

Probability distributions of GC content reflect the evolution of primate species

Marco V. José ^{1,2,*}, Qi Lu², Juan R. Bobadilla ^{1,2}

¹Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, México

²Centro Internacional de Ciencias A. C. Cuernavaca, Morelos, México

BIOMAT 2013

The Fields Institute for Research in Mathematical Sciences

Toronto, Canada, November 4-8, 2013

13th BIOMAT

International Symposium on Mathematical and Computational Biology

Background



LIFE WAS BORN AS A COMPLEX SYSTEM. YET NEITHER THE CLASIC THEORY OF EVOLUTION NOR THE GENETICAL THEORY OF NATURAL SELECTION (FISHER, HALDANE, WRIGHT) DEAL WITH COMPLEX SYSTEMS. THEY ARE CONCERNED WITH VARIABILITY WHICH NATURAL SELECTION ACTS ON. THE NEUTRAL THEORY OF MOLECULAR EVOLUTION (KIMURA) STATES THAT AT THE MOLECULAR LEVEL THE EVOLUTIONARY PROCESSES ARE ESSENTIALY RANDOM.

Background



Almost three decades ago, Yunis and Prakash (1982) demonstrated that comparison of Giemsa-banded karyotypes showed a very high degree of similarity between man, chimpanzee, gorilla, and orangutan. The G-banded late-prophase chromosomes of these four species showed an extensive homology.

Since Bernardi et al. [2], three major genomic fragments with low, median and high GC content were formally defined in the human genome and are now called **isochores**.

Many relationships between GC content and genomic properties have been unveiled [3-7]. These relationships demand that special attention should be paid to the evolution of GC content itself.

[1] Yunis JJ, Prakash O (1982). *Science* **215**, 1525-1530.

[2] Bernardi, G., et al. (1985). *Science*, **228**, 953–958.

[3] Duret L, Mouchiroud DG C (1995). *J. Mol. Evol.*, **40**, 308–317.

[4] Smith ZE, Higgs DR (1999). *Hum. Mol Genet*, **8**, 1373–1386.

[5] Lander ES, et al. (2001). *Nature*, **409**, 860–921.

[6] Khelifi, A. et al. (2006). *J Mol Evol*, **62**, 745–752.

[7] Duret L, Arndt PF (2008). *PLoS Genetics*, **4**(5) 1-19.

Background



Comparisons of DNA sequences between humans and the great apes showed that the African apes, especially the chimpanzees and the bonobos, but also the gorillas, are more closely related to humans than are the orangutans in Asia [8]. Thus, from a genetic standpoint, humans are essentially African apes. Our sense of uniqueness as a species was further shattered by the revelation that human DNA sequences differ by, on average, only 1.2% from those of the chimpanzees [9]. Humans and apes share a recent common ancestry.

[8] Miyamoto MM, et al. (1987). *Science* 238: 369-373.

[9] Chen FC, Li WH, (2001). *Am. J Hum. Genet.* 68: 444-456

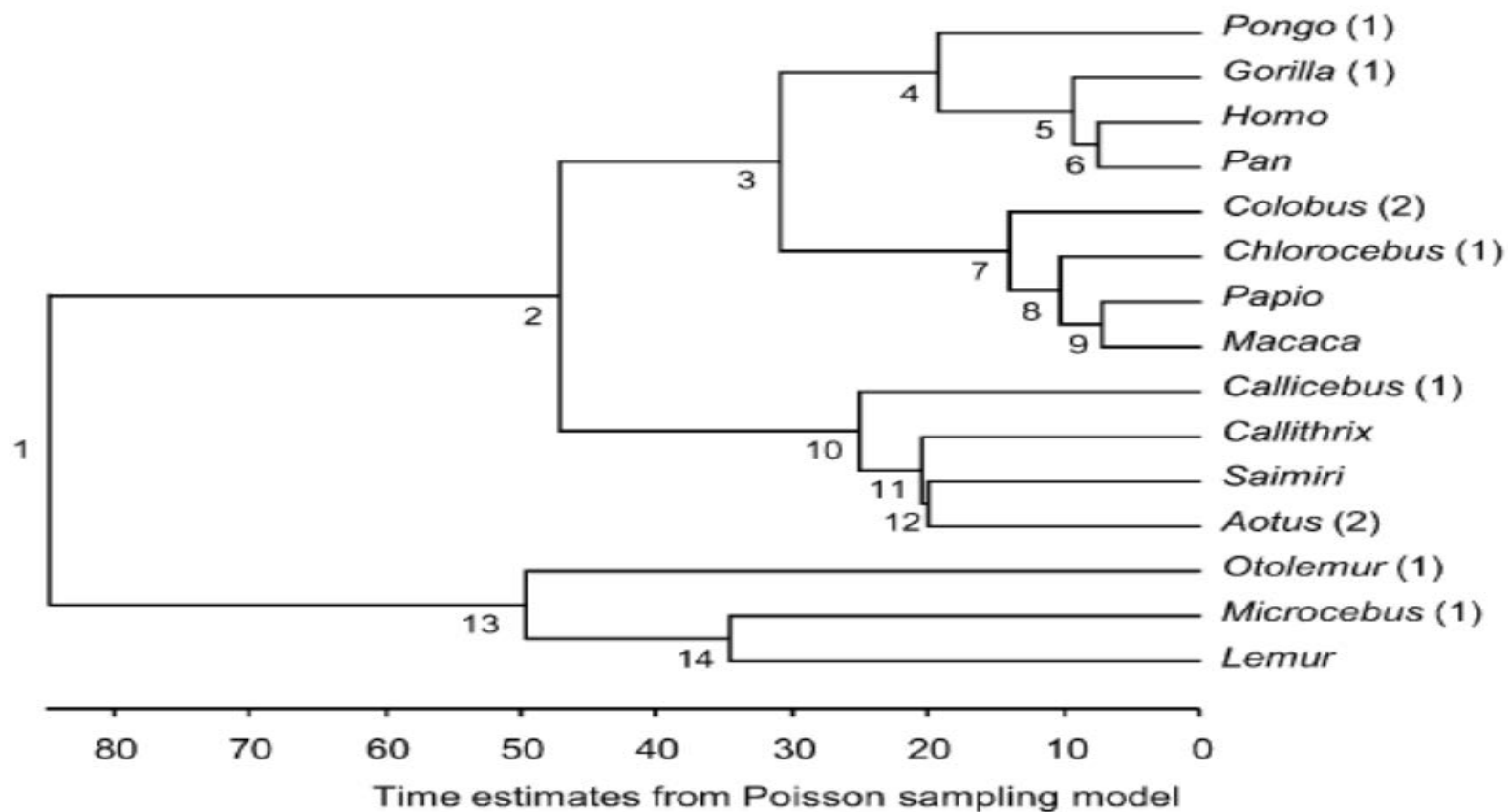



FIGURE 2. Rooted tree showing mean estimates of node ages t_1, \dots, t_{14} obtained when using the Poisson sampling model to provide prior distributions for the t_1 and t_2 node ages. The numbers in parentheses after the species name indicate that the sequence for that species is only available at that locus. Species without labels have data for both loci.



Explaining the evident biological traits that separate modern humans from our closest relatives, the chimpanzees, demands an explanation given the often cited 1.2-1.5% difference between orthologous nucleotide sequences [9,10]. This genetic distance is too small to account for their substantial differences. Regulatory changes [11], amino acid changes (e.g. [10, 12]) and the detection of gene gain and loss since the split of humans from chimpanzees indicate that humans differ from chimpanzees by at least ~6% [13]. Another counterintuitive finding is that there seems to be more genes that have undergone positive selection in chimpanzee evolution than in human evolution [14].

The history that different chromosomes experienced during evolution may not be the same [15, 16] and the structure of a given chromosome can be very different from others [17, 18]. In this context, the GC content evolution in a specific chromosome may be the result of several factors. Despite there are disputes about what leads to the variation of GC content in mammalian genomes, the overall consensus is that GC content in mammals is becoming homogenized [19-22].

Compared with investigations on GC content at large-scales ($10^3 \sim 10^4$ kb), the GC content at fine-scales ($10^0 \sim 10^1$ kb) comes into sight just recently [23, 24].

- [10] Mikkelsen TS, et al. (2005). *Nature* **437**: 69–87.
- [11] Hahn MW, et al. (2004). *Genetics* **167**: 867–877.
- [12] Bustamante CD, et al. (2005). *Nature* **437**: 1153–1157.
- [13] Demuth JP, et al. (2006). *PLoS ONE* **1(1)**: e85.
- [14] Bakewell MA, et al. (2007). *PNAS* **104(18)**: 7489-7494.
- [15] Patterson N, et al. (2006). *Nature* **441**: 1103–1108.
- [16] Hobolth A, et al. (2007). *PLoS Genetics* **3**: 0294-0304.
- [17] Skaletsky H, et al. (2003). *Nature* **423**, 825–837.
- [18] Rozen, S., et al. (2003). *Nature* **423**, 873–876.
- [19] Duret L, et al. (2002). *Genetics* **162**, 1837–1847.
- [20] Smith NG, et al. (2002). *Genome Res.*, **12**, 1350–1356.
- [21] Webster MT, et al. (2003). *Mol Biol Evol* **20**, 278–286.
- [22] Webster MT et al. (2004). *Trends Genet* **20**, 122–126.
- [23] Duret L, et al. (2008). *PLoS Genetics* **4(5)** 1-19.
- [24] Bullaughey KL, et al. (2008). *Genome Res.* pages gr.071548.107+.

