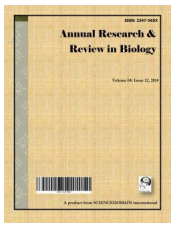


Annual Research & Review in Biology
7(4): 206-221, 2015, Article no.ARRB.2015.123
ISSN: 2347-565X

SCIENCEDOMAIN *international*
www.sciencedomain.org

Mean Variance Relationships of Genome Size and GC Content

Sunil Kanti Mondal[1], Rabindra Nath Das[2], Sudip Kundu[3], Jinseog Kim[4*], Gurprit Grover[5] and Shamim Akhtar Ansari[6]

[1]Department of Biotechnology, University of Burdwan, Burdwan, W. B., India.
[2]Department of Statistics, University of Burdwan, Burdwan, West Bengal, India.
[3]Department of Biophysics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata, W.B., India.
[4]Department of Statistics and Information Science, Dongguk University, Gyeongju, Korea.
[5]Department of Statistics, University of Delhi, Delhi, India.
[6]Tropical Forest Research Institute, RFRC, Jabalpur, India.

Authors' contributions

This work was carried out in collaboration between all authors. Author SKM collected the data and managed the literature searches. Author RND analyzed the data and wrote the paper. Authors SK, GG and SAA gave some biological explanations of the findings. Author JK reanalyzed the data. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/ARRB/2015/16709
Editor(s):
(1) George Perry, Dean and Professor of Biology, University of Texas at San Antonio, USA.
Reviewers:
(1) Ilham Zahir, Department of Biology, Sidi Mohamed Ben Abdellah University, Morocco.
(2) Anonymous, University of Jos, Nigeria.
(3) Anonymous, University of Kansas, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=976&id=32&aid=9755

Original Research Article

Received 11th February 2015
Accepted 9th June 2015
Published 13th June 2015

ABSTRACT

The present article focuses how the genome size and GC content are explained based on codon and amino-acid usage. This current study aims to identify the statistically significant factors of genome size and GC content using statistical modeling. The present analyses show that habitat (P = 0.08), taxonomy (P = 0.02), genome GC content (P < 0.01), isolation temperature (P< 0.01), GC% of the 2nd position within a codon for protein coding part (P< 0.01), number of total tRNA genes within genome (P< 0.01), lower (P< 0.01) and upper (P = 0.01) boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic (P< 0.01), aromatic (P<

**Corresponding author: E-mail: jinseog.kim@gmail.com, rabin.bwn@gmail.com;*

0.01), and positively charged r group containing amino acids (P< 0.01) are statistically significant effects of entire genome size. On the other hand, taxonomy (P = 0.03), genome size (P< 0.01), isolation temperature (P = 0.02), GC% of protein coding part of total genome (P< 0.01), GC% of the 1st (P< 0.01), 2nd (P< 0.01), and 3rd position (P< 0.01) within a codon for protein coding part, number of total tRNA genes within genome (P< 0.01), lower (P< 0.01) and upper (P< 0.01) boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic (P< 0.01), aromatic (P< 0.01) and negatively charged r group containing amino acids (P = 0.01) are statistically significant effects of entire genome GC content. These analyses support, and also try to remove some conflicts of many earlier research findings. However, the present analyses also have identified *all* new causal factors in the variance models, and many additional causal factors in the mean models of genome size and genome GC content, which was not reported by the earlier investigators.

Keywords: Amino acid; codon; genome size; genome GC content; joint generalized linear models; log-normal model; gamma model; non-constant variance.

1. INTRODUCTION

The relationship researchers noticed that the association between genome GC content (also genome size), codon and amino-acid usage is ahistorical. In the three domains of living organisms, it is observed that genes and genomes at mutation / selection equilibrium reproduce a unique relationship between nucleic acid and protein composition. An association between the species specificity in synonymous codon choice and amino acid usage was identified. In this correlation, proteins with species-specific amino acid usage were also coded with species-specific synonymous codon choice. Correlations between genome composition (in terms of GC content) (also genome size) and usage of particular codons and amino acids have been widely used, but it is still unclear and inconclusive [1-4]. For a long time, the central issue of evolutionary genomics was to find out the adaptive strategy of nucleic acid molecules of various microorganisms having different optimal growth temperatures (Topt). Long-standing controversies exist regarding the correlations between genomic G+C content and Topt, and this debate has not been yet settled [5].

The average genome size of microorganisms differs significantly between and within biomes. Aquatic microbiomes also showed large variation in average genome sizes, ranging from 1.5 to 5.5 Mb for Bacteria and Archaea. Microbial average genome lengths in the terrestrial biome were significantly higher than in the host-associated and aquatic biomes [6]. The presence of nucleotides (guanine and cytosine), known as `GC content' varies from among genomes of different species and phyla [7,8]. The genomic GC-content of bacteria varies dramatically, from less than 20% to more than 70% [9,10]. This variation may be due to the differences in the patterns of mutation between bacteria, or may be due to intrinsic, or extrinsic factors; or whether it is the result of neutral processes or selection [8,11].

Different organisms have idiosyncratic, and sometimes extremely biased, preferences for one synonymous codon over another. The distributions of genome size and GC-content for environmental microbial communities show a distinct pattern. The observed GC patterns are not a result of differing species compositions in each environment, as simulations of these compositions using sequenced genomes with the same phylogenetic distribution results in distinct GC patterns. Even closely related sequences, when they are from different environments, show a marked difference in GC content, more so than when they are from the same environment. The correlation between genome size and GC content is very small, as there is one possible environmental impact that the genomes in aquatic microorganism are smaller than in soil [12]. It has been known for some time that the frequencies of some codons and amino acids correlate with genome size and GC content [13], the causality has remained unclear and inconclusive: Correlations could exist because selection for a particular codon or amino-acid usage produces a particular genome size and GC content determines codon and amino-acid usage according to combinatorial principles. In this article, it is examined how the genome size and genome GC content are associated with codons and amino-acids usage.

In the evolutionary theory of synonymous codon usage, some researches sought to explain interspecific variation in overall sequence composition, and noted correlations between GC content and amino acid content across different species. Earlier researches have suggested that the genomes were at equilibrium with respect to mutation, and they have also explained how directional mutation could affect the composition of coding sequences [7,13,14]. Although it has not been explained why species with similar genome composition have recognizably distinct sequences for individual genes? The genome GC content (also genome size) has been shown to correlate with cross-species differences in frequencies of codons [15,16] and amino acids [17,18]. The frequency of some amino acids is generally low in the low GC content bacterium but it increases in the high GC bacterium. It clearly shows that the amino acid usage of a protein can be very different between high GC and low GC content bacteria [19,20]. There is a tendency of large genomes to be GC rich and small genomes to be GC poor [19]. The reason for this may be that large genomes are generally found in more complex environments, as there may be an indirect link between GC content and niche complexity. Another factor could be the preferred growth temperature of an organism, which has been proposed to correlate with GC content [21], but this is under debate [2,22].

