Computational and Structural Biotechnology Journal 13 (2015) 352–357

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Homology-Independent Metrics for Comparative Genomics

Tarcisio José Domingos Coutinho [a], Glória Regina Franco [a], Francisco Pereira Lobo [b,*]

[a] Departamento de Bioquímica e Imunologia, Programa de pós-graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Antonio Carlos Avenue, 6627, Belo Horizonte, Minas Gerais CEP 31270-901, Brazil

[b] Laboratório Multiusuário de Bioinformática, Embrapa Informática Agropecuária, André Tosello Avenue, 209, Barão Geraldo, Campinas, São Paulo CEP 13083-886, Brazil

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form 6 April 2015
Accepted 18 April 2015
Available online 5 May 2015

Keywords:
Comparative genomics
Homology-independent metrics
Genomic signatures

ABSTRACT

A mainstream procedure to analyze the wealth of genomic data available nowadays is the detection of homologous regions shared across genomes, followed by the extraction of biological information from the patterns of conservation and variation observed in such regions. Although of pivotal importance, comparative genomic procedures that rely on homology inference are obviously not applicable if no homologous regions are detectable. This fact excludes a considerable portion of "genomic dark matter" with no significant similarity — and, consequently, no inferred homology to any other known sequence — from several downstream comparative genomic methods. In this review we compile several sequence metrics that do not rely on homology inference and can be used to compare nucleotide sequences and extract biologically meaningful information from them. These metrics comprise several compositional parameters calculated from sequence data alone, such as GC content, dinucleotide odds ratio, and several codon bias metrics. They also share other interesting properties, such as pervasiveness (patterns persist on smaller scales) and phylogenetic signal. We also cite examples where these homology-independent metrics have been successfully applied to support several bioinformatics challenges, such as taxonomic classification of biological sequences without homology inference. They where also used to detect higher-order patterns of interactions in biological systems, ranging from detecting coevolutionary trends between the genomes of viruses and their hosts to characterization of gene pools of entire microbial communities. We argue that, if correctly understood and applied, homology-independent metrics can add important layers of biological information in comparative genomic studies without prior homology inference.

Contents

Abbreviations: DOR, dinucleotide odds ratio; GC3S, frequency of G + C at the third position of synonymous codons; HI, homology-independent; HD, homology-dependent; ORF, open reading frame; RSCU, relative synonymous codon usage; NC, effective number of codons; NC′, effective number of codons considering GC3S.

* Corresponding author. Tel.: +55 19 32115843.
E-mail addresses: coutinho.tarcisio@gmail.com (T.J.D. Coutinho), gfrancoufmg@gmail.com (G.R. Franco), franciscolobo@gmail.com, francisco.lobo@embrapa.br (F.P. Lobo).

http://dx.doi.org/10.1016/j.csbj.2015.04.005

1. Introduction

Most computational methods available for comparative genomics rely on initial similarity searches to infer homology relationships and, consequently, analyze the wealth of genomic data currently available. Among others, current methods for comparative genomic analysis based on the detection of homologous sequences allow the 1) determination of the time of divergence between taxa through the theory of molecular clock [1]; 2) automatic annotation of new genomes based on orthology inference [2]; 3) estimation of the rates of evolution of protein families [3]; 4) analysis of the overall evolution of genomes through genome-scale analysis of patterns of gain/loss of genomic elements [4]; 5) searching for higher-order layers of positional genomic information (haplotypic blocks, synteny, etc.) [5]; and 6) genomic-scale search for patterns of positive Darwinian selection [6], among many others.

The computational methods for comparative genomic analysis based in the detection of homologous regions, from now on referred as homology-dependent (HD) methods, although crucial for several bioinformatic pipelines, are limited to genomic sequences with detectable homologous regions, usually identified through computationally intensive software that contains somewhat arbitrary cut-offs to define groups of homologous sequences [7,8]. The failure to detect such regions prevents the application of virtually any HD method and excludes several interesting classes of DNA sequences from further analysis. ORFans — orphans Open Reading Frames (ORFs) without any detectable similarity to other sequences — are commonly found in complete genomes and in environmental sequences and constitute a true "dark matter" of biological data that cannot be surveyed using traditional HD methods [9,10]. In some newly discovered taxa, such as the large DNA viruses from *Mimiviridae*, the vast majority of coding sequences do not share significant similarity with known proteins [11].

