Laws of genome nucleotide composition

Zhang Zhang[1,2,3,*]

[1] National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[2] China National Center for Bioinformation, Beijing 100101, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author (zhangzhang@big.ac.cn).

Running head: Laws of genome nucleotide composition

Abstract

Genome nucleotide composition is of fundamental significance in molecular evolution, genome organization and synthetic biology. Albeit studied for decades, it remains unclear whether there is any theoretical law underlying variable genome nucleotide composition across different species. From the mathematical viewpoint, here we propose three laws of genome nucleotide composition and validate their effectiveness on a large collection of empirical genome sequences across three domains of life. The three laws together provide a unifying framework that is able to unravel the mystery of genome-wide nucleotide composition variation across diverse species, paving the way towards a new era with quantitative insights for deciphering basic principles of life and further advancing theoretical biology.

One-Sentence Summary: The three laws of genome nucleotide composition provide a fundamental framework for molecular evolution, genome organization and synthetic biology.

Introduction

Genome nucleotide composition, one of the most important characteristics at the genome-wide level, is usually expressed in terms of the proportions of four bases as well as their combinations in DNA molecule. It has been studied for decades (*1-3*) that genomes of different species are highly variable in their nucleotide composition (*4*, *5*), as demonstrated that guanine-plus-cytosine content varies widely with a broader range from ~20% to ~80% (*6-9*). A body of empirical evidence has further accumulated that heterogeneity of genome-wide nucleotide composition in different species associates closely with a variety of factors, such as, genome size (*10*), phylogeny (*11*, *12*), growth temperature (*13*), environment (*14*), origin of replication (*15*), transposable element (*16*), bacterial land colonization (*17*), codon/amino acid usage (*18-20*), cell fate control (*21*), and natural selection (*22-25*), etc. Although theoretical efforts have also been made in modeling the genome-wide nucleotide composition (*26-29*), it remains unclear whether there is any law underlying such variable genome nucleotide composition across different species. Theoretically, such law(s) would be desirable to be used as a fundamental framework for better understanding genome composition dynamics, molecular evolution, genome organization and synthetic biology. Built upon previous findings, here we propose three laws of genome nucleotide composition in a mathematical manner and demonstrate their effectiveness to formulate diverse genome nucleotide compositions in a large collection of 17,873 complete genome sequences across three domains of life.

First law: the law of base pairing

The first law is Chargaff's rules (*30-32*) that adenine (A) pairs with thymine (T) and guanine (G) pairs with cytosine (C), leading to $P(\mathrm{A}) = P(\mathrm{T})$ and $P(\mathrm{G}) = P(\mathrm{C})$ (Eqs. 1 and

2), where P is the proportion (probability) of any base as well as their combination. Such base pairing symmetry in any double-strand genome and each single strand corresponds to the Chargaff's first parity rule and second parity rule, respectively. With biological, chemical and physical significances, the rules played a crucial role in the discovery of the double helix structure of DNA in 1953 (*33*) and laid profound foundations in advancing molecular biology and genomics (*34*). Despite the debate on the Chargaff's second parity rule (*35, 36*), the first law holds valid as testified by nearly perfect linear regression in a wide diversity of genomes across the three domains of life (Fig. 1A and 1B).

$$P(\mathrm{A}) = P(\mathrm{T}) \tag{1}$$

$$P(\mathrm{G}) = P(\mathrm{C}) \tag{2}$$

Second law: the law of base equality

The second law states the base equality between purines (A and G) and pyrimidines (T and C). As a consequence of the base-pairing nature of DNA double helix, a 1:1 stoichiometric ratio of purines (R) and pyrimidines (Y) can be deduced. In other words, the proportion of R approximates the proportion of Y, namely, $P(\mathrm{R}) = P(\mathrm{Y})$. Additionally considering guanine-plus-cytosine (S or GC) and adenine-plus-thymine (W) contents, $P(\mathrm{A}) + P(\mathrm{G}) + P(\mathrm{T}) + P(\mathrm{C}) = 100\%$ can be further expressed as $P(\mathrm{R}) + P(\mathrm{Y}) = 2P(\mathrm{R}) = 2P(\mathrm{Y}) = P(\mathrm{S}) + P(\mathrm{W}) = 100\%$ (Eqs. 3 and 4).

