SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

GENOMICS

BIOINFORMATICS

INFORMATION THEORY AND COMPUTATION

Received
11 September 2012

Accepted
20 November 2012

Published
8 January 2013

Correspondence and requests for materials should be addressed to S.A. (sakhter@sciences.sdsu.edu) or R.A.E. (redwards@mail.sdsu.edu)

Applying Shannon's information theory to bacterial and phage genomes and metagenomes

Sajia Akhter[1], Barbara A. Bailey[2], Peter Salamon[2], Ramy K. Aziz[3,4,5] & Robert A. Edwards[1,3,6]

[1]Computational Science Research Center, San Diego State University, San Diego, CA, [2]Department of Mathematics and Statistics, College of Sciences, San Diego State University, San Diego, CA, USA, [3]Department of Computer Science, College of Sciences, San Diego State University, San Diego, CA, USA, [4]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt, [5]Systems Biology Research Group, University of California San Diego, La Jolla, CA, USA, [6]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA.

All sequence data contain inherent information that can be measured by Shannon's uncertainty theory. Such measurement is valuable in evaluating large data sets, such as metagenomic libraries, to prioritize their analysis and annotation, thus saving computational resources. Here, Shannon's index of complete phage and bacterial genomes was examined. The information content of a genome was found to be highly dependent on the genome length, GC content, and sequence word size. In metagenomic sequences, the amount of information correlated with the number of matches found by comparison to sequence databases. A sequence with more information (higher uncertainty) has a higher probability of being significantly similar to other sequences in the database. Measuring uncertainty may be used for rapid screening for sequences with matches in available database, prioritizing computational resources, and indicating which sequences with no known similarities are likely to be important for more detailed analysis.

The extraordinary advances in speed and throughput of sequencing technologies in the past decade have generated an unprecedented wealth of complete or near complete genome sequences, and have also allowed the emergence of the technology of metagenomics or random community genomics, which aims at sequencing DNA from environmental microbial communities without culturing or isolating individual microbes. Today thousands of fully sequenced genomes and over 7,000 metagenomes have been deposited in public repositories, e.g., GenBank[1], Genomes Online Database (GOLD)[2], the SEED database[3], and the Metagenomics RAST (MG-RAST) server[4]. To be annotated and analyzed, metagenome sequences are compared to genes, proteins, protein domains, protein families, and genomes in known databases. It was shown a few years ago that approximately 19 hours were needed to analyze one megabase of DNA sequence (if linear compute complexity is assumed), and each data set required about one month of computing time (Unpublished data, Edwards, R. A., 2008). However, MG-RAST and other public services handle the analysis by using large compute clusters dedicated to sequence searching. Because of the deluge of sequence data, new efficient tools and methods are required for analyzing and comparing sequences, and for prioritizing the sequences to be analyzed when comprehensive analysis is not feasible.

One approach to prioritizing the analysis of unknown genomic or metagenomic sequences is examining the information content of known genes, proteins, and genomes to explore possible patterns or trends that might help in predicting informative sequences, i.e., those sequences likely to encode proteins or to provide new rather than redundant knowledge about the sample to which they belong. In the cell, the information flows from DNA to amino acid sequences, as DNA is transcribed into RNA then translated into amino acids to make proteins. Depending on the different combinations of bases in the deoxyribonucleotides of the DNA sequence, different amino acids are added to the nascent, growing polypeptide chain. Complex proteins consist of different combinations of amino acids and therefore are encoded by various combinations of the four sequence bases. Homopolymeric tracts like AAAAAAAAC or TTTTTTTCCCCC can only code one or few different amino acids and encode for proteins with amino acid repeats. Therefore, we hypothesize that they are much less likely to encode functional proteins than DNA containing equimolar mixtures of bases (e.g., AGCTAGCTAGCT).

Statistical approaches derived from information theory can quantify the amount of information in a DNA sequence. Several investigators have examined different aspects of information content of genomes, including Shannon's uncertainty[5–9] and symmetry[10,11]. For example, Chang and coworkers calculated Shannon's uncertainty index for all the complete prokaryotic and eukaryotic genomes available in 2005. They found that Shannon's information in complete genomes is greater than that in matching random sequences and they described a coarse-grain model for genome growth and evolution that allows a genome to diverge at any stage during its growth[6,7].