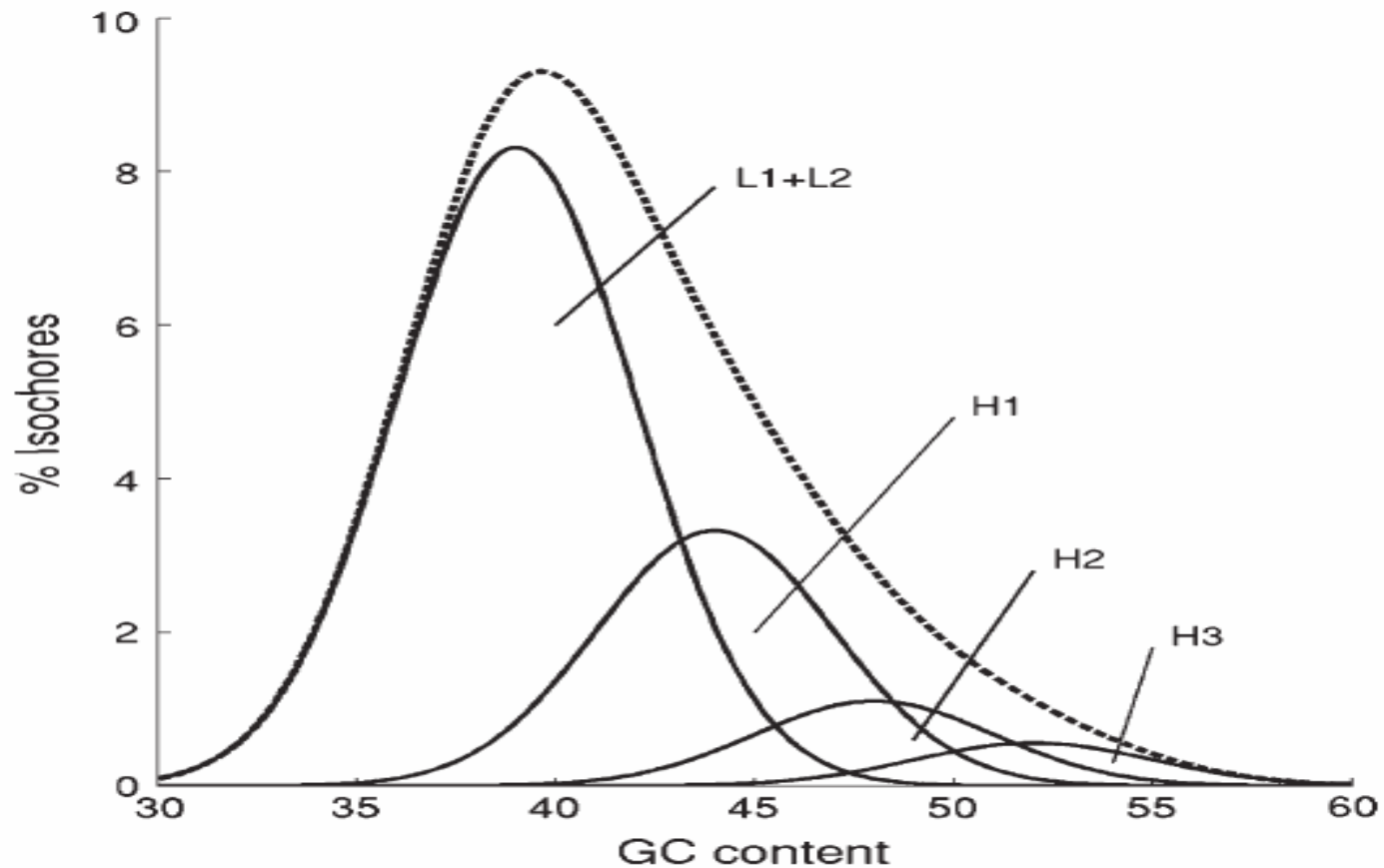


FIG. 1.—Illustration of the traditional five-Gaussian description of isochores families in the human genome. The Gaussians corresponding to the two GC-poor families (L1 and L2) are customarily merged into a single Gaussian. The superposition of the four remaining Gaussians is plotted in a dashed line. Modified from Pavlicek et al. (2002).

From: Cohen N., et al., *Mol. Biol. Evol.* (2005) **22**(5):1260–1272.



GOAL

To analyze the GC content of primate species via a novel indicator called *Local Average Distance of GC Dinucleotides (LADGC)*.

To model the distribution of *LADGC* assuming a stochastic process using the Black and Scholes model (Fokker-Planck equation).

METHODS



Negative correlations between LADGD and GC content

Instead of using the classic “windows strategy” [27, 28], which moves a fixed window size along the sequence and measures the GC content inside the window, we here propose to use the *LADGC*. First, we determine the distance series of the duplet GC along any chromosome. Then we fix the number N of GC duplets at which we will perform the analysis. Once we found N GC dinucleotides, we record the distance L_i (bps) of every two neighbor GC dinucleotides. Thus,

$$\text{LADGC} = \langle GC \rangle_{LD} = \frac{\sum_{i=1}^{N-1} L_i}{N} \quad (1)$$

According to (1), $\langle GC \rangle_{LD}$ is a measure of the local average distance (bps) of N GC dinucleotides. On the other hand, if the local length $\sum_{i=1}^{N-1} L_i$ is given, it is also straightforward to calculate the local GC content. We have performed this calculation at small and large-scale ($N = 2,000$) for human, chimpanzee and macaque.



MODEL ASSUMPTIONS

In order to investigate the GC content evolution in different species following the foregoing concept, we employed the probability density function (PDF) of *LADGC* for chromosomes of three primate species and found that the comparison of the PDFs among these species show very similar patterns and all PDFs can be approximately fitted by a log-normal distribution (Figures).

We use a normal distribution to approximate the histogram of real data, which means we neglect the effect of skewness and kurtosis, then, we can propose the following general stochastic equation to describe the evolutionary process:

$$dS = \mu(S, T)dt + \sigma(S, T)dW \quad (1)$$

where s represents the local average distance of GC. The idea behind equation (1) is that the evolutionary process is a random process and it can be modeled by two factors, one is the mean μ , which is the factor related to the trend, and the other is the variance σ , which means noise.

the simplest assumption we can choose is that three species had different evolution speed. Simplifying equation (1) with our assumption, we get,

$$dS = V_S(t)tdt + \sigma dW \quad (2)$$

We then do the exponential transform to get the real LADGC(Y), apply Ito's law to $Y = \exp(S)$, and we get,

$$\frac{dY}{Y} = (V_S(t)t + \frac{\sigma^2}{2})dt + \sigma dW \quad (3)$$

Stochastic behavior of distance series in both human and chimpanzee genomes: the use of Fokker-Planck equation

We propose the **Fokker-Planck equation (Chapman-Kolmogorov forward equation)** as a candidate for describing human-genome's "hierarchy diffusion". Alternatively, we suggest the Chapman-Kolmogorov backward equation as a candidate for describing chimpanzee's genome diffusion. The forward and backward equations are, respectively,

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial D_1(x,t)p(x,t)}{\partial x} + \frac{\partial^2 D_2(x,t)p(x,t)}{\partial x^2}$$

and

$$\frac{\partial p(x,t)}{\partial t} = -D_1(x,t)\frac{\partial p(x,t)}{\partial x} + D_2(x,t)\frac{\partial^2 p(x,t)}{\partial x^2} \quad [1]$$

The method for checking the Fokker-Planck equation

The Fokker-Planck equation is equivalent to the Chapman-Kolmogorov equation. For simplicity, we use $Y_i = Y(S_i)$ to quote the different value of Y when choose the different values of S .

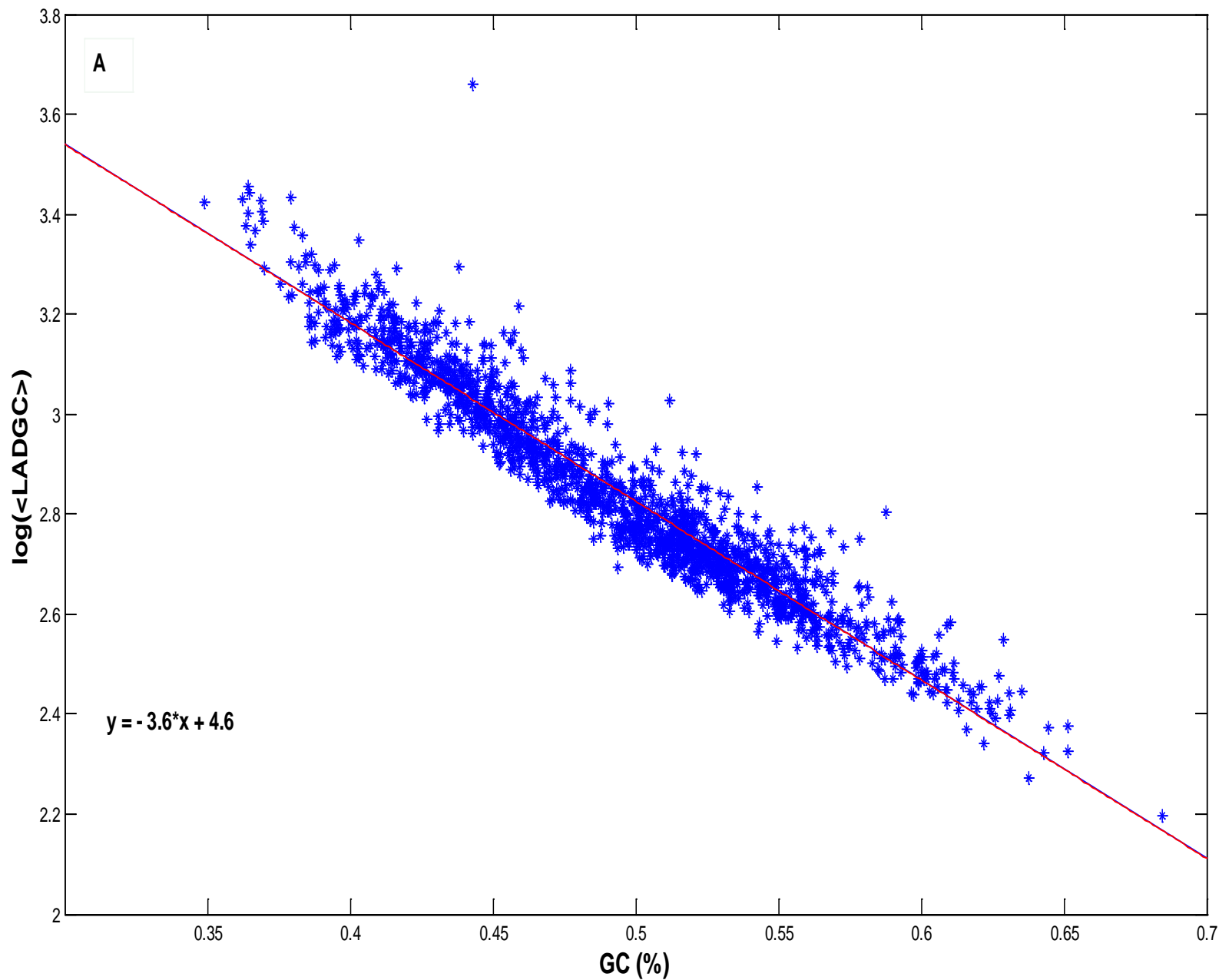
$$p(Y_0, S_0 | Y_2, S_2) = \int p(Y_0, S_0 | Y_1, S_1) p(Y_1, S_1 | Y_2, S_2) dY_1 \quad [2]$$

This equation makes it possible to check that the Fokker-Planck equation actually satisfies the human case. All the work we have to do is just to estimate the transfer probability (conditional probability) with fixed initial condition:

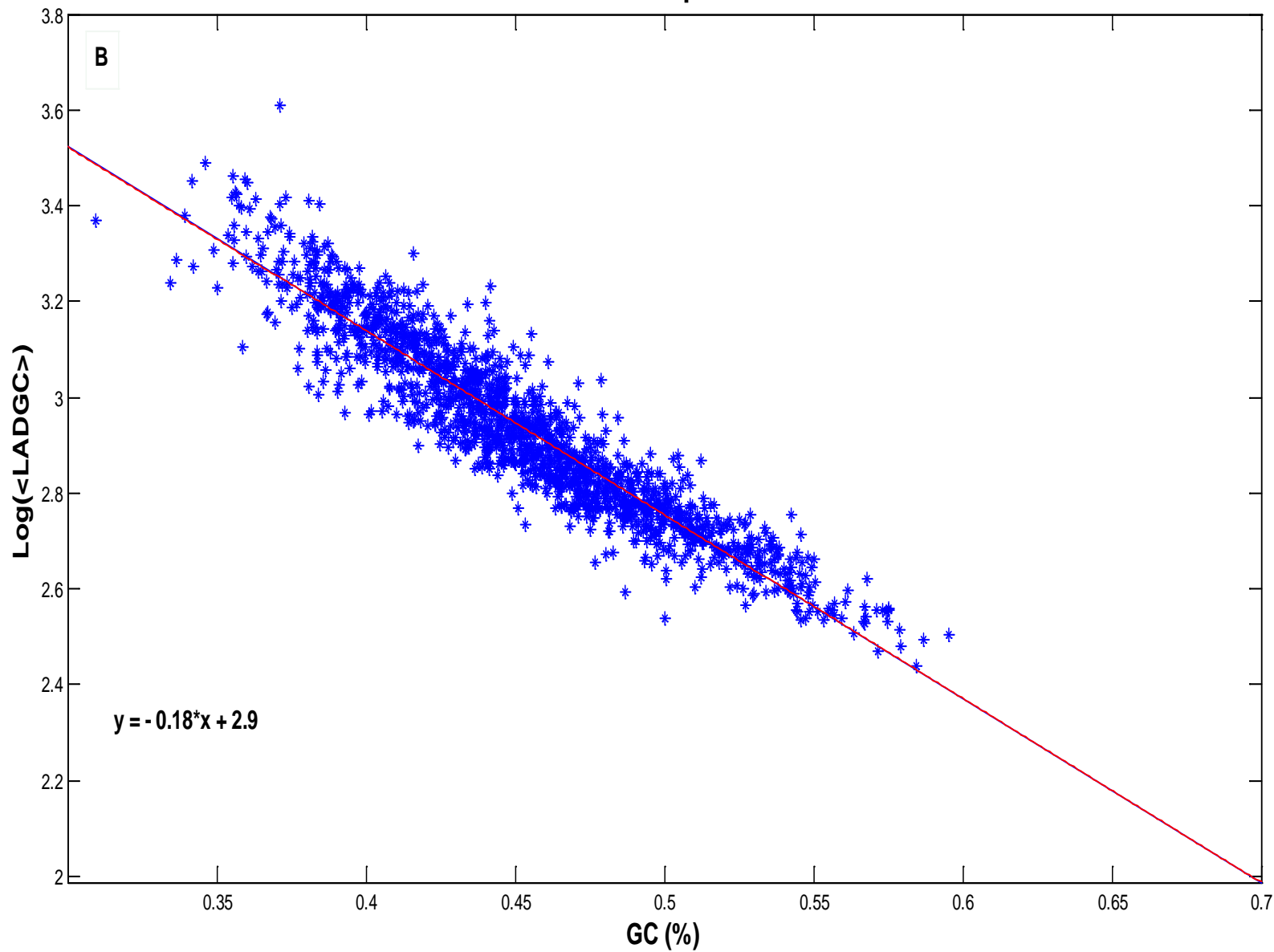
$$p(Y_f, S_f | Y_i, S_i) = \frac{p(Y_f, S_f; Y_i, S_i)}{p(Y_i, S_i)} \quad [4]$$

to create the estimator. Here, $p(Y_f, S_f; Y_i, S_i)$ is the joint pdf, i denotes the initial value whereas f stands for the final value.

