Journal of Theoretical Biology 288 (2011) 92–104

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

On parameters of the human genome

Wentian Li

The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, 350 Community Drive, NY 11030, USA

ARTICLE INFO

Article history:
Received 13 April 2011
Received in revised form
28 June 2011
Accepted 21 July 2011
Available online 3 August 2011

Keywords:
Genome size
Karyotype
Human genes
Transcription factors
Single nucleotide polymorphisms

ABSTRACT

There are mathematical constants that describe universal relationship between variables, and physical/chemical constants that are invariant measurements of physical quantities. In a similar spirit, we have collected a set of parameters that characterize the human genome. Some parameters have a constant value for everybody's genome, others vary within a limited range. The following nine human genome parameters are discussed here, number of bases (genome size), number of chromosomes (karyotype), number of protein-coding gene loci, number of transcription factors, guanine–cytosine (GC) content, number of GC-rich gene-rich isochores, density of polymorphic sites, number of newly generated deleterious mutations in one generation, and number of meiotic crossovers. Comparative genomics and theoretical predictions of some parameters are discussed and reviewed. This collection only represents a beginning of compiling a more comprehensive list of human genome parameters, and knowing these parameter values is an important part in understanding human evolution.

1. Introduction

Mathematics, physics, and chemistry all have their standard sets of basic constants, as expected for any quantitative science. Mathematical constants are non-dimensional quantities that numerically characterize relationship between variables. Examples include golden ratio ϕ, circle circumference-to-diameter ratio π, and natural logarithmic base e. A total of 136 mathematical constants are listed and discussed in Finch (2003). Fundamental constants in physics are dimensional invariant values with various physical contents (Fritzsch, 2004). Examples include the speed of light c, Planck constant h, Newtonian constant of gravitation G in physics (Mohr et al., 2008). A total 326 physical/chemical constants are listed in National Institute of Standards and Technology (NIST)'s database.

Compared to these mathematical and physics sciences, biology, in particular molecular biology, is less quantitative, and generally speaking, constants are almost non-existent. We of course recognize some branches of biology with strong quantitative tradition, such as Mendel's genetics, Fisher–Haldane–Wright's population genetics, and the modern day bioinformatics. Even as biology becomes more quantitative in recent years, the numerical value of some baseline measurement is still typically less focused on than relative changes of some quantitative measurements. The term constant implies that the value will never change. Although such assumption has been challenged concerning fundamental physical constants, in particular in the cosmology context (Dirac, 1937; Gamow, 1967; Varshalovich and Potekhin, 1995; Uzan, 2003), the suggested magnitude of variation is extremely small and the proposed time for such a change is very long.

We are used to the notion that at the biological level, there could be structures or processes that are universal for all currently living species (e.g., genetic code), or some branches of species (e.g., division of sexes for most animals and plants). On the other hand, any attempt to quantify biological organisms numerically is to take a historical snapshot at a specific time and space; and due to evolution, these quantities may change over time. Due to this reason, the term *parameter* is better than the term *constant* in describing numerical features in biology, with the admission that these quantitative measures could eventually be different.

Why do we need to know these parameter values, if they are supposed to change in the future? Why could it be possibly interesting? First, for many genetic and genomic analyses, familiarity with human genome parameter makes it possible to do quick back-of-the-envelope calculation. In other words, these numbers are useful in practice. Second, comparing parameter values in human genome with that in other genomes reveal rich information about evolution. It is worth repeating Dobzhansky's line that "nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1964), and these parameter values included. Third, in the scope of human species, these parameter

E-mail address: wli@nslij-genetics.org

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.07.021

Table 1
Abbreviations used: TF: transcription factor; GC%: guanine–cytosine content; F: female; M: male; kb: 10^3 bases; Mb: 10^6 bases; Gb: 10^9 bases; NS: non-synonymous substitution; S: synonymous substitution; NC: substitution in non-coding region. Notes. n1: person-to-person variation. n2: based on the ~70–80 copy number variations (CNV) with average size of ~250 kb (Perry et al., 2008). n3: see Doolittle and Sapienza (1980), Orgel and Crick (1980), Charlesworth et al. (2002)). n4: see Cavalier-Smith (1982a), Cavalier-Smith (1982b). n5: 22 pairs are autosomal. n6: people who carry extra chromosome of chr21, chr18, chr13, chrX, etc. may still survive, though with abnormal symptoms (Down syndrome, Edwards syndrome, Patau syndrome, triple-X syndrome). n7: see Spuhler (1948a). n8: see King and Jukes (1969) and Ohno (1972). n9: see Bernardi (2007) n10: see Lynch (2010). n11: see Watterson (1975). n12: per diploid per generation.