$$P(\mathrm{R}) = P(\mathrm{Y}) = 50\% \tag{3}$$

$$P(\mathrm{S}) + P(\mathrm{W}) = 100\% \tag{4}$$

Conforming with the second law as well as previous findings (*4, 7, 8*), $P(\mathrm{R})$ is observed to approximate $P(\mathrm{Y})$ and fluctuate around 50% in diverse genomes belonging to Archaea, Bacteria and Eukarya (Fig. 1C and 1D). Meanwhile, considering that independence of two variables means that the occurrence of one variable does not affect the probability of the

other, S, varying from ~20% to ~80%, is believed to be statistically independent from R, as indicated by linear regression slope very close to the optimum value of zero and intercept near 0.5 (Fig. 1D).

Third law: the law of base composition

The third law expresses the principle of base composition. Suppose that the universal set of the four bases is $\Omega=\{A, T, G, C\}$ and X and Y are two subsets of Ω, there are three common operations in set theory: (1) union of X and Y, viz., $X \cup Y$, is the set of all elements that are a member of X or Y or both; (2) intersection of X and Y, viz., $X \cap Y$, is the set of all elements that are members of both X and Y; (3) complement of X relative to Ω, viz., X^c, is the set of all members that are not members of X. Thus, GC and purine contents can be denoted as $S = G \cup C$ and $R = A \cup G$, respectively. Likewise, $W = A \cup T = S^c$ and $Y = T \cup C = R^c$. Because S and R form an independent pair as mentioned above, thus $P(S \cap R)$ can be quantitatively expressed as $P(S)$ multiplied by $P(R)$, namely, $P(S \cap R) = P(S) \times P(R)$, which is also applicable to the other three pairs: S^c and R, S and R^c, and S^c and R^c (for details see our previous study (*28*)). As a result, the proportion of each base can be quantitatively formulated as

$$P(A) = P(S^c \cap R) = P(S^c) \times P(R), \tag{5}$$

$$P(T) = P(S^c \cap R^c) = P(S^c) \times P(R^c), \tag{6}$$

$$P(G) = P(S \cap R) = P(S) \times P(R), \tag{7}$$

$$P(C) = P(S \cap R^c) = P(S) \times P(R^c), \tag{8}$$

where $P(S^c) = P(W) = 1 - P(S)$ and $P(R^c) = P(Y) = 1 - P(R)$ according to Eqs. 3-4. Strikingly, if $P(R) = 0.5$, then $P(A) = P(T) = P(S^c)/2$ and $P(G) = P(C) = P(S)/2$, which is a special case equivalent to the first law or Chargaff's rules. Based on our empirical datasets,

the third law is effective to quantify the base proportions very close to the observed ones, as signified by squared correlation coefficients very close to the optimum value of 1.0, linear regression slopes near the optimum value of 1.0 and intercepts approaching the optimum value of zero (Fig. 1E to 1H).

Concluding thoughts

The first law of base pairing uncovers the complementary nature of DNA that is essential for its structure, stability and function, the second law of base equality reveals the equality relationship between purines and pyrimidines as well as the independence relationship between GC content and purine content, and the third law deduces the mathematical principle of each base composition that is of great utility for designing genomes with any specified composition. Together, the three laws provide a unifying framework for studying genome nucleotide composition (Fig. 2), which is not only able to unravel the mystery of various genome-wide nucleotide compositions across diverse organisms as validated in large-scale empirical genome sequences across the three domains of life, but also to detect organisms with unusual nucleotide composition that potentially experience complex evolutionary processes adapted to extreme environments (a case example is *Candidatus* Chazhemtobacterium aquaticus Ch65 (*37*) with strong disparity of 14.97% G and 29.83% C in Fig. 1B).

