*Bioinformatics*, 34(21), 2018, 3624–3630
doi: 10.1093/bioinformatics/bty392
Advance Access Publication Date: 14 May 2018
Original Paper

OXFORD

Sequence analysis

# Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA)

**Le Zhang[1,2,*,†], Ming Xiao[2,3,†], Jingsong Zhou[1] and Jun Yu[4,5,*]**

[1]College of Computer Science, Sichuan University, Chengdu 610065, China, [2]School of Computer and Information Science, Southwest University, Chongqing 400715, China, [3]College of Mobile Telecommunications, Chongqing University of Posts and Telecommunications, Chongqing 400000, China, [4]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and [5]University of Chinese Academy of Sciences, Beijing 100049, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally.

Associate Editor: John Hancock

Received on January 3, 2018; revised on April 25, 2018; editorial decision on May 7, 2018; accepted on May 9, 2018

## Abstract

**Motivation:** This study addresses several important questions related to naturally underrepresented sequences: (i) are there permutations of real genomic DNA sequences in a defined length (k-mer) and a given lineage that do not actually exist or underrepresented? (ii) If there are such sequences, what are their characteristics in terms of k-mer length and base composition? (iii) Are they related to CpG or TpA underrepresentation known for human sequences? We propose that the answers to these questions are of great significance for the study of sequence-associated regulatory mechanisms, such cytosine methylation and chromosomal structures in physiological or pathological conditions such as cancer.

**Results:** We empirically defined sequences that were not included in any well-known public databases as lineage-associated underrepresented permutations (LAUPs). Then, we developed a Jellyfish-based LAUPs analysis application (JBLA) to investigate LAUPs for 24 representative species. The present discoveries include: (i) lengths for the shortest LAUPs, ranging from 10 to 14, which collectively constitute a low proportion of the genome. (ii) Common LAUPs showing higher CG content over the analysed mammalian genome and possessing distinct CG*CG motifs. (iii) Neither CpG-containing LAUPs nor CpG island sequences are randomly structured and distributed over the genomes; some LAUPs and most CpG-containing sequences exhibit an opposite trend within the same *k* and *n* variants. In addition, we demonstrate that the JBLA algorithm is more efficient than the original Jellyfish for computing LAUPs.

**Availability and implementation:** We developed a Jellyfish-based LAUP analysis (JBLA) application by integrating Jellyfish (Marçais and Kingsford, 2011), MEME (Bailey, *et al.*, 2009) and the NCBI genome database (Pruitt, *et al.*, 2007) applications, which are listed as Supplementary Material.

**Contact:** zhangle06@scu.edu.cn or junyu@big.ac.cn

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

## 1 Introduction

For a given length of DNA comprising the four natural nucleotides A, T, G and C, all possible sequences of the same length but not in the same nucleotide order are referred to as sequence permutations (Bujnicki, 2002; Gill and Machattie, 1976; Jeltsch, 1999). Due to the existence of covalent nucleotide modifications, such as the methylation and hydroxymethylation of cytosine, these permutations are limited only to sequences composed of the four primary nucleotides (Koskinen, 2012). However, when we insist on a true distribution of all possible permutations in the current best collections comprising gigabase-magnitude genomic sequences that include the most dominant public DNA databases, such as GenBank, EMBL and DDBJ (Ouellette, 1998; Stoesser *et al.*, 1999; Tateno *et al.*, 2002), a subset of DNA sequences appears to be underrepresented; they are neither sequence defects nor errors since their abundance is above both thresholds. Such underrepresentation is expected to have broad indications, including sequence-specific regulatory motifs and chromosomal DNA or RNA structural landmarks, such as telomeric repeats of GGATTT. The underrepresented sequences have further demonstrated to be lineage-associated when examined over broad taxa.

Hampikian and Anderson (Hampikian and Andersen, 2007) initially defined these aforementioned sequences as nullomers. Subsequently, many scientists (Acquisti *et al.*, 2007; Herold *et al.*, 2008; Vergni and Santoni, 2016) continued Hampikian and Anderson's study and focused on the origin of nullomers. Here, we empirically define such sequences that have never existed in any well-known public databases as lineage-associated underrepresented permutations (LAUPs). Notably, the widely used public databases for this study include GenBank, EMBL and DDBJ (Ouellette, 1998; Stoesser *et al.*, 1999; Tateno *et al.*, 2002).

This study defines these aforementioned sequences as LAUPs, but not nullomers (Hampikian and Andersen, 2007), because of four considerations. First, nullomers or LAUPs are similarly defined if both sequences are limited to a particular lineage, but the former emphasizes the absence of a sequence and the latter emphasizes the characteristics of a sequence. The nullomers from mammals or even vertebrates are similar and may be classified as well as traced to some ancestral groups; however, if genomes of other lineages are analysed in a similar manner, then the nullomers are different, such as between vertebrates and arthropods. Second, when sequence variations in a population background are categorized, an allelic definition must be introduced. In such a context, a nullomer and its variants may be classified into major or minor alleles; if a nullomer becomes a minor allele and one of its variants becomes the major allele, then naming the nullomer is inappropriate. Third, since sequencing platforms consistently have error rates, nullomer sequences become empirically defined because a fraction of the nullomers may result from sequencing errors. Fourth, since the goals of this study are not limited to only nullomer sequences, we may have to define closely related sequences based on structural and conformational similarities. Thus, it is rather precise to use an underrepresented permutation notation to define these sequences.