In the present article, responses genome size and genome GC content are modelled based on codons and amino-acids usage. It is identified that both the responses are *non-normal* are *heteroscedastic*. Accordingly, both the responses are modelled using joint generalized linear models. In the present analysis, habitat, genome GC content, isolation temperature, GC% of the 2nd position within a codon for protein coding part, number of total tRNA genes within genome, lower and upper boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic, aromatic and positively charged r group containing amino acids are identified as the significant factors for the mean of genome size, whereas its variance is explained by taxonomy and number of total tRNA genes within genome. On the other hand, mean genome GC content is explained by statistically significant factors genome size, isolation temperature, GC% of protein coding part of total genome, GC% of the 1st, 2nd, and 3rd position within a codon for protein coding part, lower and upper boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic and negatively charged r group containing amino acids, while the variance of genome GC content is explained by statistically significant factors taxonomy, GC% of the 1st and 2nd position within a codon for protein coding part, number of total tRNA genes within genome, lower and upper boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic and aromatic r group containing amino acids.

Some earlier findings about the genome size and GC content are cited as in the above. This literature invites some doubts and debates about the causal factors of the genome size and GC content. What are the backgrounds of these doubts and debates of the earlier findings? Some of the defects of the earlier studies are described in Section 2.

2. BACKGROUND

In earlier researches, linear correlation and simple regression lines [1,5,22] have been fitted to derive the relationships between genome GC content, codons and amino-acids usage. Based on classical assumptions (which are *not* valid for any positive data set), these relationships have been derived. As a result, the predictions (drawn from these analyses) relating these responses have thus far had limited success. This can be remedied by taking into account an appropriate statistical technique and the differential effect of selection on the different positions within codons. Recently, some simple models have been provided, based solely on purifying selection and mutation at the nucleotide level, that quantitatively predicts both codon and amino-acid usage trends across archaea, bacteria and eukaryotes on the basis of the genome GC content [23,24]. In earlier researches, it has been identified that the response variances of genome size and genome GC content are *non-constant*, distributions are *non-normal*, and many factors *may* effect on these responses. Under these situations, classical simple and multiple regression analyses are completely *inappropriate*.

Many of the relationships researchers have sought to identify between genome GC con-tent (also genome size), codons and amino acids usage are still unclear and inconclusive. The reason is that evidences are insufficient or conflicting. Generally, validated relationships are established based on an appropriate statistical analysis. Some previously reported statistical

analyses indicate that certain relationships between genome GC content, codons and amino acids usage are inconsistent. For a better understanding of these relationships, further studies are indispensable. The functional relationship is considered a probabilistic (regression or generalized linear model (GLM)) model that provides an approximation to relatively more complex phenomenon [25-28]. If the univariate response data sets are independent or dependent, heteroscedastic and belong to exponential family, both the mean and the variance need to be modelled simultaneously, using link functions for natural mean and variance. This modelling approach is known as joint generalized linear model (JGLM) [29].

For non-constant variance (heteroscedastic) data, log-transformation is often recom-mended to stabilize the variance [30]. However, in practice, the variance is not always stabilized by an appropriate (seems to be suitable) transformation [27]. For heteroscedastic response, classical regression technique gives inefficient analysis, often resulting in an error so that significant factors are classified as insignificant. For instance, the analysis by Myers et al. [27] missed many important factors of the process. This is a serious error in any data analysis. It is well known that the positive data sets are analyzed either by the log-normal or the gamma models [26,31-34]. The present authors have noticed that the original data set is positive, the response variance is non-constant, distribution is non-normal, and the model fit criteria measure values are inconsistent. In earlier analyses, these features of the data sets were not counted. As a result, the earlier findings invite some doubts and debates. These observations have motivated us to take up this present study.

Generally, some continuous positive response variables belong to the exponential family of distributions, and their variances may or may not be constant, as the variance may or may not have relation with the mean. The problem of non-constant variance (for the response variable y) in linear regression is a departure from the standard least squares assumptions. This problem of inequality of variance occurs often in practice, frequently in conjunction with a non-normal response variable. To minimize the problem, an appropriate method is to transform the response variable to stabilize the variance. This makes the distribution of the response variable closer to the normal distribution, and it improves the fit of the model to the data. However, in practice, a suitable transformation may not always stabilize the variance [27,33]. Thus, for analysis of positive data with non-constant variance, it is crucial to use joint generalized linear models (JGLMs) (modelling of mean and variance simultaneously) to identify the significant factors of the process [29,33]. Joint GLMs (with relevant references) for log-normal and gamma models are described in Section 3.

3. METHODOLOGY: JOINT MEAN AND VARIANCE MODELS UNDER LOG-NORMAL AND GAMMA DISTRIBUTION

The class of generalized linear models includes distributions useful for the analysis of some continuous positive measurements in practice which have non-normal error distributions. The problem of non-constant variance in the response variable *y* in linear regression is due to departure from the standard least squares assumptions. Transformation of the response variable is an appropriate method to stabilize the variance. For heteroscedastic data, the log-transformation is often recommended [30]. However, in practice the variance may not always be stabilized despite a proper transformation [27; Table 2.7, p. 36]. Box [35] proposed the use of linear models with data transformation.

For example, when

$$E(Y_i) = \mu_i \text{ and } Var(Y_i) = \sigma_i^2 \mu^2_i ;$$

the transformation $Z_i = \log(Y_i)$ gives stabilization of variance $Var(Z_i) \approx \sigma_i^2$. However, if a parsimonious model is required, a different transformation is needed. Thus, a single data transformation may fail to meet various model assumptions. Nelder and Lee [36] proposed using joint generalized linear models (GLMs) for the mean and dispersion.

When the response Y_i is constrained to be positive log transformation $Z_i = \log Y_i$ is used. Under the log-normal distribution, a joint modelling of the mean and dispersion is such that

$$E(Z_i) = \mu_i \text{ and } Var(Z_i) = \sigma_i^2;$$
$$\mu_i = x_i^t \beta \text{ and } \log(\sigma_i^2) = g_i^t \gamma; \quad (1)$$

where x_i^t and g_i^t are the row vectors for the regression coefficients β and γ in the mean and dispersion model, respectively.