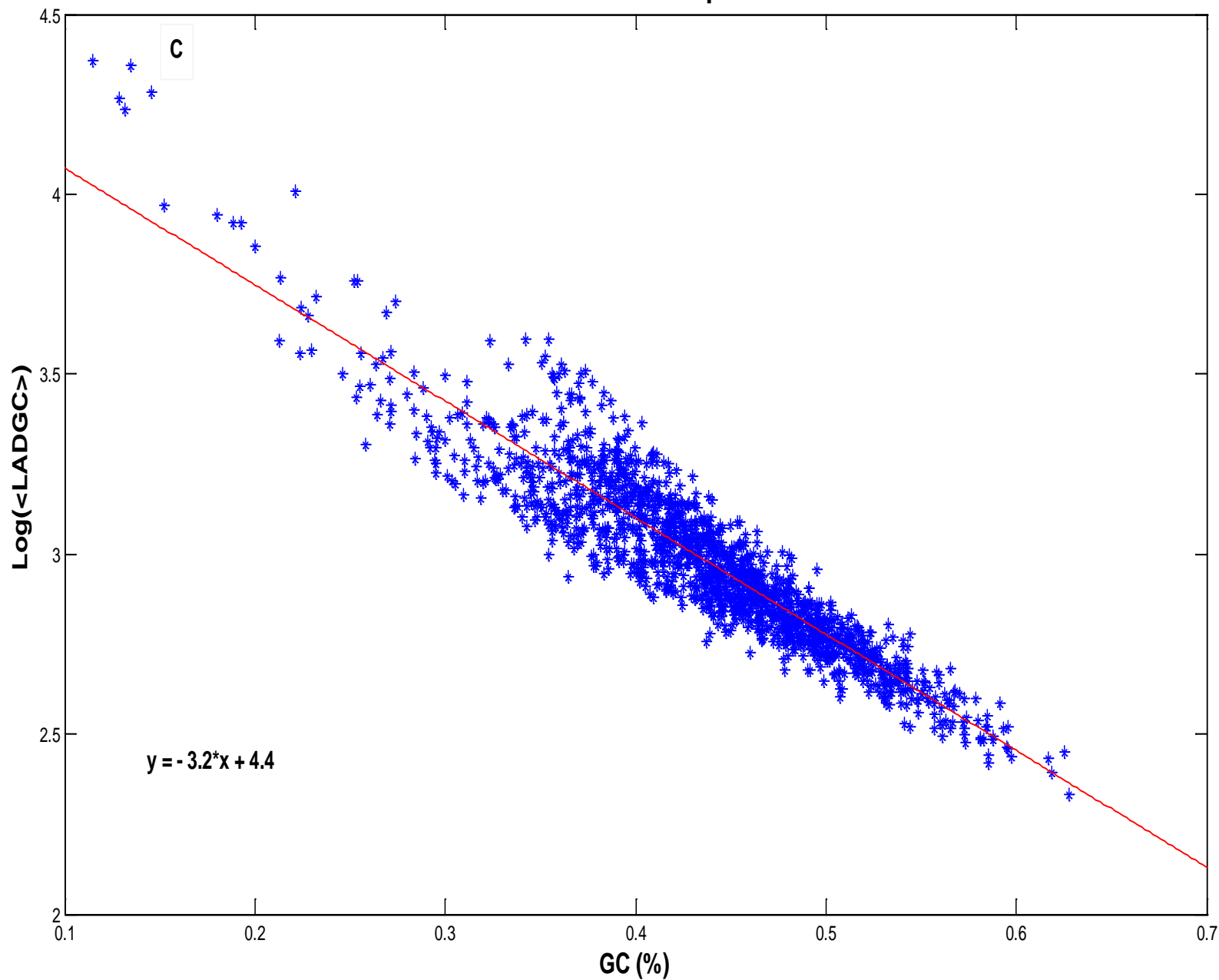
LADGC versus % GC Human chromosome 19

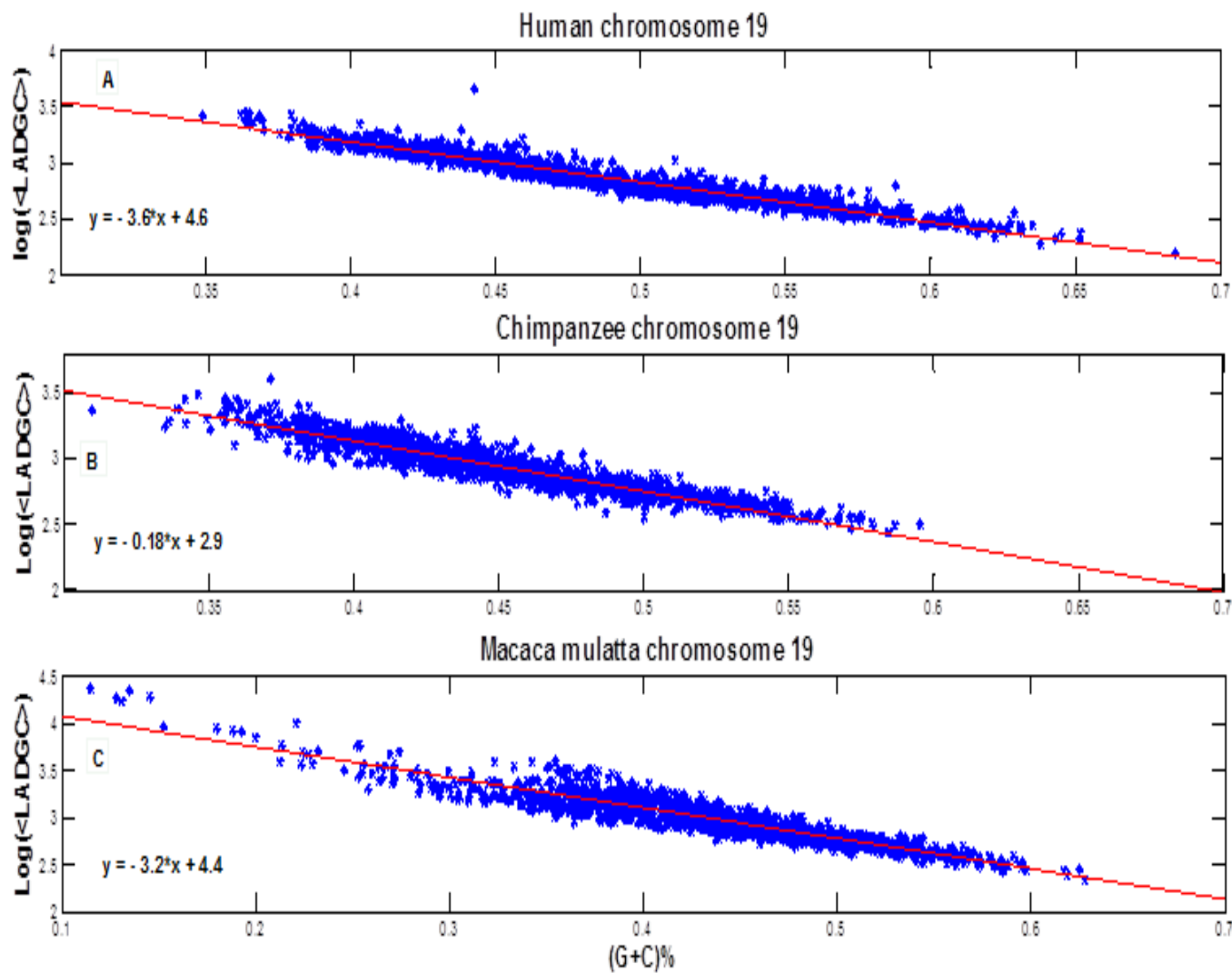


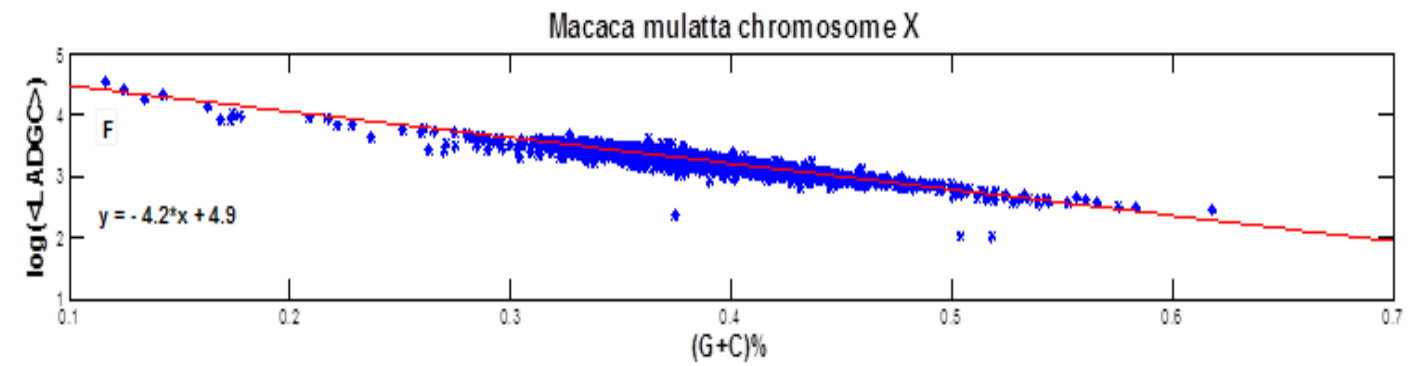
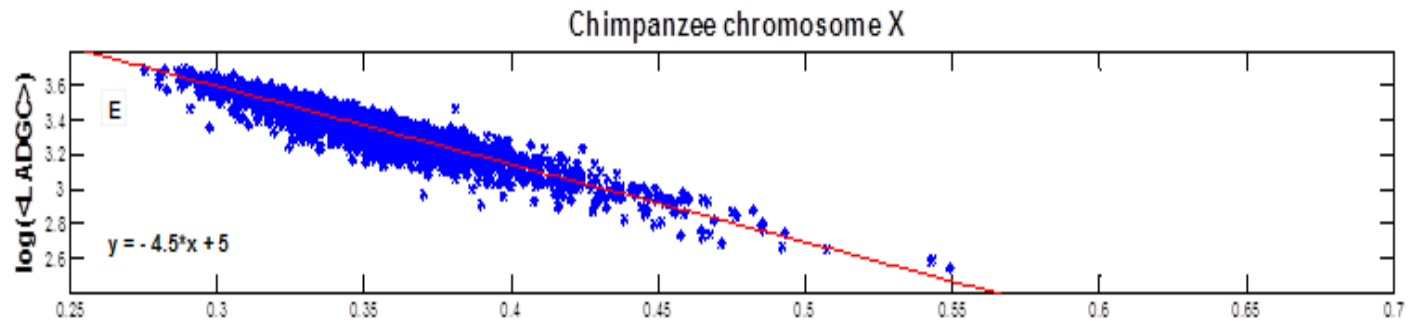
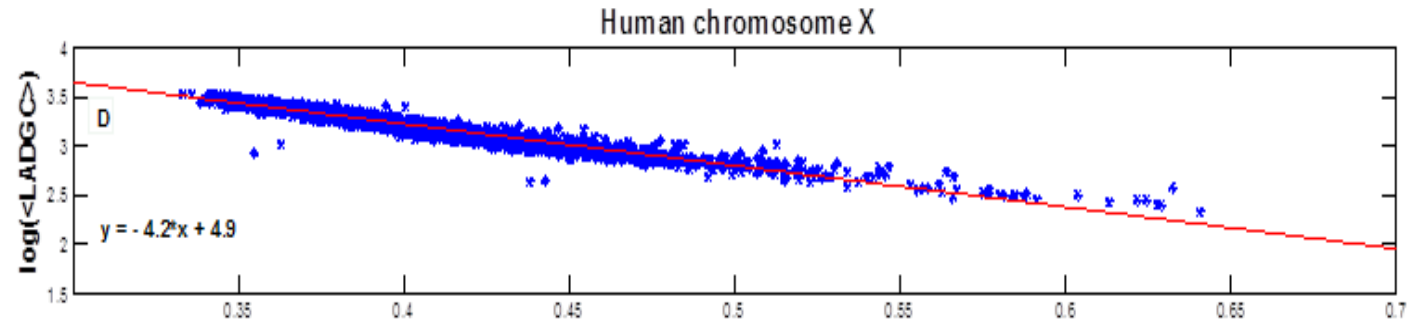
LADGC versus % GC Chimpanzee chromosome 19



LADGC vs % GC Rhesus macaque chromosome 19

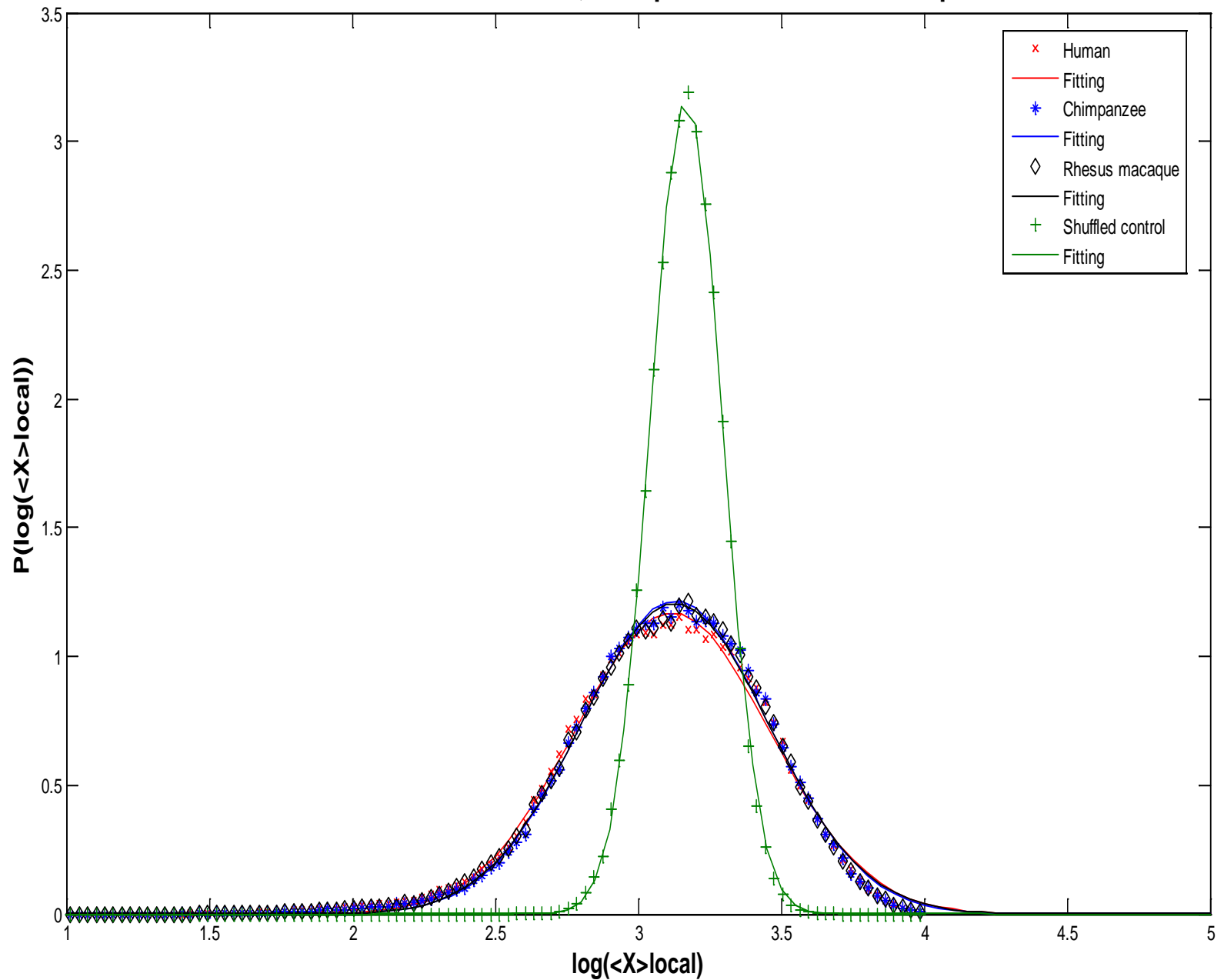




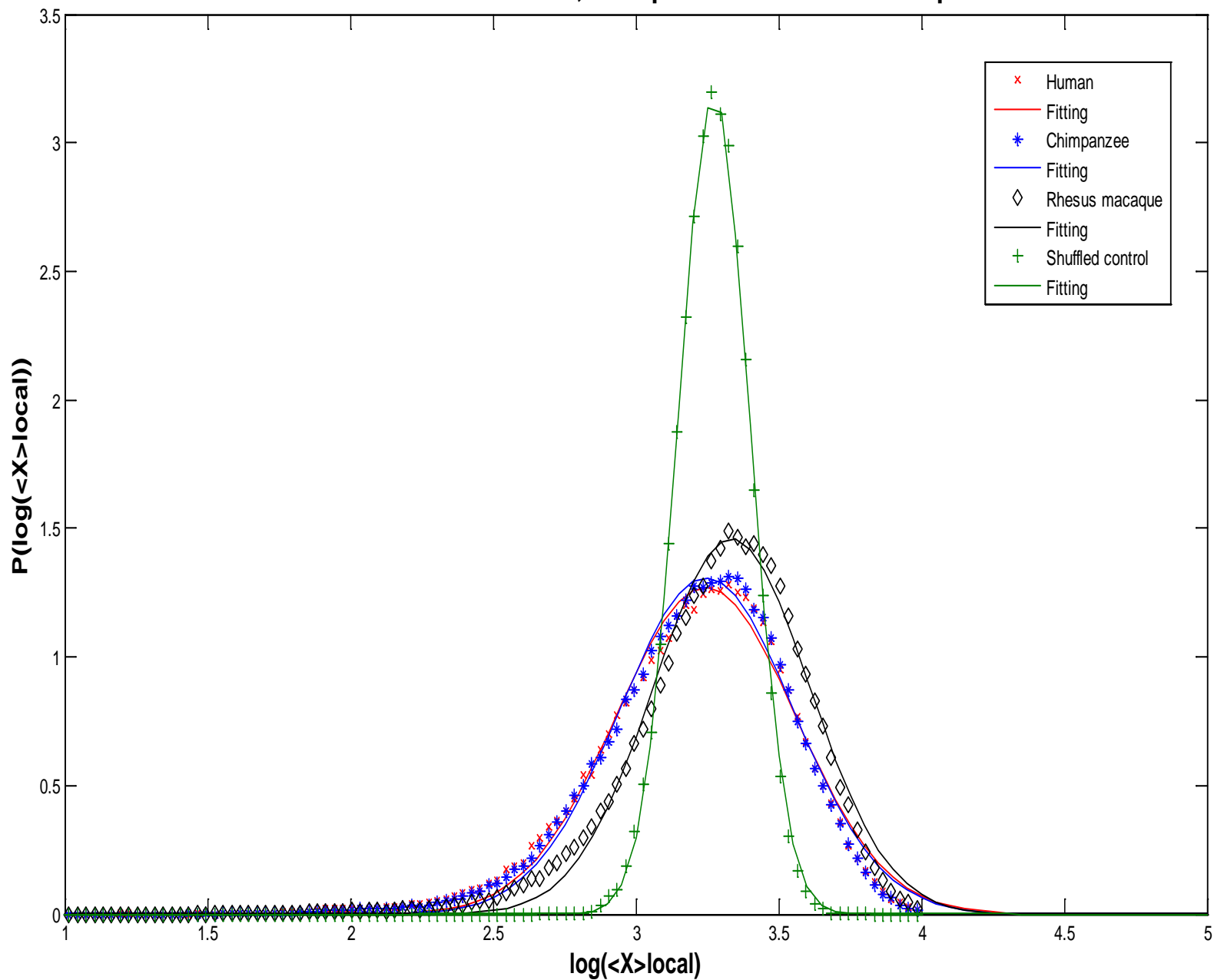


Chromosome	Human	Chimpanzee	Rhesus macaque
1	0.417565491	0.41590226	0.41851042
2	0.402448476	a 0.40852114 b 0.39462015	0.3948904
3	0.396901712	0.39589135	0.40590444
4	0.382399205	0.38198768	0.39588178
5	0.395205245	0.39420969	0.37881057
6	0.396202798	0.39465033	0.39476164
7	0.407569212	0.40579139	0.41413379
8	0.40170985	0.40002187	0.40039024
9	0.413059146	0.41214013	0.4164439
10	0.415860442	0.41417815	0.45544863
11	0.415761285	0.4141031	0.40938363
12	0.408039024	0.40673039	0.39255893
13	0.385315555	0.38454186	0.41032843
14	0.408871807	0.4078842	0.41628924
15	0.422396851	0.42142682	0.41455268
16	0.447942254	0.44540159	0.4556974
17	0.455847751	0.45282653	0.38492053
18	0.397850229	0.3968446	0.39872069
19	0.483413743	0.47926593	0.48233426
20	0.44125543	0.43933282	0.44715777
21	0.408443278	0.41006246	
22	0.479414783	0.47863044	
X	0.395161295	0.39181972	0.39252717
Y	0.39965144	0.40089052	
Mean	0.4158	0.4139	0.4133
Standard deviation	0.027	0.0261	0.0263

