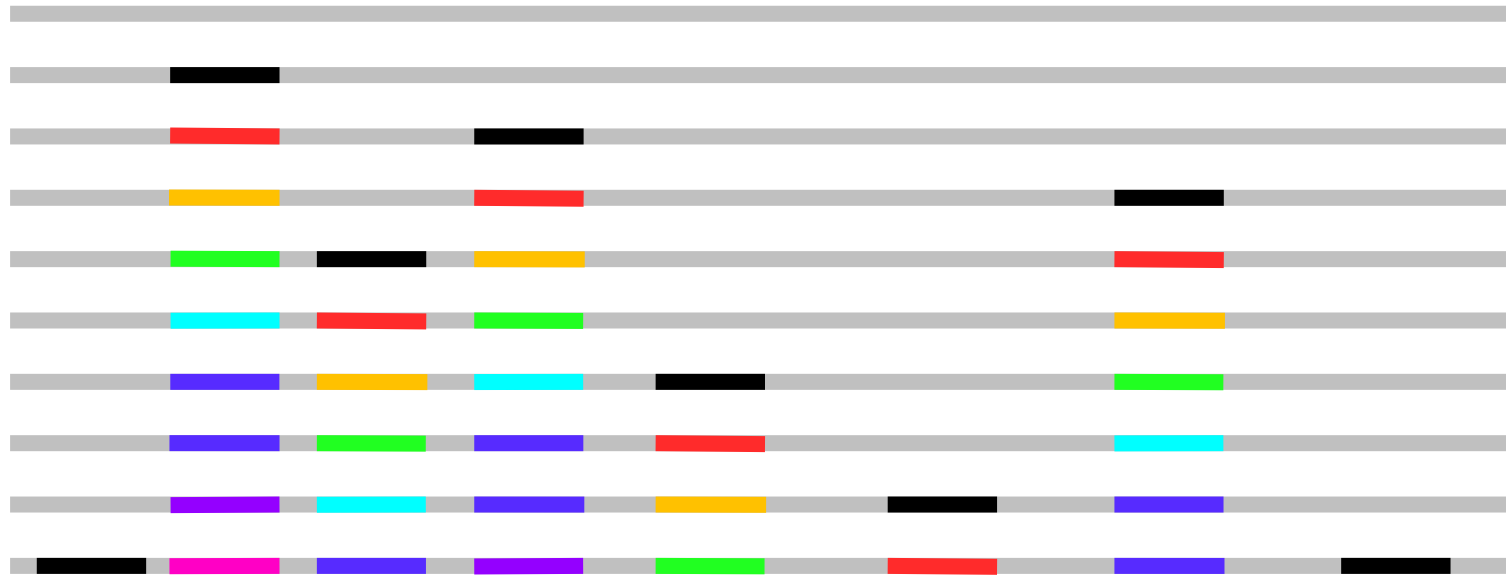
$$\frac{\partial}{\partial t} m(r, t) = 0$$

and given by

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

Duplication-Mutation Model

- starting with a random sequence:

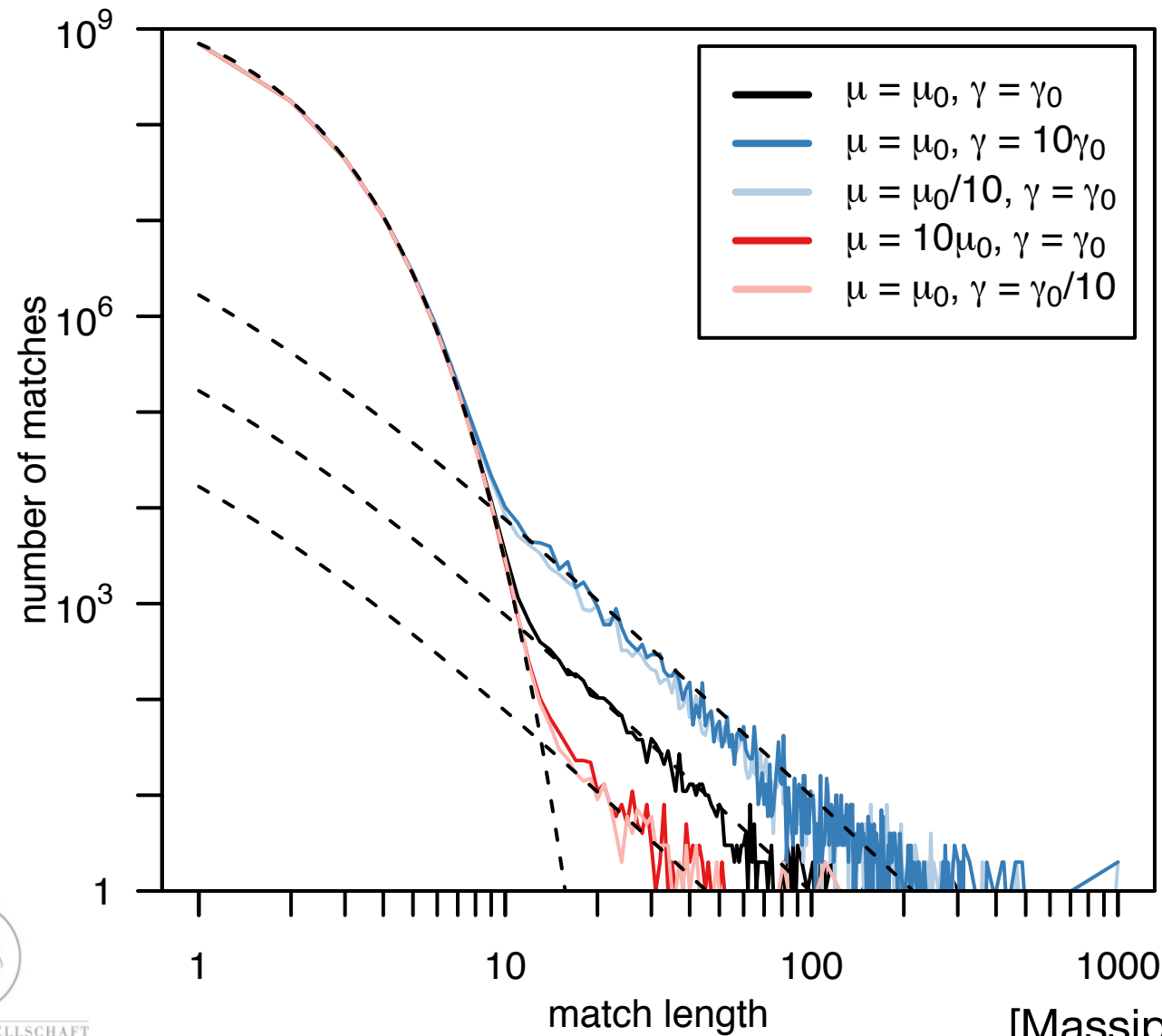


- one reaches a stationary state:

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



Simulated Sequences

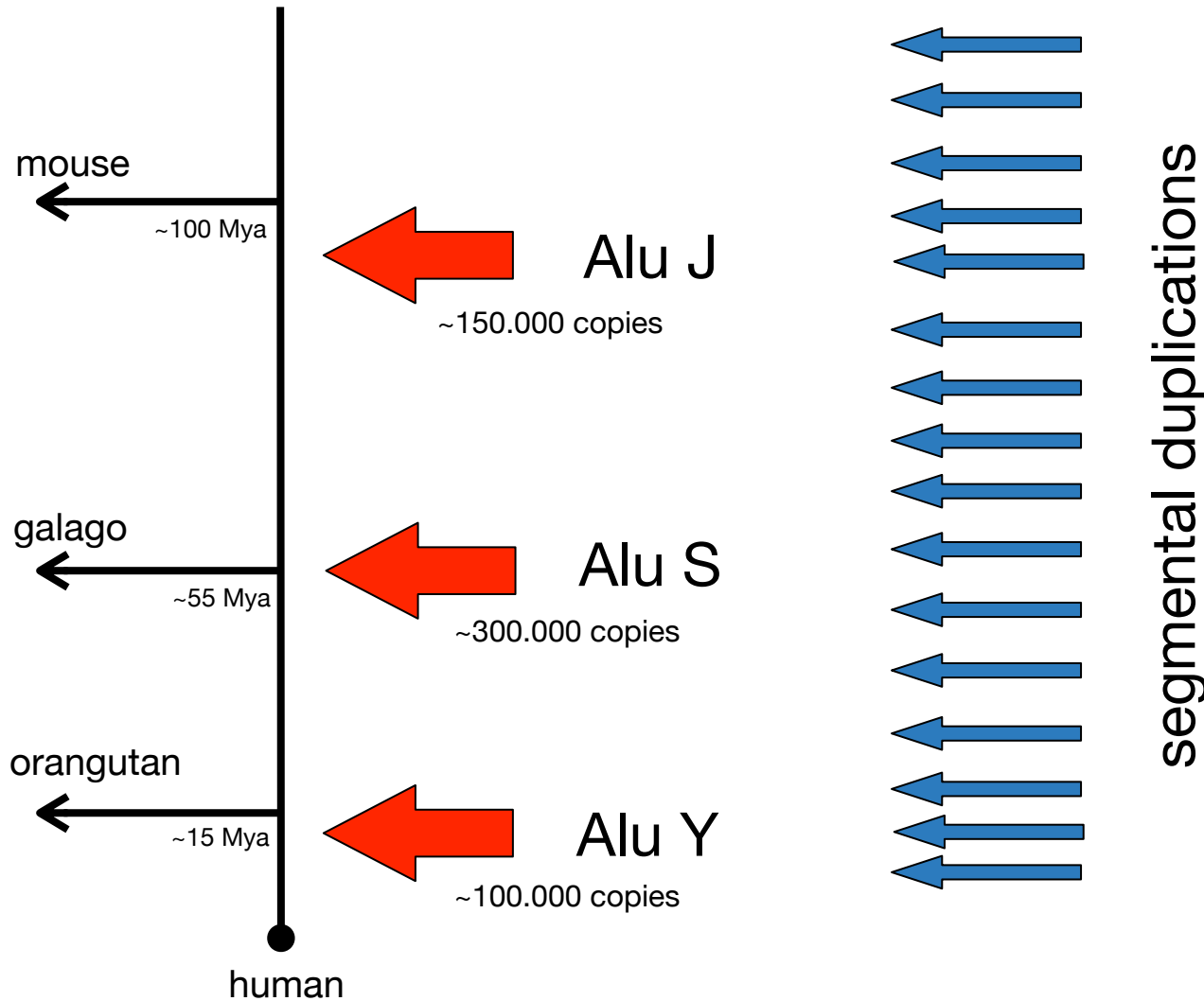


MAX-PLANCK-GESELLSCHAFT

[Massip & PA, PRL 2013]

