PLOS ONE

RESEARCH ARTICLE

Randomness in Sequence Evolution Increases over Time

Guangyu Wang[1,2,3], Shixiang Sun[1,2,3], Zhang Zhang[1,2]*

1 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing 100101, China, **2** BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing 100101, China, **3** University of Chinese Academy of Sciences, Beijing 100049, China

* zhangzhang@big.ac.cn

Abstract

The second law of thermodynamics states that entropy, as a measure of randomness in a system, increases over time. Although studies have investigated biological sequence randomness from different aspects, it remains unknown whether sequence randomness changes over time and whether this change consists with the second law of thermodynamics. To capture the dynamics of randomness in molecular sequence evolution, here we detect sequence randomness based on a collection of eight statistical random tests and investigate the randomness variation of coding sequences with an application to *Escherichia coli*. Given that core/essential genes are more ancient than specific/non-essential genes, our results clearly show that core/essential genes are more random than specific/non-essential genes and accordingly indicate that sequence randomness indeed increases over time, consistent well with the second law of thermodynamics. We further find that an increase in sequence randomness leads to increasing randomness of GC content and longer sequence length. Taken together, our study presents an important finding, for the first time, that sequence randomness increases over time, which may provide profound insights for unveiling the underlying mechanisms of molecular sequence evolution.

Citation: Wang G, Sun S, Zhang Z (2016) Randomness in Sequence Evolution Increases over Time. PLoS ONE 11(5): e0155935. doi:10.1371/journal.pone.0155935

Editor: Yu Xue, Huazhong University of Science and Technology, CHINA

Received: March 31, 2016

Accepted: May 6, 2016

Published: May 25, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The co-author Zhang Zhang is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

Introduction

The second law of thermodynamics states that a system tends to progress in the direction of increasing entropy [1], where a system in this context includes engineered devices as well as biological organisms and entropy is a measure of randomness; that is to say, a system naturally progresses from nonrandomness to randomness [2]. Consistently, evidence has accumulated that the diversity and complexity in biology tend to increase in any evolutionary system, agreeing well with the second law of thermodynamics [3–7] that randomness never decreases over time. At the molecular level, genome sequences during evolution evolve toward incorporating more intricate mechanisms, indicative of increasing entropy and complexity. Additionally, aging is at least partially due to an accumulation of errors in DNA [8], which can be also explained by an increase in randomness. Considering that cancer can be considered as an

evolutionary process [9, 10], mutations and epigenetic imbalances during cancer progression can lead to randomness increase [11, 12], which also consists with the second law of thermodynamics. Therefore, characterizing the dynamics of molecular sequence randomness is of great significance for providing profound insights in unveiling the underlying mechanisms in molecular sequence evolution.

Over the past several years, efforts have been devoted to detecting randomness on molecular sequences primarily at the protein level [13–20]. However, it remains unknown whether DNA sequence randomness changes over time and whether this change consists with the second law of thermodynamics. Specifically, previous studies converted amino acid sequences into bit sequences, based on different groupings of amino acids according to their physicochemical properties, such as size, hydrophobicity, charge, polarity, mass, etc. However, they adopted different physicochemical properties for conversion of amino acid sequences into bit sequences, thus lacking a widely accepted conversion that can be used for randomness detection. In addition, previous studies ignored the degeneracy of the genetic code, that is, amino acids are encoded by different *n*-fold degenerate codons that often have completely different features. For example, CGN (N = A, T, G, C) and AGR (R = A, G) encode Arg, but the former presents higher GC content than the latter.

Based on our previous studies [21–25], codons are not randomly allocated in the genetic code, which can be divided into two halves in a more straightforward and informative manner (Table 1), viz., pro-robustness half (PRH) and pro-diversity half (PDH) that represent robustness and diversity, respectively. Specially, codons in PRH are robust to nucleotide changes at the 3rd codon position (cp3) since they do not provoke the amino acid change (e.g., CCN codes for Pro, where N represents any nucleotide). Conversely, codons in PDH are sensitive to nucleotide changes at cp3; nearly most changes between purines and pyrimidines at cp3 lead to amino acid change (e.g., GAR codes for Glu and GAY codes for Asp, where R = purines and Y = pyrimidines). Although there are three amino acids (Arg, Leu and Ser) encoded by six-fold degenerate codons, they are distributed across the two halves, playing important balancing roles for error minimization [25]. Considering that robustness and diversity are two important features, therefore, it would be desirable to detect sequence randomness based on PDH and PRH and investigate whether a sequence is able to keep a balance between robustness and diversity. As molecular sequences accumulate mutations during evolutionary process, will sequences change the degree of randomness over time? Is this change consistent with the second law of thermodynamics, that is, sequence randomness increases over time?

To address these issues, here we investigate molecular sequence randomness based on a collection of eight statistical random tests. The availability of multiple strains' genome sequences

Table 1. The content-centric re-organization of the genetic code.