For the constant coefficient of variation (i.e., variance increases with the mean), we have

$$\mathrm{Var}(Y) = \sigma^2\{E(Y)\}^2 = \sigma^2\mu_y^2.$$

Further, if the systematic part of the model is multiplicative on the original scale, and hence additive on the log scale, then

$$Y_i = \mu y_i \varepsilon_i \quad (i = 1,2,3,\ldots..,\ n)$$

with

$$\eta_i = \log \mu_{Yi} = \mathbf{x_i^t\beta} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ...+x_{ip}\beta_p \qquad (2)$$

and ε_i's are independent identically distributed (IID) with $E(\varepsilon_i^2) = 1$. In generalized linear models (GLMs), μ_{Yi} is the scale parameter and $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is the shape parameter.

For non-constant variance response, Nelder and Lee [36] proposed a modelling approach for the multiplicative model (2). These researchers advocated the use of joint generalized linear models (JGLMs):

$$E(y_i) = \mu_i \text{ and } \mathrm{Var}(y_i) = \sigma^2\ V(\mu_i);$$

with

$$\eta_i = \log(\mu_{Yi}) = x_i^t\beta \text{ ;and } \varepsilon_i = \log(\sigma^2_{Yi}) = g_i^t\gamma_Y; \qquad (3)$$

where x_i and g_i are the row vectors used in the mean and the dispersion models, respectively. The regression coefficients (β_y) of the mean model and (γ_y) of the dispersion model are estimated, respectively, by the maximum likelihood (ML) and the restricted ML (REML) method [33,37]. The restricted likelihood estimators have proper adjustment of the degrees of freedom by estimating the mean parameters, which is important in the analysis of data from quality engineering because the number of parameters of β is often relatively large compared with the total sample size.

In GLMs, the variance consists of two components, one is $V(\mu_i)$, which depends on the mean (μ_i), and the other is σ_i^2, which is independent of the mean adjustment. The variance function ($V(\cdot)$) characterizes the distribution of GLMs family. For example, if $V(\mu) = 1$, the distribution is Normal, Poisson if $V(\mu) = \mu$, gamma if $V(\mu) = \mu^2$, etc. Some detailed discussion on GLMs approaches is given in [29,33,37-41].

4. GENOME DATA, ANALYSIS AND INTERPRETATION

A. Data: Genome data set under the present study contains 576 microorganisms (observations) on 17 variables. There are 158 aquatic, 68 terrestrial and 350 hosts. The present data set is collected by the following method.

Primary data collection:

1. The kingdom, taxonomical classification, habitat, temperature range, size and GC% of the genomes of a large number of microorganisms whose genomes have been completely sequenced and retrieved from NCBI (www.ncbi.nlm.nih.gov/genomes/lproks.cgi).
2. The entire codon usage table, coding GC%, GC1%, GC2% and GC3% which are available have been retrieved from `Codon Usage Database' (http://www.kazusa.or.jp/codon/), and for the rest organisms we have retrieved all the cDNA sequences from NCBI (ftp://ftp.ncbi.nih.gov/genomes/) or PATRIC (http://brcdownloads.patricbrc.org/patric2/).
3. The entire genome sequences have been retrieved from NCBI and PATRIC.
4. The three letter code (if it is available) for the individual organism has been retrieved from KEGG, and if is not available we have put an alpha numeric three character code started by A (for aquatic)/ T (for terrestrial)/ H (for host).

Secondary data generation:

1 For all of those microorganisms, using a PERL script written in house we have calculated `amino-acid frequencies' (within 100) of each amino acid, and the RSCU values from codon usage tables or cDNA sequence file following translation table 11 (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi/).

2 Using a PERL script written in house, coding GC%, GC1%, GC2% and GC3% have been calculated for those if the CUTG table was unavailable.

3 Using ARAGORN (http://130.235.46.10/ARAGORN/) and in

house PERL script we have calculated the codon specific tRNA number, GC% range of the tRNA encoding part from the entire genome sequences.

Arrangement of data under different parameter:

1. These organisms are classified into three major groups based on their habitat type-organisms isolated from terrestrial, aquatic and host.
2. The organisms have been classified into different groups based on different taxonomical groups (like: Proteobacteria alpha, beta, gama, firmicutes, cyanobacteria etc) within the same habitat.
3. The microorganisms have been marked as psychrophilic, mesophilic, thermophilic and hyperthermophilic according to their temperature range of living.
4. On the basis of physiochemical properties amino acids have been grouped into non-polar aliphatic r group containing amino acids (NPA), aromatic r group containing amino acids (ARO), polar uncharged r group containing amino acids (PUC), positively charged r group containing amino acids (PCH), negatively charged r group containing amino acids (NCH) and group wise average frequencies have been calculated using MS-Excel.

B. Variables

Dependent variables:

The dependent variables in the present study are the genome size and the genome GC content.

Independent variables:

There are two sets of independent variables, qualitative and quantitative. Three independent variables (habitat, taxonomy, temperature range) are qualitative and the remaining thirteen are continuous variables. Table 1 presents a description of each set of item and how they are operationalized for the present study. The present data set is not displayed here, as it would substantially increase the length of the paper. However, we may submit our data set on request for verification of our analysis.

4.1 Genome Size Data Analysis and Interpretations

In the present subsection, the dependent variable genome size is analyzed, treating it as the response variable, in relation to the 16 covariates as explanatory variables (Table 1). Table 1 displays the independent variables and their respective levels. There are three factors and fourteen continuous variables (Table 1). For factors, the constraint that the effects of the first levels are zero is accepted. Therefore, it is taken that the first level of each factor as the reference level by estimating it's as zero. Suppose that α_i for i = 1, 2, 3 represents the main effect of A. It is taken $\hat{\alpha}_1 = 0$, so that $\hat{\alpha}_2 = \hat{\alpha}_2 - \hat{\alpha}_1$. For example, the estimate of the effect $A2$ means the effect of difference between the second and the first levels in the main effect A, i.e., $\hat{\alpha}_2 - \hat{\alpha}_1$. Note that the factors habitant, temperature and taxonomy have respectively, three, four and eighteen levels (Table 1). As taxonomy has more levels, it is treated here as a variable, and the other two are treated as factors for the present analysis.

In the present subsection, it is aimed to identify the factors which have significant effects on genome size (response variable). It is identified that the genome size is a non-constant variance response. Thus, we have fitted the data set with both the joint log-normal and gamma models in Section 3. It is found that the joint log-normal models fit is better than the gamma fit (based on Akaike information criterion (AIC) and graphical analysis), so only the results of log-normal models fit are displayed in Table 2. The selected models have the smallest AIC value in each class. It is well known that AIC selects a model which minimizes the predicted additive errors and squared error loss [42; p. 203-204). The value of AIC of the selected model (Table 2) is 1601 + 2*17 = 1635.0.

Fig. 1(a) displays the histogram of residuals. It does not show any lack of fit (due to missing variables or influential observations). Fig. 1(b) presents the absolute residuals plot with respect to fitted values. This is a flat diagram with the running means, indicating that the variance is constant under the joint GLM log-normal fitting. Fig. 2(a) and Fig. 2(b), respectively, display the normal probability plot for the mean and the variance model in Table 2. Neither figure shows any systematic departures, indicating no lack of fit of the selected final models.