Shannon's uncertainty[12–14] was originally designed for encoding and decoding data transmitted and received through a digital communication system. Since sequence data can also be represented as a system where DNA is transformed into amino acids, this theory can be used to calculate the amount of information or uncertainty of a sequence. For each sequence, the uncertainty measurement per base pair generates a score from 0 to $2n$, where n is word length. The greater the uncertainty, the more even is the distribution of each word. For example, the sequence AAAA can only be read using two letter words as AA regardless of the register and has little uncertainty. In contrast, the sequence ACGT can be read as AC, CG and GT, depending on the register and has more inherent uncertainty and information.

Here, the information content was examined for complete bacterial and phage genomes, and the analysis was extended to the calculation of Shannon's uncertainty for each sequence within metagenomic libraries. The effects of word size, genome length, and GC% on Shannon's uncertainty have also been examined. We demonstrate that the information content of sequences from metagenomes correlates with the number of similar sequences that is found by comparison to databases of known sequences. Using this approach may speed up the processing time for analyzing metagenomic data and allow prioritization of computational resources.

Results

Shannon's uncertainty in complete bacterial and phage genomes. Shannon's index was calculated for 600 complete phage genomes and 94 complete bacterial genomes (listed in Supplementary Table S1) using word lengths ranging from 1 to 12 nucleotides (nt). Shannon's indices of phage and bacterial genomes were similar up to word length 7 nt (Figure 1), implying an even distribution of all possible sequence words in phage and bacterial genomes. From word length 8 to 12 nt, the rate of increase of Shannon's index is higher in bacterial genomes than in phage genomes. Moreover, for word lengths greater than 10 nt, Shannon's index can differentiate bacteria and phages (Figure 1).

Factors influencing differences in shannon's uncertainty between complete bacterial and phage genomes. The difference between Shannon's indices of phage and bacterial genomes for word length greater than 8 nt suggested that either word length, genome size or a combination of both might influence this uncertainty value.

Shannon's index vs. word length. Word length is reportedly an important factor influencing the value of Shannon's index[6,7]. A high Shannon's index (close to the maximum possible index, i.e., for word length n, the maximum index will be $2n$) depends on the presence of all possible combinations of words in the genome. Consequently, the longer the genome the higher the probability of having different word variations. For a given word length of n, there are 4^n possible word combinations for DNA sequences. The length of most phage genomes (585 out of 600) ranges from 4^7 bp to less than 4^9 bp (Figure 2). Therefore, for word size greater than 8 nt, many words will only be represented zero or one times, which will result in a lower Shannon's index for most of these genomes. In contrast, the average length of the 94 bacterial genomes used in this analysis is about 3 million bp (between 2×4^{10} and 4^{11} bp). Therefore, bacterial genomes have a higher Shannon's index than phage genomes using word lengths