Nowadays (and in the foreseeable future), we are drowning in the deluge of multi-omics data growing at exponential rates, thirsting for fundamental theories to explain or predict a range of biological phenomena and to derive general laws from a number of empirical observations and experiments. From this perspective, the three laws, presented here as the union of theoretical and empirical work, potentially pave the way towards a new era with

quantitative insights for studying genome organization and evolution, driving synthetic genome engineering and further advancing theoretical biology as a long-term objective on a par with theoretical physics.

Authors' contributions

ZZ conceptualized this idea, analyzed the data, plotted the figures, wrote the manuscript, acquired the funding support, and supervised the research. The author read and approved the final manuscript.

Acknowledgments

I sincerely thank Haipeng Li, Yalong Guo, Yong Zhang, Jun Yu, Songnian Hu, Jingfa Xiao and Lina Ma for their valuable discussions on this work as well as Shuai Jiang and Zhao Li for their assistance in data collection. This work was supported by the National Natural Science Foundation of China (32030021) and International Partnership Program of the Chinese Academy of Sciences (153F11KYSB20160008).

References

1. N. Sueoka, On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* **48**, 582-592 (1962).
2. S. Karlin, A. M. Campbell, J. Mrazek, Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**, 185-225 (1998).
3. A. Nekrutenko, W. H. Li, Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res* **10**, 1986-1995 (2000).
4. E. Haywood-Farmer, S. P. Otto, The evolution of genomic base composition in bacteria. *Evolution* **57**, 1783-1792 (2003).
5. S. Karlin, J. Mrazek, Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci U S A* **94**, 10227-10232 (1997).
6. A. Muto, S. Osawa, The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* **84**, 166-169 (1987).
7. J. Hu, X. Zhao, Z. Zhang, J. Yu, Compositional dynamics of guanine and cytosine content in prokaryotic genomes. *Res Microbiol* **158**, 363-370 (2007).
8. X. Zhao, Z. Zhang, J. Yan, J. Yu, GC content variability of eubacteria is governed by the pol III alpha subunit. *Biochem Biophys Res Commun* **356**, 20-25 (2007).

9. H. Wu, Z. Zhang, S. Hu, J. Yu, On the molecular mechanism of GC content variation among eubacterial genomes. *Biol Direct* **7**, 2 (2012).
10. H. Nishida, Evolution of genome base composition and genome size in bacteria. *Front Microbiol* **3**, 420 (2012).
11. A. O. Mooers, E. C. Holmes, The evolution of base composition and phylogenetic inference. *Trends Ecol Evol* **15**, 365-369 (2000).
12. E. R. Reichenberger, G. Rosen, U. Hershberg, R. Hershberg, Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol Evol* **7**, 1380-1389 (2015).
13. H. Musto *et al.*, Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett* **573**, 73-77 (2004).
14. K. U. Foerstner, C. von Mering, S. D. Hooper, P. Bork, Environments shape the nucleotide composition of genomes. *EMBO Rep* **6**, 1208-1213 (2005).
15. J. R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660-665 (1996).
16. S. Boissinot, On the Base Composition of Transposable Elements. *International journal of molecular sciences* **23**, (2022).
17. H. Wu, Y. Fang, J. Yu, Z. Zhang, The quest for a unified view of bacterial land colonization. *The ISME journal* **8**, 1358-1369 (2014).
18. R. Hershberg, D. A. Petrov, Selection on codon bias. *Annu Rev Genet* **42**, 287-299 (2008).
19. G. A. Singer, D. A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581-1588 (2000).
20. A. Garcia, S. Gonzalez, F. Antequera, Nucleosomal organization and DNA base composition patterns. *Nucleus* **8**, 469-474 (2017).
21. R. Pantier *et al.*, SALL4 controls cell fate in response to DNA base composition. *Mol Cell* **81**, 845-858 e848 (2021).
22. F. Hildebrand, A. Meyer, A. Eyre-Walker, Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **6**, e1001107 (2010).
23. H. Long *et al.*, Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol* **2**, 237-240 (2018).
24. M. Bulmer, The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897-907 (1991).
25. S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, H. H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci U S A* **101**, 3480-3485 (2004).
26. R. A. Elton, Theoretical models for heterogeneity of base composition in DNA. *J Theor Biol* **45**, 533-553 (1974).
27. G. A. Churchill, Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* **51**, 79-94 (1989).
28. Z. Zhang, J. Yu, Modeling compositional dynamics based on GC and purine contents of protein-coding sequences. *Biol Direct* **5**, 63 (2010).
29. D. R. Forsdyke, Complementary Oligonucleotides Rendered Discordant by Single Base Mutations May Drive Speciation. *Biological Theory* **16**, 237-241 (2021).
30. E. Chargaff, Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* **6**, 201-209 (1950).
31. E. Chargaff, Some recent studies on the composition and structure of nucleic acids. *J Cell Physiol Suppl* **38**, 41-59 (1951).

