Bioinformatics, 34(12), 2018, 2129–2131
doi: 10.1093/bioinformatics/bty068
Advance Access Publication Date: 6 February 2018
Applications Note

OXFORD

Genome analysis

GCevobase: an evolution-based database for GC content in eukaryotic genomes

Dapeng Wang

Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK

Associate Editor: John Hancock

Received on December 11, 2017; revised on January 20, 2018; editorial decision on February 3, 2018; accepted on February 5, 2018

Abstract

Summary: How to comprehend the underlying mechanism behind the origin and evolution of genome composition such as GC content has been regarded as a long-standing crucial question, highlighting its biological significance and functional relevance. To varying extents, several systematically identified patterns of GC content variations are shown to be linked to a set of genomic features in the events of replication, transcription, translation and recombination, with strong contrasts between diverse phylogenetic or taxonomical groups. In this situation, we develop a repository—GCevobase—which houses compositional and size related data presented in various forms from 1118 genomes including 5 major clades of eukaryotic species such as vertebrates, invertebrates, plants, fungi and protists. It analyzes the cautiously selected sequences with clearly-defined bases and structures them under the taxonomical classification system (kingdom, phylum, class, order and family) at the genome and gene scales. It uses the diversified and intelligible graphs to show the statistical measurements of GC content in the sequence, at the three codon positions and at 4-fold degenerate sites and CDS length and their genome-wide correlations and display the evolutionary pathways of GC content by taking into account between-species orthologs and within-species paralogs for each annotated gene. In addition, a lot of internal and external links have been created, making it an effective communication between the data from individual genomes and the raw data are downloadable.

Availability and implementation: https://github.com/NextGenBioinformatics/GCevobase

Contact: dapeng.wang@plants.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The variability in GC content presents a striking property of genome composition in nucleotide sequences, with patterns of GC-rich and GC-poor regions being formed at the two levels such as genomes and genes. Variation in GC richness is remarkably characterized by two appreciably distinctive classes of taxonomical groups in both animals and plants, showing both lineage-specificity and gene-specificity (Cammarano *et al.*, 2009; Costantini *et al.*, 2009; Glemin *et al.*, 2014). Two large-scale studies derived from sophisticated sampling of species/genomes have identified that GC content is not only correlated to the genome features including gene expression and local recombination rate but also associated with some particular phenotypes such as body mass (Romiguier *et al.*, 2010; Serres-Giardi *et al.*, 2012). In the aspect of application, GC content proves its usefulness for correcting for the experimental bias and improving the accuracy of measurement for Next-Generation-Sequencing data analysis (Benjamini and Speed, 2012). A number of hypotheses have been proposed to explain the characters of GC content within species and between species but unfortunately none of them could fully and consistently interpret all the observations (Eyre-Walker and Hurst, 2001). Though extensive studies have been carried out on the interplay between GC content and other potentially relevant genomic characteristics, a dedicated resource for GC content in a more complete taxonomical sampling is still lacking. We report a database that stores the data for GC content and the in-depth analyses from an evolutionary point of view for the three purposes. First, we use

the identical framework and standard to process the protein-coding sequences and perform calculations on high-quality sequences in order to obtain reliable compositional parameters and size parameters such as effective length (the total length of the nucleotides in the unambiguous codons), which makes all the data comparable. Second, we organize the data in multiple layers of taxonomical classifications to investigate the similarity between closely-related species and the difference between distantly-related species in terms of genome composition. Third, we choose various forms of textual and graphical presentation in the unit of genomes and genes to unravel the trajectory of GC content evolution by integrating nucleotide compositional data with homolog data. Ultimately, the database shall provide a comprehensive map on how GC content evolves throughout the entire phylogeny of eukaryote organisms.

