Bioinformatics, 31(8), 2015, 1293–1295
doi: 10.1093/bioinformatics/btu803
Advance Access Publication Date: 4 December 2014
Applications Note

OXFORD

Genome analysis

GenomeCons: a web server for manipulating multiple genome sequence alignments and their consensus sequences

Tetsuya Sato[1,2] and Mikita Suyama[1,2,*]

[1]Medical Institute of Bioregulation, Kyushu University and [2]CREST, Japan Science and Technology Agency, 812-8582 Fukuoka, Japan

[*]To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 30, 2013; revised on November 19, 2014; accepted on November 27, 2014

Abstract

Summary: Genome sequence alignments provide valuable information on many aspects of molecular biological processes. In this study, we developed a web server, GenomeCons, for manipulating multiple genome sequence alignments and their consensus sequences for high-throughput genome sequence analyses. This server facilitates the visual inspection of multiple genome sequence alignments for a set of genomic intervals at a time. This allows the user to examine how these sites are evolutionarily conserved over time for their functional importance. The server also reports consensus sequences for the input genomic intervals, which can be applied to downstream analyses such as the identification of common motifs in the regions determined by ChIP-seq experiments.

Availability and implementation: GenomeCons is freely accessible at http://bioinfo.sls.kyushu-u.ac.jp/genomecons/

Contact: mikita@bioreg.kyushu-u.ac.jp

1 Introduction

The progress of genome sequencing projects for various species has now made it possible to compare genome sequences over a wide range of related species. Genome sequence alignments provided by the UCSC Genome Browser (Karolchik *et al.*, 2014) greatly facilitate the identification of the regions under strong selective pressure by measuring the conservation in related species.

Genome sequence alignments in the UCSC Genome Browser are useful for the visual inspection of a genomic region of interest, but they are not suitable for downloading to perform further detailed analyses. Moreover, the browser can handle only a single locus at a time, making it difficult to survey many loci, for example, a set of the genomic regions identified by ChIP-seq experiments.

To conduct a large-scale analysis of genome sequence alignments, it is necessary to download and manipulate the data stored in the Multiple Alignment Format (MAF) from the UCSC Genome Browser (Karolchik *et al.*, 2014). As expected based on the length of an entire genome and the number of species available, the amount of data in these genome sequence alignments is huge. Furthermore, genome sequences of any two species are not always collinear because of the genome rearrangements that occur during the course of evolution. Thus, genome alignments consist of a huge set of small alignment blocks because of the inclusion of species with various levels of diversity. Therefore, it is not a trivial task to retrieve a genome sequence alignment for a region of interest from a huge amount of MAF files.

A possible solution to this problem is to use a set of specialized tools for manipulating genome sequence alignments (Blankenberg *et al.*, 2011), which have been implemented on the Galaxy web server (Goecks *et al.*, 2010). Recently, another program, MafFilter, was developed for this task (Dutheil *et al.*, 2014). These programs facilitate the highly flexible manipulation of MAF files, but they also require command-line and/or multi-step operations, which are