The compositional dynamics, or degree of sequence variability, of DNA sequences, particularly human DNA sequences, are slightly biased in TpA (Yomo and Ohno, 1989) and CpG (Gardiner-Garden and Frommer, 1987). The best known CpG enrichments are the so-called CpG islands (CGI) sequences that are not only partitioned among half of the housekeeping genes (Byun *et al.*, 1999; Daniel Eller, 2007; Farré, 2007; Gardiner-Garden and Frommer, 1987; Lawson and Zhang, 2008; Yang *et al.*, 2015) but also regulate ~70% of the mammalian genome (Han *et al.*, 2008). Most CGI sequences are associated with the 5′ ends of house-keeping genes (Tykocinski and Max, 1984). CGI sequences are defined as those with the observe/expectation values >0.6 and whose GC content is >50% (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002).

Recent studies have indicated (Pan *et al.*, 2017; Pongor *et al.*, 2017; Yu *et al.*, 2013) that CpG sequence underrepresentation is largely due to the instability of cytosine methylation (Thellin *et al.*, 1999). Since both LAUPs and CGI sequences are underrepresented, we question if a connection exists between LAUPs and CGI sequences.

To investigate the connection between LAUPs and CGI sequences, we developed a Jellyfish-based LAUPs analysis application (JBLA) for efficient LAUPs computation by integrating several relevant software tools (Bailey *et al.*, 2009; Rozenberg *et al.*, 2008). Thus, JBLA not only investigates the CG content and motif of LAUPs as a bridge for studying CGI sequences but also for examining CG permutations in both LAUPs and CGI sequences.

This study generated the following interesting findings: (i) length ranges of the shortest LAUPs from 24 representative species, (ii) higher CG content and enriched CG*CG motifs among common LAUPs found in the mammalian genome and (iii) relatedness between GC-rich LAUPs and CGI sequences. In general, this study focused on the motifs and sequence composition patterns between LAUPs and CGI sequences since both LAUPs and CGI sequences are CG-enriched and underrepresented.

## 2 Materials and methods

### 2.1 Data source

The sequence data for this study included those of 24 representative species (Supplementary Table S1) from NCBI sites (Pruitt *et al.*, 2007) and were the latest versions of the GenBank GBFF (GenBank Flat File) format (Ouellette, 1998). Supplementary Table S1 lists the names, data sizes and genome data versions for all 24 species. We divided the 24 species into 5 lineages: (i) Bacteria, (ii) Plants, (iii) Human, (iv) Primates (other than Human) and (v) Mammals (excluding Primates). We used a Perl script to remove unnecessary annotations and extract the genome sequence data, and we used the sequences and their reverse-complement sequences (Segerstéen *et al.*, 1986) in this analysis. Notably, this study focused on lineage-associated species and analysed common LAUPs for the same lineage to avoid the impact of different genome sizes because it is difficult to normalize the genome sizes for different lineages.

### 2.2 LAUP analysis algorithm

The LAUPs analysis algorithm is shown in Figure 1, which comprises LAUPs computing (Supplementary Fig. S1A) and data analysis components (Supplementary Fig. S1B and C). We detail the components and define several important terms for the analysis in Table 1.

#### 2.2.1 LAUPs computing component

This component was used to obtain a set of LAUPs for each species, such as *Homo sapiens* and *Pan troglodytes* (Pruitt *et al.*, 2007). The related algorithm is detailed in Supplementary Method 1.1.

#### 2.2.2 Data analysis component

This component consists of content analysis process, motif analysis and CpG-containing sequences analysis scenarios.