32. E. Chargaff, R. Lipshitz, C. Green, Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *J Biol Chem* **195**, 155-160 (1952).
33. J. D. Watson, F. H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
34. D. R. Forsdyke, J. R. Mortimer, Chargaff's legacy. *Gene* **261**, 127-137 (2000).
35. D. Mitchell, R. Bridge, A test of Chargaff's second rule. *Biochem Biophys Res Commun* **340**, 90-94 (2006).
36. Y. Guo, An overview on the DNA nucleotide compositions across kingdoms. *Gene Reports* **8**, 45-48 (2017).
37. V. V. Kadnikov, A. V. Mardanov, A. V. Beletsky, O. V. Karnachuk, N. V. Ravin, Complete Genome of a Member of a New Bacterial Lineage in the Microgenomates Group Reveals an Unusual Nucleotide Composition Disparity Between Two Strands of DNA and Limited Metabolic Potential. *Microorganisms* **8**, 320 (2020).

Figure Legends

Fig. 1. Genome-wide proportions of nucleotide composition across three domains of life.

(A) Proportion of adenine (A) vs. proportion of thymine (T); **(B)** Proportion of guanine (G) vs proportion of cytosine (C); **(C)** Proportion of guanine-plus-cytosine content (S) vs. difference between proportion of purine content (R) and proportion of pyrimidine content (Y); **(D)** Proportion of S vs. proportion of R; **(E-H)** Observed vs. estimated proportions for A, T, G, and C, respectively. Observed proportions of the four bases were derived from a large collection of 17,873 complete genome sequences in NCBI RefSeq, including 313 in Archaea, 17,289 in Bacteria, and 271 in Eukarya. Estimated proportions of the four bases were quantified according to Eqs. 5-8. In all panels, each point represents the composition for a specific genome. All detailed information were summarized in Supplementary Table S1.

Fig. 2. Schematic representation of laws of genome nucleotide composition.

The laws are illustrated by equations and operations of intersection ($\cap$), union ($\cup$) and complement (^{c}) on four bases—adenine (A), thymine (T), guanine (G) and cytosine (C), where P is the proportion (probability) of any base as well as their combination. The first law of base pairing states the complementary nature of DNA that leads to $P(A)=P(T)=P(W)/2$ and $P(G)=P(C)=P(S)/2$, where $W=A\cup T$, $S=G\cup C$, $R=A\cup G$, and $Y=T\cup C$. The second law of base equality reveals the quantitative relationships of $P(R)=P(Y)=50\%$ and $P(S)+P(W)=100\%$, where $R\cup Y$ or $S\cup W$ is the universal set $\Omega=\{A, T, G, C\}$, $R\cup W=\{A, T, G\}=C^c$ and $Y\cup S=\{T, G, C\}=A^c$. The third law deduces the

mathematical principle of each base composition given the independence relationship between S and R, viz., $P(G)=P(S \cap R)=P(S)P(R)$, $P(A)=P(W \cap R)=P(W)P(R)$, $P(C)=P(S \cap Y)=P(S)P(Y)$, and $P(T)=P(W \cap Y)=P(W)P(Y)$, where $W=S^c$ and $Y=R^c$.