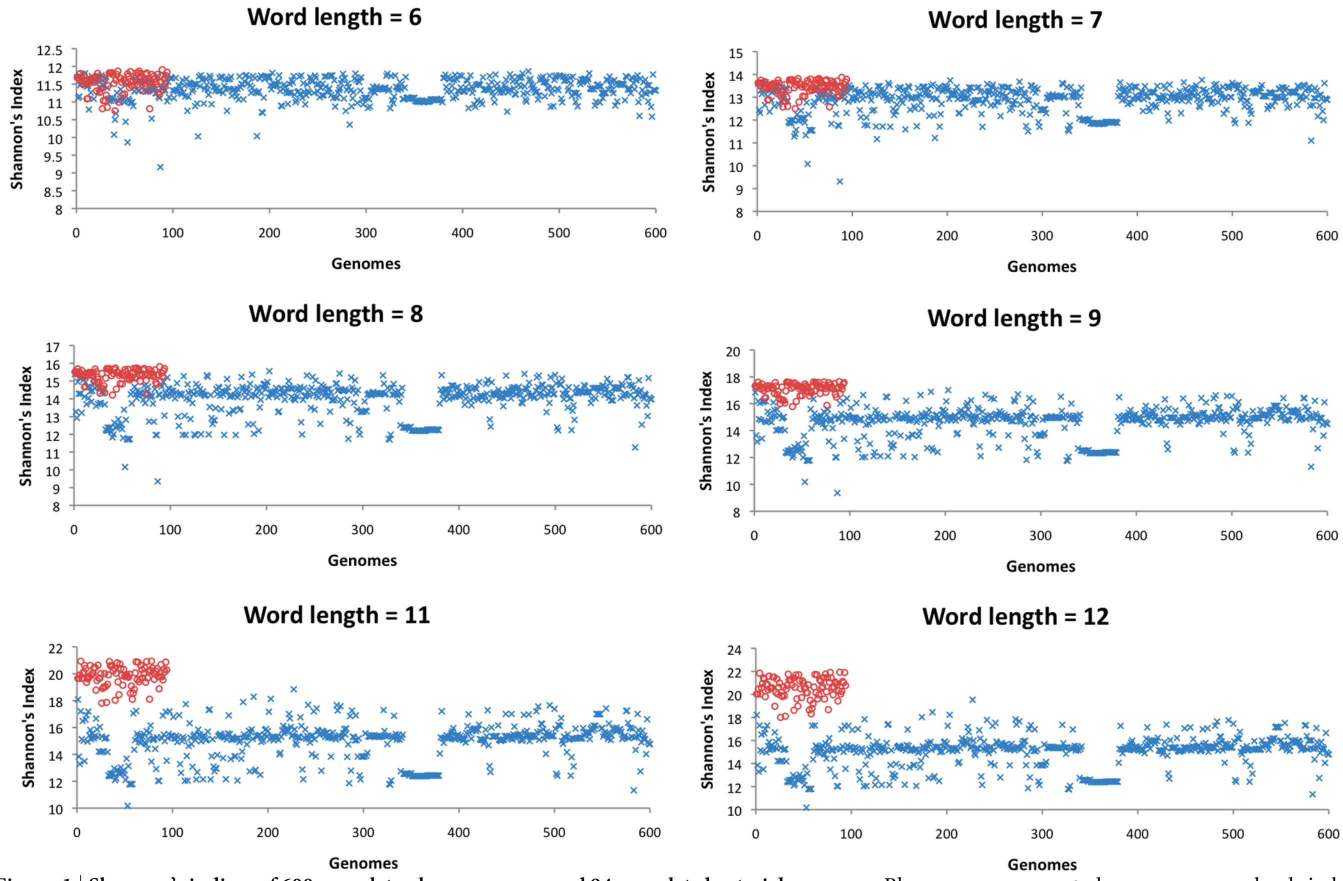

Figure 1 | Shannon's indices of 600 complete phage genomes and 94 complete bacterial genomes. Blue crosses represent phage genomes and red circles represent bacterial genomes. As the word length increases, Shannon's index is more discriminatory between phage and bacterial genomes.

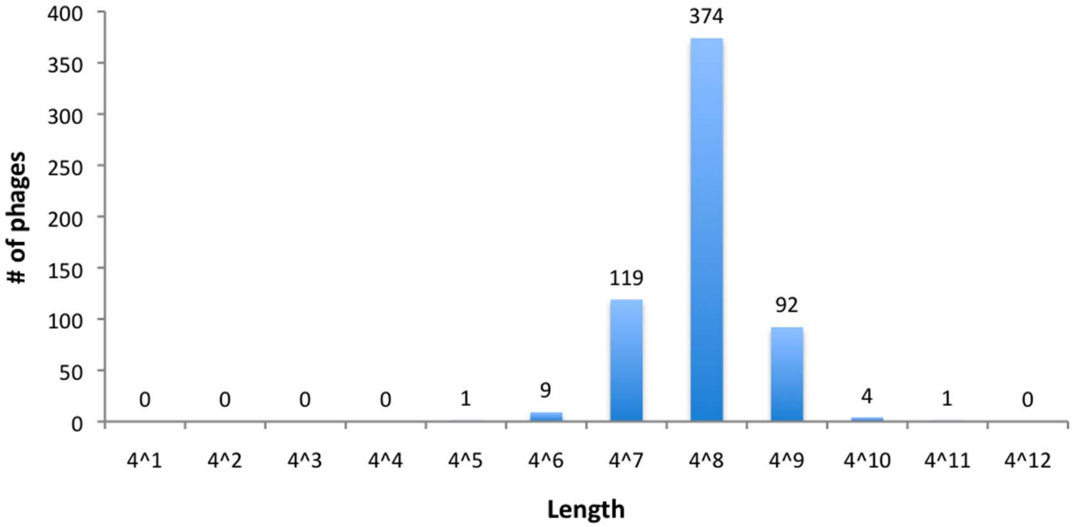

Figure 2 | Length distribution of 600 complete phages.

smaller than 12 nt. For word lengths 11 nt or 12 nt, Shannon's index can distinguish phage and bacterial genomes (Figure 1) although this is likely because phage genomes are too short to generate sufficiently high Shannon's indices for words of this size.