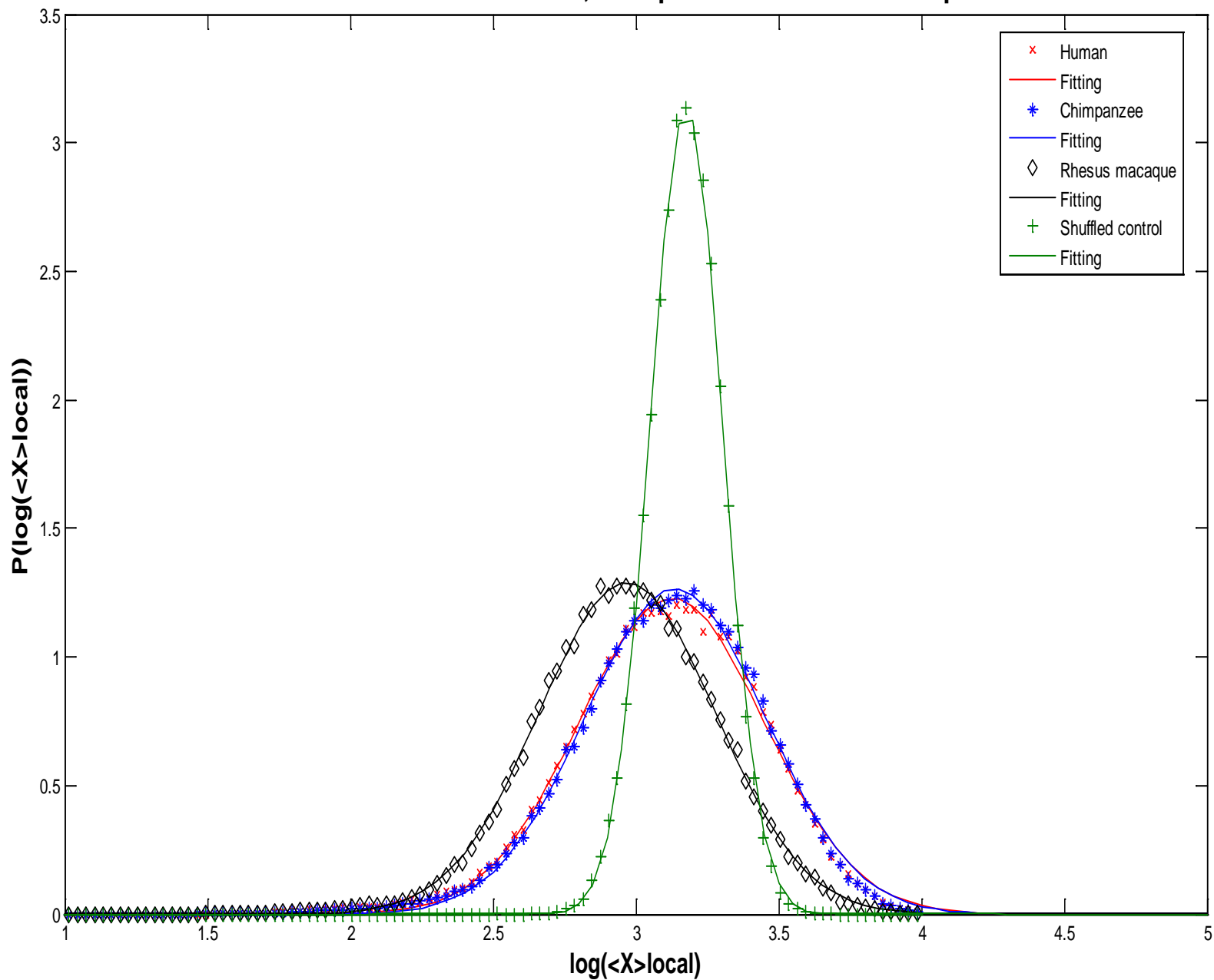
Chromosome 1: Human, Chimpanzee & Rhesus macaque



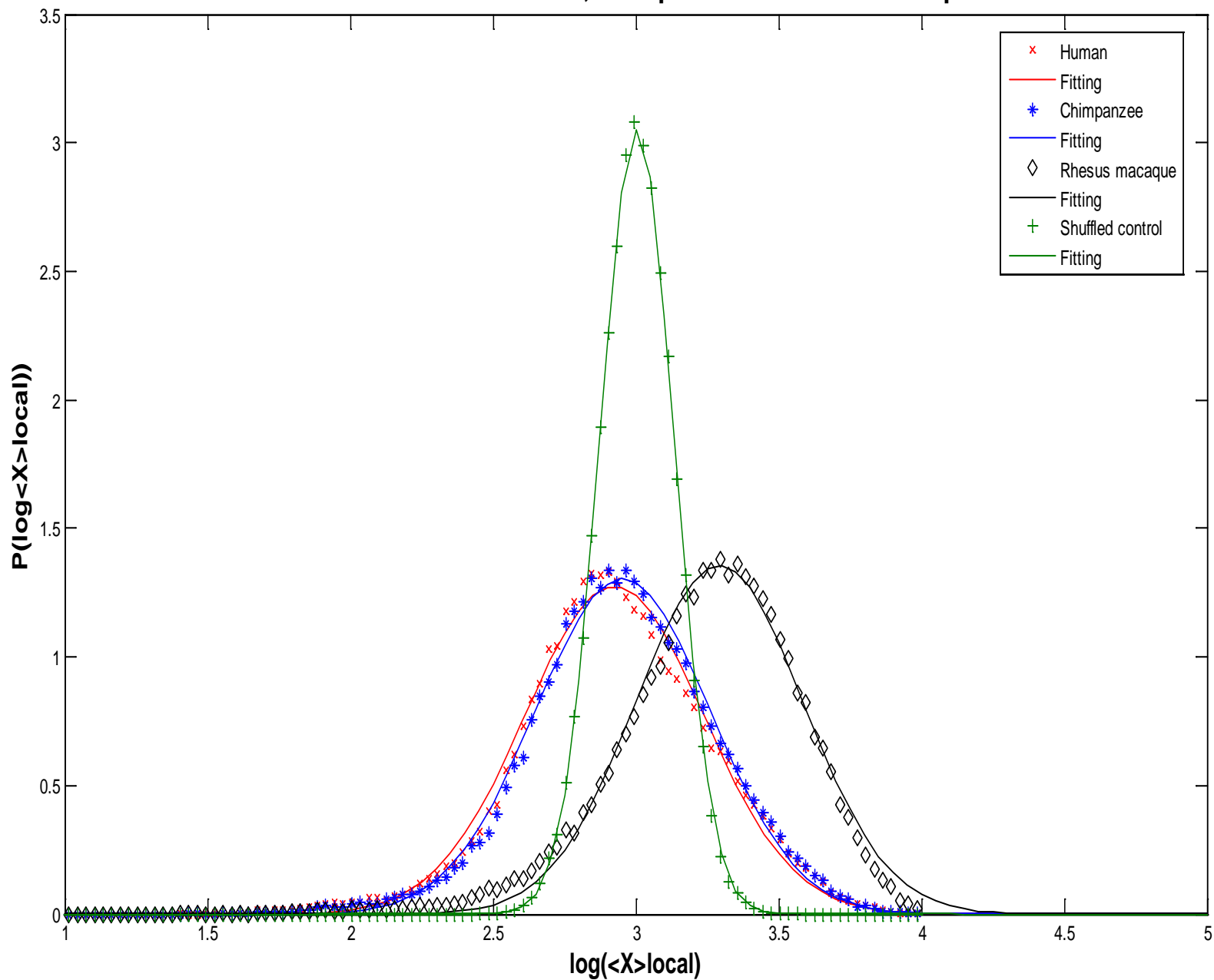
Chromosome 5: Human, Chimpanzee & Rhesus macaque



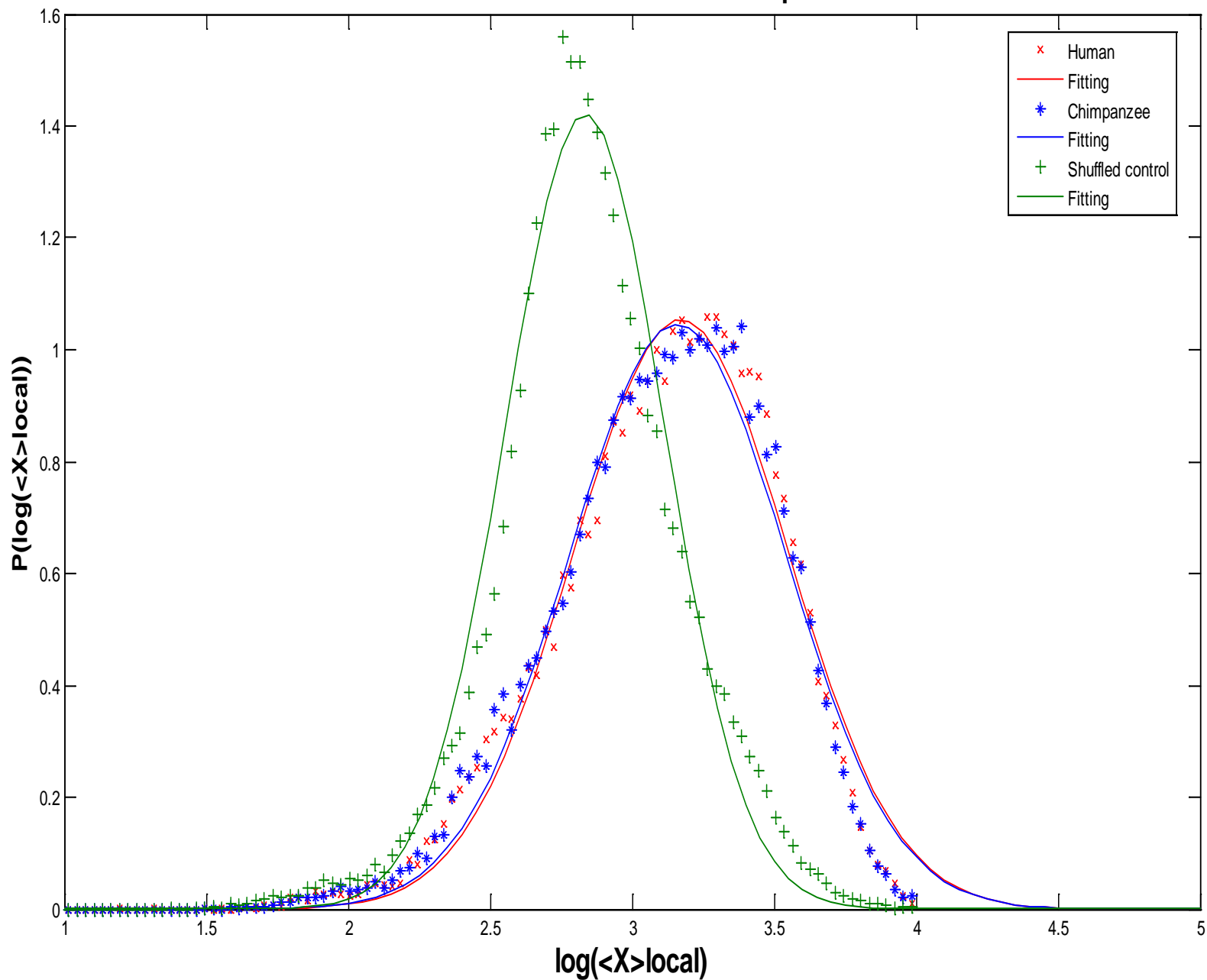
Chromosome 10: Human, Chimpanzee & Rhesus macaque