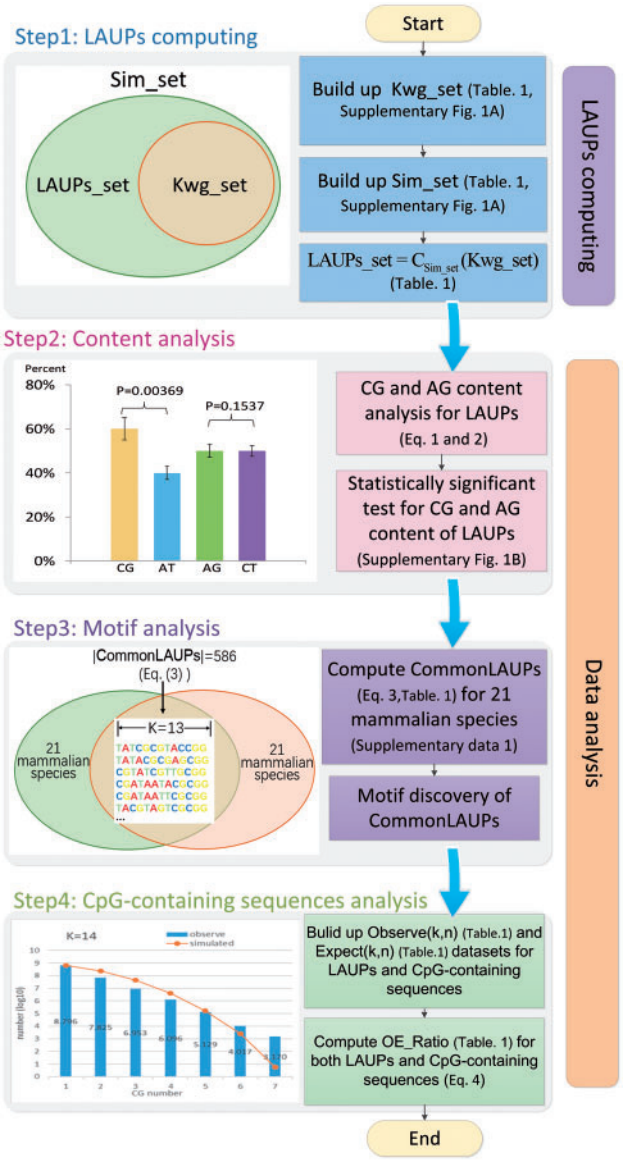


**Fig. 1.** Workflow of the LAUPs analysis algorithm. Here, the terms of Kwg_set, Sim_set and LAUPs_set are detailed in Table 1 (Color version of this figure is available at *Bioinformatics* online.)

(1) Content analysis: we computed the CG and AT content proportion as well as the purine (AG) and pyrimidine (TC) content proportion of LAUPs. We employed a well-developed statistical testing process (Zhang *et al.*, 2017a, b) to validate the content statistical significance between CG and AT as well as the difference between purine (AG) and pyrimidine (TC) for LAUPs. The related algorithm is detailed in Supplementary Method 1.2.

(2) Motif analysis process: sequence motifs are short and recurring patterns in DNA sequences with potential biological function (D'Haeseleer, 2006). The motif search procedure is called motif discovery (Bailey and Elkan, 1994). This study employed a motif search tool, MEME (Bailey *et al.*, 2009), to search for frequently occurring or common LAUP patterns of various species.

(3) CpG-containing sequences analysis: this analysis investigated the CG frequency between commonLAUPs and CpG-containing human sequences. The related algorithm is detailed in Supplementary Methods 1.3 and 1.4.

## 3 Results

### 3.1 The shortest LAUPs for the representative species

To define sequence permutations of different lengths or k-mers in this study, we first built a dataset that contained all possible sequence permutations, and subsequently searched for sequence permutations at a given k-mer length of a genomic sequence to investigate their representations, CG contents, motifs and proportions among the permutations. Although we conducted this analysis for most species, this study focused on mammalian genome sequences, which are further partitioned into three categories: humans, primates (non-human primates) and mammals (non-primate mammals); the genomes of plants and bacteria were used as controls (Supplementary Table S1).

We show that the length of the shortest LAUPs lies between 10 and 14 for whole genomes (WG) (Fig. 2A), and the shortest LAUPs average proportion rapidly and non-linearly increases with increasing length of the sequences (Fig. 2B and Supplementary Fig. S2A).

### 3.2 GC and AG content analysis for LAUPs

After determining the length range of the shortest LAUPs, we used Equations (1) and (2) to explore the GC and AG (purine) content of the LAUPs in 24 representative species (Supplementary Table S1).

**Table 1.** Terms for sequence analysis

| Definition | Description |
|---|---|
| LAUPs | The sequences that never exist in major public databases with respect to a given lineage |
| Sim_set | Comprises all possible $4^K$ k-mers. |
| Kwg_set | Collects the k-mers from the existing public databases of the species |
| LAUPs_set | The complement of Kwg_set with respect to Sim_set |
| The shortest LAUPs' average proportion (Fig. 2) | The number of the shortest LAUPs over that of all permutations |
| CpG-containing sequence | The sequence comprising CG*, where * indicates any nucleotides |
| CommonLAUPs | LAUPs intersection set for 21 mammalian species |
| Observe ($k$, $n$) | The number of k-mer(s) that consist of an $n$ number of CpGs for each $k$. Here, $n$ is the number of CpGs in a k-mer, and $k$ is the length of a k-mer |
| Expect ($k$, $n$) | The number of k-mer(s) that consist of an $n$ number of CpGs for each $k$ in the random condition (Supplementary Method 1.4) |
| OE_Ratio ($k$, $n$) | The ratio between Observe ($k$, $n$) and Expect ($k$, $n$), detailed in Equation (4) |
| Positive state ($k$, $n$) | If OE_Ratio ($k$, $n$) is >1 |
| Negative state ($k$, $n$) | If OE_Ratio ($k$, $n$) is ≤1 |

