Prokaryotes, Metagenomics, and GC-Content

Erin R. Reichenberger*, Gail L. Rosen†, Uri Hershberg*‡, and Ruth Hershberg§
*Department of Biomedical Engineering, Science & Health Systems, Drexel University, Philadlephia, PA 19104 USA
†Department of Computer and Electrical Engineering, Drexel University, Philadelphia, Pennsylvania 19104 USA
‡Department of Microbiology and Immunology, Drexel University, Philadelphia, Pennsylvania 19104 USA
§Rachel and Menachem Mendelovitch Evolutionary Processes of Mutation and Natural Selection Research Laboratory
Department of Genetics, the Ruth and Bruce Rappaport Faculty of Medicine, Technion University, Haifa, 31096 Israel

***Abstract*—The degree of variation in nucleotide content across all prokaryotic genomes is expansive and ranges from ∼15% to ∼75% guanine and cytosine (GC). There is an ongoing debate as to the causes of this extensive variation, however, since variation in nucleotide content is a genome-wide trait that affects the genome as a whole, it is highly interesting to understand what drives such variation. Employing 183 metagenomic datasets (959G) from numerous types of environments, a Unix environment pipeline of command-line bioinformatics tools, scripting languages, and statistical programs was employed to investigate the influence of environment on GC-content. Using several statistical approaches, we show that each type of environment has a distinct GC-signature that cannot be entirely explained by disparities in phylogenetic composition. Further, our results indicate that environment *and* phylogeny impact nucleotide composition.**

Keywords–GC-Content; Prokaryotes; Metagenomics; Mutation; Genomic Variation; Big Data, Environmental Influence.

I. INTRODUCTION

The causes of the great variation in nucleotide composition of prokaryotic genomes have long been disputed [1]–[3]. In our previous work, we used extensive metagenomic and whole-genome data containing over 31 million sequences to demonstrate that both phylogeny and the environment shape prokaryotic nucleotide content [4]. The GC-content – which is the percentage of guanine and cytosine in a genome or fragment of DNA is important as it can describe the makeup of an organism, provide insight into an organism's evolution, and expand our understanding of gene expression.

II. METHODOLOGY

Shotgun-sequenced fasta files (183 datasets) from 14 environments were obtained from MG-Rast [5]. The details of each project's methodology, metadata and geographic location can be found utilizing a mapping API we created (http://simlab.biomed.drexel.edu/maps/map.php) [6]–[17].

After screening each dataset (e.g., ambiguous/short reads), the remaining sequences were extracted and classified according to phylogeny [18]. The GC-content was calculated for each classified read, followed by a mean GC-content calculation for each phylum, each sample (there were multiple samples in an environmental category), and each environmental category.

III. RESULTS

After calculating the mean GC for all environments, we found that each environment carried a distinct GC-content signature. We found a similarly distinct GC-level trend in 111 samples that comprised a single type of environment. To rule out the possibility that variation in GC-composition between environments could be explained by differences in phylogenetic composition, each environment's prokaryotic community was investigated from two standpoints; the microbial composition and the phylum pair-wise correlation level in GC-content in each environmental category.

A. Microbial Composition

The relative abundance of each phylum in an environment was calculated. Additionally, to assess whether phyla differed at the genus-level, a taxonomic list of the genus names present in each environment was compiled. Using the intersection and union of the lists, the level of similarity (Jaccard similarity coefficient) in the genera contained within two environments was calculated.

B. Phyla and GC-Content

In the process of looking at phylogenic distribution, we found that different phyla were characterized by different mean GC-contents. Additionally, some phyla were characterized by a much broader GC-content range than others. These averages and possible ranges of nucleotide compositions for each taxonomic classification (phylum-level) were, to a large extent, maintained across different environments and were in accord with the GC-levels of fully-sequenced prokaryotic genomes. Phylogeny therefore seems to impose a clear limit on the range of nucleotide content a prokaryote can adopt.

C. Hypergeometric Distribution, Phyla, and GC-Content

The GC-content variation seen in prokaryotes provided an opportunity to observe the behavior of a phylum. Using our largest environmental dataset (111 samples), we found that the GC-content of a phylum with a high range of variability would be at its upper bounds in a high GC sample and the lower bounds in a low GC sample.

D. Correlations, Phyla, and GC-Content

The correlative relationship between the GC-content of each phylum was assessed using the Spearman correlation coefficients. Our analysis showed a number of statistically significant correlations which appeared at a frequency much greater than expected by chance. A significant correlation would indicate that whatever force influenced the nucleotide content in one phylum, had a similar effect on the nucleotide content of the remaining phyla.

E. Assessing Correlations: Phyla, GC-Content, and the 3rd Codon Position of 4-fold Redundant Amino Acids

We confirmed our results and ensured that our findings were not related to artifacts due to amino acid usage by annotating the classified sequences and re-running the correlative analysis on them [19]. The annotated sequences were

 ISBN: 978-1-61208-409-1

examined for the location of those amino acid with four-fold redundancies (Alanine, Arginine, Glycine, Leucine, Proline, Serine, Threonine, Valine) and the 3rd codon positions of these codons were extracted for GC-content calculations. As the third codon positions of fourfold degenerate codons do not affect the amino acid sequence of a protein, their nucleotide content should not be affected by selection at the level of amino acid usage. We found that the GC-content of the 3rd codon position of fourfold degenerate codons within protein-coding genes was correlated between phyla across environments far more frequently than expected by chance.

IV. Conclusion

Employing numerous shotgun-sequenced datasets as well as data from all currently available fully-sequenced genomes, we show that both phylogeny and environment influence prokaryotic nucleotide composition. We demonstrate that, across environments, different phyla have distinct nucleotide compositions. We then show that GC-levels vary by environment in a manner that can not be explained solely by differences in phylogenetic composition. Combined, our results demonstrate that both phylogeny and the environment significantly affect nucleotide composition and that the environmental differences affecting nucleotide composition are far subtler than previously appreciated.