		1st base			
		A	T	G	C
	A	AAR(K)	TAR(St)	GAR(E)	CAR(Q)
		AAY(N)	TAY(Y)	GAY(D)	CAY(H)
2nd base	T	ATR(I, M)	TTR(L)	**GTN(V)**	**CTN(L)**
		ATY(I)	TTY(F)		
	G	AGR(R)	TGR(St, W)	**GGN(G)**	**CGN(R)**
		AGY(S)	TGY(C)		
	C	**ACN(T)**	**TCN(S)**	**GCN(A)**	**CCN(P)**

Note: N represents any nucleotide. R represents A and G. Y represents T and C. St indicates stop codon.

doi:10.1371/journal.pone.0155935.t001

for a given species provides opportunity to systematically track sequence randomness over time as genes presenting in all related strains are believed to be evolutionarily ancient and those presenting in individual strains are relatively young [26, 27]. Therefore, we collect a total of 61 *Escherichia coli* strains and explore the sequence randomness in the context of pan-genome where genes are classified into different groups according to their presence in different number of strains. As essential genes are more evolutionarily conservative and ancient than non-essential genes [27], we also perform similar analysis by grouping genes based on gene essentiality. We further investigate GC content and sequence length that are in close association with sequence randomness.

Methods

Conversion of coding sequences into bit sequences

Following by previous studies [14, 19, 20], biological sequences are converted into bit sequences, which is of practical significance for making randomness detection doable that can rely on many empirical statistical tests (such as The Runs Test, The Random Walker Test and The Serial Test). According to our previous studies [21–24], the genetic code can be re-organized based on both GC and purine contents and accordingly divided into two halves (Table 1), viz., PRH and PDH. Based on these two halves, coding sequences can be converted into bit sequences, where '0' represents a codon in PRH and '1' represents a codon in PDH.

Randomness testing of bit sequences

A bit sequence is composed of a series of '0' and '1' [28]. Various statistical tests have been proposed to test a null hypothesis that biological bit sequences are random [13, 14, 16, 17, 20, 28–30]. Among them, the National Institute of Standards and Technology (NIST) 800–22 Statistical Test Suite is widely used for random sequence testing. The NIST Statistical Test Suite includes sixteen tests to assess the randomness of binary sequences and each test focuses on a particular characteristic of binary random sequence (S1 Table). Since some tests require sequences longer than 10^5 (which cannot be always satisfied for sequences in prokaryotes) and thus are inapplicable in biological sequences, we adopt a total of 8 statistical tests (viz., the Frequency Test, the Cumulative Sums Test, the Cumulative Sums Test Reverse, the Runs Test, the Discrete Fourier Transform Test, The Non-overlapping Template Matching Test, The Serial Test, The Approximate Entropy Test; see details in S1 Table), to examine the randomness of coding sequences.

As there are 8 statistical tests used for randomness detection, an 8-dimension vector is employed to describe a sequence, where each dimension represents a *P*-value that is derived from a randomness test. For any given coding sequence *X*, its general randomness vector R_x is formulated as

$$R_x = (S_x^1, S_x^2, \cdots, S_x^8), \tag{1}$$

where S_x^i is the rounded value of negative e natural logarithm of *P*-value in the i^{th} random test.

Since any sequence can be represented as an 8-dimension randomness vector, we developed a two-step clustering algorithm [30] based on randomness vectors to cluster sequences into different groups. The first step is to measure the similarity of different sequences using log-likelihood distances and then to cluster sequences into multiple groups with a maximized log-likelihood function. The second step is to further cluster groups by a standard agglomerative clustering method, i.e., comparing their distances to a threshold, and then to determine the best number of clusters based on Schwarz's Bayesian Inference Criterion (BIC) [31].

Data collection

All coding sequences of 61 *E. coli* strains were downloaded from NCBI (National Center for Biotechnology Information) [32]. Essential genes of *E. coli* were retrieved from DEG (Database of Essential Genes; http://www.essentialgene.org) [33]. To avoid stochastic errors, sequences that are less than 100bp were removed from analysis. Detailed information can be found at S2 Table.