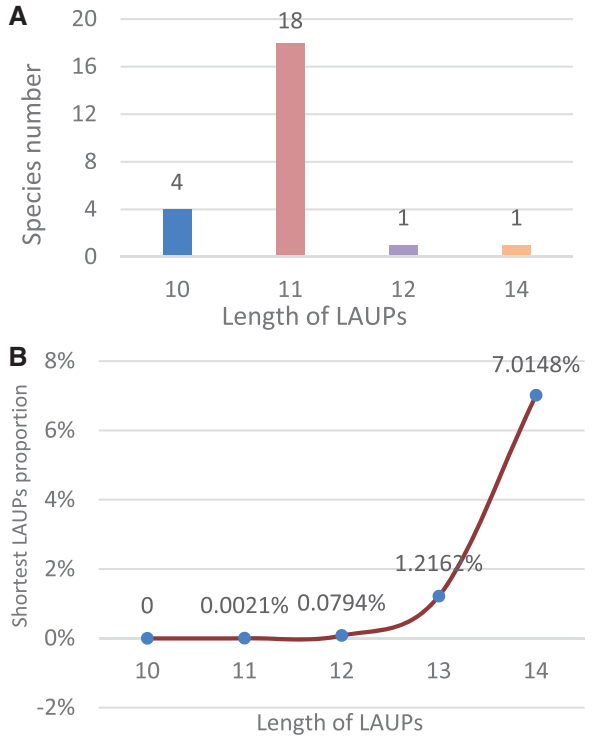


**Fig. 2.** The shortest LAUPs for the representative species. (**A**) The number of species for the shortest LAUPs. (**B**) The average proportion of the shortest LAUPs (denoted in Table 1) (Color version of this figure is available at *Bioinformatics* online.)

$$content\_GC(LAUP_K) = \frac{num(G_K) + num(C_K)}{k} \quad (1)$$

$$content\_AG(LAUP_K) = \frac{num(A_K) + num(G_K)}{k} \quad (2)$$

Here, $k$ is the length of LAUPs, where function num() represents the number of the bases; content_GC() and content_AG() are used to compute the CG and AG (purine) contents in LAUPs, respectively. As shown in Figure 3, the GC content appeared to be over-represented among LAUPs, ~60%, albeit a constant AG (purine) content of 50% was also observed (Supplementary Table S2). The GC content of LAUPs was significantly greater than that of the AT, and no significant difference was observed between AG (purine) and CT (pyrimidine) (left figure of Step 2 in Fig. 1 and Supplementary Table S3). We detail the statistical test procedure (Zhang *et al.*, 2017a, b) in Supplementary Figure S1B, and the sample data are shown in Supplementary Table S1.

### 3.3 Motif analysis for LAUPs

We used Equation (3) to compute the LAUPs intersection set (CommonLAUPs) for 21 mammalian species (Supplementary Table S1).

$$CommonLAUPs = \cap\left(\sum_{i=1}^{21} LAUPs_k^i\right) \quad (3)$$

Here, $LAUPs_k^i$ defines the LAUPs set for each mammalian species, and $i$ is the index of species.

The length of the shortest LAUPs for the mammalian lineage is 13, and CommonLAUPs consists of 586 LAUPs (left figure of Step 3 in Fig. 1). We also investigated putative motifs among the common

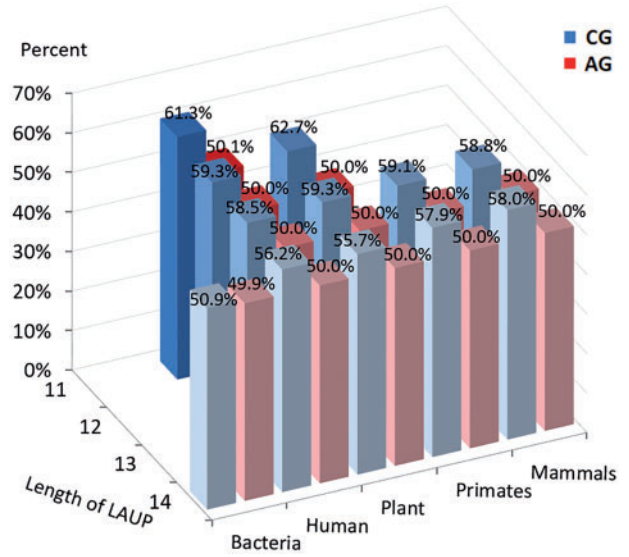


**Fig. 3.** AG (purine) and GC contents of some common LAUPs (Color version of this figure is available at *Bioinformatics* online.)