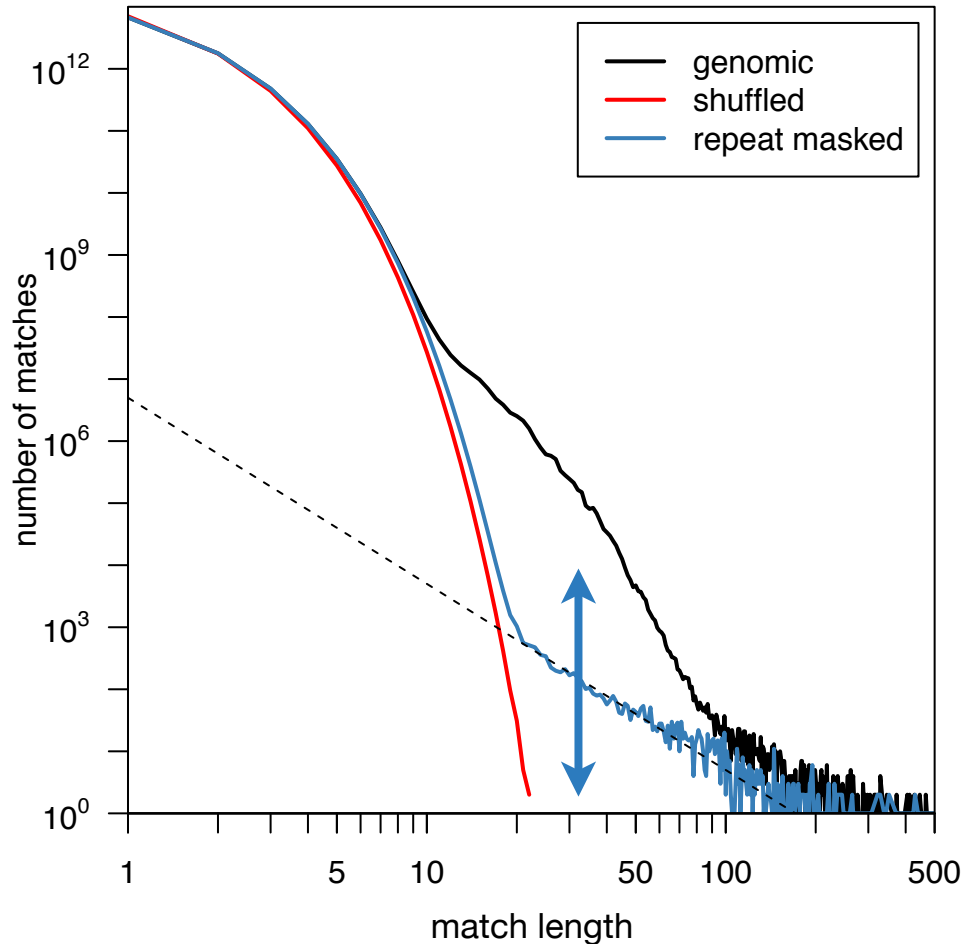
Segmental Duplications

- continuous generation of new stick due to SDs



[Bailey & Eichler, 2006]