In this mini-review we compile a list of metrics used in comparative genomic studies that share an unusual property for this purpose: they do not rely on initial homology inference, and can be calculated from individual sequence data alone. Such metrics, from now on referred as homology-independent (HI) metrics, can be easily calculated for virtually any fragment of any genome that fulfills a few criteria, such as minimum length and complexity. These metrics usually detect biases by comparing the observed frequencies of nucleotide words, especially dinucleotide and codons, with expected frequencies for the same words. The dinucleotide usage patterns in a given genome are commonly referred in the scientific literature as genomic signatures, since they are also taxon-specific and highly conserved in a given genome [12–14]. Here we also highlight the relative strengths and weaknesses of such metrics and report comparative genomic studies that applied such metrics to extract biologically meaningful information that would be otherwise impossible to obtain using common HD comparative genomic methods.

2. Homology-Independent Metrics: Causes and Properties

2.1. Causes of Variation in Homology-Independent Metrics

Most explanations for the biased values observed in HI metrics are due to a complex interplay between three broad groups of phenomena that shape together the use of nucleotide words in genomes. One of these groups is composed of mutational pressures where a given nucleotide word is significantly more (or less) used than its expected frequency due to mutational events. Possible sources of mutational pressures are distinct transition/transversion ratios [15], CpG underrepresentation in vertebrate genomes due to methylation/deamination processes occurring in this dinucleotide [16] and distinct nucleotide incorporation efficiency by polymerases during genome replication [17], among many others.

A second group of phenomena responsible for HI biases in genomes is composed of selection pressure events, in which natural selection shapes the differential usage of nucleotide motifs. In fact, several nucleotide motifs (such as the "TATA box") interact with the transcription/translation machinery and are classic examples of conservation of nucleotide words in genomic sequences due to selection pressure [18]. Another classic cause of variation in nucleotide words (codons) induced by selection pressure is the more-than-expected usage of synonymous codons that corresponds to the more abundant aminoacyl-tRNAs in cell cytoplasm in order to increase translation speed/efficiency, a selection pressure particularly strong in single-celled organisms [19] and in highly-expressed genes [20]. A final broad class of factors known to influence the use of nucleotide words in genomes is the occurrence of neutral processes such as genetic drift during the course of evolution [21]. Therefore, if properly modeled and interpreted, the results obtained through HI metrics in comparative genomic studies can highlight broad patterns of mutational and selective forces as well as random variations acting in the genomes under analysis.

2.2. General Properties of Homology-Independent Metrics

Besides not requiring previous homology relationships to analyze genomic data, HI metrics contain other general properties shared by most or all of them. An interesting property is that most HI metrics contain null models that take into account major factors already known to influence the frequencies of nucleotide words. Different HI metrics consider factors such as GC content, observed frequencies of smaller words that compose the word under analysis, degeneracy of the genetic code and amino acid usage, among others, when calculating null models. Therefore, any bias detected using HI metrics with proper null models is not explainable by factors already taken into account when computing null models, and represent biological phenomena that require further explanation/investigation. Also, several HI metrics are pervasive in the sense that values for whole genomes should persist at smaller scales. Some HI metrics remain reasonably constant for fragments with as few as 125 base pairs when compared with values calculated for entire genomes, or even when comparing coding and non-coding regions of genomes, making HI metrics a robust option to develop procedures to classify nucleotide sequences to taxonomic units, as in the case of genomic signatures [13].

Another useful aspect of HI metrics when applied to comparative genomics is the fact that some of them generate results that contain phylogenetic signal and are able to represent phylogenetic relationships, arguably with a more global view of the evolutionary process [22]. The patterns of DOR and codon usage bias in complete genomes of prokaryotes present a strong correlation with phylogenetic trees of 16S ribosomal RNA and housekeeping genes [14,23]. Comparative genomics using HI metrics may better reflect the global phylogenetic relationships between complete genomes by considering, for instance, events of horizontal transfer (HGT, one of the many factors known to change local frequencies of nucleotide words in genomic sequences) as part of the evolutionary signal in opposition to the reductionist analysis of single genes as proxies to faithfully represent the phylogenetic history of the entire genome.