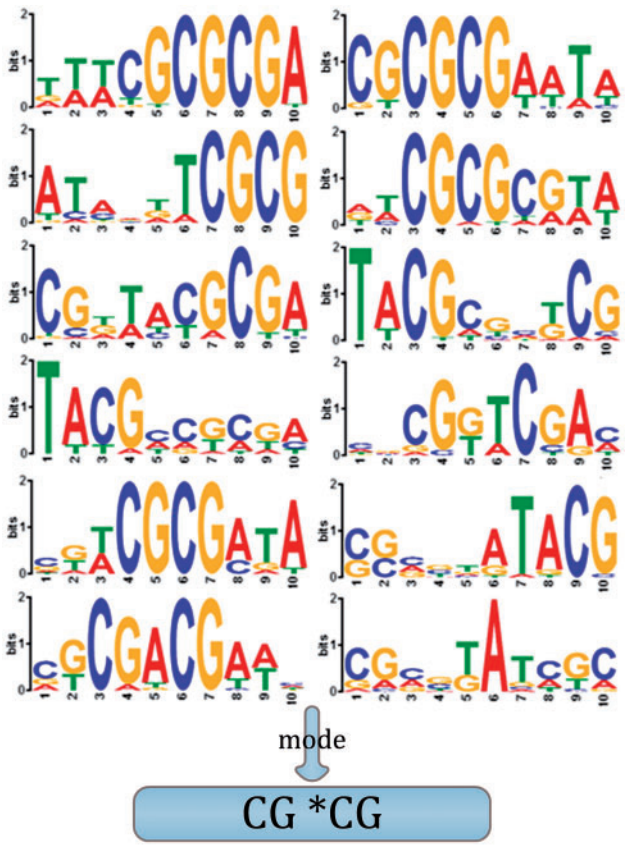


**Fig. 4.** Motif discovery logos for 21 typical mammalian species. The horizontal and vertical axes represent the position of the motif and its 'bits per position' (D'Haeseleer, 2006), respectively (Color version of this figure is available at *Bioinformatics* online.)

LAUPs (Bailey *et al.*, 2009) by Motif Discovery logos (Schneider and Stephens, 1990). Figure 4 reveals a striking common characteristic: the CpG dinucleotide sequence appeared biased towards a series of CG*CG sequences.

### 3.4 CpG-containing sequences analysis

The LAUPs among mammals had a higher GC content, consistent with that of CGI sequences (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). Further examination revealed a pattern of CpG*CpG in most high-GC LAUPs (Fig. 3). Since CGI sequences may play regulatory roles (Yu *et al.*, 2013; Zhu *et al.*, 2008), we examined the relatedness of the two GC-rich sequence groups, LAUPs versus CGI, in terms of structural and conformational features. Here, we investigated the CGI definition (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002) and explored permutation patterns for the CGI sequences, considering only the CpG density but not the CpG permutation. However, CGI sequences are typically too long and too abundant to directly examine their permutation patterns; therefore, we selected short LAUPs as a bridge for the present comparative analysis.

Since the shortest length of CGI sequence is still 500 bp (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002), we decomposed CGIs into several short sub-subsequences and used 14-mers as the upper length limit because the length of the shortest *CommonLAUPs* [Equation (3)] for mammalian species was 13 (left figure of Step 3 in Fig. 1). The analysis was also subsequently extended to an upper limit of 14 to investigate common features for CpG-containing sequences (Fig. 5A).

$$OE\_Ratio(k,n) = \frac{Observe(k,n)}{\exp ect(k,n)} \tag{4}$$

Equation (4) defines OE_Ratio ($k$, $n$) for the description of differences between Observe ($k$, $n$) and Expect ($k$, $n$) (left figure of Step 4 in Fig. 1). The computation of Observe ($k$, $n$) and Expect ($k$, $n$) is detailed in Supplementary Methods 1.3 and 1.4, respectively. In Figure 5A, the OE_Ratio ($k$, $n$) appears positively related to both k and n (Supplementary Table S4). As indicated by Equation (4), the function OE_Ratio ($k$, $n$) describes the ratio between functions Observe ($k$, $n$) and Expect ($k$, $n$). If OE_Ratio is >1, then we set OE_Ratio as a positive state; otherwise, this function was set as a negative state. Interestingly, Figure 5B reveals that when commonLAUPs have a positive state, most human CpG-containing sequences show a negative state and vice visa for the same $k$ and $n$.

### 3.5 Algorithm performance comparison

In this study, we developed a JBLA to analyse LAUPs sequences with two major components (Fig. 1): LAUPs computation and data analysis. To compute LAUPs more efficiently, JBLA has improved Jellyfish (Marçais and Kingsford, 2011) in the following three aspects.

First, a novel inputting function based on Jellyfish is used to receive a length of k-mer for computing multiple LAUPs instead of one LAUP at a time. This innovation decreases the cost of file input and output (I/O), especially for large datasets. The second and third innovations are based on two important parameters of Jellyfish (Marçais and Kingsford, 2011), which determine the memory size and computing efficiency for sequence analysis with Equation (5). One parameter is the length of k-mer, and the other parameter is the size of the counting field (Marçais and Kingsford, 2011). Since previous results have demonstrated that $k$ value ranges from 10 and 14, the second innovation initializes $k$ as 14 to reduce computing costs. Since it is important to determine whether k-mer appears or not instead of precisely how many times k-mer appears, the third innovation sets the counting field size to one for saving memory.