Table 2 shows the parameters habitat, genome GC content, isolation temperature, GC% of the 2nd position within a codon for protein coding part, number of total tRNA genes within genome, lower and upper boundary of GC% for tRNA encoding genes, average frequency (within 100)

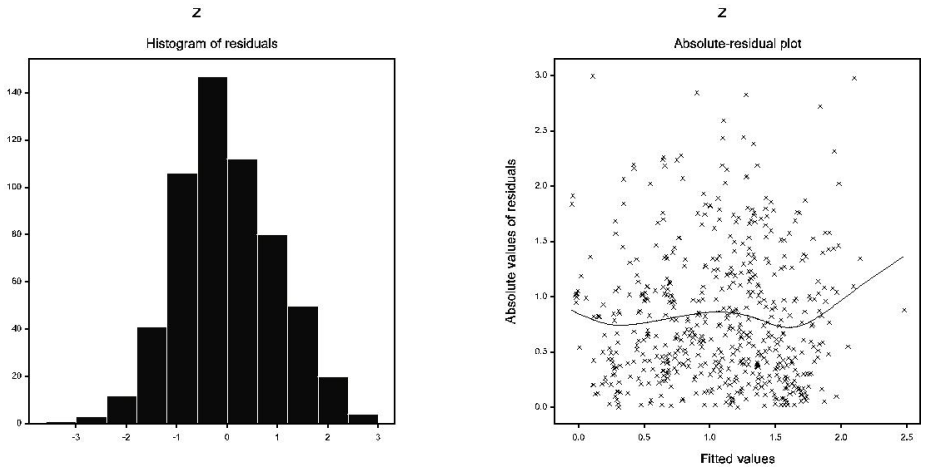
of non-polar aliphatic, aromatic and positively charged r group containing amino acids are statistically significant factors of mean genome size. Mean genome size is positively associated with genome GC content, habitant `terrestrial', isolation temperature `mesophilic' and `psychrophilic', GC% of the 2nd position within a codon for protein coding part, number of total tRNA genes within genome, upper boundary of GC% for tRNA encoding genes, and is negatively associated with lower boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic, aromatic and positively charged r group containing amino-acids usages. Note that the habitant \ host" and the isolation temperature \ hyper-thermophilic" are *insignificant*, and the isolation temperature `psychrophilic' is *partially* (0.05< *P* <0.15) positively significant.

Table 1. Operationalization of variables in the analysis

Domain/ Variable name	Operationalization
HABITAT (A)	Habitat(1 = Aquatic, 2 = Terrestrial, 3 = Host)
TAXONO (x1)	Taxonomy(1= Actinobacteria, 2= Alpha Proteobacteria, 3= Beta Proteobacteria, 4= Gamma Proteobacteria, 5= Delta Proteobacteria, 6= Cyanobacteria, 7= Bacteroidetes, 8= Firmicutes, 9= Deinococcus, 10= Thermotogae, 11= Planctomycetes, 12= Crenarchaeota, 13= Euryarchaea, 14= Chlamydia, 15= Fuso bacteria, 16= Nano archaea, 17= Epsilon Proteobacteria , 18= Spirochete)
SIZE (y)	Entire genome size in MB
GC%(y1)	Entire genome GC%
TEMP (B)	Temperature(1=Thermophilic, 2=Mesophilic, 3=Psychrophilic, 4= Hyper-Thermophilic)
COD GC% (x6)	GC% of protein coding part of entire genome
GC1% (x7)	GC% of the 1st position within a codon for protein coding part
GC2% (x8)	GC% of the 2nd position within a codon for protein coding part
GC3% (x9)	GC% of the 3rd position within a codon for protein coding part
tRNA (x10)	Number of total tRNA genes within genome
tRNA GC1%(x11)	Lower boundary of GC% for tRNA encoding genes
tRNA GC2% (x12)	Upper boundary of GC% for tRNA encoding genes
AVG NPA(x20)	Average frequency (within 100) of non-polar aliphatic r group containing amino acids
AVG ARO(x24)	Average frequency (within 100) of aromatic r group containing amino acids
AVG PUC (x30)	Average frequency (within 100) of polar uncharged r group containing amino acids
AVG PCH (x34)	Average frequency (within 100) of positively charged r group containing amino acids
AVG NCH (x37)	Average frequency (within 100) of negatively charged r group containing amino acids

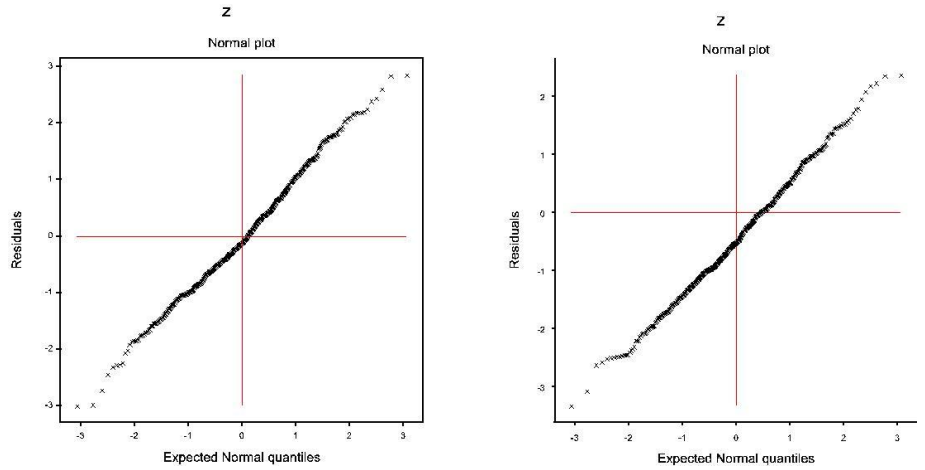
Table 2. Results for mean and dispersion models of genome size data from log-normal fit