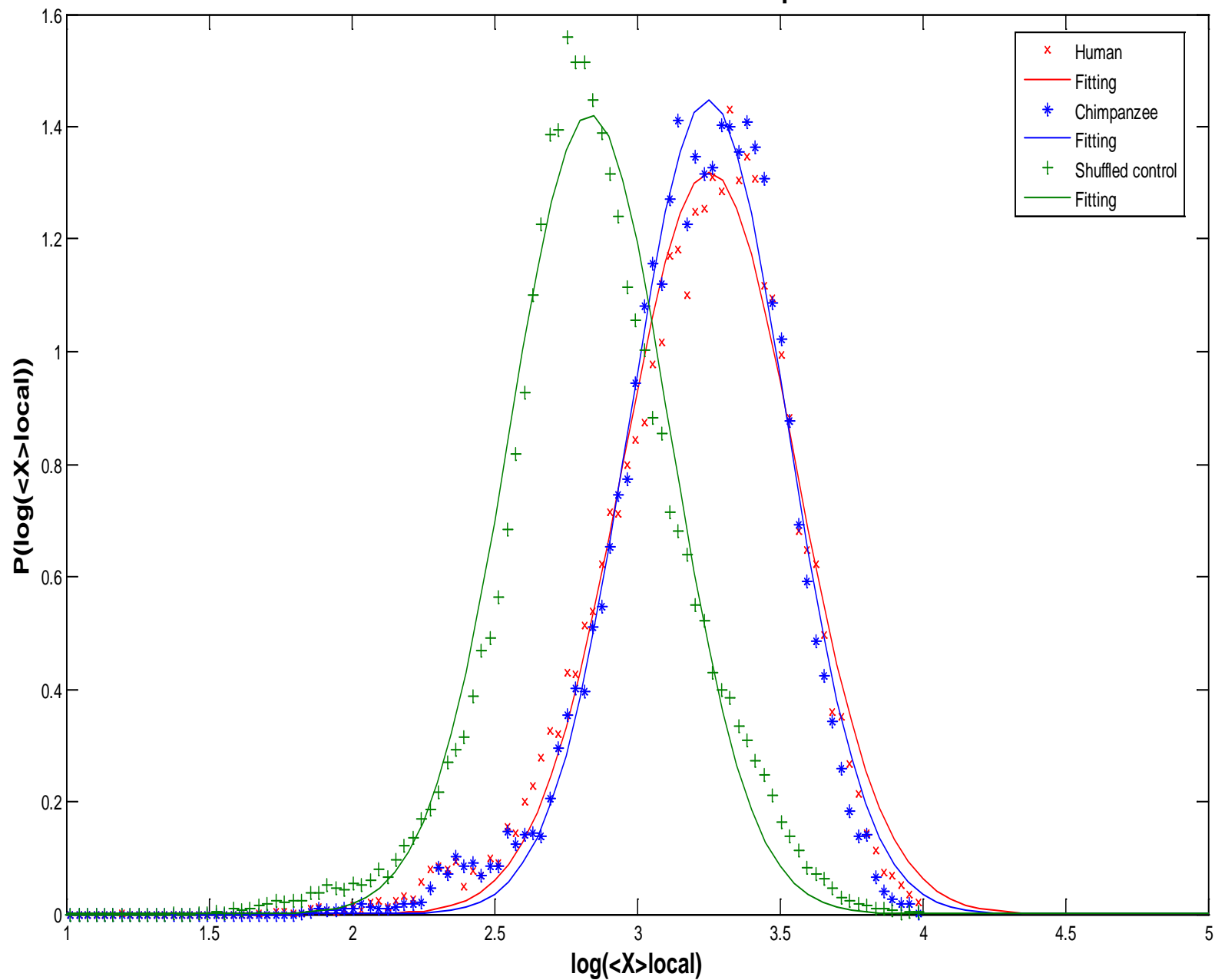
Chromosome 17: Human, Chimpanzee & Rhesus macaque

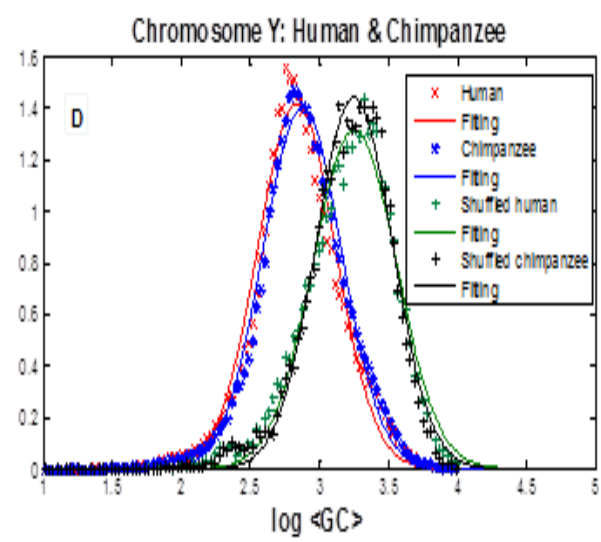
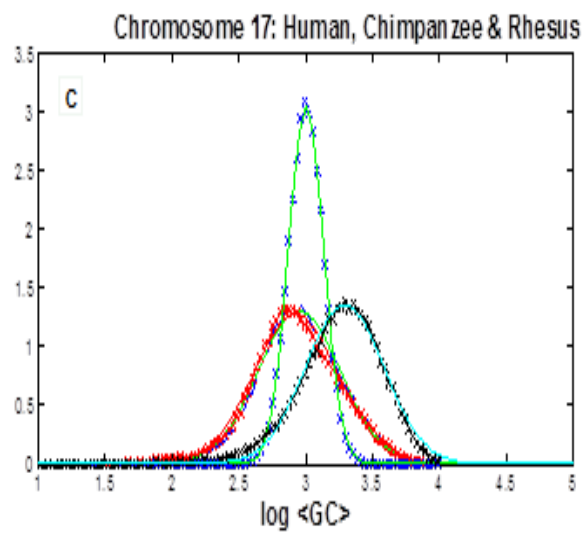
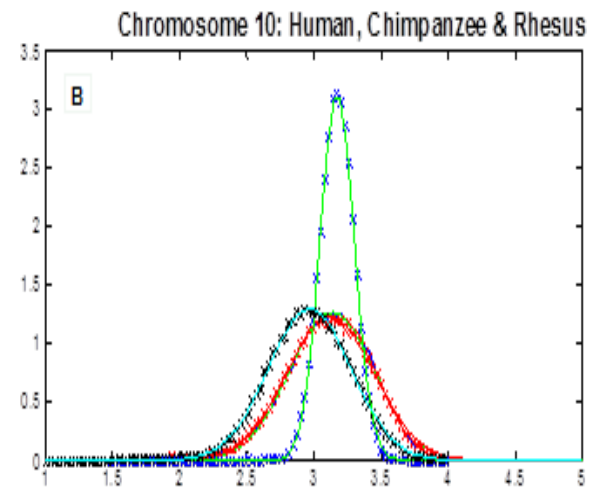
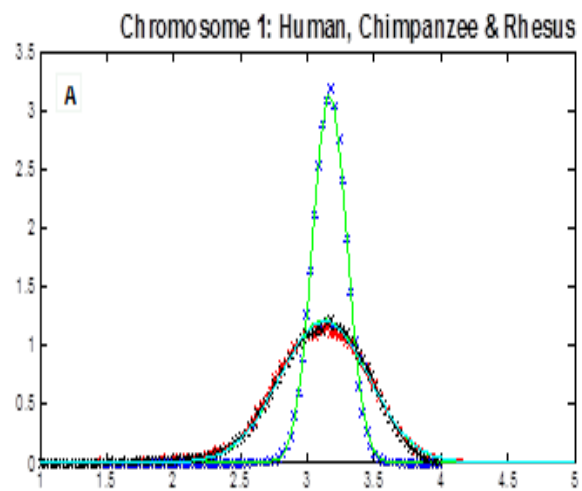