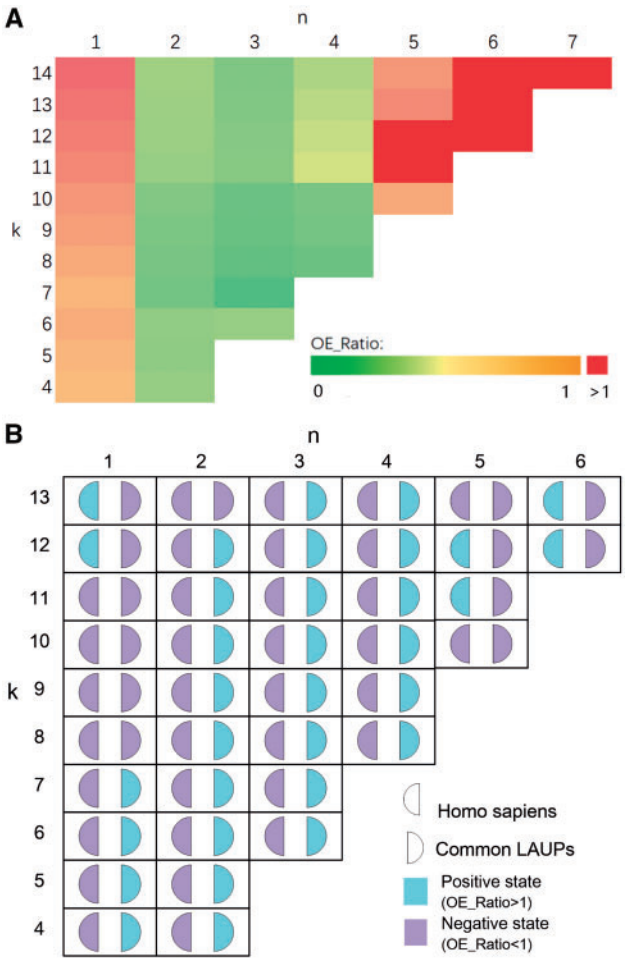


**Fig. 5.** OE_Ratio. (**A**) An OE_Ratio heat map for CpG-containing human sequences; (**B**) comparison between human CpG-containing sequences and common LAUPs (Color version of this figure is available at *Bioinformatics* online.)

$$Size(HashTable) = 1.25 * (G + k * b) \tag{5}$$

Here, $G$ is the size of a dataset, $k$ is the length of k-mer and $b$ is the length of counting field (unit: bit).

We compared the LAUPs computing efficiency between JBLA and Jellyfish by selecting three species (*Oryza sativa*, *Microcebus murinus* and *Miniopterus natalensis*) as the test datasets, employing both JBLA and Jellyfish to compute LAUPs 10 times when $k$ is from 9 to 11, and by using statistical tests (Fig. 6) to demonstrate statistically significant differences between JBLA and Jellyfish.

## 4 Discussion

Since CGI and LAUP sequences are both underrepresented among mammalian genomes, we analysed short LAUPs and explored their CpG-containing sequence permutations by using JBLA between the two GC-rich sequences.

We first revealed the length of the shortest LAUPs ranges from 10 to 14 (Fig. 2A and Supplementary Fig. S2). There are two factors considered relevant to this size range. One factor is the uniqueness or complexity of genome sequences; as far as mammalian genomes are concerned, $k = 15$ is the lower bound of such complexity, $\sim 3\times10^9$ bp ($4^{15}$). The other factor is the physical force for base pairing. The force of the shortest LAUPs should be between 10*14 = 140 pN and 14*20 = 280 pN,

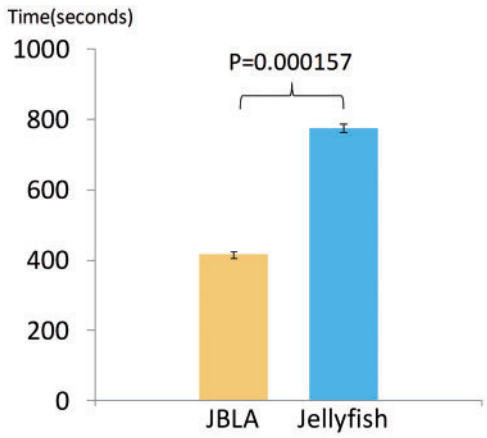


**Fig. 6.** The computing efficiency between JBLA and Jellyfish (Color version of this figure is available at *Bioinformatics* online.)

consistent with Zhang *et al.* (2015) that the base-pairing strength of single dG/dC and single *dA/dT* are 20.0 ± 0.2 pN and 14.0 ± 0.3 pN, respectively. Similarly, Essevaz-Roulet *et al.* (Essevaz-Roulet *et al.*, 1997) and Clausen-Schaumann *et al.* (Clausen-Schaumann, *et al.*, 2000) indicate that the maximum rupture forces of DNA duplexes are 300 pN, which are slightly greater than the low limit of the shortest mechanical force of LAUPs. Therefore, we consider that the length of the shortest LAUPs is mechanistically meaningful and reasonable, which could be used as a threshold for further research. Moreover, Figure 2 shows that the lengths of the shortest LAUPs for most species are ∼11 for most species. We consider that this interesting length range should be related to the length of a complete double helix, which is ∼10 (Chen *et al.*, 2008; Worning *et al.*, 2000).