3. Main Metrics for Homology-Independent Analyses

Several HI metrics have been used for comparative genomics. Supplementary Table 1 contains a non-exhaustive list of software to calculate HI metrics to allow users to analyze their own data. Below follows a formal description of the most common metrics:

3.1. Genomic Signatures

Genome signature is an umbrella term used to refer to similar concepts, but to different HI metrics. A genome signature refers to any HI metric that can be calculated from a DNA sequence with sufficient length and compositional complexity that enables the correct classification of the sequence to its source genome [12,24]. An ideal genomic signature should satisfy three major criteria: 1) it should be species-specific; 2) it should reflect phylogenetic history and 3) it should be pervasive [12].

3.1.1. GC Content

The simplest known HI metric used as genomic signature is the GC content of genomic sequences (percentage of G + C in a sequence). Despite being a simple metric, GC presents a huge variation across genomes, ranging from approximately 20% in *Plasmodium falciparum* [25] to 70% in some actinobacteria [26]. GC content is reasonably constant within a given genome, and was already found to be correlated with several universal factors of microbial lifestyles such as temperature [27], niche complexity [28] and aerobiosis [29].

3.1.2. Dinucleotide Odds Ratio

Dinucleotide Odds Ratio (DOR) is defined as the ratio between observed and expected frequencies of a dinucleotide in a sequence, and is perhaps the canonical example of both an HI metric and a genomic signature [24]. The expected frequency of a dinucleotide (null model) is defined as the product of the observed frequencies of its two nucleotides in the sequence under analysis, therefore removing background nucleotide frequency as a possible source of bias. DOR values close to one indicate observed frequencies close to expectation, and values significantly above or below one indicate the presence of mutational and/or selection bias actively shaping the frequency of dinucleotides. As arbitrary cutoff, DOR values observed outside the range of 0.78–1.25 are commonly considered to have low or high relative abundance, respectively [30]. Eq. (1) describes the DOR calculation for dinucleotide *xy* for a single-stranded sequence.

$$P_{xy} = \frac{f_{xy}}{f_x f_y}. \tag{1}$$

Eq. (1) describes the calculation of dinucleotide odds ratio (P_{xy}) for single-stranded genomes. f_x and f_y denote the frequency of mononucleotides *x* and *y* in a given sequence, and f_{xy} denotes the observed frequency of dinucleotide *xy* in the same sequence.

The calculation shown in Eq. (1) is valid only for single-stranded sequences that do not obey the first Chargaff's parity rule. For double-stranded genomes the frequency of each nucleotide is calculated in a symmetrical way to take into account the complementary nucleotide located at the opposite strand. If we denote the nucleotide frequencies in double stranded genomes with "°", f°(T) = f°(A) = (f(A) + f(T)) / 2 and f°(C) = f°(G) = (f(C) + f(G)) / 2. Based on the above equation and the first Chargaff's parity rule, the DOR for double-stranded sequences is calculated as follows:

$$P_{xy} = \frac{2\left(f_{xy} + f_{zw}\right)}{\left(f_x + f_y\right)\left(f_z + f_w\right)}. \tag{2}$$

Eq. (2) describes the calculation of dinucleotide odds ratio (P_{xy}) for double-stranded genomes. f_x and f_y denote any two nucleotides, and f_z and f_w denote the frequencies of nucleotides *z* and *w*, complementary to *y* and *x*, respectively.

3.1.3. Relative Synonymous Codon Usage (RSCU)

This metric is commonly used to estimate bias in the use of synonymous codons and removes the differences in the frequencies of amino acids as a possible bias factor (null model) [31]. Observed values are the counts of the *jth* codon for the *ith* amino acid, and expected values are calculated by counting the total of codons encoding the amino acid *i* in a given sequence divided by the degeneracy class of amino acid *i* (two, three, four or six). RSCU values for each codon are calculated according to Eq. (3):

$$RSCUij = \frac{X_j^i}{E_j^i}. \tag{3}$$

Eq. (3) describes the calculation of RSCU for codon *j*. X_j^x — observed number of occurrences of the *jth* codon for the *ith* amino acid. E_j^i — expected number of occurrences of *jth* codon for the *ith* amino acid. Expected values are calculated by counting all synonymous codons coding for amino acid *i* in a sequence divided by the number of synonymous codons that code for this amino acid (amino acid degeneracy class).