Self-Alignment of the Human Genome



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

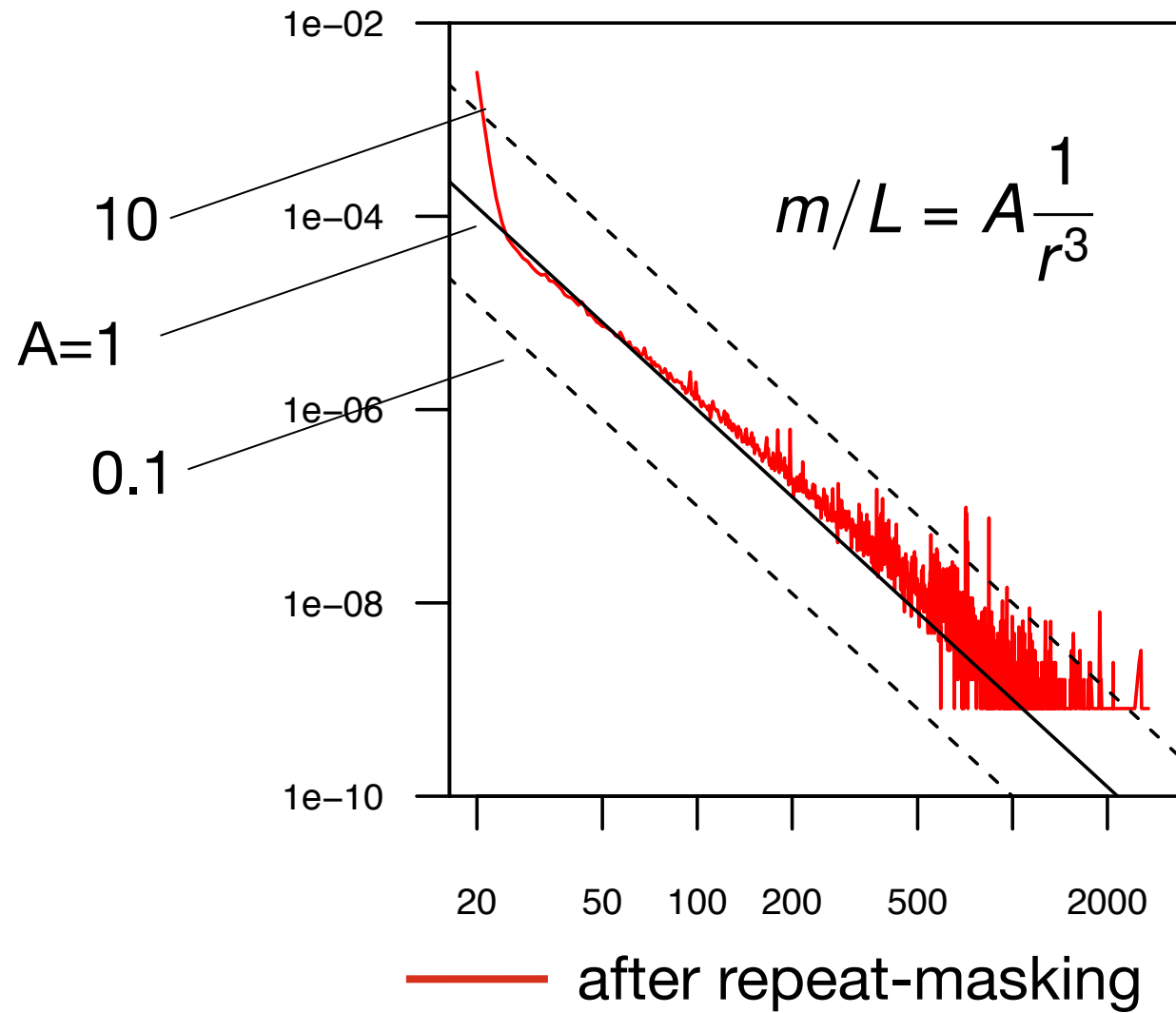
$$A = \frac{\gamma K}{\mu}$$