Figures

Fig. 1

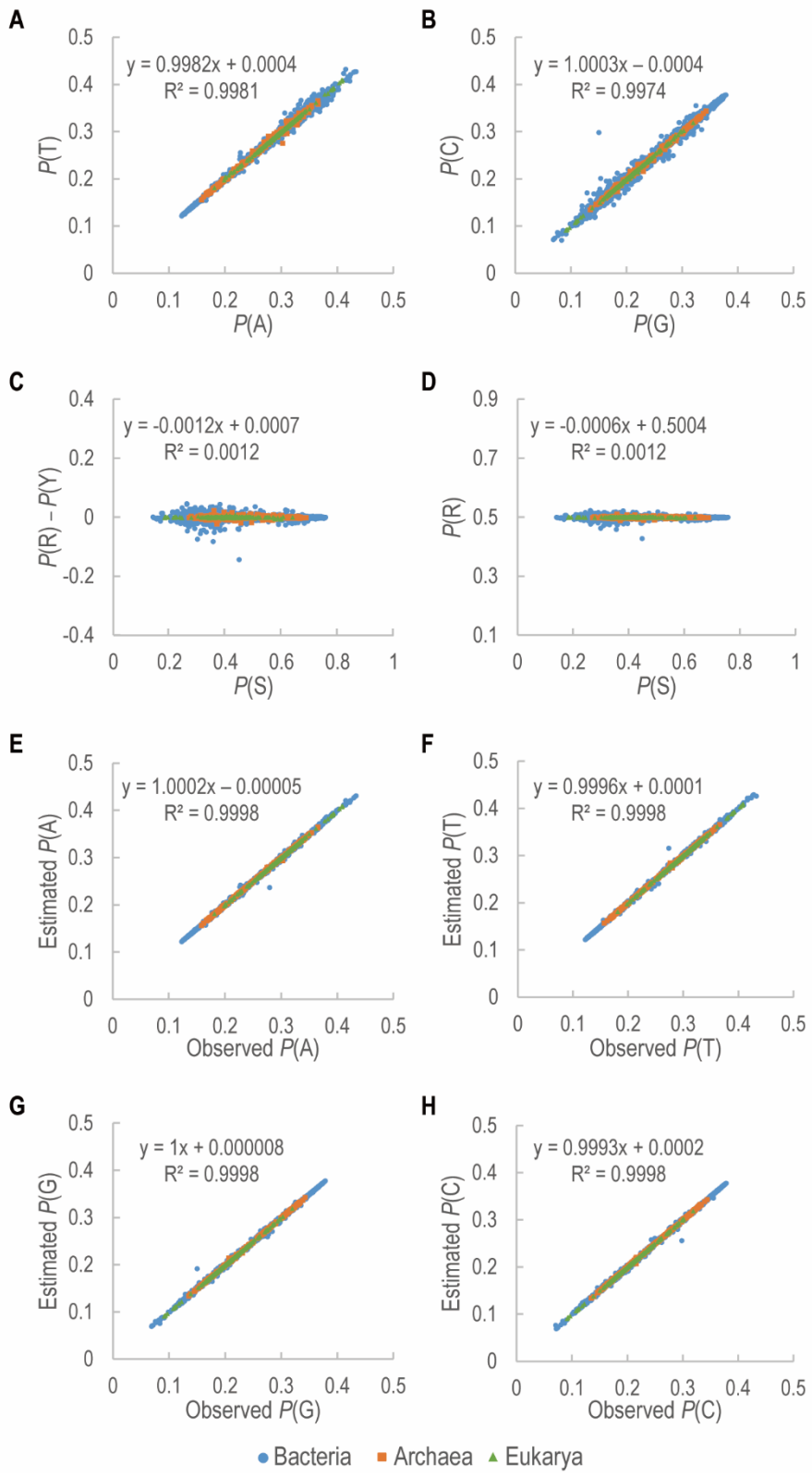