Results and Discussion

Detection of randomness in molecular sequences

To fully capture sequence randomness, we integrate a collection of 8 statistical tests to detect randomness in molecular sequences according to a content-centric organization of the genetic code that splits codons into PDH and PRH (Table 1; see Methods). Based on these 8 tests, we devise an 8-demension vector, where each dimension represents a *P*-value derived from a randomness test. As a result, any sequence can be denoted as an 8-dimension randomness vector. We further develop a two-way clustering algorithm based on randomness vector and apply it to all sequences in *E. coli* MG1655, leading to two clusters with distinct statistical properties of randomness (Fig 1): the random cluster (n = 2,892) and the nonrandom cluster (n = 1,069). Detailed information of statistical testing on these two clusters is tabulated into S1 and S2 Tables. Considering the significance levels of 8 statistical tests, the random cluster has a higher percentage (>89.42%) of sequences whose statistical significance levels are larger than 0.1, clearly showing that the majority of sequences in this cluster have random patterns. Contrastingly, the nonrandom cluster contains a larger proportion of sequences that have significance levels less than 0.1 (Fig 1). Intriguingly, the runs test performs very similar in both clusters. This result is in agreement with a previous finding that the runs test is unable to detect

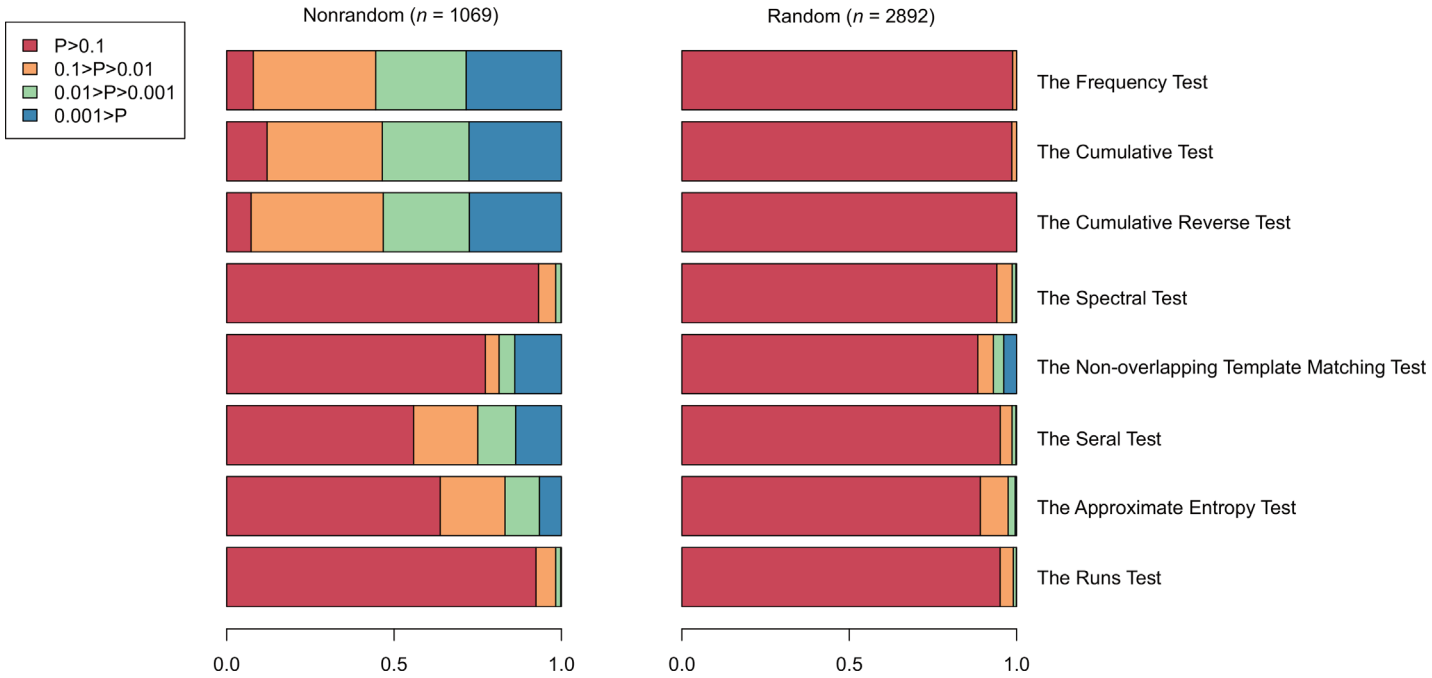

Fig 1. Random and non-random clusters based on 8 statistical tests. Bars are color-coded by different ranges of *P*-value.

doi:10.1371/journal.pone.0155935.g001

randomness in biological sequences [18]. Likewise, the spectral test yields similar performances in both clusters, indicating its incapability in detecting randomness biological sequences as well.

Investigation of sequence randomness over time

A pan-genome represents the union of all gene sets in all available strains of a species, which includes core genes that are present in all strains and dispensable genes that are present in multiple but not all strains [34]. As core genes are believed to be more ancient [26], therefore, we hypothesize that sequence randomness increases over time and core genes most likely contain more randomness.