Parameter	Value	Variability[n1]	Theory, model, prediction
# bases (genome size)	~3 Gb	$>\pm 20$ Mb[n2]	Selfish-DNA[n3], cell size[n4]
# chromosomes (karyotype)	23 pairs[n5]	0[n6]	Fusion/fission
# genes	~20,000	?	Comparative genomics[n7], mutation load[n8]
# TF gene	~2000	?	?
GC%	~40%	?	Neoselectionist[n9], mutation–selection balance[n10]
# GC-/gene-rich isochores	~120	?	?
Density of polymorphic sites	~1/kb	~1–10/kb ?	Infinite-site model[n11]
# deleterious mutations	~1[n12]	~0.5–2 ?	Ratio of #NS/(#S+#NC)
# crossovers	~80 (F), ~50 (M)	~50–100 (F), ~30–70 (M)	Müller's ratchet ?

values are relatively stable. As soon as some of these parameters start to change, we are arguably in a process of evolving towards a post-human species.

Constancy is at the other end of spectrum from variability and polymorphism. Each human being's genome is slightly different from another, with the exception of identical twins. Even that claim has been challenged (Bruder et al., 2008; Kaminsky et al., 2009), though any genomic difference between identical twins should be small, if not zero (Baranzini et al., 2010; Ono et al., 2010). The variability has its limit though: if the change in a person's genome is too far beyond the norm, e.g., one extra copy of chromosome 1, that person will not survive. If polymorphism is essential in identifying genetic basis of phenotypes and medical conditions (Buchanan and Higley, 1921), then constancy is essential in identifying us human as human.

Table 1 lists the human genome parameters or quantities to be discussed in this paper. The selection of these quantities is in some sense arbitrary. Some parameter values do not change from person to person, other vary among individuals within a range (e.g., number of crossovers). Some are measured precisely, others are still work-in-progress (e.g., number of transcription factor genes). Most parameters are well defined, whereas for others the definition itself is still under debate (e.g., gene-rich isochores). With the human genome sequence in public database, such as NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/human/), UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/), and Ensembl (http://useast.ensembl.org/Homo_sapiens/), many statistics can be obtained easily. Differing from other efforts of providing a summary of the biological information from the human genome (Scherer, 2008 and http://www.humangenomeguide.org/), this review focuses more on numerical and theoretical aspects of the human genome.

2. Human haploid genome size: 3×10^9 basepairs

The February 2009 version of the human genome (GRCh37/hg19) contains 3.098 Gb (b=basepair, kb=10^3b, Mb=10^6b, Gb=10^9b) including autosomal, X, Y chromosomes and mitocondrial DNA, sequenced and unsequenced (thus labeled by "N"), aligned and unaligned (thus in the "random" segment files). Since the number of unsequenced bases in heterochromatic regions is an estimation, its value (234 Mb) can be less reliable. Although we now use *b* to measure genome size, before the genome age, genome size was measured by weight of DNA molecules in the unit of picogram ($=10^{-12}$ g), called *C*-value (Swift, 1950). The conversion between the two units is: 1 pg=978 Mb or 1 Mb=1.022×10^{-3} pg (Gregory et al., 2007). The 3 Gb human genome size is roughly equal to 3 pg *C*-value.

Previously, it was thought that the larger a genome, the more complex the organism. Refusing the statement that plants are more complex than human because they have much higher *C*-values led to the realization that plants have higher proportion of repetitive sequences than human (Flavell et al., 1974), and genome size by itself is not a perfect measure of complexity. This is the "junk DNA explanation" of the "*C*-value paradox" (Pagel and Johnstone, 1992).

If genome size cannot really measure the organism complexity, it can be however used to infer evolutionary processes and rates. It is estimated that due to the difference of retrotransposition rate between human and other primates, there is a 15–20% expansion of human genome size in the past 50 million years (Liu et al., 2003). Another study has shown that genome size was reduced in mammals around the Cretaceous-Tertiary (KT) boundary ~65 millions years ago (Rho et al., 2009)—a mass extinction period that killed the dinosaurs.

There are two schools of thoughts in understanding genome size: the selfish-DNA hypothesis and the bulk-DNA hypothesis (Lynch, 2007). The selfish-DNA school considers the mobile elements in non-coding region as the driving force of genome size changes (Doolittle and Sapienza, 1980; Orgel and Crick, 1980; Charlesworth et al., 2002). A more complete description would consider all forces that affect genome size (besides mobile/transposable elements, there are also insertion, deletion, duplication, and chromosome translocation), characterize whether the mutation experienced neutral random drift or selective advantage/disadvantage, then run a computer simulation (Petrov, 2001). In particular, deletion or DNA loss has been suggested as an important factor in determining the genome size (Petrov et al., 2000).

The bulk-DNA hypothesis is based on the observation that genome size is highly correlated with the cell volume, nuclear volume (Cavalier-Smith, 1982, 1985), and cell division rate (Bennett, 1972). It has been hypothesized that the cell size increase from prokaryotes and eukaryotes was crucially due to the added mitochondria DNA/genes (Lane and Martin, 2010).

The fact that genome size is correlated with the cell volume has been used to infer the genome size of long-extinct species such as dinosaurs (Organ et al., 2007). In the bulk-DNA hypothesis, a larger genome is necessary to maintain a large cell volume with more complex cellular structures (Cavalier-Smith, 2005).