Shannon's index vs. genome length. Shannon's indices for all phage genomes have been plotted against their lengths. For word length 12 nt, Shannon's index highly correlates with the logarithm of the genome length (Figure 3). For word length 9 nt, there is still a

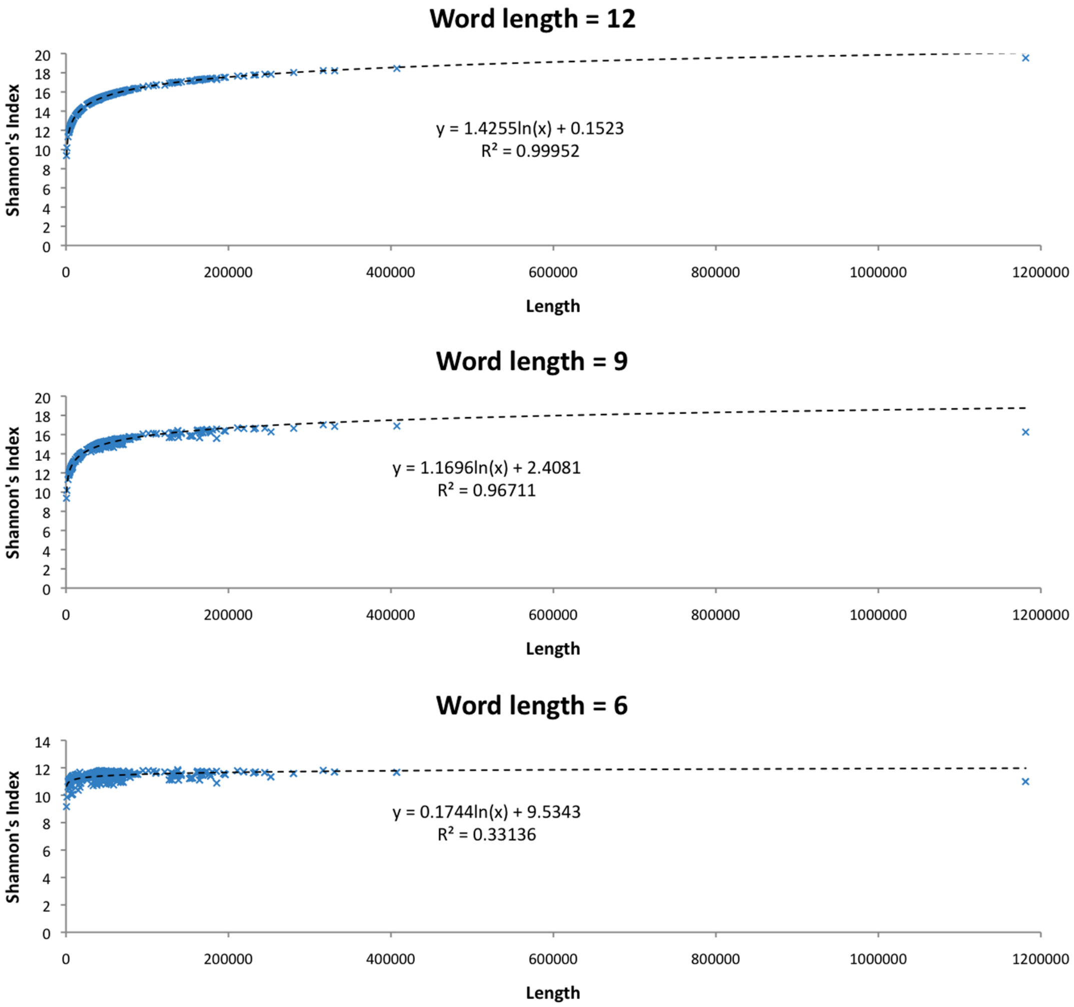

Figure 3 | Shannon's index vs. length for 600 complete phage genomes using word length 9 and 12.

significant correlation; however, for shorter word lengths, no significant correlation was observed between genome length and Shannon's index. For shorter word lengths, most of the genomes have almost all combination of words in their genomes, so there is no strong correlation between Shannon's index and genome length. In contrast, for longer words, the bigger genomes have more combinations of words than the smaller genomes; so Shannon's index correlates with genome length.

Calculations with irrelevant word lengths may give the wrong impression and create false differences between genomes (Figure 1). To compare genomes based on Shannon's index, the word length (n) should be chosen in a way that allows the possibility of having all combination of words (4^n) in all the genomes. Therefore, for a given genome of length L, the possible word length (n) to calculate Shannon's index should be (Equation 1)

$$4^n < L \Rightarrow n \le \left\lfloor \frac{1}{2}\log_2 L \right\rfloor \qquad (1)$$

Shannon's index vs. GC%. For most phage genomes, the maximum word length that should be used to calculate Shannon's index (Equation 1) is 7 nt. When word lengths from 1 to 7 nt were used to calculate Shannon's index, GC-rich and GC-poor genomes were found to have lower Shannon's index since these genomes tend to have less diverse word combinations than genomes with 50% GC content (Figure 4a). The strong relationship between Shannon's index and |GC% − 0.5| for word length 1 to 5 nt suggests that Shannon's index is strongly influenced by the GC composition of the DNA sequence (Figure 4b). For word lengths above 6 nt, the relationship is not strongly supported. Different sequences may have the same GC%, but Shannon's index depends on the distribution of the different word combinations. Therefore two different sequences having the same GC% may have different Shannon's indices, and the probability of this happening increases with the word length. Thus, as word length is increased, the correlation between Shannon's index and GC content becomes weaker (Figure 4b).