To test this hypothesis, we collect 61 publically available *E. coli* genomes from [35] (S2 Table), perform the pangenome analysis and classify genes of *E. coli* MG1655 into five groups according to their presence in these 61 strains: Specific (that are genes presenting in 1–15 strains; $n = 111$), Medium-Specific (that are genes presenting in 16–30 strains; $n = 126$), Medium (that are genes presenting in 31–45 strains; $n = 315$), Medium-Core (that are genes presenting in 46–60 strains; $n = 1{,}347$) and Core (are genes presenting in all 61 strains; $n = 2{,}060$). Consistent with our expectations, the proportion of random genes is significantly different in these five groups (Chi-square test, $P<0.0001$; Fig 2) and grows gradually from specific genes to core genes, exhibiting 47.75% in specific genes and reaching the highest at 76.02% in core genes. As core genes are more ancient whereas specific genes are relatively young [26], these results clearly show that sequence randomness increases over time.

To further validate our results, we perform similar analysis by considering gene essentiality since essential genes that are critical for an organism's survival are thought to be more ancient [26, 36, 37]. We retrieve 527 essential genes and 2,956 non-essential genes from DEG

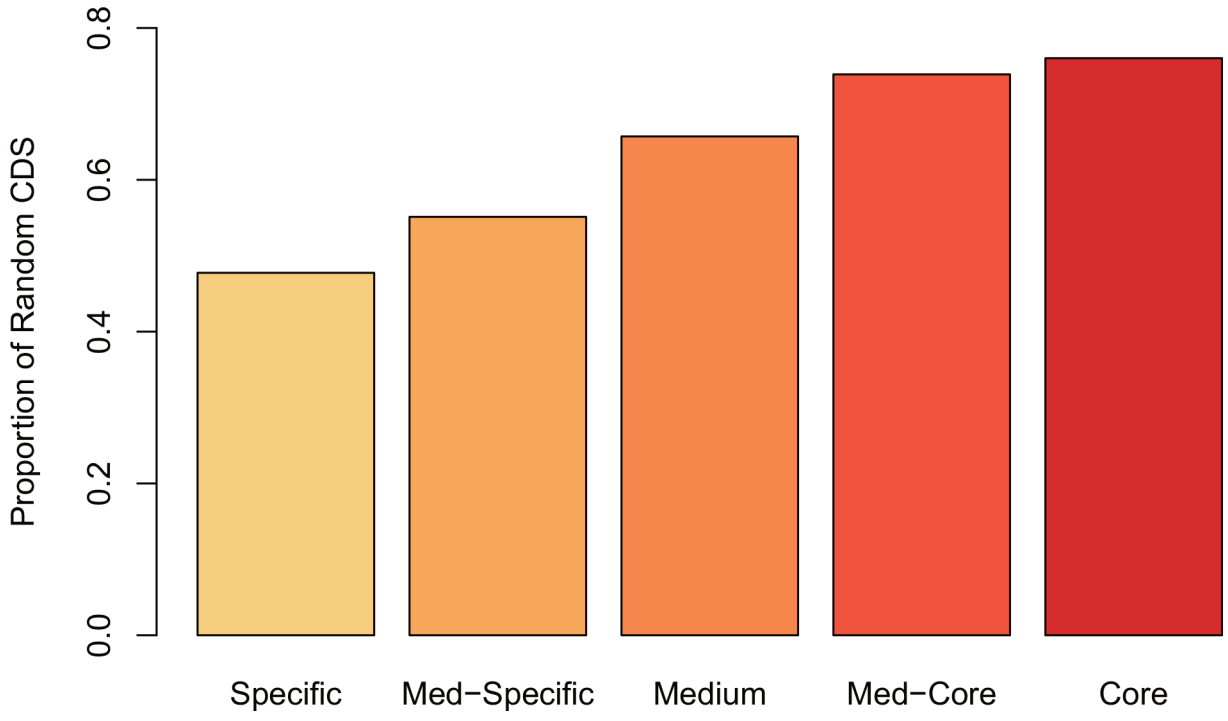

Fig 2. Proportion of random sequences in *E. coli*.

doi:10.1371/journal.pone.0155935.g002

Table 2. Statistical test of randomness between essential and non-essential genes.

Cluster	Essential Genes	Non-essential Genes	2×2χ^2*P*-value
Random	418	2120	<0.0001
Nonrandom	109	836	

doi:10.1371/journal.pone.0155935.t002

(Database of Essential Genes) [38]. In contrast to core genes that are derived from computational analysis, essential genes derived from DEG are identified by experimental approach. Consistently, a chi-square test of independence demonstrates that essential genes have a significant excess of random genes compared with non-essential genes ($P<0.0001$; Table 2). Ribosome proteins play a significant role in translation machinery and are believed to be more ancient than others [39]. We find that the majority of ribosome proteins (74%; S3 Table) are random, consisting well with our results that old genes are more random. Taken together, these results collectively demonstrate that randomness in molecular sequence increases over time. As randomness is detected based on grouping codons into PRH and PDH, an increase in sequence randomness during evolution leads to a uniform usage of codons in these two halves (Table 3), suggesting that sequences evolve toward achieving a good balance between robustness and diversity.