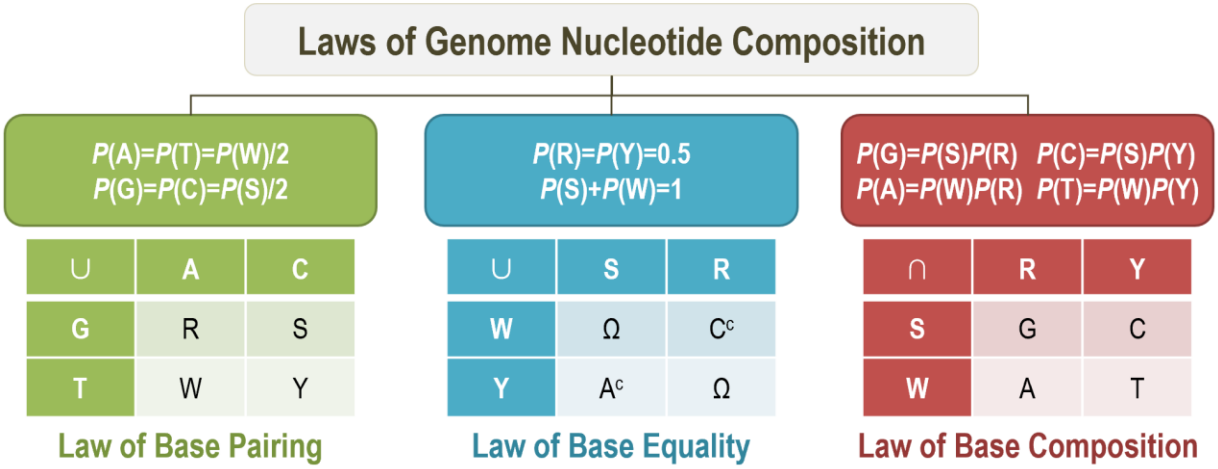
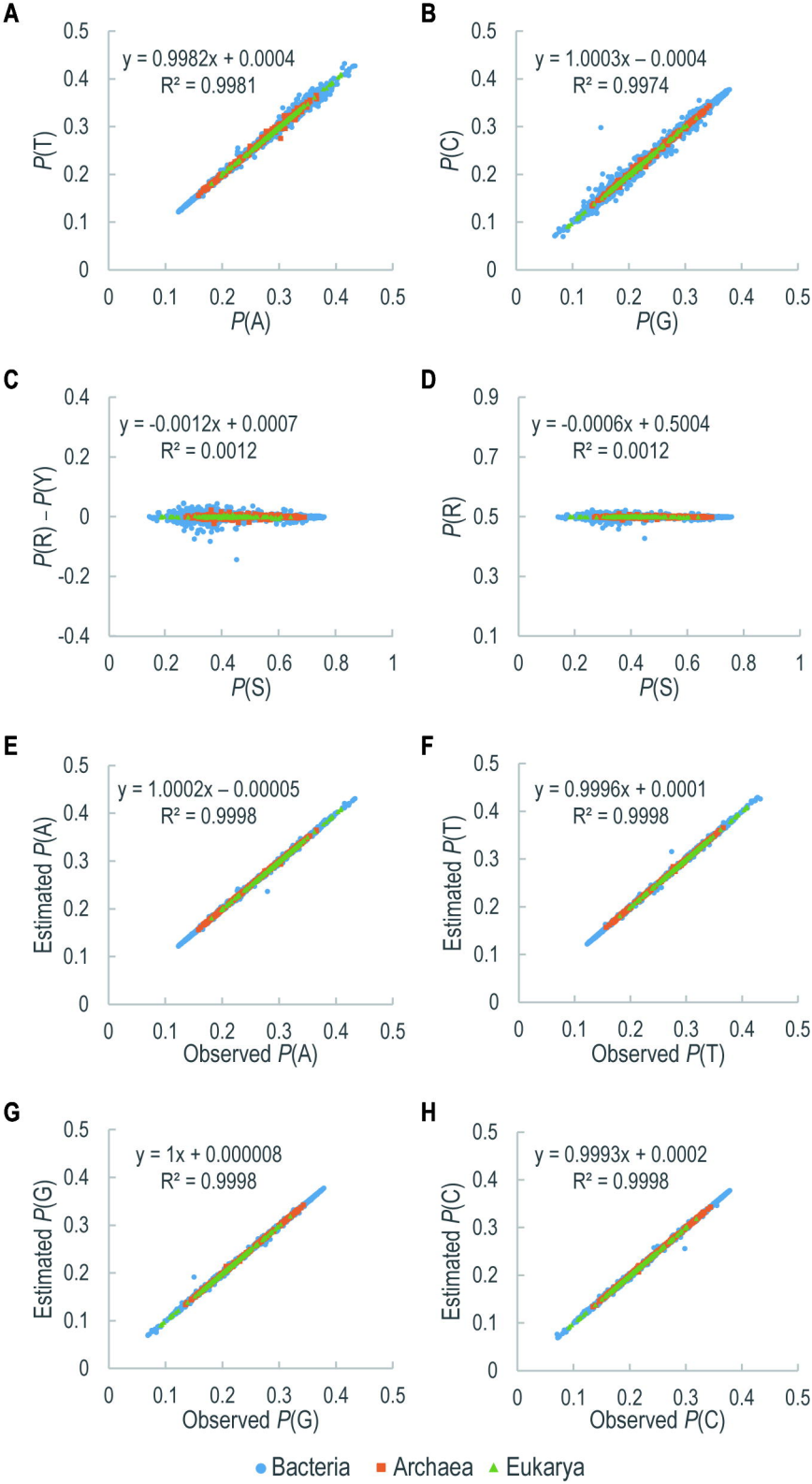