Shannon's uncertainty in metagenomes. Shannon's uncertainty was calculated for different metagenomic data sets. The maximum uncertainty equates to a sequence that has equal frequencies of each word (e.g. A, G, C, T for word length one) and the majority of reads in a metagenome have an uncertainty greater than 1.8 per nt (Figure 5a) suggesting an even distribution of bases in the reads, although the relative information content of the reads varies by sample.

To investigate whether information content correlates with functional content, we compared the frequencies with which each sequence matched an entry in the known databases. The similarities

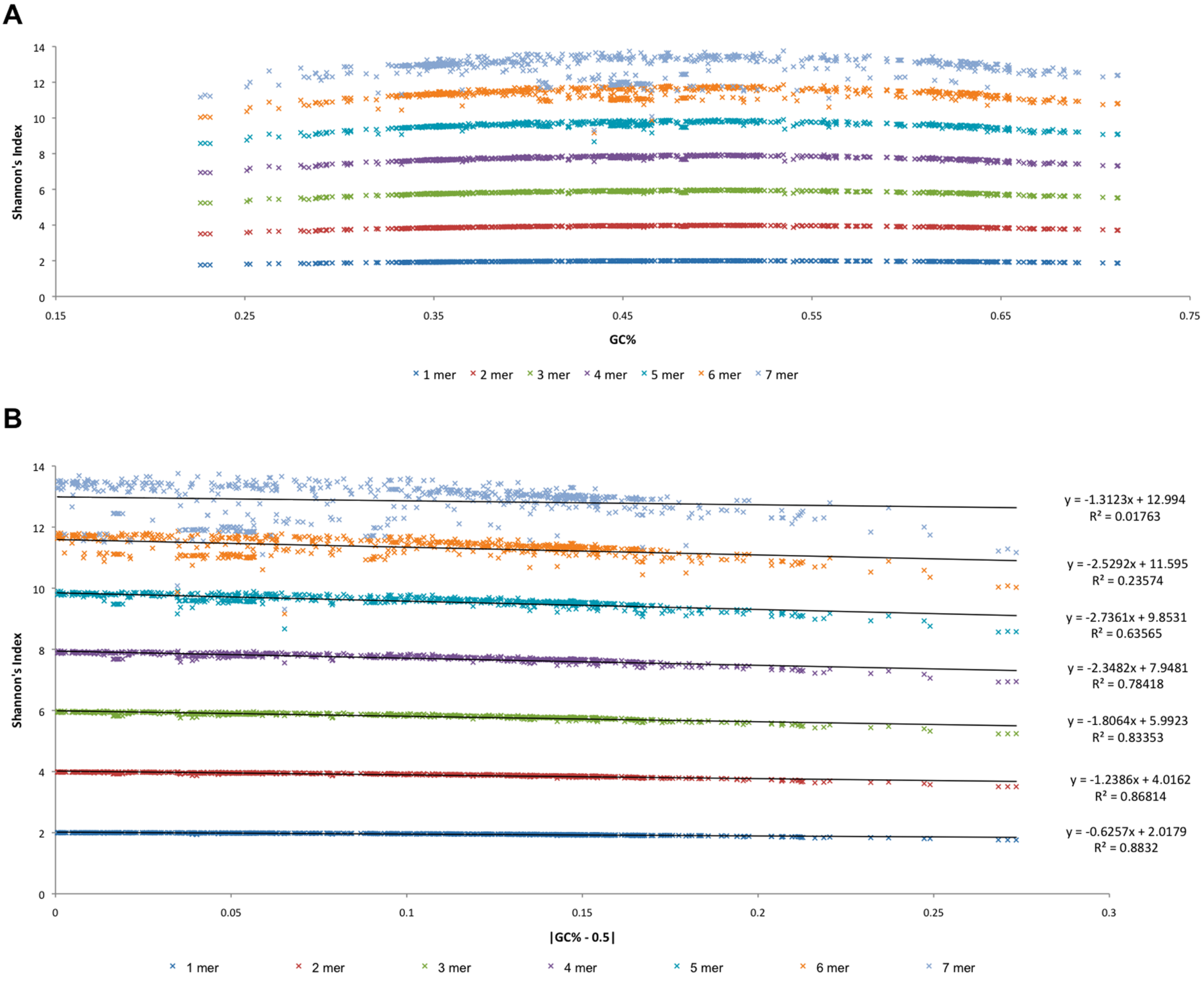

Figure 4 | (a) Shannon's index vs. GC% for 600 complete phage genomes using word length 1 nt to 7 nt. (b) The relationship between Shannon's index and |GC% −0.5| for 600 complete phage genomes using word length 1 nt to 7 nt.

between the metagenomic sequences and the SEED non-redundant protein database had been pre-calculated using BLASTX[15,16] as a part of the annotation and analysis procedure performed by the MG-RAST server[4]. For a set of reads with a given uncertainty, the fraction of reads that were similar to sequences in the SEED non-redundant database was extracted from these pre-calculated similarities (Figure 5b). A read with more information (higher uncertainty) was more likely to be similar to sequences in the database than a read with less information. Different metagenomes varied in the fraction of reads that are similar to known sequences, but this likely reflects the sampling limitations that have thus far limited the breadth of the known sequences[17].

Discussion

Since the publication of the first complete genome sequences, genome composition has been appealing to mathematicians, statisticians, and computer scientists. Base distribution statistics, skews and biases[18–22], sequence symmetries[10,11], and information content[5–8] have all been examined in the hope of deciphering hidden codes within the genomes[11] and better understanding genome growth and evolution[7,23–27].

Among the mathematical methods used, Shannon's uncertainty has previously been considered as a genome analysis strategy[5–8]. In the work of Chang and colleagues[6,7], Shannon's uncertainty was calculated for complete prokaryotic and eukaryotic genomes available at that time, and it was found that genomes belonged to a universality class that could be mathematically represented by a simple formula, yet *Plasmodium* genomes stood out as an intriguing exception, still unexplained[6,7]. Additionally, the variation of Shannon's index with sequence word length and genome length was examined[6]. Here, our findings confirmed and advanced that study by establishing the relationship between word size and genome length for calculating Shannon's index.

We also found that at a certain word lengths, Shannon's index can be used to differentiate phage and bacterial sequences. Although this differentiation is sensitive to genome length, with some modification, this observation can help find phage genes embedded in bacterial genomes. As an application, we calculated Shannon's index for a group of DNA sequences using a word size of 12 nt (four consecutive