on average: as many back-up copies as mutations



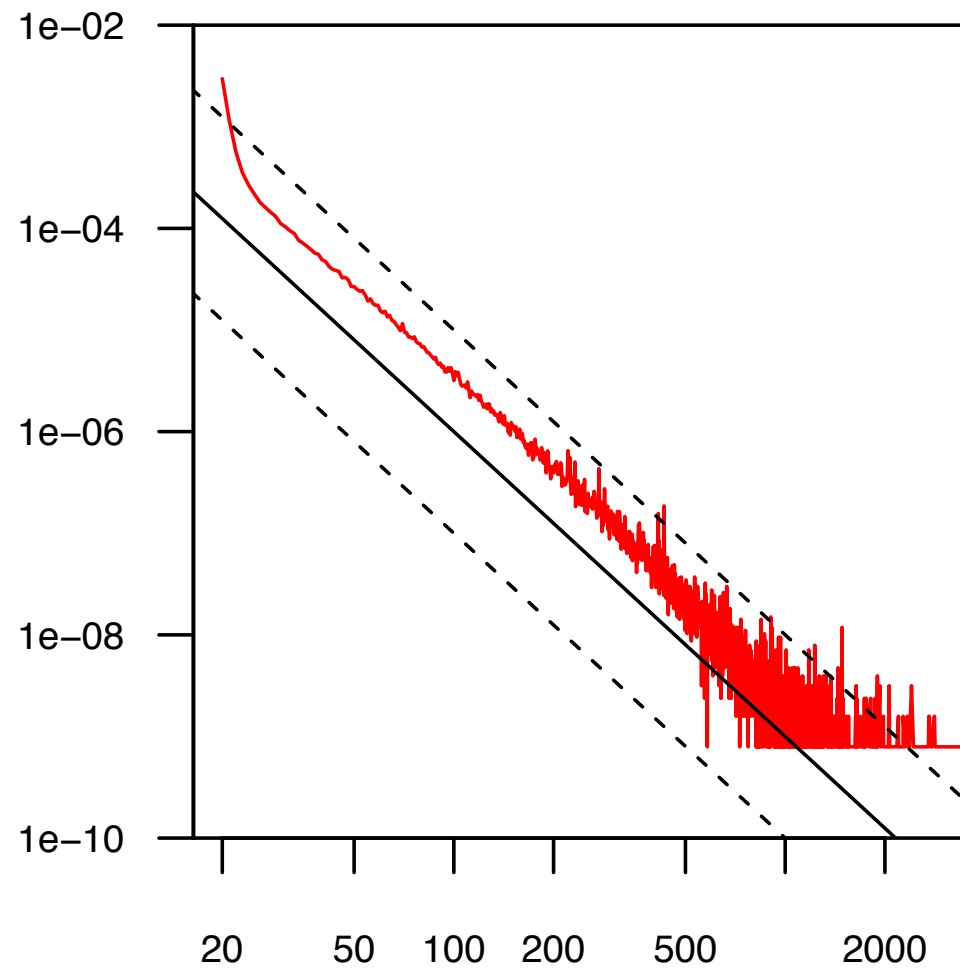
Human

homo_sapiens – all



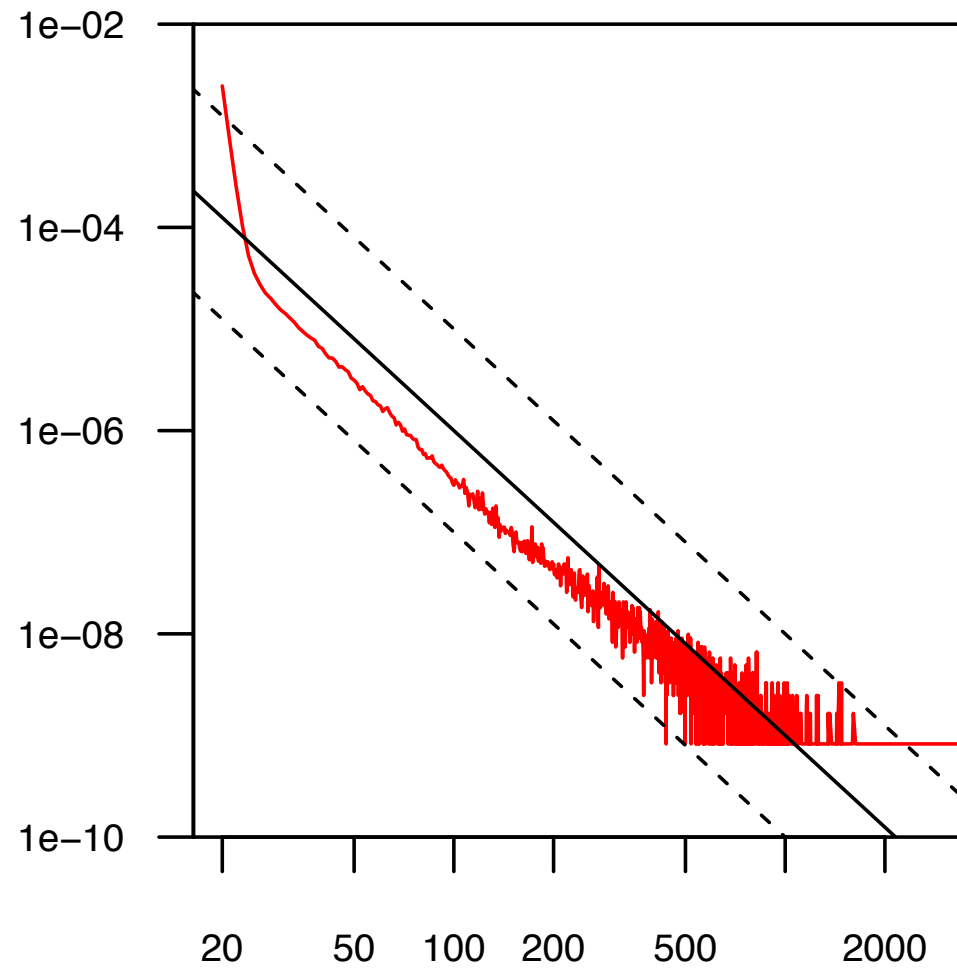
Mouse

mus_musculus – all