3.1.4. Genomic Signatures Using Longer Words

Genomic signatures based on nucleotide words with length two and three are commonly used as HI metrics due to the immediate biological relevance of words of these lengths for pivotal biological processes involving nucleic acids, such as DNA modification mechanisms in vertebrates which recognize dinucleotides as modification sites [16] and translation, which is indissociable of the codon concept. However, using longer DNA words as genomic signatures adds more dimensions to compare and stratify sequence data (e.g. there are, respectively, 16, 64, 256 and 1024 different DNA words of lengths 2, 3, 4 and 5, respectively). Although not possessing a clear, intuitive biological relevance such as words of length two or three, the length increase arguably improves classification performance of genomic signatures metrics [32–35]. Null models for words of these lengths often involve more complex procedures, such as zero- or higher-order Markov models to account for the frequencies of smaller words that compose each DNA word [35].

3.2. Effective Number of Codons (NC) and Variations

3.2.1. Effective Number of Codons (NC)

The effective number of codons (NC) is calculated for coding regions and represents the overall bias of preferential use of synonyms codons in a given gene/genome. The equation to calculate NC is conceptually similar to the calculation of the effective population size used in population genetics. It generates a single number that ranges from 20 (extreme codon usage bias where each amino acid is coded by only one codon) to 61 (absence of bias in the choice of synonyms codons, indicating equal usage for all codons) [36]. Extreme values of CG content in coding regions restrict the number of codons effectively available and, consequently, NC values are heavily influenced by CG content [37]. This metric is one of the most sensitive to detect biases in codon usage, and is calculated using the following equations (from [38]):

$$\theta_a = \frac{n_a \sum_{i=1}^{k} p_i^2 - 1}{n_a - 1}. \tag{4}$$

Eq. (4) describes the calculation of $\boldsymbol{\theta}_a$ (homozygosity of amino acid *a*). p_i — frequency of codon *i*; *k* — number of synonymous codons for amino acid *a* (amino acid degeneracy class); n_a — observed number of

codons for amino acid a (only for amino acids with degeneracy class greater than one).

From the values of θ_a for each amino acid one should compute the average values θ_r for each amino acid degeneracy class (e.g. two, three, four or six-fold degeneracy) (Eq. (5)):

$$\theta_r = \frac{1}{n_{RC}} \sum_{a \in RC} \theta_a. \tag{5}$$

Eq. (5) describes the calculation of average values $\boldsymbol{\theta}_r$ for each class of codons r. n_{RC} — number of amino acids in a degeneracy class; RC — set of all amino acids belonging to a degeneracy class.

Finally, NC is computed as described in Eq. (6):

$$NC = 2 + \left(\frac{9}{\theta_2}\right) + \left(\frac{1}{\theta_3}\right) + \left(\frac{5}{\theta_4}\right) + \left(\frac{3}{\theta_6}\right). \tag{6}$$

Eq. (6) describes the calculation of NC. Each θ value in equation corresponds to an average value (θ_r) calculated for each amino acid degeneracy class (two, three, four and six).

3.2.2. NC-Plot

Since NC values are strongly influenced by GC content [36] it is useful to compare observed NC values against theoretical values calculated in function of GC content (null model) in order to exclude GC content as a possible source of variation. An approach in this direction is the calculation of NC plots, which consists of a theoretical curve correlating expected values of NC as a function of GC content at third bases of synonymous codons (GC3S). GC3S are commonly assumed to be a good proxy of "true" background genome composition regarding nucleotide frequencies since these positions are supposed to be under a more relaxed selection pressure (although there are controversies, see e.g. [39]). By including in the chart the observed values of GC3S and NC for a given coding sequence and finding points located above or below the theoretical curve it is possible to detect coding sequences with biased observed NC values after considering GC3S values. Eq. (7) generates the theoretical curve of NC plots:

$$NC_{theoretical} = 2 + f_{GC3S} + \left(\frac{29}{f_{GC3S}^2 + (1 - f_{GC3S})^2}\right). \tag{7}$$

Eq. (7) describes the theoretical values of NC as a function of GC3S. f_{GC3S} — frequency of GC3S. The resolution of this equation with f_{GC3S} parameter ranging from zero to one generates the theoretical curve of NC plot.