	Covariate	Estimate	s.e.	t	P-value	95%	C.I.
Mean model	Constant	1.2750	0.4471	2.852	0.01	0.4	2.151
	GC%(y1)	0.0129	0.0027	4.728	<0.01	0.007	0.018
	TEMP2(B2)	0.3903	0.1193	3.271	<0.01	0.156	0.624
	TEMP3(B3)	0.2343	0.1490	1.572	0.12	-0.057	0.526
	TEMP4(B4)	0.0934	0.1849	0.505	0.61	-0.269	0.455
	HABITAT2(A2)	0.0848	0.0488	1.738	0.08	-0.011	0.18
	HABITAT3(A3)	-0.0420	0.0880	-0.477	0.63	-0.214	0.13
	GC2%(x8)	0.0124	0.0016	7.722	<0.01	0.009	0.015
	tRNA(x10)	0.0101	0.0007	15.548	<0.01	-0.004	0.024
	tRNA GC1%(x11)	-0.0080	0.0021	-3.894	<0.01	-0.012	-0.004
	tRNA GC2%(x12)	0.0107	0.0039	2.763	0.01	0.003	0.018
	AVG NPA(x20)	-0.2357	0.0523	-4.505	<0.01	-0.338	-0.133
	AVG ARO(x24)	-0.0740	0.0225	-3.282	<0.01	-0.118	-0.03
	AVG PCH(x34)	-0.1559	0.0413	-3.774	<0.01	-0.237	-0.075
Variance model	Constant	-1.1442	0.2257	-5.068	<0.01	-1.587	-0.702
	TAXONO(x1)	0.0302	0.0132	2.298	0.02	0.004	0.056
	tRNA (x10)	-0.0221	0.0035	-6.331	<0.01	-0.029	-0.015



(a) (b)

Fig. 1. (a) The histogram plot of residuals and (b) the absolute residuals plot with respect to fitted values for genome size data (Table 2)



(a) (b)

Fig. 2. The normal probability plot of the (a) mean and (b) variance for genome size data (Table 2)

Table 2 shows that taxonomy and the number of total tRNA genes within genome are statistically significant with the variance of genome size. The variance of genome size is positively associated with the taxonomy, and is negatively associated with the number of total tRNA genes within genome, indicating that the variance of genome size decreases with the increasing of the number of total tRNA genes within genome.

4.2 Genome GC Content Data Analysis and Interpretations

In the present subsection, genome GC content is considered as the response variable, and the remaining other variables are treated as explanatory variables. Genome GC content data set is identified as a non-constant variance response. Therefore, it has been fitted using both the joint log-normal and gamma models (Section 3). It is observed that joint gamma models fit is better than the log-normal fit (based on AIC and graphical analysis), so only the results of gamma fit are presented in Table 3. The selected models have the smallest AIC value (2701.922 + 2*23 = 2747.922; Table 3) in each class.

Fig. 3(a) and Fig. 3(b) display respectively, the histogram of residuals and the absolute residuals plot with respect to the fitted values. The

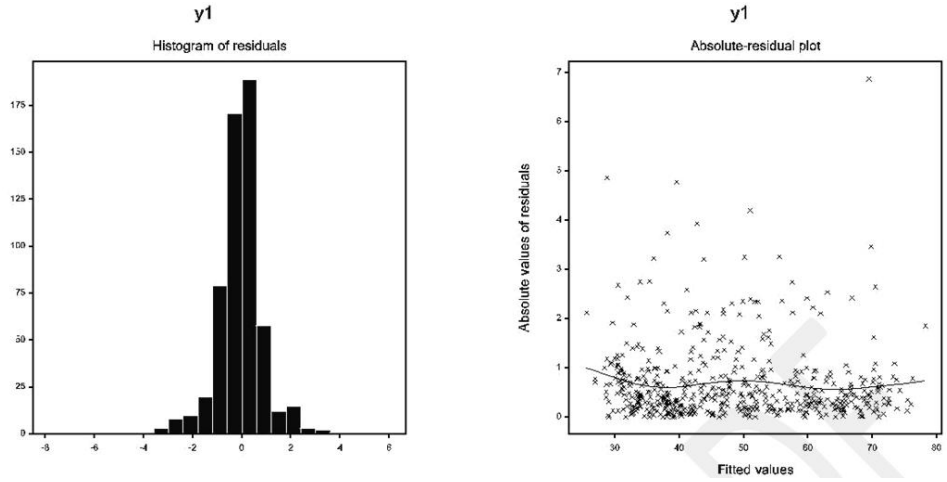
histogram plot (Fig. 3(a)) does not show any lack of fit. Fig. 3(b) is a flat diagram with the running means, indicating that variance is constant under the joint GLM gamma fitting. Fig. 4(a) and Fig. 4(b) display respectively, the normal probability plot for the mean and the variance model in Table 3. There does not exist any systematic departure in any one of these two figures. So, there is no lack of fit of the final selected models.

Table 3 shows that the genome size, isolation temperature `hyper thermophilic', GC% of protein coding part of total genome, GC% of the 1st, 2nd, and 3rd position within a codon for protein coding part, number of total tRNA genes within genome, lower boundary of GC% for tRNA encoding genes are positively (significant) associated with the mean of genome GC content, indicating that if these effects increase, genome mean GC content will increase. Also upper boundary of GC% for tRNA encoding genes, average frequency (within 100) of non-polar aliphatic and negatively charged r group containing amino acids are negatively (significant) associated with the mean of genome GC content, indicating that if these effects decrease, genome mean GC content will increase, and vice-versa. Again, isolation temperature `psychrophilic' is also partially negatively associated with the mean of genome GC content. This implies that genome GC content is low at the isolation temperature `psychrophilic' and is indifferent at the mesophilic level.

Table 3 shows that the variance of genome GC content is positively associated with GC% of the 1st position within a codon for protein coding part, upper boundary of GC% for tRNA encoding genes and average frequency (within 100) of aromatic r group containing amino acid, indicating that if these effects increase, the variance of genome GC content will increase. Again, taxonomy, GC% of the 2nd position within a codon for protein coding part, number of total tRNA genes within genome, lower boundary of GC% for tRNA encoding genes and average frequency (within 100) of non-polar aliphatic r group containing amino acid are negatively associated with the variance of genome GC content, indicating that if these effects increase, variance will decrease.

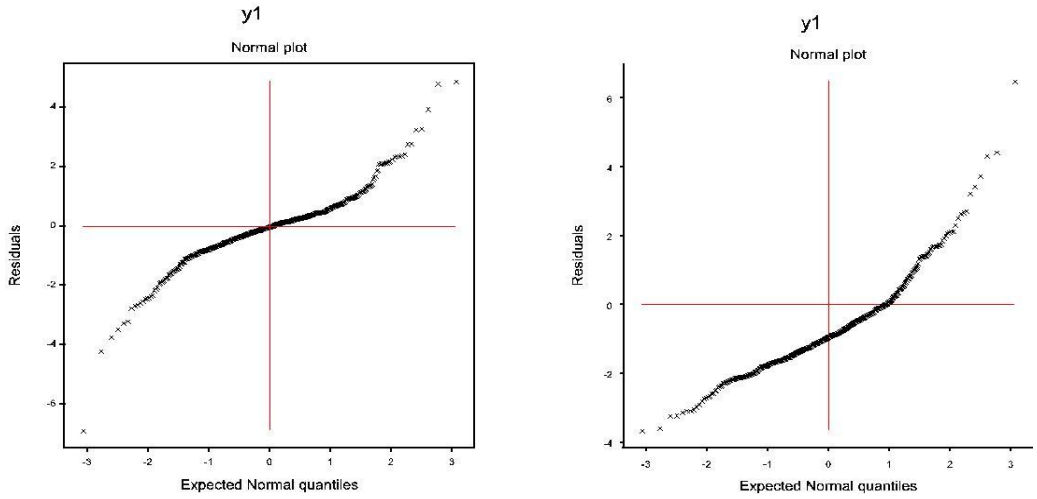
Table 3. Results for mean and variance models of GC content data from gamma fit