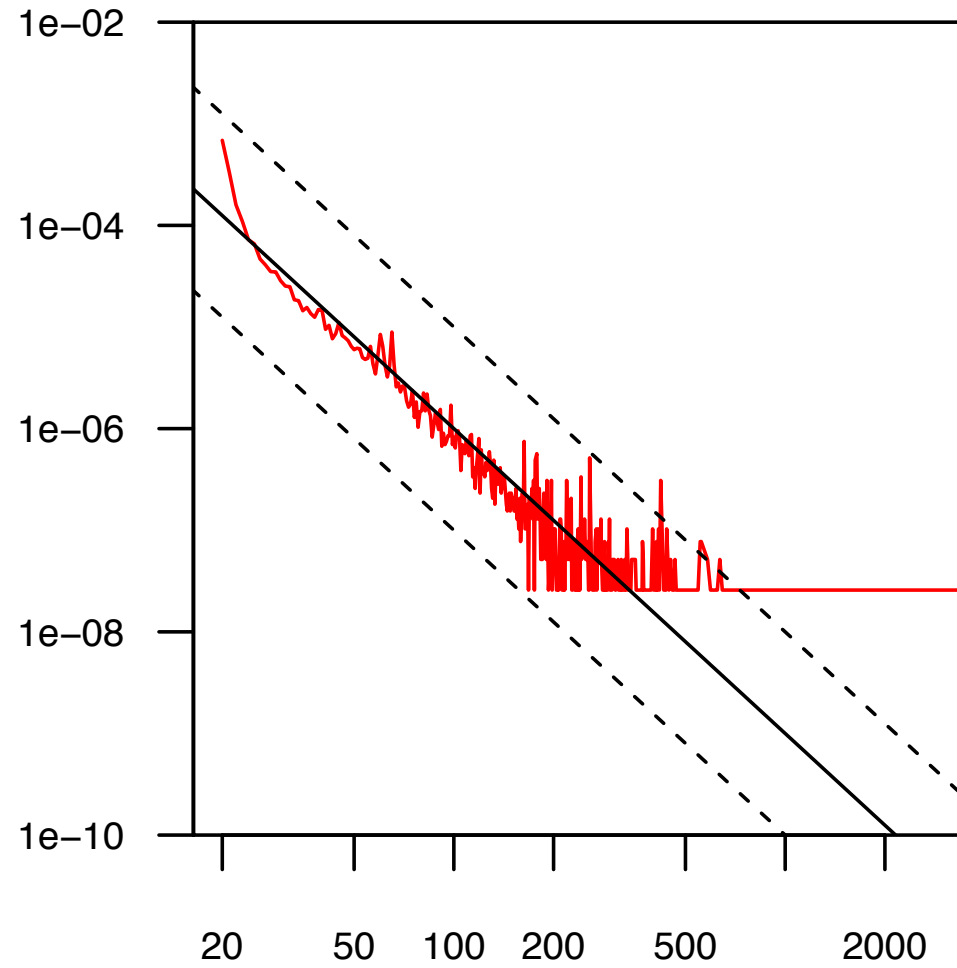
Dog

canis_familiaris – noScaffold



Worm

caenorhabditis_elegans – all



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