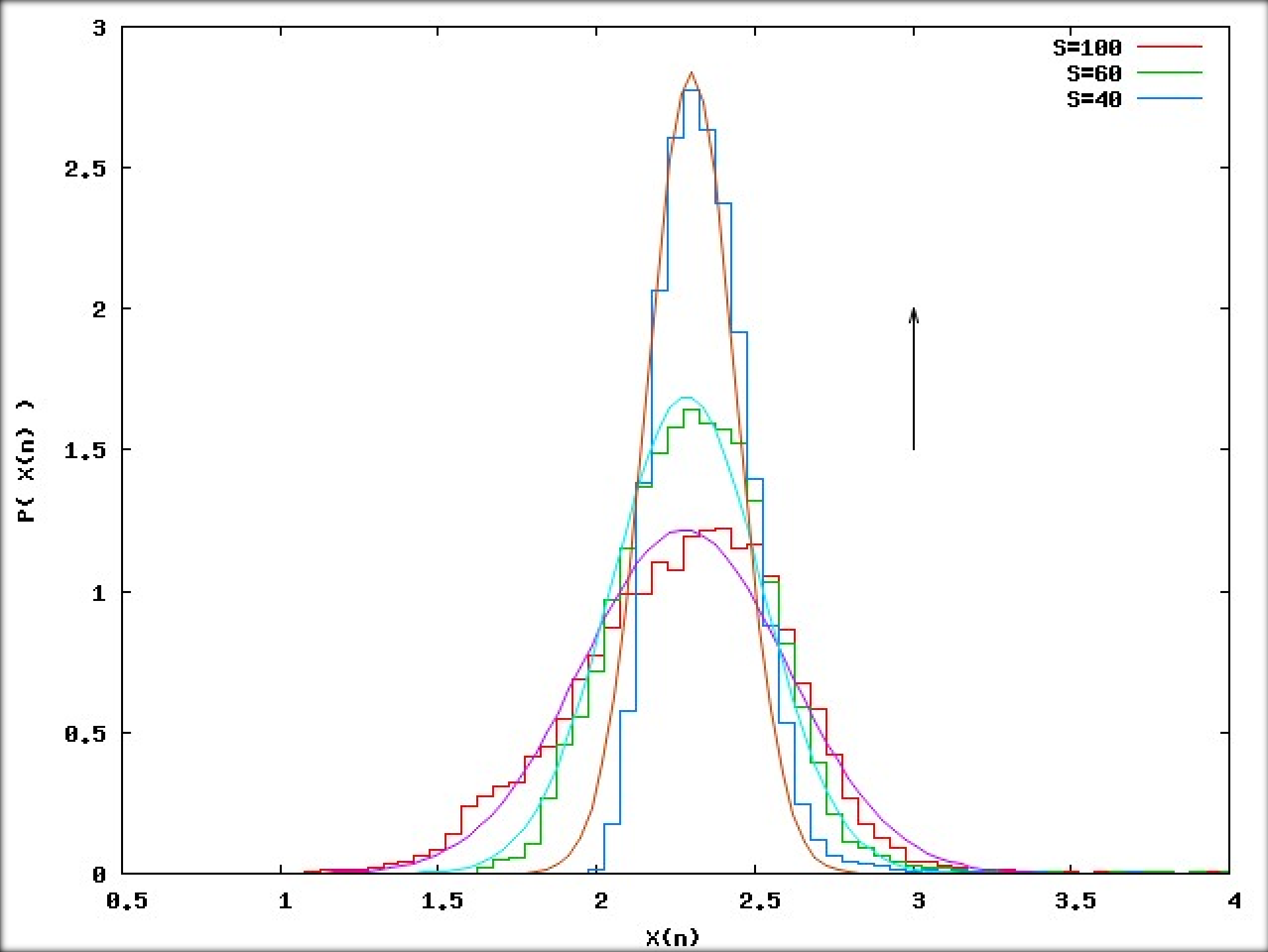
Chromosome 21: Human & Chimpanzee

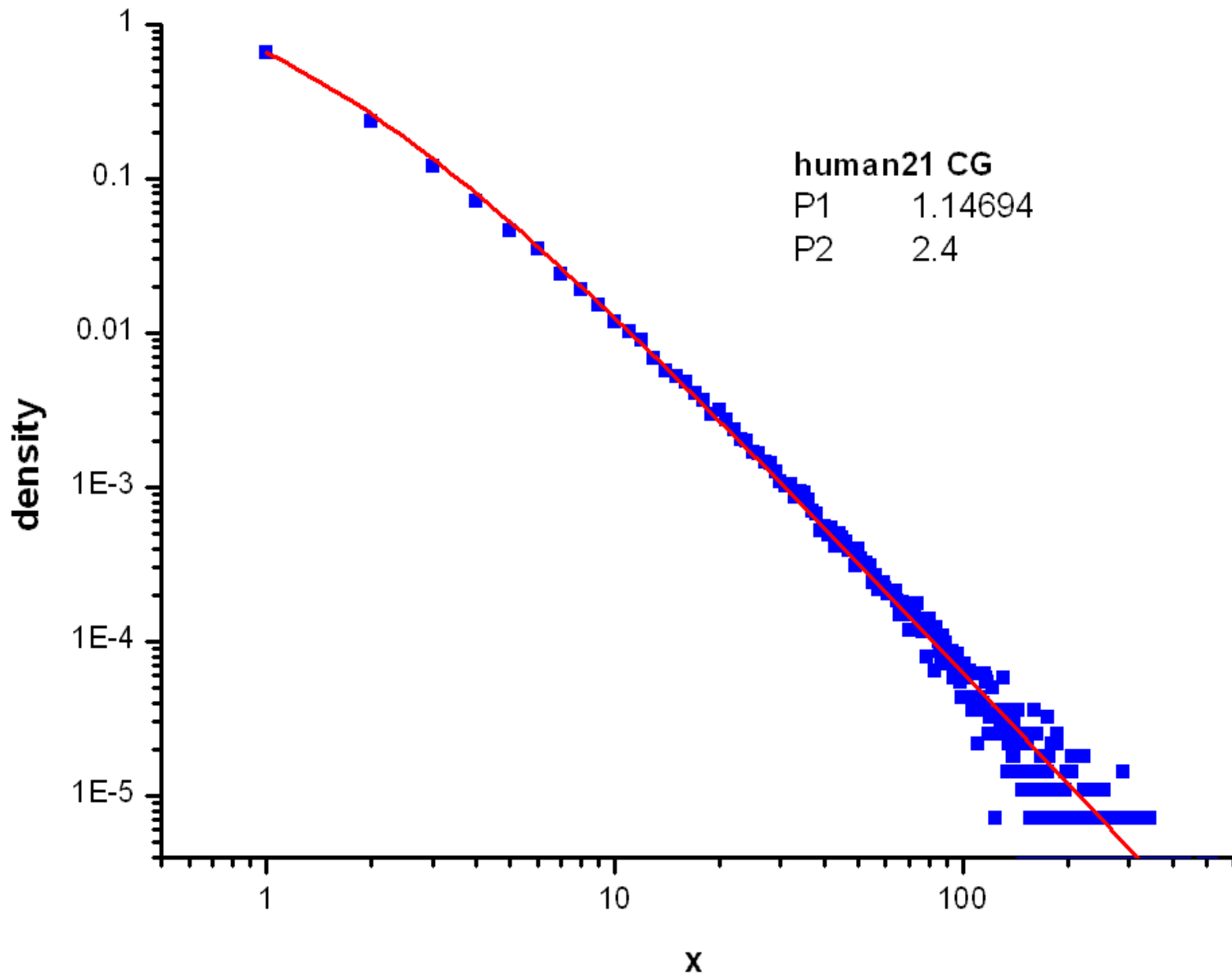


Chromosome Y: Human & Chimpanzee

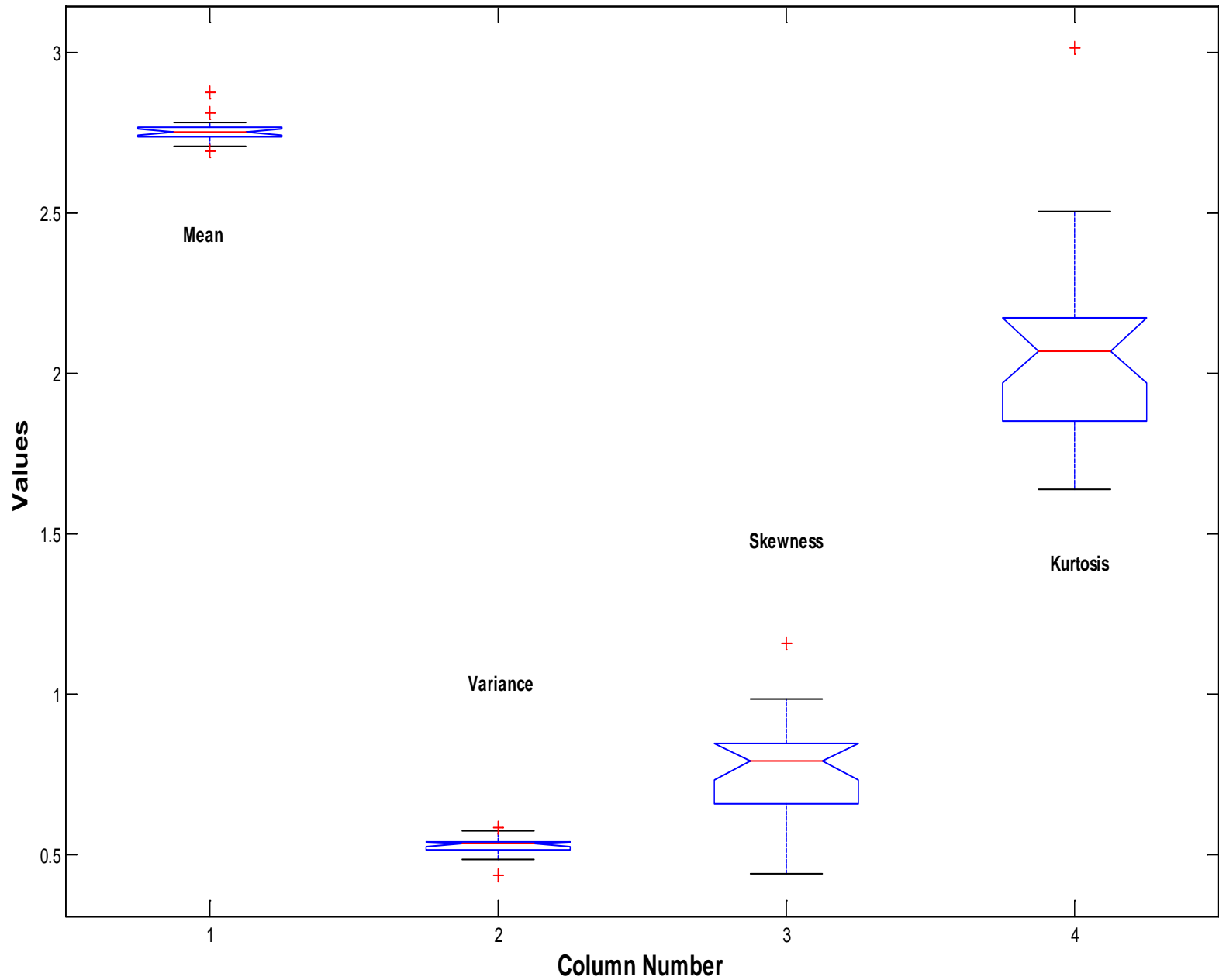




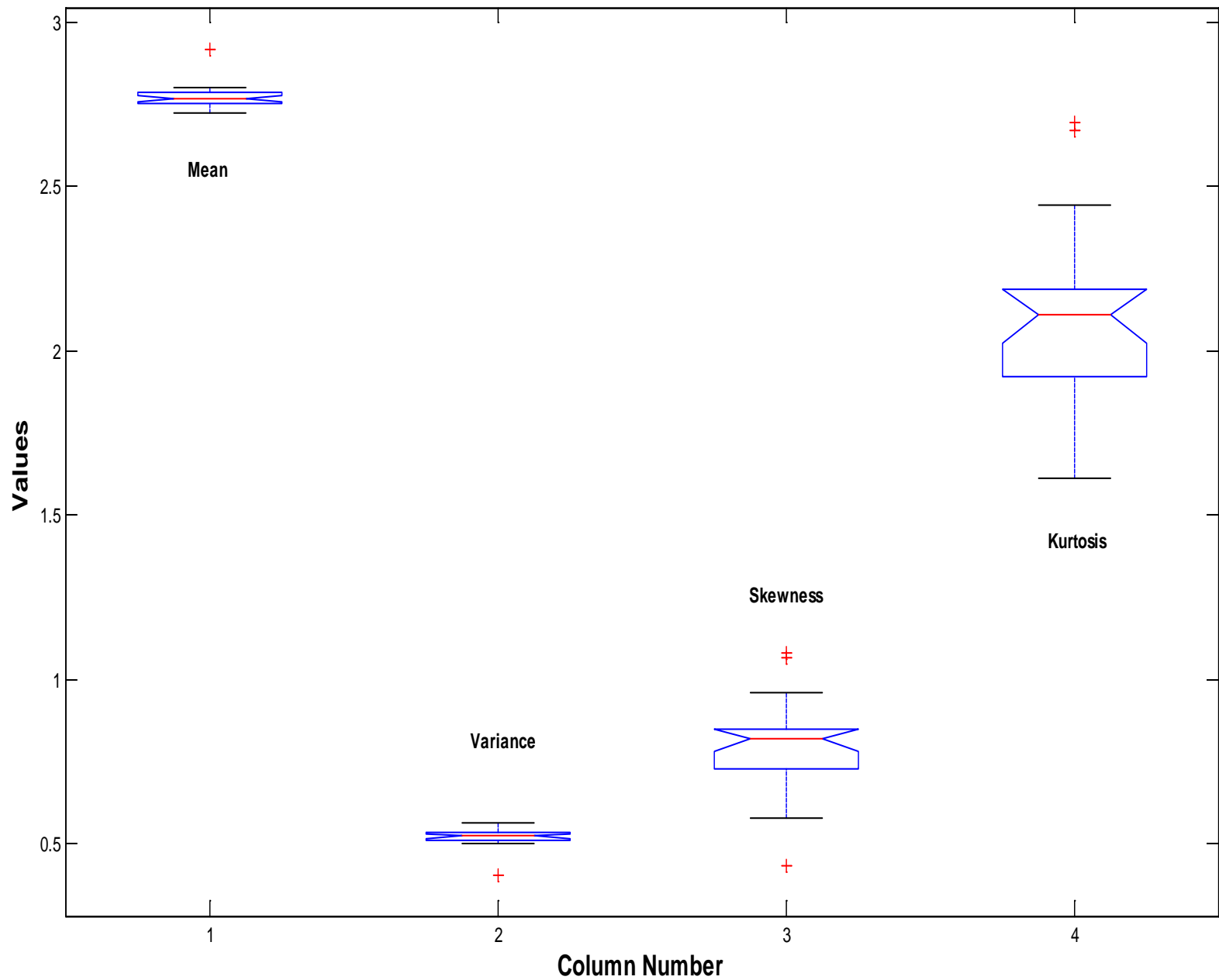




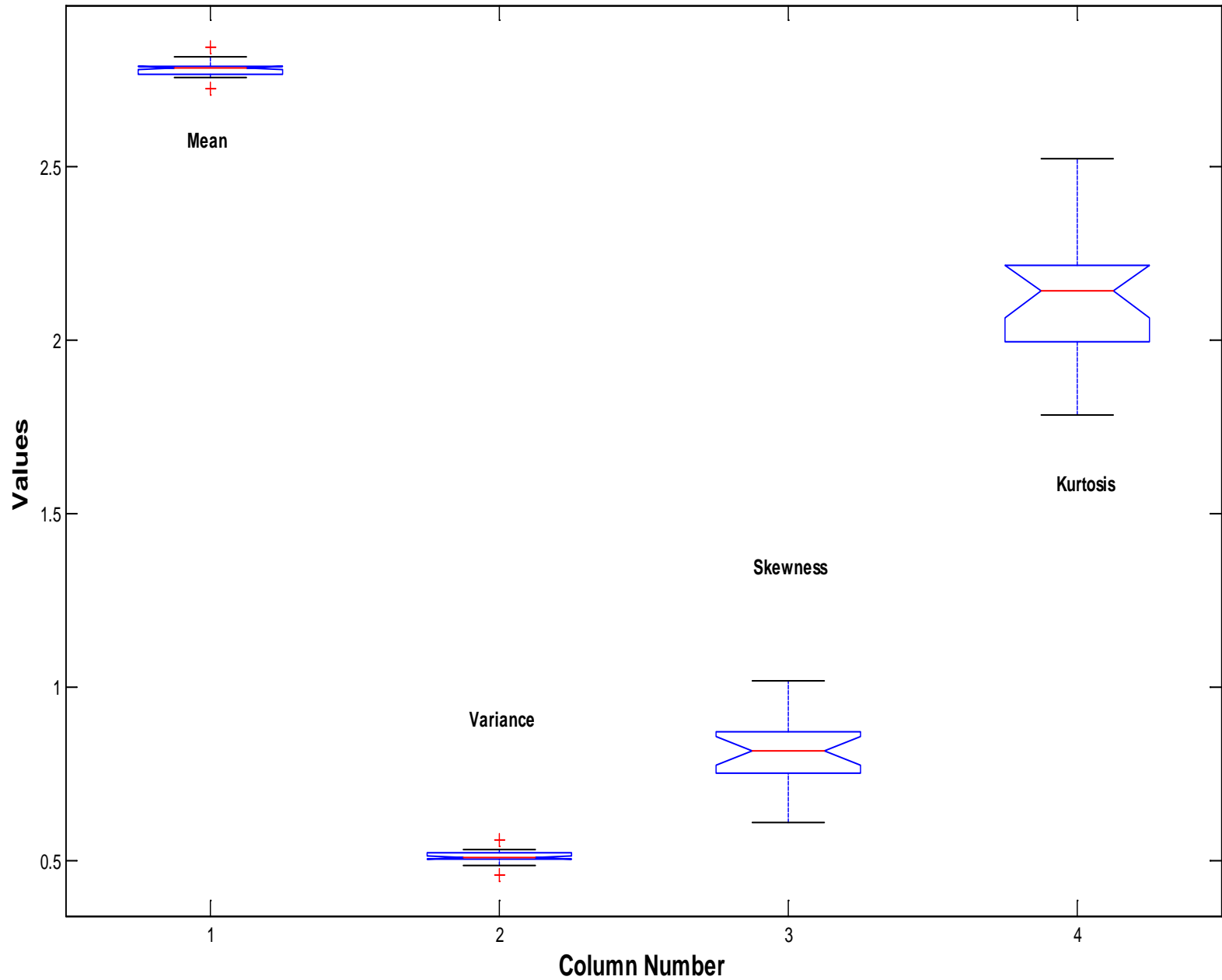
Statistics GC content Homo sapiens all chromosomes

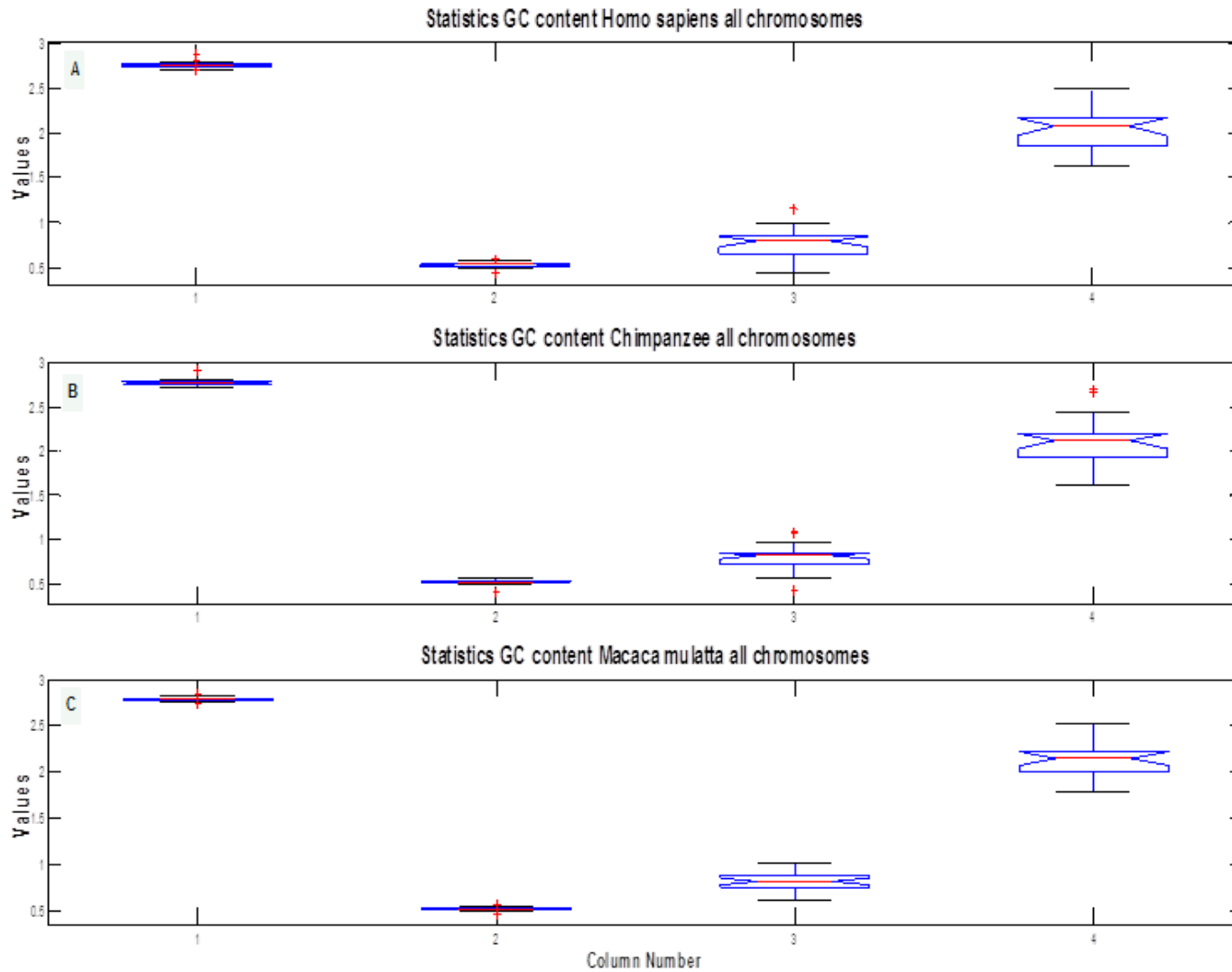


Statistics GC content Chimpanzee all chromosomes



Statistics GC content Macaca mulatta all chromosomes







RESULTS AND CONCLUSIONS

- ® There are not statistical differences between chromosomes of a given primate species.
- ® There is a log-linear negative correlation between *LADGC* and the GC content (%).
- ® The probability distribution of GC content in all primate chromosomes follows a quasi-lognormal distribution that can be modeled by a Black and Scholes model (Fokker-Planck equation).
- ® Then a unique type of distribution summarizes GC-rich, GC-poor and GC-intermediate regions in a given chromosome.
- ® This type of distribution would be in agreement with Neutral Molecular Evolution but the actual deviations from the lognormal as captured by long tails, skewness and kurtosis indicate strong positive natural selection at these sites. Differences among chromosomes of different species lie mostly in the skewness, kurtosis and less often in the mean of the distributions.
- ® In turn this indicates that natural selection favors nonlinear relationships among the distance series of GC content.
- ® According to the shuffled lognormal control, the GC-intermediate regions of chromosomes are subjected mostly to neutral mutations.

Final comment

In the XIX century, the hand was one of the symbols of the perfection of the human body, such as God has conceived it on the sixth day of Genesis. Although the Theory of Evolution finally prevailed, acknowledging man's simian lineage, some naturalists retained their view of the human being, merely shifting him from the status of the masterwork of Creation to that of the summit of evolution, its natural culmination. Evolution ceased to operate in the human species! Primates are all pentadactyls like their distant ancestors, but they present a derived feature: an opposable thumb on each limb. Evolution is not necessarily an improvement but simply a transformation. When human ancestors became bipedal, several MYA, the big toe evolved rapidly and lost its original mobility. The loss of its opposability is a derived characteristic that appeared late, linked together with the acquisition of bipedalism. Among other characteristics, it distinguishes man from his closest relative, the chimpanzee. Our most evolved finger is not our marvelous, archaic, opposable thumb, but our clumsy, recent big toe⁶⁰.