	Covariate	Estimate	s.e.	t	P-value	95%	C.I.
Mean model	Constant	3.0800	0.0594	51.86	<0.01	2.9635	3.1964
	TEMP2(B2)	-0.0059	0.0137	-0.43	0.67	-0.0327	0.0209
	TEMP3(B3)	-0.0317	0.0198	-1.61	0.11	-0.0705	0.0071
	TEMP4(B4)	0.0402	0.0164	2.45	0.02	0.0081	0.0723
	SIZE(y)	0.0046	0.0015	3.02	<0.01	0.0016	0.0075
	COD GC%(x6)	0.0078	0.0008	10.28	<0.01	0.0062	0.0093
	GC1%(x7)	0.0068	0.0004	17.77	<0.01	0.006	0.0075
	GC2%(x8)	0.0053	0.0002	24.53	<0.01	0.0049	0.0056
	GC3%(x9)	0.0038	0.0005	7.93	<0.01	0.0028	0.0048
	tRNA (x10)	0.0002	0.0001	1.57	0.12	0.0001	0.0004
	tRNA GC1%(x11)	0.0019	0.0002	8.48	<0.01	0.0015	0.0023
	tRNA GC2%(x12)	-0.0031	0.0005	-6.41	<0.01	-0.0041	-0.0021
	AVG NPA(x20)	-0.0360	0.0081	-4.46	<0.01	-0.0519	-0.0201
	AVG NCH(x37)	-0.0104	0.0041	-2.52	0.01	-0.0184	-0.0023
Variance model	Constant	-2.6671	1.3263	-2.011	0.04	-5.2666	-0.0675
	TAXONO(x1)	-0.0327	0.0147	-2.229	0.03	-0.0615	-0.0039
	GC1%(x7)	0.0757	0.0079	9.554	<0.01	0.0602	0.0911
	GC2%(x8)	-0.0171	0.0060	-2.841	0.01	-0.0289	-0.0053
	tRNA(x10)	-0.0154	0.0042	-3.694	<0.01	-0.0236	-0.0072
	tRNA GC1%(x11)	-0.0841	0.0088	-9.525	<0.01	-0.1013	-0.0669
	tRNA GC2%(x12)	0.0726	0.0161	4.516	<0.01	0.041	0.1041
	AVG NPA(x20)	-1.2744	0.1869	-6.817	<0.01	-1.6401	-0.9081
	AVG ARO(x24)	0.5206	0.0438	11.898	<0.01	0.4347	0.6064



(a) (b)

Fig. 3. (a) The histogram plot of residuals and (b) the absolute residuals plot with respect to fitted values for genome GC content data (Table 3)



(a) (b)

Fig. 4. The normal probability plot of the (a) mean and (b) variance for genome GC content data (Table 3)

5. DISCUSSIONS AND CONCLUDING REMARKS

This article focuses on the determinants of genome size and genome GC content based on codons and amino-acids usage. The present response data set is positive, so the possible probability model is log-normal or gamma [26,31]. Both the responses genome size and genome GC content are identified as non-constant variances (Tables 2, 3). Thus, the joint models of mean and variance are derived from log-normal and gamma distributions. The present data set has been examined using both the joint log-normal and gamma models [33]. It is observed that the joint log-normal models fit is much better than the gamma models for genome size, while for genome GC content, the situation is quite reverse, therefore, only the appropriate results of JGLMs are reported.

The results (in Table 2) related to genome size can be interpreted in the following ways.

1. It has been pointed that there is a weak correlation between genome size and GC content [12]. From Table 2 and Table 3, it is clear that the genome size and GC content are positively (statistically significant) correlated (a strong association as $P = 2.87e^{-6}$) with each other. It implies that if a new group of bacterial is studied, we would expect that those with larger genomes would have a larger average GC content than those with smaller genomes. Therefore, the present analysis supports the finding of [19], and it may be restated that for a new group of bacterial, as the large genomes to be GC rich and small genomes to be GC poor. Earlier researchers have explained this situation as that the large genomes are generally observed in more complex environments, as there may be an indirect link between GC content and niche complexity.
2. From Table 2, it is clear that the mean genome size is highly associated (P= 0.0011) (significant) with isolation temperature. It is positively significant at isolation tempera-ture level 2 i.e., at mesophilic, and partially at level 3 i.e., at psychrophilic, and insignifi-cant at level 4 i.e., at hyper-thermophilic. These results indicate that the genome size is higher at mesophilic and psychrophilic than thermophilic, and it is indifferent at hyper thermophilic. In earlier researches, some controversies exist regarding the association be-tween genome size and different optimal growth temperatures [21], but the present analysis gives a clear information.
3. It is observed that the type of habitat is associated with the genome size (Table 2). Habitat type 2, i.e., terrestrial is positively partially significant, and habitat type 3, i.e., host is insignificant with the mean genome size. These results indicate that the average genome size of the terrestrial is significantly higher than the aquatic (supports [16]), and it is indifferent for the host.
4. Mean genome size is positively (significant) associated with GC% of the 2nd position within a codon for protein coding part (Table 2). This indicates that for a new group of bacterial, genome size will increase if the GC% of the 2nd position within a codon for protein coding part will increase, and vice-versa.
5. Average genome size is positively (statistical significant) associated each with tRNA and tRNA GC2% (Table 2). These imply that the genome size will increase separately with the increase of number of total tRNA genes within genome and upper boundary of GC% for tRNA encoding genes.
6. Average genome size is negatively (significant) associated with the lower boundary GC% for tRNA encoding genes (Table 2). This implies that as the genome size increases, the lower boundary GC% for tRNA encoding genes decreases.
7. Average genome size is negatively (statistical significant) associated each with the average frequency (within 100) of non-polar aliphatic, aromatic and positively charged r group containing amino acids usages (Table 2). These indicate that the genome size will be large if each of the average frequency (within 100) of non-polar aliphatic, aromatic and positively charged r group containing amino acids usages will be low, and vice-versa. These present results are a little bit different from the earlier findings [19,20].
8. Variance of genome size is negatively associated (significant) with the number of total tRNA genes within genome. This indicates that if the total tRNA genes increase, variance of genome size decreases. Consequently, genome size increases.
9. Taxonomy is also associated with the variance of genome size, indicating that the variance of genome size changes with the type of taxonomy of the organisms. That is the variation of genome size exists within the different types of taxonomy. This result agrees with the findings of earlier researches [7,8].