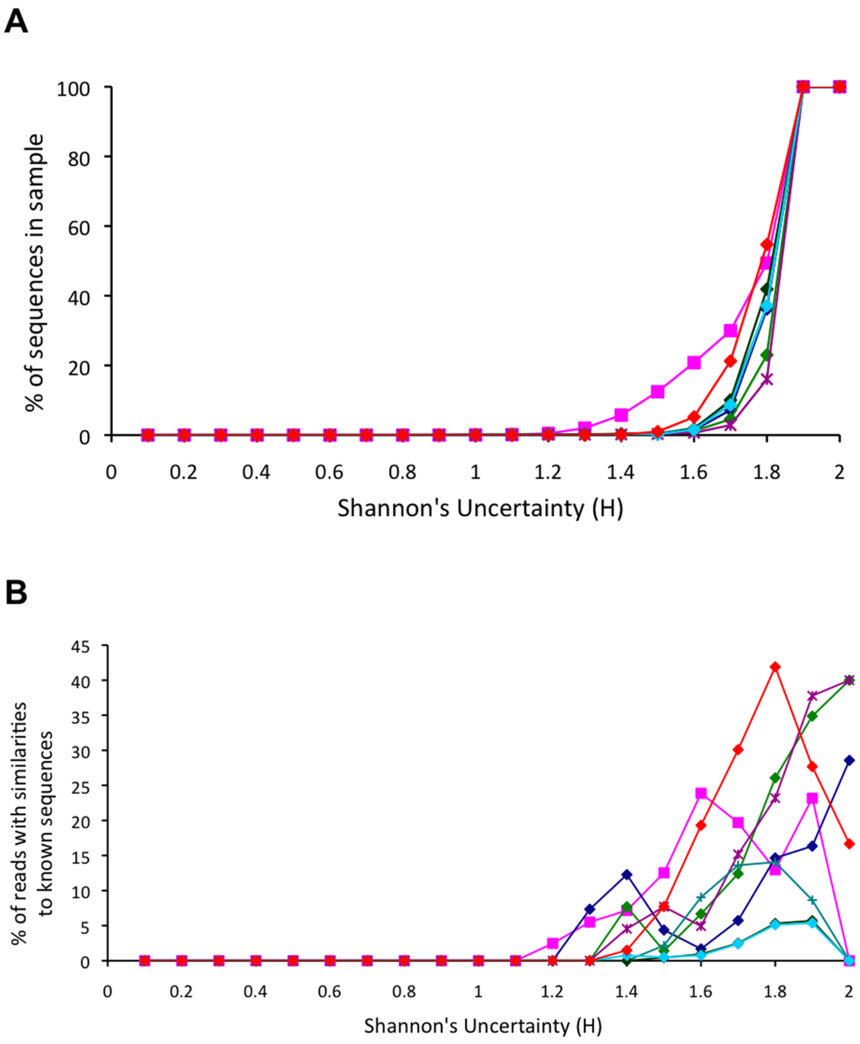

Figure 5 | (a) Cumulative comparison of the uncertainty (for word length 1) in DNA sequences in metagenome samples. Eight samples representative of the 24 used in this study are shown here: Soudan Mine Black Stuff (pink[35]), Line islands Kingman reef phage (light green[36]), Line islands Tabuaren phage (light blue[36]), Marine phages from the Gulf of Mexico (blue[30]), Marine samples supplemented with DMSP (magenta[37]), Line islands Palmyra Phage (dark green[36]), Line islands Christmas Reef phage (red[36]), Marine samples supplemented with vanillate (green[37])). The uncertainty is greater than 1.7 for 85% to 90% sequences of all samples. (b) Comparison of Shannon's uncertainty and the observed similarity to known sequences. Shannon's uncertainty (H) was calculated for word length one, and is compared with similarity to the SEED no-redundant protein database. Samples are coloured the same as in Fig. 5. Word lengths up to 11 letters were also used to calculate (H) and all cases confer same results (data not shown).

amino acids) and we were able to use this group of sequences to detect prophages in bacterial genomes[28].

Finally, our findings show that the information content of metagenomic sequences varies from sample to sample, but about 85% of those sequences have high levels of uncertainty, suggesting that they are comprised of approximately equal numbers of each of the four bases (Figure 5). In addition, the information content in metagenomic sequences was found to correlate with the likelihood that the sequence would be similar to a previously characterized sequence (in the non-redundant database). This suggests that the large numbers of metagenomic sequences could be rapidly sorted based on their information content to prioritize similarity searches and other common computations. It is to be noted, however, that those metagenomic sequences have to be preprocessed and cleared of potential repeats or homopolymeric runs, sometimes introduced by sequencing methods (e.g., the introduction of runs of nucleotides during high-throughput