A
y = 0.9982x + 0.0004
R² = 0.9981
P(T)
P(A)
B
y = 1.0003x – 0.0004
R² = 0.9974
P(C)
P(G)
C
y = -0.0012x + 0.0007
R² = 0.0012
P(R) - P(Y)
P(S)
D
y = -0.0006x + 0.5004
R² = 0.0012
P(R)
P(S)
E
y = 1.0002x – 0.00005
R² = 0.9998
Estimated P(A)
Observed P(A)
F
y = 0.9996x + 0.0001
R² = 0.9998
Estimated P(T)
Observed P(T)
G
y = 1x + 0.000008
R² = 0.9998
Estimated P(G)
Observed P(G)
H
y = 0.9993x + 0.0002
R² = 0.9998
Estimated P(C)
Observed P(C)
Bacteria
Archaea
Eukarya

Fig. 2



Supplementary information

Table S1 Statistics of genome-wide nucleotide composition as well as other related meta information of 17,873 complete genome sequences across three domains of life.


A
y = 0.9982x + 0.0004
R² = 0.9981
P(T)
P(A)
B
y = 1.0003x – 0.0004
R² = 0.9974
P(C)
P(G)
C
y = -0.0012x + 0.0007
R² = 0.0012
P(R) – P(Y)
P(S)
D
y = -0.0006x + 0.5004
R² = 0.0012
P(R)
P(S)
E
y = 1.0002x – 0.00005
R² = 0.9998
Estimated P(A)
Observed P(A)
F
y = 0.9996x + 0.0001
R² = 0.9998
Estimated P(T)
Observed P(T)
G
y = 1x + 0.000008
R² = 0.9998
Estimated P(G)
Observed P(G)
H
y = 0.9993x + 0.0002
R² = 0.9998
Estimated P(C)
Observed P(C)
Bacteria
Archaea
Eukarya

Laws of Genome Nucleotide Composition

$P(A)=P(T)=P(W)/2$
$P(G)=P(C)=P(S)/2$

∪	A	C
G	R	S
T	W	Y

Law of Base Pairing

$P(R)=P(Y)=0.5$
$P(S)+P(W)=1$

∪	S	R
W	Ω	C^c
Y	A^c	Ω

Law of Base Equality

$P(G)=P(S)P(R)$ $P(C)=P(S)P(Y)$
$P(A)=P(W)P(R)$ $P(T)=P(W)P(Y)$

∩	R	Y
S	G	C
W	A	T

Law of Base Composition