The present results (Table 3) of genome GC content may be interpreted below.

1. From Table 3 (also in Table 2) genome GC content is positively (significant) associated with the genome size. Therefore, the *same* interpretation as in serial no. 1 (for genome size) is valid here.
2. In earlier researches, the factor growth

temperature has been proposed to correlated with GC content [21], but this is under debate [2,5]. In the present analysis (Table 3), it is clear that the genome GC content is highly associated (significant) with isolation temperature. It is *insignificant* at isolation temperature level 2 i.e., at mesophilic and *partially negatively* at level 3 i.e., at psychrophilic and *positively significant* at level 4 i.e., at hyper-thermophilic. Therefore, it is concluded that the genome GC content is higher at hyper-thermophilic than at thermophilic, lower at psychrophilic and is *indifferent* at mesophilic. These present findings are more specific.

3. GC% of protein coding part of entire genome (COD GC%) is positively (highly significant) associated with the genome GC content (Table 3). This implies that GC content is large or small according as COD GC% is rich or poor.
4. Each of the GC% of the 1st, 2nd and 3rd position within a codon for protein coding part is directly (highly significant) associated with the genome GC content (Table 3). This indicates that the genome GC content will be large if each of the GC% of the 1st, 2nd and 3rd position within a codon for protein coding part is rich.
5. Genome GC content is directly associated each with tRNA (partially significant) and tRNA GC1% (highly significant, $P = 2.22e^{-16}$), but inversely with tRNA GC2% (highly significant, $P = 3.10e^{-10}$) (Table 3). These imply that the genome GC content will be large separately with the increase of number of total tRNA genes within genome, lower boundary of GC% for tRNA encoding genes and with the decrease of upper boundary of GC% for tRNA encoding genes.
6. Genome GC content is inversely (highly statistically significant, $P = 9.92e^{-6}$) associated each with the average frequency (within 100) of non-polar aliphatic and negatively charged r group containing amino acids usages (Table 3). These indicate that the genome GC content will be large if each of the average frequency (within 100) of non-polar aliphatic and negatively charged r group containing amino acids usages will be low, and vice-versa. These present results are completely different from earlier findings [19,20].
7. Taxonomy is also associated with the variance of genome GC content (Table 3), indi-cating that the genome GC content changes with the type of taxonomy of the organisms. That is the variation of genome GC content exists within the different types of taxonomy (supports the findings of [7,8]).
8. Variance of genome GC content is associated positively (significant) with GC1% of the 1st and negatively with GC2% of the 2nd position within a codon for protein coding part (Table 3), respectively. This indicates that genome GC content variance will be small if the GC1% is small and GC2% is large.
9. Variance of genome GC content is inversely associated each with tRNA and tRNA GC1%, but directly with tRNA GC2% (each is highly significant, $P < 0.001$) (Table 3). The relation of tRNA, tRNA GC1% and tRNA GC2% with the variance of genome GC content is completely reverse to its mean. These imply that the variance genome GC content will be small separately with the increase of number of total tRNA genes within genome, lower boundary of GC% for tRNA encoding genes and decrease with the upper boundary of GC% for tRNA encoding genes.
10. Genome GC content variance is associated (highly significant, $P < 0.0001$) negatively with the average frequency (within 100) of non-polar aliphatic and positively with aro-matic r group containing amino acids usages (Table 3), respectively. These indicate that the variance of genome GC content will be small if the average frequency (within 100) of non-polar aliphatic r group containing amino acids usages will be high and aromatic r group containing amino acids usages will be low.

In early researches, it has been pointed that the variations of genome size and GC content are non-constant [6,12], yet only the mean models have been derived based on constant variance assumption. In the present study, however, both the mean and the variance models of genome size and GC content have been derived (Sections 4.1, 4.2). Some of the present results are little cited in the literature. For example, the present analysis has first derived the

determinants of the variances of both the genome size and GC content (Sections 4.1, 4.2). This article presents a clear interpretation about the determinants of genome size and GC content. It tries to remove some conflicts of earlier findings (as in above). Most of the estimated effects are highly statistically significant. Only a few partially significant effects are included in the model for better fitting. Standard deviations of all the estimated effects are very small, indicating that the estimates are stable [29]. The present study may provide substantial new information to explain both the mean and the variance models of genome size and GC content.

The findings in Section 4 can be explained in the biological path-way. A few possible explanations on the relationships between GC content, genome size and survival in terrestrial environment are given below.

¤ In Tables 2 and 3, it is identified that the large genomes to be GC rich and small genomes to be GC poor. Biologically, this may be explained as follow.

DNA is the double helical master molecule carrying information for expression of life through transcription and translation. The building blocks, i.e. four nucleotides (A,T, G,C) stack one over the other providing extension of genome size commensurate with the biological requirement of different organisms as well as of their horizontal pairings as AT and GC for stabilization of DNA molecule. Of these, GC by virtue of triple hydrogen bindings provides more stability than AT with only double hydrogen binding. Thus, it is expected that GC% needs to increase with the increase of genome size for structural stability. The same logic can be extended for preponderance of GC at 1,2 and 3 position of codon for avoiding/ reducing mismatching chances between mRNA codon and tRNA anticodon vis-a-vis possible translational error during protein synthesis due to stability caused by triple hydrogen binding. This also explains high GC content in coding region and genome size. However, it may be reiterated that the nature appears to have given equal weight to all four nucleotides as for as creation of genetic code for different amino acids is concerned. During the course of selection, chemical stability of GC over AT (U) has probably succeeded.

¤ The relationship of genome size and GC content can be explained in other way. Preponderance of GC in large size genome is also important from the viewpoint of or-chestrated expression, i.e. switch on and off, of genes as per requirement of situation for survival of organisms. This is achieved through methylation of cytosine in GC pair. The GC methylation occurs both in gene promoter sequences and sequences of gene per se.

¤ In Table 2, it is identified that the mean genome size is significantly higher in the terrestrial than in aquatic and is indifferent in host. Biologically, this may be viewed as follow.

The terrestrial habitat harbours extreme and diverse conditions in comparison to aquatic and host conditions, thereby requiring large genome size, which can remain stable only if GC content increases.

¤ In Table 2, it is identified that the mean genome size is negatively associated each with the average frequency (within 100) of non-polar aliphatic, aromatic and positively charged r group containing amino acids usages. Also in Table 3, it is identified that the mean genome GC content is negatively associated each with the average frequency (within 100) of non-polar aliphatic and negatively charged r group containing amino acids usages. Biologically, this may be illustrated as follow.