3.2.3. Effective Number of Codons Considering the GC Content (NC′)

Another approach to deal with the strong influence of GC content over NC values was the development of a new metric, conceptually similar to NC, but that takes into account GC content, called NC′ [38]. Therefore, any bias evidenced by NC′ must be interpreted excluding GC content as a possible source of variation. However, other biases were introduced in this new metric. For instance, while NC values range from 20 to 61 and have clear and intuitive biological meaning, the NC′ values range from 0 to 61, which are not readily interpretable [40].

NC′ metric uses the chi-square test (X^2) to calculate the deviation of observed frequencies of use of each codon i (p_i, Eq. (8)) when compared with expected values (e_i). Expected values may be calculated in various ways, such as from the frequencies of mono, di or trinucleotides comprising the codon, therefore accounting for distinct null model scenarios. The expected deviation value for each amino acid (X_a^2) is calculated as follows:

$$X_a^2 = \sum_{i=1}^{k} \frac{n_a (p_i - e_i)^2}{e_i}. \tag{8}$$

Eq. (8) describes the calculation of expected deviation values for each amino acid a. i — codon under analysis; k — amino acid degeneracy class; n_a — number of observed codons for amino acid a; p_i — observed frequency of codon i; e_i — expected frequency of codon i.

Having the X_a^2 values one can compute $\theta_a^{'}$ (conceptually similar to θ_a used in NC calculation) as shown in Eq. (9):

$$\theta_a^{'} = \frac{X_a^2 + n_a - k}{k(n_a - 1)}. \tag{9}$$

Eq. (9) describes the calculation of modified homozygosity values ($\boldsymbol{\theta}_{\boldsymbol{a}}^{'}$) for amino acid a. n_a — number of observed codons for amino acid a; k — amino acid degeneracy class.

From the values of $\theta_a^{'}$ one can calculate NC′ in a manner similar to NC computation as described in Eqs. (5) and (6).

4. Current Applications of Homology-Independent Metrics in Comparative Genomics

The first and still the most popular field in comparative genomics with extensive application of HI metrics is the taxonomic classification of biological sequences or subsequences without prior phylogenetic tree reconstruction, such as in the case of genomic signatures. For instance, several pipelines for classification of sequences from environmental genomics to taxonomic space (binning procedures in metagenomics studies) rely on HI metrics [41–43]. Additionally, other classification methods to detect subsequences with biased distribution of HI metrics within longer genomic sequences could be used to detect several cases of exogenous DNA in a given genome. Several tools already make use of HI metrics for this purpose aiming at detecting important classes of evolutionary events, such as general cases of horizontal gene transfer (HGT) [44–46] and the detection of particular cases of HGT, such as genomic/pathogenicity islands [47] and phage integration sites [48]. For some groups of organisms such as large DNA viruses, HI metrics are sometimes the only class of tools available to study how these genomes evolved such large repertories of ORFans and to demonstrate that they arrived through multiple HGT events [11].

Nowadays, HI metrics have been used to answer questions in comparative genomics far beyond their initial use as genomic signatures for taxonomical classification of sequences. Such metrics were recently used to detect community-specific signatures of synonymous codon usage biases in metagenomic samples from different ecological niches. Such biased codons correlate with expression levels and occur regardless of individual phylogeny of organisms [49]. Additionally, such community-specific codon usage biases also predict lifestyle-specific genes, detecting coding sequences relevant for adaptation of organisms to specific ecological niches. These studies revealed a higher level of organization of metagenomic samples, where entire microbial communities share gene pools optimized for cross-genome translation and behave as a single meta-genome. Codon usage biases alone are also capable to predict other features for single microorganisms and/or microbial communities such as growth speed [50] and can be used to infer gene function based solely on the evolutionary changes for translation efficiency [51].