We also consider that shortest LAUPs may start as unique sequences or become distributed over only certain limited sequence contexts because proportionally, some shortest LAUPs must be generally rare (Fig. 2B and Supplementary Fig. S2C), and these rare and unique sequences should be related to sequence-associated regulation mechanisms under physiological or pathological conditions.

Sequence characteristics are partitioned into two simple compositions, the GC and purine (A + G or simply AG) contents. We not only observe obvious higher GC content in the shortest LAUPs (Fig. 3; Supplementary Tables S2 and S3; Supplementary Fig. S1B) but also the higher GC content highlights CpG sequences rather GpC (Schweitzer and Kool, 1995). We further examined the sequence patterns among the LAUPs discovers a CG*CG mode from the common LAUPs of mammalian genomes (Fig. 4). Thus, consecutive CpG dinucleotides immediately prompted us to examine CGI sequences.

Since most CGI sequences are too long to directly compare with LAUPs for permutation patterns, we use short LAUPs and their CpG-containing sequences as a bridge to investigate those of CGI sequences (left figure of Step 3 in Fig. 1). Here, we use expected CpG-containing sequences to represent the sum of CGI sequences and non-CGI CpG-containing sequences. Obviously, only a true and *de novo* CGI sequence analysis provided an answer for the relatedness between CGI and LAUP sequences.

The results of this study are multi-fold. First, since both observed and expected CpG-containing sequences have frequency information attached, we determined that these sequences are not random (Fig. 5A) and their CpG*CpG modes and motifs are related to the two parameters (*k* and *n*). Second, when LAUPs have a positive state, most CpG-containing sequences show a negative state (Fig. 5B). Thus, their CpG*CpG motifs and modes appear to mirror each other, suggesting the existence of strong sequence or sequence motif selections, preserving or excluding some of the permutations and patterns within, which may serve as unique motifs for both trans and cis regulations, which are typically involved in protein-based recognition and/or chromosomal structural conformation, respectively. More detailed analysis is certainly expected, focusing on both CGI and LAUPs directly to search for meaningful sequence motifs and patterns, followed by experimental evaluation. A final notion is attributable to JBLA, which is much more efficient than Jellyfish for LAUPs computing (Fig. 6) and secures future analysis based on larger datasets.

Although this study explores several interesting findings based on the characteristics of LAUPs, further investigation is still needed. For example, we examined LAUPs and their variants in the context of pan-cancer genomes for regulatory sequence changes. In addition, genome-wide DNA sequences are abundant, but the cost to analyse these sequences are computing-intensive and time-consuming. Also, we do not have a database that hosts both species- and lineage-associated LAUPs from other sequenced genomes, which can offer a platform for other data analysis activities, such as data distribution and visualization. Therefore, our further study will employ high performance computing (Jiang *et al.*, 2015; Jiang *et al.*, 2011) and related data mining algorithms(Gao *et al.*, 2017; Jiang *et al.*, 2015; Jiang *et al.*, 2011; Peng *et al.*, 2014; Zhang *et al.*, 2017a, b; Zhang *et al.*, 2016) to speed up the Jellyfish software and build up a LAUPs warehouse in the distant future.

## Acknowledgements

This work was supported by the General Program from National Natural Science Foundation of China, Chongqing excellent youth award and the Chinese Recruitment Program of Global Youth Experts as well as by the Fundamental Research Funding of the Chinese Central Universities.

## Funding

This work was supported by National Natural Science Foundation of China [61372138], the National Science and Technology Major Project [2018ZX10201002], Chongqing Research Program of Basic Research and Frontier Technology [cstc2015jcyjA40026, No. cstc2016jcyjA0568] and Chinese Chongqing Distinguish Youth Funding [cstc2014jcyjjq40003].

*Conflict of Interest*: none declared.

## References

Acquisti,C. *et al.* (2007) Nullomers: really a matter of natural selection? *PLos One*, **2**, e1022.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bujnicki,J.M. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. *BMC Evol. Biol.*, **2**, 3.

Byun,R. *et al.* (1999) Evolutionary relationships of pathogenic clones of Vibrio cholerae by sequence analysis of four housekeeping genes. *Infect. Immun.*, **67**, 1116–1124.

Chen,K. *et al.* (2008) A novel DNA sequence periodicity decodes nucleosome positioning. *Nucleic Acids Res.*, **36**, 6228–6236.

Clausen-Schaumann,H. *et al.* (2000) Mechanical stability of single DNA molecules. *Biophys. J.*, **78**, 1997–2007.

Daniel Eller,C. *et al*. (2007) Repetitive sequence environment distinguishes housekeeping genes. *Gene*, **390**, 153–165.