Variation of GC content and sequence length over time

As sequence randomness increase may provoke random nucleotide composition, we further test whether GC content becomes more random over time. If nucleotide composition in one gene is random, its GC content is expected to be around 0.514 (≈ (96–2) / (64×3–3×3) after removal of three stop codons). Therefore, we compare GC contents of random and nonrandom sequences and investigate their variations in the pan-genome context (Fig 3). Our results show that random sequences present GC contents significantly different from nonrandom sequences (t-test, $P<10^{-14}$; Fig 3); GC content in random sequences fluctuates around 0.51, always higher than that in nonrandom sequences, and intriguingly, such pattern is strikingly apparent in specific genes. This result is consistent well with a previous study that GC content in old human genes is around 0.51 [40]. With the increasing presence in more *E. coli* strains, the difference of GC content between random and nonrandom genes is radically reduced. These results show that GC content indeed goes random over time; GC content in random sequences varies within a very narrow range around 0.51, strongly indicating that random sequences achieve robustness-diversity balance.

It has been extensively reported that GC content is correlated positively with sequence length [41–43]. Therefore, we wonder whether sequence length varies over time (Fig 4).

Table 3. Percentage of genes that equally use codons in PDH and PRH.

Pan-genome group	Percentage*
Core	77.6%
Medium-Core	75.9%
Medium	65.4%
Medium-Specific	53.5%
Specific	46.8%

* *P*-value<0.05 (The frequency test)

doi:10.1371/journal.pone.0155935.t003

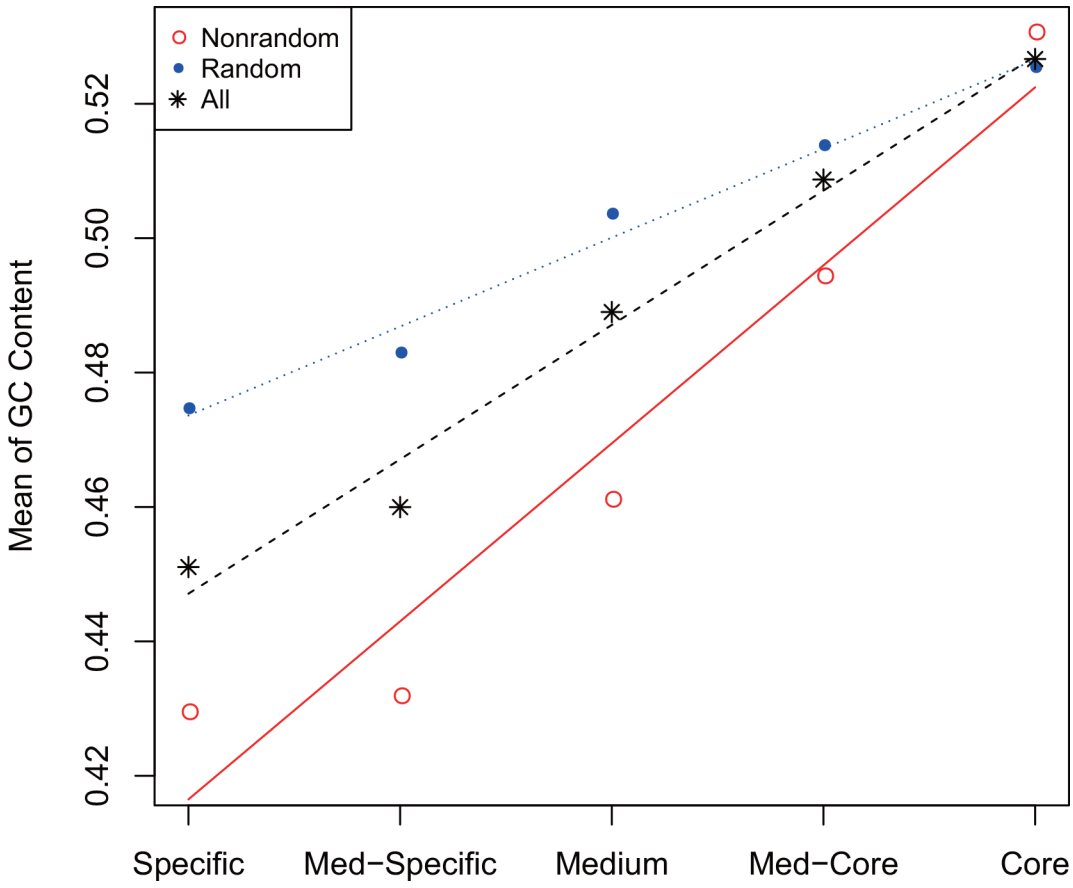

Fig 3. Variation of GC contents in the *E. coli* pan-genome. Random and nonrandom sequences are examined separately and each dot represents the average of GC content across a specific gene set.

doi:10.1371/journal.pone.0155935.g003

Agreeing with expectations, core genes are longer than specific genes and therefore, sequence length increases over time. In addition, random genes tend to be always longer than nonrandom genes. Collectively, with the increase of sequence randomness during evolution, sequences evolve toward higher GC content fluctuating at random and possess longer length, which is more pronounced in random sequences.