Acknowledgment

The authors would like to thank Calvin Morrison for his assistance, and Yemin Lan for gathering the AAI genomic data.

This short paper is an advance of the publication *Prokaryotic nucleotide composition is shaped by both phylogeny and the environment*. This collaborative project is supported by the Louis and Bessie Stein Foundation. ERR is supported by a Ford Foundation; RH is supported by ERC FP7 CIG grant [321780], a Yigal Allon Fellowship, and by the Robert J. Shillman Career Advancement Chair. Research reported in this publication is supported by the NIH [P01AI106697], NSF [0845827, 1120622] and the Department of Energy [DE-SC0004335].

References

[1] K. Foerstner, C. von Mering, S. Hooper, and P. Bork, "Environments shape the nucleotide composition of genomes," EMBO Reports, vol. 6, no. 12, Dec 2005, pp. 1208–1213.

[2] F. Hildebrand, A. Meyer, and A. Eyre-Walker, "Evidence of Selection upon Genomic GC-Content in Bacteria," PLOS Genetics, vol. 6, no. 9, Sep 2010.

[3] R. Raghavan, Y. D. Kelkar, and H. Ochman, "A selective force favoring increased G plus C content in bacterial genes," Proceedings of the National Academy of Sciences of the United States of America, vol. 109, no. 36, Sep 4 2012, pp. 14 504–14 507.

[4] E. R. Reichenberger, G. Rosen, U. Hershberg, and R. Hershberg, "Prokaryotic nucleotide composition is shaped by both phylogeny and the environment," Genome Biology and Evolution, 2015 (Accepted).

[5] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards, "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes," BMC Bioinformatics, vol. 9, Sep 19 2008.

[6] F. E. Angly, D. Willner, A. Prieto-Davo, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, E. A. Dinsdale, M. Furlan, M. Haynes, M. R. Henn, Y. Hu, D. L. Kirchman, T. McDole, J. D. McPherson, F. Meyer, R. M. Miller, E. Mundt, R. K. Naviaux, B. Rodriguez-Mueller, R. Stevens, L. Wegley, L. Zhang, B. Zhu, and F. Rohwer, "The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes," PLOS Computational Biology, vol. 5, no. 12, Dec 2009.

[7] P. Belda-Ferre, L. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simon-Soro, M. Pignatelli, and A. Mira, "The oral metagenome in health and disease," ISME Journal, vol. 6, no. 1, Jan 2012, pp. 46–56.

[8] C. Desnues, B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer, "Biodiversity and biogeography of phages in modern stromatolites and thrombolites," Nature, vol. 452, no. 7185, Mar 20 2008, pp. 340–U5.

[9] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White, and F. Rohwer, "Functional metagenomic profiling of nine biomes," Nature, vol. 452, no. 7187, APR 3 2008, pp. 629–U8.

[10] R. A. Edwards, B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer, "Using pyrosequencing to shed light on deep mine microbial ecology," BMC Genomics, vol. 7, Mar 20 2006.

[11] V. Kunin, J. Raes, J. K. Harris, J. R. Spear, J. J. Walker, N. Ivanova, C. von Mering, B. M. Bebout, P. N. R., P. Bork, and P. Hugenholtz, "Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat," Molecular Systems Biology, vol. 4, 2008, p. 198.

[12] X. Mou, S. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran, "Bacterial carbon processing by generalist species in the coastal ocean," Nature, vol. 451, no. 7179, Feb 7 2008, pp. 708–U4.

[13] D. T. Pride, J. Salzman, M. Haynes, F. Rohwer, C. Davis-Long, R. A. White, III, P. Loomer, G. C. Armitage, and D. A. Relman, "Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome," ISME Journal, vol. 6, no. 5, MAY 2012, pp. 915–926.

[14] B. Rodriguez-Brito, L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, J. Buchanan, C. Desnues, E. Dinsdale, R. Edwards, B. Felts, M. Haynes, H. Liu, D. Lipson, J. Mahaffy, A. Belen Martin-Cuadrado, A. Mira, J. Nulton, L. Pasic, S. Rayhawk, J. Rodriguez-Mueller, F. Rodriguez-Valera, P. Salamon, S. Srinagesh, T. F. Thingstad, T. Tran, R. V. Thurber, D. Willner, M. Youle, and F. Rohwer, "Viral and microbial community dynamics in four aquatic environments," ISME Journal, vol. 4, no. 6, Jun 2010, pp. 739–751.

[15] B. K. Swan, C. J. Ehrhardt, K. M. Reifel, L. I. Moreno, and D. L. Valentine, "Archaeal and Bacterial Communities Respond Differently to Environmental Gradients in Anoxic Sediments of a California Hypersaline Lake, the Salton Sea," Applied and Environmental Microbiology, vol. 76, no. 3, Feb 2010, pp. 757–768.

[16] L. Wegley, R. Edwards, B. Rodriguez-Brito, H. Liu, and F. Rohwer, "Metagenomic analysis of the microbial community associated with the coral Porites astreoides," Environmental Microbiology, vol. 9, no. 11, Nov 2007, pp. 2707–2719.

[17] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, "Human gut microbiome viewed across age and geography," Nature, vol. 486, no. 7402, Jun 14 2012, pp. 222+.

[18] A. Brady and S. Salzberg, "PhymmBL expanded: confidence scores, custom databases, parallelization and more," Nature Methods, vol. 8, no. 5, May 2011, pp. 1208–1213.

[19] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," Nucleic Acids Research, vol. 38, no. 20, Nov 2010.

 ISBN: 978-1-61208-409-1