D'Haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotechnol.*, **24**, 423–425.

Essevaz-Roulet,B. *et al*. (1997) Mechanical separation of the complementary strands of DNA. *Proc. Natl. Acad. Sci. USA*, **94**, 11935–11940.

Farré,D. *et al*. (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.*, **8**, R140.

Gao,H. *et al*. (2017) Developing an agent-based drug model to investigate the synergistic effects of drug combinations, *Molecules*, **22**, 2209.

Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.

Gill,G.S. and Machattie,L.A. (1976) Limited permutations of the nucleotide sequence in bacteriophage T1 DNA. *J. Mol. Biol.*, **104**, 505.

Hampikian,G. and Andersen,T. (2007) Absent sequences: nullomers and primes. *Pac. Symp. Biocomput.*, **12**, 355–366.

Han,L. *et al*. (2008) CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol.*, **9**, R79–R12.

Herold,J. *et al*. (2008) Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, **9**, 167.

Jeltsch,A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.*, **49**, 161–164.

Jiang,B.N. *et al*. (2015) Novel 3D GPU based numerical parallel diffusion algorithms in cylindrical coordinates for health care simulation. *Math. Comput. Simulat.*, **109**, 1–19.

Jiang,B.N. *et al*. (2011) Employing graphics processing unit technology, alternating direction implicit method and domain decomposition to speed up the numerical diffusion solver for the biomedical engineering research. *Int. J. Numer. Meth. Bio.*, **27**, 1829–1849.

Koskinen,A.M.P. (2012) Nucleosides, nucleotides, and nucleic acids. In: *Asymmetric Synthesis of Natural Products*. John Wiley & Sons, New York, pp. 175–185.

Lawson,M. and Zhang,L. (2008) Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5 '-UTR region. *Gene*, **407**, 54–62.

Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Ouellette,B.F. (1998) The GenBank sequence database. *Methods Biochem. Anal.*, **39**, 16.

Pan,H. *et al*. (2017) CpG and methylation-dependent DNA binding and dynamics of the methylcytosine binding domain 2 protein at the single-molecule level. *Nucleic Acids Res.*, **45**, 9164–9177.

Peng,H. *et al*. (2014) Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. *Bioinformatics*, **30**, 1899–1907.

Pongor,C.I. *et al*. (2017) Optical trapping nanometry of hypermethylated CPG-island DNA. *Biophys. J.*, **112**, 512.

Pruitt,K.D. *et al*. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Rozenberg,J.M. *et al*. (2008) All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC Genomics*, **9**, 67.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Schweitzer,B.A. and Kool,E.T. (1995) Hydrophobic, non-hydrogen-bonding bases and base pairs in DNA. *J. Am. Chem. Soc.*, **117**, 1863.

Segerstéen,U. *et al*. (1986) Frequent occurrence of short complementary sequences in nucleic acids. *Biochem. Biophys. Res. Commun.*, **139**, 94.

Stoesser,G. *et al*. (1999) The EMBL nucleotide sequence database. *Mol. Biotechnol.*, **33**, 29–33.

Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA*, **99**, 3740–3745.

Tateno,Y. *et al*. (2002) DNA data bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.*, **30**, 27.

Thellin,O. *et al*. (1999) Housekeeping genes as internal standards: use and limits. *J. Biotechnol.*, **75**, 291–295.

Tykocinski,M.L. and Max,E.E. (1984) CG dinucleotide clusters in MHC genes and in 5' demethylated genes. *Nucleic Acids Res.*, **12**, 4385–4396.

Vergni,D. and Santoni,D. (2016) Nullomers and high order nullomers in genomic sequences. *PLoS One*, **11**, e0164540.

Worning,P. *et al*. (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in Thermotoga maritima. *Nucleic Acids Res.*, **28**, 706.

Yang,Y. *et al*. (2015) Sequence analysis of housekeeping genes and virulence-related genes and antimicrobial susceptibility testing of bordetella pertussis strains isolated in China. In: *International Symposium on Antimicrobial Agents and Resistance*, pp. S119–S120.

Yomo,T. and Ohno,S. (1989) Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. *Proc. Natl. Acad. Sci. USA*, **86**, 8452–8456.

Yu,D.H. *et al*. (2013) Developmentally programmed 3' CpG island methylation confers tissue- and cell-type-specific transcriptional activation. *Mol. Cell. Biol.*, **33**, 1845.

Zhang,L. *et al*. (2016) Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. *Nanoscale*, **8**, 14877.

Zhang,L. *et al*. (2017a) EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. *J. Mol. Cell Biol.*, **9**, 477–488.

Zhang,L. *et al*. (2017b) Building up a robust risk mathematical platform to predict colorectal cancer. *Complexity*, **2017**, 1–14.

Zhang,T.B., *et al*. (2015) Determination of base binding strength and base stacking interaction of DNA duplex using atomic force microscope.. *Sci Rep.*, **5**, 9143.

Zhu,J. *et al*. (2008) On the nature of human housekeeping genes. *Trends Genet.*, **24**, 481.