HI metrics have also been used to detect coevolutionary trends in biological systems composed of viruses and their hosts. The genomes of such disparate organisms coexist in the same cellular space and compete for the same resources. Therefore, it is reasonable to assume viral and host genomes to share some common compositional features due to constraints induced by host factors, such as the molecular mechanisms for the detection of foreign nucleic acids and for the translation of coding sequences.

The work of Lobo et al. 2009 used HI metrics (DOR and RSCU) to detect coevolutionary trends in a virus-host biological system [52]. The *Flaviviridae* family is composed of monophyletic viruses that infect vertebrate (mammals and birds) and/or invertebrate (ticks and

mosquitoes) organisms. This work assumes that *Flaviviridae* that infect a single host lineage would be subjected to specific host-induced pressures and, therefore, selected by them. The authors observed that the two host groups possess very distinctive dinucleotide and codon usage patterns. A pronounced CpG under-representation was found in the vertebrate group, possibly induced by the methylation-deamination process, exclusive of vertebrate genomes, as well as a prominent TpA decrease. The invertebrate group displayed only a TpA frequency reduction bias, with no CpG bias. *Flaviviridae* viruses mimicked host nucleotide motif usage in a host-specific manner. Vertebrate-infecting viruses possessed under-representation of CpG and TpA, and insect-only viruses displayed only a TpA under-representation bias. Especially, single-host *Flaviviridae* members which persistently infect mammals or insect hosts (*Hepacivirus* and insect-only *Flavivirus*, respectively) were found to possess a codon usage profile more similar to that of their hosts than to phylogenetically related *Flaviviridae*. Vertebrates and mosquito genomes are under very distinct lineage-specific constraints, and *Flaviviridae* viruses which specifically infect these lineages appear to be subject to the same evolutionary pressures that shaped their host coding regions, evidencing lineage-specific coevolutionary processes between viral and host genomes that could not be surveyed using HD metrics.

5. Summary and Outlook

Since Karlin et al. 1994 described DOR and demonstrated its utility as a genomic signature [24,30], several compositional parameters calculated from nucleotide sequence data alone have been developed that fulfill and even surpass the original requisites to be classified as genomic signatures. Besides being used for their original purpose (classification of biological sequences to a taxonomic space), several other layers of biologically meaningful information are available through these metrics. In common, all these metrics are: 1) calculated from genomic sequence data alone; 2) do not rely on homology inference and 3) add a new layer of biologically meaningful data to the genomes under analysis, such as taxonomic information in the case of genomic signatures, but also other information such as putative gene expression profiles and convergent coevolutionary patterns.

The pervasiveness observed in HI metrics that made them useful for sequence classification into taxonomic space also permeates through other biological dimensions and could be used to describe and model biological entities in other classification systems besides biological taxonomy, with a utility far broader than the original scope of genomic signatures. For this reason, we argue that these metrics are better described as homology-independent (HI) metrics for comparative genomics. Genomic sequences from disparate groups that contain some niche superposition, such distinct species of microorganisms belonging to the same ecological community or viruses and their respective hosts coexisting in the same environment, were found to possess surprising patterns of coevolutionary events not known before HI metrics were applied to investigate them. The use of HI metrics as tools for comparative genomic analysis, if correctly understood and applied, can reveal important pieces of biological information in genomic sequences without priorhomology inference.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.csbj.2015.04.005.