Conclusion

To fully picture the dynamics of randomness in molecular sequence evolution, here we detected sequence randomness in *E. coli* and explored randomness variation over evolutionary time based on the fact that in the context of pan-genome core genes are more ancient. Consistent with the second law of thermodynamics, we found that core genes are more random than specific genes, indicating that randomness in molecular sequence increases over time. Moreover, this conclusion still holds true when we considered gene essentiality, given that essential genes are more conservative and ancient than non-essential genes. To our knowledge, our study presents an important finding, for the first time, that randomness in sequence evolution increases over time, coupled with an increase in randomness of GC content and longer sequence length, which needs further validation in a wide range of species across three domains of life.

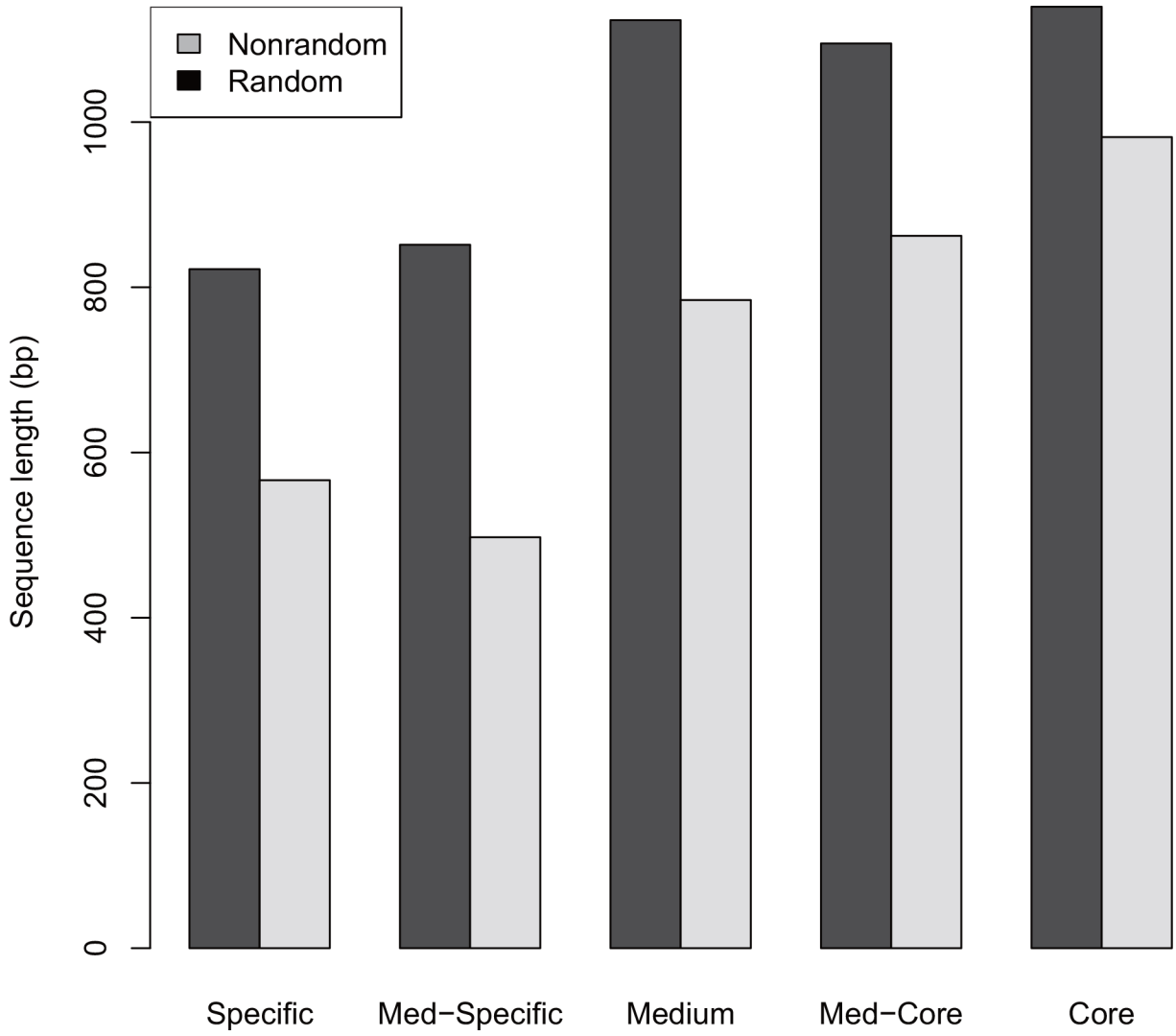

Fig 4. Length of coding sequences in the *E. coli* pan-genome. Random and nonrandom sequences are examined separately and each bar represents the average of sequence length across a specific gene set.

doi:10.1371/journal.pone.0155935.g004

Supporting Information

S1 Table. Characteristics of the NIST Statistical Tests.
(XLS)

S2 Table. 61 publically available *E. coli* genomes.
(XLS)

S3 Table. Ribosomal proteins in Pan-genome group and Random group.
(XLS)

S4 Table. Proportion of Each Test in Random Group.
(XLS)

S5 Table. Proportion of Each Test in Nonrandom Group.
(XLS)

Acknowledgments

We thank Dr. Lina Ma and Dawei Huang for valuable comments on this work.