sequencing)[29]. For this purpose, tools such as PRINSEQ[29] MG-RAST[4] can be used prior to sequence analysis of metagenomic data sets. Moreover, the correlation between information content and similarity may provide a rapid mechanism to screen for either false positive matches (sequences matching the database that should not) or false negative matches (sequences with no match in the database, but that should). Of course, the extremely large numbers of sequences with high uncertainty but no similarity in the databases might be influenced by the lack of sampling in the known databases[30].

Methods

Retrieval of genomic and metagenomic data sets. All genomes used in this analysis were retrieved from the SEED database and servers[3] (http://servers.theseed.org), where they have been consistently annotated and classified into subsystems[31,32] in the RAST server[33] (http://rast.nmpdr.org) Likewise, metagenomic sequence data sets were retrieved from the MG-RAST server[4] (http://metagenomics.theseed.org).

For the calculation and analysis of Shannon's uncertainty, a subset of 24 metagenomes was selected from the previously studied SCUMS data set[34], most of which were created by pyrosequencing. The metagenomes were chosen to represent the range of data sets available from sequences sampled in simple and well-characterized environments to more complex environments with multiple species present. The raw data were used without assembly, and the samples included in the data set cover both viral and microbial metagenomes, sampled from such diverse biomes as mines, marine environment, soils, and animals[34–37]. The shortest sequence in the data set was 31 bp and the longest was 362 bp.

Calculation of Shannon's uncertainty. Shannon's uncertainty was calculated using **Equation 2**[14],

$$-\sum p_i \; \log(p_i) \tag{2}$$

where p_i is the frequency of the i-th word in a sequence. For example, for word length one, p_i is calculated from the frequencies of the nucleotides{A, G, C, T}. If each word is equally frequent, $p_i = 0.25$. In general, for all words of length n being equally likely, p_i is $1/4^n$.

1. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **38**, D46–51 (2010).
2. Liolios, K. *et al.* The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **38**, D346–354 (2011).
3. Aziz, R. K. *et al.* SEED Servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* **7**, e48053 (2012).
4. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
5. Zeeberg, B. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. *Genome Res.* **12**, 944–955 (2002).
6. Chang, C. H. *et al.* Shannon information in complete genomes. *Proc. IEEE Comput. Syst Bioinform. Conf.* 20–30 (2004).
7. Chang, C. H. *et al.* Shannon information in complete genomes. *J. Bioinform. Comput. Biol.* **3**, 587–608 (2005).
8. Chen, H. D., Chang, C. H., Hsieh, L. C. & Lee, H. C. Divergence and Shannon information in genomes. *Phys. Rev. Lett.* **94**, 178103 (2005).
9. Arias-Gonzalez, J. R. Entropy involved in fidelity of DNA replication. *PLoS ONE* **7**, e42272 (2012).
10. Kong, S. G. *et al.* Inverse symmetry in complete genomes and whole-genome inverse duplication. *PLoS ONE* **4**, e7553 (2009).
11. Poptsova, M. S. *et al.* Hidden chromosome symmetry: in silico transformation reveals symmetry in 2D DNA walk trajectories of 671 chromosomes. *PLoS ONE* **4**, e6396 (2009).
12. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Techn. J.* **27**, 379–423; 623–656 (1948).
13. Shannon, C. E. The mathematical theory of communication. 1963. *MD Comput.* **14**, 306–317 (1997).
14. Schneider, T. Information Theory Primer, Available at http://www.ccrnp.ncifcrf.gov/~toms/paper/primer (1999).
15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
16. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
17. Eisen, J. A. & Fraser, C. M. Phylogenomics: intersection of evolution and genomics. *Science* **300**, 1706–1707 (2003).
18. Grigoriev, A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **26**, 2286–2290 (1998).
19. Grigoriev, A. Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res.* **60**, 1–19 (1999).
20. Ma, X. R., Xiao, S. B., Guo, A. Z., Lv, J. Q. & Chen, H. C. DNAskew: statistical analysis of base compositional asymmetry and prediction of replication boundaries in the genome sequences. *Acta Biochim Biophys Sin (Shanghai)* **36**, 16–20 (2004).
21. Guy, L., Karamata, D., Moreillon, P. & Roten, C. A. Genometrics as an essential tool for the assembly of whole genome sequences: the example of the chromosome of Bifidobacterium longum NCC2705. *BMC Microbiol.* **5**, 60 (2005).
22. Collyn, F., Guy, L., Marceau, M., Simonet, M. & Roten, C. A. Describing ancient horizontal gene transfers at the nucleotide and gene levels by comparative pathogenicity island genometrics. *Bioinformatics* **22**, 1072–1079 (2006).
23. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
24. Roten, C. A., Gamba, P., Barblan, J. L. & Karamata, D. Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res.* **30**, 142–144 (2002).
25. Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3**, e130 (2005).
26. Ochman, H. & Davalos, L. M. The nature and dynamics of bacterial genomes. *Science* **311**, 1730–1733 (2006).
27. Raghavan, R., Kelkar, Y. D. & Ochman, H. A selective force favoring increased G+C content in bacterial genes. *Proc. Natl Acad. Sci. USA* **109**, 14504–14507 (2012).
28. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126 (2012).
29. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
30. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol* **4**, e368 (2006).
31. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
32. Aziz, R. K. Subsystems-based servers for rapid annotation of genomes and metagenomes. *BMC Bioinformatics* **11**, O2 (2010).
33. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
34. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
35. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57 (2006).
36. Dinsdale, E. A. *et al.* Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* **3**, e1584 (2008).
37. Mou, X., Sun, S., Edwards, R. A., Hodson, R. E. & Moran, M. A. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**, 708–711 (2008).

Acknowledgements

This work was supported by grant NSF DBI-0850356 from the NSF Division of Biological Infrastructure to RAE (PhAnToMe project).

Author contributions

SA, RKA, RAE conceived the study and designed the experiments; SA conducted the experiments and prepared all figures; SA, BB, PS, RKA analyzed and interpreted the data; BB supervised the statistical analyses; SA, RKA and RAE wrote the manuscript. All authors read and approved the final version.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Akhter, S., Bailey, B.A., Salamon, P., Aziz, R.K. & Edwards, R.A. Applying Shannon's information theory to bacterial and phage genomes and metagenomes. *Sci. Rep.* **3**, 1033; DOI:10.1038/srep01033 (2013).