Amino acids with polar side groups carry more information than amino acids with non-polar aliphatic side groups for secondary and tertiary folding and performance of diverse physiological functions of the protein. This may be the reason that the genome size and GC content have retained exceedingly high codons for polar amino acids, resulting in a positive correlation with polar amino acids.

Because of the above reasons, GC content finds positive correlations with the genome size of different organism thriving in diverse terrestrial habitat. Further, GC content pro-vides stability and integrity to large fragile DNA molecules commensurate with genome size, its preponderance in coding (gene) region responsible for regulation (expression / sup-pression / silencing) of functional genes/ transposable elements in different situations such as diverse environmental conditions,

organisms, organs and tissues and aging (Juvenility vs maturity gradient). In codons, GC content avoids translational errors of functional genes.

To fill the gaps in the genetic research literature, this study derives the relationships of genome size, GC content, codon and amino-acid usage. The mathematical models (in Tables 2 and 3) in this report show the relationships of genome size and GC con-tent. The models reported here illuminate the complex relationships. Fortunately, a true mathematical model can open the truth that is covered by the complex relationships.

Our results, though not completely conclusive, are revealing:

¤ Our findings confirm many previous research findings (Section 4).

¤ An important conclusion has to do with the use of earlier used statistical models. While further research is called for, we find that a joint log-normal and gamma models are much more effective than either traditional simple, multiple regression and Log-Gaussian models (with constant variance), because they better fit the data. In short, research should have greater faith in these results than those emanating from the simple, multiple regression and Log-Gaussian (with constant variance) models.

ACKNOWLEDGEMENTS

The authors are very much indebted to the Editor and the referees who have provided valuable comments to improve this paper.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Wang HC, Susko E. Roger AJ. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. Biochem. Biophys. Res. Commun. 2006;342:681-684.
2. Marashi SA, Ghalanbor Z. Correlation between genomic GC levels and optimal growth temperatures are not robust. Biocem. Biophys. Res. Commun. 2004; 325:381-383.
3. Basak S, Ghosh TC. On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biocem. Biophys. Res. Commun. 2005;330:629-632.
4. Basak S, Mandal S, Ghosh TC. Correlation between genomic GC levels and optimal growth temperature: Some comments. Biocem. Biophys. Res. Com-mun. 2005; 327:969-970.
5. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. Genome GC level, optimal growth temperature, and genome size in prokaryotes. Biochem. Biophys. Res. Commun. 2006;347:1-3.
6. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes, Proc. Natl. Acad. Sci. USA. 2004;101(9): 3160-3165.
7. Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA. 1962;48:582-592.
8. Hildebrand1 F, Meyer A, Eyre-Walker A. Evidence of Selection upon Ge-nomic GC-Content in Bacteria. PLoS Genetics. 2010; 6(9):1-9.
9. Bernardi G, Bernardi G. Codon usage and genome composition. J. Mol. Evol. 2010;22:363-365.
10. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. Annu Rev Genet, 2004;38:771-792.
11. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, et al. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science. 2006; 314:267-272.
12. Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S. A universal trend of amino acid gain and loss in protein evolution. Nature. 2005;433:633-638.
13. Sueoka N. Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harb Symp Quant Biol. 1961;26:35-43.
14. Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA. 1988;85:2653-2657.

15. Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA. 1987;84:166-169.
16. Osawa S, Ohama T, Yamao F, Muto A, Jukes TH, Ozeki H, Umesono K. Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. Proc. Natl. Acad. Sci. USA. 1988;85:1124-1128.
17. Foster PG, Jermiin LS, Hickey DA. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 1997;44:282-288.
18. Wilquet V, Van de Casteele M. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res Microbiol. 1999;150:21-32.
19. Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18:291-294.
20. Gu X, Hewett-Emmett D, Li WH. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica. 1988; 102:383-391.
21. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 2004;573:73-77.
22. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: A reply to Marashi and Ghalanbor. Biochem. Biophys. Res. Commun. 2005;330:357-360.
23. Collins DW, Jukes TH. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J. Mol. Evol. 1993;36:201-213.
24. Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biology. 2001;2(4):1-14.
25. Chatterjee S, Price B. Regression Analysis by Examples, 3rd ed., Wiley and Sons, New York; 2000.
26. McCullagh P, Nelder JA. Generalized Linear Models, Chapman & Hall, London; 1989.
27. Myers RH, Montgomery DC, Vining GG. Generalized Linear Models with Applications in Engineering and the Sciences. John Wiley & Sons, New York; 2002.
28. Palta M. Quantitative Methods in Population Health: Extensions of Ordinary Regression. Wiley and Sons, New York; 2003.
29. Lee Y, Nelder JA, Pawitan Y. Generalized Linear Models with Random Effects (Unified Analysis via H-likelihood), London: Chapman and Hall; 2006.
30. Box GEP, Cox DR. An analysis of transformations. J. R. Statist. Soc. B. 1964; 26:211-252.
31. Firth D. Multiplicative errors: Log-normal or gamma? J. R. Statist. Soc. B, 1988;50: 266-268.
32. Das RN. Discrepancy in fitting between log-normal and gamma models: An illustration. Model Assisted Statistics and Applications. 2012;7(1):23-32.
33. Das RN, Lee Y. Log normal versus gamma models for analyzing data from quality-improvement experiments. Quality Engineering. 2009;21(1):79-87.
34. Das RN, Park JS. Discrepancy in regression estimates between Log-normal and Gamma: Some case studies. J. Applied Statistics. 2012;39(1):97-111.
35. Box GEP. Signal-to-Noise Ratios, Performance Criteria and Transformations (with discussion). Technometrics. 1988;30: 1-40.
36. Nelder JA, Lee Y. Generalized linear models for the analysis of Taguchi-type experiments. Applied Stochastic Models and Data Analysis. 1991;7:107-120.
37. Lee Y, Nelder JA. Generalized Linear models for the analysis of quality improvement experiments. Can. J. Statist. 1998;26:95-105.
38. Lee Y, Nelder JA. Robust design via generalized linear models. J. Qual. Tech. 2003;35:2-12.
39. Lesperance ML, Park S. GLMs for the analysis of robust designs with dynamic characteristics. J. Qual. Tech. 2003;35: 253-263.

40. Qu Y, Tan M, Rybicki L. A unified approach to estimating association measures via a joint generalized linear model for paired binary data. Commun. Statist. Theory Meth. 2000;29:143-156.

41. Wolfinger RD, Tobias RD. Joint estimation of location, dispersion, and random effects in robust design. Technometrics. 1998;40: 62-71.

42. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer-Verlag, USA; 2001.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=976&id=32&aid=9755