Author Contributions

Conceived and designed the experiments: GW ZZ. Performed the experiments: GW. Analyzed the data: GW SS. Contributed reagents/materials/analysis tools: GW SS. Wrote the paper: GW ZZ.

References

1. Saunders PT, Ho MW. On the Increase in Complexity in Evolution .2. The Relativity of Complexity and the Principle of Minimum Increase. J Theor Biol. 1981; 90(4):515–30. doi: 10.1016/0022-5193(81)90303-9 PMID: WOS:A1981LV69600006.
2. Jaakkola S, Ei-Showk S, Annila A. The driving force behind genomic diversity. Biophys Chem. 2008; 134(3):232–8. doi: 10.1016/J.Bpc.2008.02.006 PMID: WOS:000255676900012.
3. McShea DW, Brandon RN. Biology's first law: the tendency for diversity and complexity to increase in evolutionary systems: University of Chicago Press; 2010.
4. Gladyshev GP. On thermodynamics, entropy and evolution of biological systems: What is life from a physical chemist's viewpoint. Entropy. 1999; 1(2):9–20.
5. Gladyshev GP. The principle of substance stability is applicable to all levels of organization of living matter. International journal of molecular sciences. 2006; 7(3):98–110. doi: 10.3390/I7030098 PMID: WOS:000237727000003.
6. Gladyshev GP. Thermodynamic self-organization as a mechanism of hierarchical structure formation of biological matter. Prog React Kinet Mec. 2003; 28(2):157–88. PMID: WOS:000183624100002.
7. Annila A, Salthe S. Physical foundations of evolutionary theory. J Non-Equil Thermody. 2010; 35(3):301–21. doi: 10.1515/Jnetdy.2010.19 PMID: WOS:000283177800009.
8. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome research. 2009; 19(7):1195–201. doi: 10.1101/gr.091231.109 PMID: WOS:000267786900006.
9. Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer. 2006; 6(12):924–35. doi: 10.1038/nrc2013 PMID: WOS:000242244400013.
10. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458(7239):719–24. doi: 10.1038/nature07943 PMID: WOS:000265193600031.
11. Gryder BE, Rood MK, Johnson KA, Patil V, Raftery ED, Yao LPD, et al. Histone Deacetylase Inhibitors Equipped with Estrogen Receptor Modulation Activity. J Med Chem. 2013; 56(14):5782–96. doi: 10.1021/jm400467w PMID: WOS:000322503000012.
12. van Wieringen WN, van der Vaart AW. Statistical analysis of the cancer cell's molecular entropy using high-throughput data. Bioinformatics. 2011; 27(4):556–63. doi: 10.1093/bioinformatics/btq704 PMID: WOS:000287246000016.
13. De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G. Do Natural Proteins Differ from Random Sequences Polypeptides? Natural vs. Random Proteins Classification Using an Evolutionary Neural Network. Plos One. 2012; 7(5). doi: ARTN e36634 doi: 10.1371/journal.pone.0036634 PMID: WOS:000305341300021.
14. Pande VS, Grosberg AY, Tanaka T. Nonrandomness in Protein Sequences—Evidence for a Physically Driven Stage of Evolution. Proceedings of the National Academy of Sciences of the United States of America. 1994; 91(26):12972–5. doi: 10.1073/Pnas.91.26.12972 PMID: WOS:A1994PY29400127.
15. Rackovsky S. "Hidden" sequence periodicities and protein architecture. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95(15):8580–4. doi: 10.1073/Pnas.95.15.8580 PMID: WOS:000075143900031.
16. Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. Bioinformatics. 2010; 26(3):310–8. doi: 10.1093/Bioinformatics/Btp660 PMID: WOS:000274342800003.
17. Munteanu CR, Gonzalez-Diaz H, Borges F, de Magalhaes AL. Natural/random protein classification models based on star network topological indices. J Theor Biol. 2008; 254(4):775–83. doi: 10.1016/J.Jtbi.2008.07.018 PMID: WOS:000260023600008.
18. White SH. The Evolution of Proteins from Random Amino-Acid-Sequences .2. Evidence from the Statistical Distributions of the Lengths of Modern Protein Sequences. J Mol Evol. 1994; 38(4):383–94. doi: 10.1007/Bf00163155 PMID: WOS:A1994NB64300007.