References

[1] Kumar S. Molecular clocks: four decades of evolution. Nat Rev Genet 2005;6:654–62.
[2] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35:W182–5.
[3] Luz H, Vingron M. Family specific rates of protein evolution. Bioinformatics 2006;22:1166–71.
[4] Lobo FP, Rodrigues MR, Rodrigues GO, Hilario HO, Souza RA, et al. KOMODO: a web tool for detecting and visualizing biased distribution of groups of homologous genes in monophyletic taxa. Nucleic Acids Res 2012;40:W491–7.
[5] Ghiurcuta CG, Moret BM. Evaluating synteny for improved comparative studies. Bioinformatics 2014;30:i9–i18.
[6] Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, et al. Patterns of positive selection in six mammalian genomes. PLoS Genet 2008;4:e1000144.
[7] Kumar S, Filipski A. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res 2007;17:127–35.
[8] Li L, Stoeckert Jr CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;13:2178–89.
[9] Yin Y, Fischer D. Identification and investigation of ORFans in the viral world. BMC Genomics 2008;9:24.
[10] Cortez D, Forterre P, Gribaldo S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol 2009;10:R65.
[11] Monier A, Claverie JM, Ogata H. Horizontal gene transfer and nucleotide compositional anomaly in large DNA viruses. BMC Genomics 2007;8:456.
[12] Dutta C, Paul S. Microbial lifestyle and genome signatures. Curr Genomics 2012;13:153–62.
[13] Jernigan RW, Baran RH. Pervasive properties of the genomic signature. BMC Genomics 2002;3:23.
[14] van Passel MW, Kuramae EE, Luyf AC, Bart A, Boekhout T. The reach of the genome signature in prokaryotes. BMC Evol Biol 2006;6:84.
[15] Keller I, Bensasson D, Nichols RA. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 2007;3:e22.
[16] Simmen MW. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics 2008;92:33–40.
[17] Berkhout B, Grigoriev A, Bakker M, Lukashov VV. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res Hum Retroviruses 2002;18:133–41.
[18] Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu Rev Biochem 2003;72:449–79.
[19] Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985;2:13–34.
[20] Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. Genes Cells 2009;14:499–509.
[21] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[22] Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 2009;106:19126–31.
[23] Coenye T, Vandamme P. Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. Microbiology 2003;149:3507–17.
[24] Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 1995;11:283–90.
[25] Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 2002;419:498–511.
[26] Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, et al. Genomics of actinobacteria: tracing the evolutionary history of an ancient phylum. Microbiol Mol Biol Rev 2007;71:495–548.
[27] Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, et al. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun 2006;347:1–3.
[28] Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO Rep 2005;6:1208–13.
[29] Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 2002;55:260–4.
[30] Burge C, Campbell AM, Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A 1992;89:1358–62.
[31] Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 1986;14:5125–43.
[32] Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res 2003;13:145–58.
[33] Yap YL, Zhang XW, Danchin A. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinformatics 2003;4:43.
[34] Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 2004;6:938–47.
[35] Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 2006;7:8.
[36] Wright F. The 'effective number of codons' used in a gene. Gene 1990;87:23–9.
[37] Comeron JM, Aguade M. An evaluation of measures of synonymous codon usage bias. J Mol Evol 1998;47:268–74.
[38] Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 2002;19:1390–4.
[39] Elhaik E, Landan G, Graur D. Can GC content at third-codon positions be used as a proxy for isochore composition? Mol Biol Evol 2009;26:1829–33.
[40] Fuglsang A. Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol Biol Evol 2006;23:1345–7.
[41] Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 2004;5:163.

[42] Mohammed MH, Ghosh TS, Reddy RM, Reddy CV, Singh NK, et al. INDUS — a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. BMC Genomics 2011;12(Suppl. 3):S4.
[43] Reddy RM, Mohammed MH, Mande SS. TWARIT: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. Gene 2012; 505:259–65.
[44] Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res 2005;33:922–33.
[45] Baran RH, Ko H. Detecting horizontally transferred and essential genes based on dinucleotide relative abundance. DNA Res 2008;15:267–76.
[46] Chaib De Mares M, Hess J, Floudas D, Lipzen A, Choi C, et al. Horizontal transfer of carbohydrate metabolism genes into ectomycorrhizal *Amanita*. New Phytol 2014; 205(4):1552–64.
[47] Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol 2010;8:373–82.
[48] Srividhya KV, Alaguraj V, Poornima G, Kumar D, Singh GP, et al. Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. PLoS One 2007;2:e1193.
[49] Roller M, Lucic V, Nagy I, Perica T, Vlahovicek K. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res 2013;41:8842–52.
[50] Vieira-Silva S, Rocha EP. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet 2010;6:e1000808.
[51] Krisko A, Copic T, Gabaldon T, Lehner B, Supek F. Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol 2014;15: R44.
[52] Lobo FP, Mota BE, Pena SD, Azevedo V, Macedo AM, et al. Virus-host coevolution: common patterns of nucleotide motif usage in *Flaviviridae* and their hosts. PLoS One 2009;4:e6282.