2 Materials and methods

Coding sequences (CDS) and functional annotation data were retrieved from a suite of Ensembl databases (https://www.ensembl.org, http://ensemblgenomes.org) such as Ensembl_release_88 (85 genomes), Ensembl_Metazoa_release_35 (68 genomes), Ensembl_Plants_release_35 (44 genomes), Ensembl_Fungi_release_35 (735 genomes) and Ensembl_Protists_release_35 (186 genomes). For each transcript, GC content for all bases together with that in three codon positions such as GC1, GC2 and GC3 as well as GC content at 4-fold degenerate sites (GC4d) were calculated with in-house Perl scripts and transcript features were extracted from the definition lines for each sequence of Fasta files. In particular, effective sequences are referred to as those codons that have clear and unambiguous bases in all three positions and both effective length and effective codon number as well as other compositional properties were computed from the effective sequences. In order to achieve a set of clean genes for each genome, the transcripts with 'protein_coding' labels in both 'gene_biotype' and 'transcript_biotype' were retained and the transcripts at the greatest effective length were chosen to represent their genes in the gene-level analysis. To define the completeness and evaluate the sequence quality, each transcript has been categorized into three groups such as 'perfect', 'complete' and 'partial'. 'Perfect' transcripts are expressed as those that have no ambiguous bases (i.e. 'N') throughout the entire sequence and have a start codon in the beginning and a stop codon in the end. In contrast, 'complete' transcripts are defined as those that have a start codon in the beginning and a stop codon in the end but include a number of ambiguous bases (i.e. 'N') in the middle codons. 'Partial' transcripts are defined as those that are not assigned to 'perfect' and 'complete' categories. Homology data including gene homologous relationship and protein identity were collected from Ensembl Compara resources and two specific categories of homologs were selected for further analysis such as 'within_species_paralog' and 'ortholog_one2one'. For ortholog data visualization, only 17 model genomes were picked due to the restriction of space occupancy of big data, which are *Anolis carolinensis*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Xenopus tropicalis*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Oryza sativa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Dictyostelium discoideum*, *Emiliania huxleyi*, *Tetrahymena thermophila* and *Thalassiosira pseudonana*. Taxonomy data were taken from The Taxonomy Database (ftp://ftp.ncbi.nih.gov/pub/taxonomy) in terms of the assignment of multi-faceted ranks for each genome such as kingdom, phylum, class, order, family, genus and species, through the guidance of taxonomy ID. To test the correlation between genome size and GC content, C-value data were extracted from Eukaryotic genome size databases (http://genomesize.com/, http://data.kew.org/cvalues/, http://www.zbi.ee/fungal-genomesize/) and integrated into this database.

3 Results

The data in the database is arranged in a three-level hierarchy such as source -> genome -> gene and the browsing functions have been offered for each of the levels. In each category of sources, all kinds of genomes are sorted according to their taxonomical categories and each taxonomical term is clickable and will be able to lead to the page showing all genomes that have been assigned to this category, making it possible to compare all genomes defined under this category. On the page of gene details, the core data are made up of a list of fundamental parameters for each gene in terms of annotation, composition and size, for instance, 'Source', 'Species Name', 'Transcript Name', 'Sequence Feature', 'Location Feature', 'Assembly Version', 'Location', 'Start', 'End', 'Strand', 'Gene Name', 'Gene Biotype', 'Transcript Biotype', 'Gene Symbol', 'Description', 'CDS Length', 'Codon Number', 'Effective Length', 'Effective Codon Number', 'First Codon', 'Last Codon', 'Completeness', 'GC (%)', 'GC1 (%)', 'GC2 (%)', 'GC3 (%)' and 'GC4d (%)'. Moreover, the gene ontology (GO) terms in three levels such as 'biological_process', 'molecular_function' and 'cellular_component' are provided to give the detailed functional annotation classifications for each gene or transcript. At genomic scale, it shows the density distributions of compositional parameters of which the shapes reveal the extents of the heterogeneity of all protein-coding genes in a complete genome (Supplementary Fig. S1A and B). For a better presentation, heatmap-like scatter plot is chosen to draw two-dimensional distributions for compositional versus size parameters, in which each data point has been placed in an appropriate bin to indicate the enrichment (Supplementary Fig. S1C and D). From an evolutionary viewpoint, color codes are used to produce the images by plotting mean against standard deviation for the key parameters and displaying the different taxonomical levels relative to the query genome (Supplementary Fig. S1E and F). More important, two similar color-coded approaches are adopted to explore the property of the query gene in the context of its other gene family members through comparing this gene with other orthologous and paralogous genes (Supplementary Fig. S1G and H). Since the empirical data shows that the isoforms in a gene might behave variably in the respect of nucleotide composition, the transcript-level calculations are also conducted, which is complementary to the gene-level analysis. The interaction of different types of functionality and data is enhanced by the existence of links between many of the related webpages. The statistical and download pages offer the high-level views of the data for all genomes and compressed tab-delimited file for processed data, respectively.

The primary objective of constructing this database is to facilitate the research centered on evolutionary dynamics of genomic composition and we are constantly maintaining its operation and staying abreast of the newly-released genomes with elevated annotation qualities.

Conflict of Interest: none declared.

References

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Cammarano,R. *et al.* (2009) The isochore patterns of invertebrate genomes. *BMC Genomics*, **10**, 538.

Costantini,M. *et al*. (2009) The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*, **10**, 146.

Eyre-Walker,A. and Hurst,L.D. (2001) The evolution of isochores. *Nat. Rev. Genet*., **2**, 549–555.

Glemin,S. *et al*. (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet*., **30**, 263–270.

Romiguier,J. *et al*. (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*., **20**, 1001–1009.

Serres-Giardi,L. *et al*. (2012) Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell*, **24**, 1379–1397.