19. Weiss O, Jimenez-Montano MA, Herzel H. Information content of protein sequences. J Theor Biol. 2000; 206(3):379–86. doi: 10.1006/Jtbi.2000.2138 PMID: WOS:000089480200006.

20. White SH, Jacobs RE. The Evolution of Proteins from Random Amino-Acid-Sequences .1. Evidence from the Lengthwise Distribution of Amino-Acids in Modern Protein Sequences. J Mol Evol. 1993; 36 (1):79–95. doi: 10.1007/Bf02407307 PMID: WOS:A1993KC50800007.

21. Zhang Z, Yu J. On the organizational dynamics of the genetic code. Genomics, proteomics & bioinformatics. 2011; 9(1):21–9.

22. Xiao J-F, Yu J. A scenario on the stepwise evolution of the genetic code. Genomics, proteomics & bioinformatics. 2007; 5(3):143–51.

23. Zhang Z, Yu J. The pendulum model for genome compositional dynamics: from the four nucleotides to the twenty amino acids. Genomics, proteomics & bioinformatics. 2012; 10(4):175–80. doi: 10.1016/j.gpb.2012.08.002 PMID: 23084772.

24. Yu J. A content-centric organization of the genetic code. Genomics, proteomics & bioinformatics. 2007; 5(1):1–6. doi: 10.1016/S1672-0229(07)60008-4 PMID: 17572358.

25. Zhang Z, Yu J. On the organizational dynamics of the genetic code. Genomics, proteomics & bioinformatics. 2011; 9(1–2):21–9. doi: 10.1016/S1672-0229(11)60004-1 PMID: 21641559.

26. Acevedo-Rocha CG, Fang G, Schmidt M, Ussery DW, Danchin A. From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. 2013; 29(5):273–9. doi: 10.1016/J.Tig.2012.11.001 PMID: WOS:000319309100001.

27. Waterhouse RM, Zdobnov EM, Kriventseva EV. Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi. Genome Biol Evol. 2011; 3:75–86. doi: 10.1093/Gbe/Evq083 PMID: WOS:000290252700008.

28. Rukhin A, Soto J, Nechvatal J, Smid M, Barker E. A statistical test suite for random and pseudorandom number generators for cryptographic applications. DTIC Document, 2001.

29. Rackovsky S. "Hidden" sequence periodicities and protein architecture. Proceedings of the National Academy of Sciences. 1998; 95(15):8580–4.

30. Shih M-Y, Jheng J-W, Lai L-F. A two-step method for clustering mixed categroical and numeric data. Tamkang Journal of science and Engineering. 2010; 13(1):11–9.

31. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1987; 6:461–4.

32. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2014; 42(D1):D7–D17. doi: 10.1093/Nar/Gkt1146 PMID: WOS:000331139800002.

33. Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic acids research. 2014; 42(D1):D574–D80. doi: 10.1093/Nar/Gkt1131 PMID: WOS:000331139800085.

34. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005; 15(6):589–94. doi: 10.1016/J.Gde.2005.09.006 PMID: WOS:000233686200004.

35. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 Sequenced Escherichia coli Genomes. Microb Ecol. 2010; 60(4):708–20. doi: 10.1007/s00248-010-9717-3 PMID: WOS:000284255700002.

36. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, et al. Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. Journal of Bacteriology. 2003; 185(19):5673–84. doi: 10.1128/jb.185.19.5673–5684.2003 PMID: 13129938

37. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome research. 2002; 12(6):962–8. doi: 10.1101/Gr.87702 PMID: WOS:000176433700013.

38. Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic acids research. 2004; 32: D271–D2. doi: 10.1093/Nar/Gkh024 PMID: WOS:000188079000062.

39. Fox GE. Origin and Evolution of the Ribosome. Csh Perspect Biol. 2010; 2(9). Artn A003483 doi: 10.1101/Cshperspect.A003483 PMID: WOS:000281575800010.

40. Alba MM, Castresana J. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 2005; 22(3):598–606. doi: 10.1093/molbev/msi045 PMID: WOS:000227163100027.

41. Xia XH, Xie Z, Li WH. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J Mol Evol. 2003; 56(3):362–70. doi: 10.1007/S00239-002-2406-1 PMID: WOS:000181329500012.

42. Oliver JL, Marin A. A relationship between GC content and coding-sequence length. J Mol Evol. 1996; 43(3):216–23. doi: 10.1007/Bf02338829 PMID: WOS:A1996VF80800007.

43. Xia XH, Wang HC, Xie Z, Carullo M, Huang H, Hickey D. Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. Mol Biol Evol. 2006; 23(7):1450–4. doi: 10.1093/Molbev/Msl012 PMID: WOS:000238545100015.