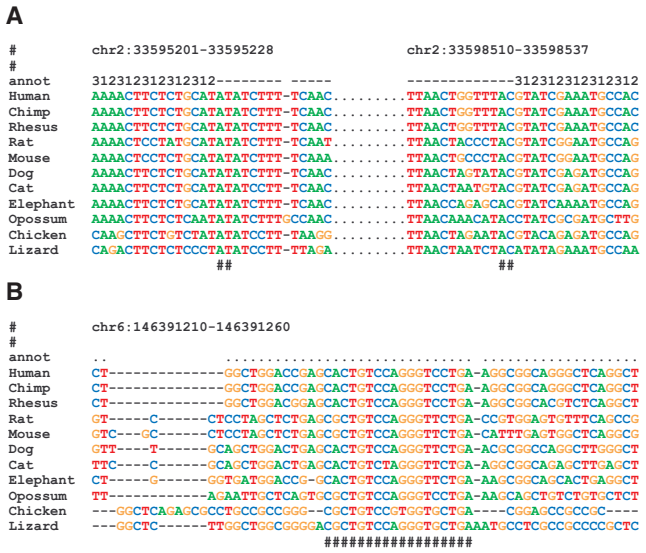

Fig. 1. Examples of genome alignment retrieved by GenomeCons. The genomic coordinates are based on the hg18 assembly. The annotation for each residue is indicated at the top of the alignment: '1–3', codon positions; '-', intron; '.', intergenic region. The conserved motifs are indicated by '#' under the alignments. (**A**) U12-type intron in the *RasGRP3* gene. The 5′ splice site (AT) and the 3′ splice site (AC) are highly conserved from humans to lizards. (**B**) Conserved transcription factor binding site for NRSF

not always straightforward for researchers who are not familiar with these tasks.

In this study, we introduce our web server, GenomeCons, a highly user-friendly, dedicated web server for manipulating genome sequence alignments and for the automatic generation of consensus sequences of the corresponding genomic intervals in one step. GenomeCons is well suited for large-scale analyses of genome sequence alignments, for example, the identification of evolutionarily conserved transcription factor binding sites in the peak regions identified by ChIP-seq experiments.

2 Implementation

We used multiple genome sequence alignments provided by the UCSC Genome Browser (Karolchik *et al.*, 2014) in the MAF format for the following four reference species: human, mouse, medaka and *Drosophila melanogaster*. We converted the MAF blocks into a consecutive alignment for each chromosome of the reference species. This concatenation approach has a clear disadvantage because it only retains the information about the genomic coordinates of the reference species, whereas it loses that for the aligned species, but the concatenated file makes it possible to access the query regions in a file very quickly via a random access function implemented during programming. To achieve even quicker access to a file, we installed a solid-state drive for storing the concatenated alignment files on our server.

As inputs, GenomeCons takes either a genomic interval of interest by specifying a chromosome and its coordinates or a set of genomic intervals in BED format, which is used widely in high-throughput genome analyses. The genomic coordinates are then converted into the coordinates in the concatenated alignment using the liftOver program (Kuhn *et al.*, 2013) on the basis of the chain files that we constructed for coordinate conversion between genomic coordinates and the positions in the concatenated alignment. This process is necessary because of the insertions and deletions in the alignments. There are also some optional settings on the web server for tweaking the output.

The server returns the output in three formats: (i) multiple genome sequence alignments of the input regions; (ii) consensus sequences where the sites with low phastCons scores (Siepel *et al.*, 2005), which is a metric that represents the conservation at each aligned position, are masked with 'N's and (iii) consensus sequences divided by low phastCons scores. The cut-off phastCons scores can be specified by the users, but the default value is set to 0.2. The users can also specify the species that will be reported in the output alignments with the 'species selection' option. The consensus sequence output has an advantage in downstream analyses, for example, for the identification of common motifs in the peak regions defined by ChIP-seq experiments. In general, *de novo* motif identification is a CPU-intensive task and the computational time depends on the amount of input residues. The use of consensus sequence, which can be thought of as the regions under selective constraints, can dramatically reduce the number of residues that need to be processed and enhance signal-to-noise ratio during motif identification.

3 Usage

Next, we provide two brief examples of the multiple genome sequence alignments retrieved by GenomeCons. One example illustrates the conservation of splice sites. Some introns are spliced out by a U12-dependent spliceosome, and these introns often have AT and AC at the 5′ and 3′ splice sites, respectively. One of these introns can be found in the *RasGRP3* gene. GenomeCons can easily be used to show that the splice sites are conserved from humans to lizards (Fig. 1A). Another example is a genome sequence alignment for the region of a transcription factor binding site identified by a ChIP-seq experiment for the neuron-restrictive silencer factor (NRSF). The data for the ChIP-seq peaks were downloaded from a previous study (Håndstad *et al.*, 2011). A representative case shows clearly that the binding motif is highly conserved among vertebrates, which contrasts with the remaining regions in the alignment (Fig. 1B).

Genome sequence alignments are valuable resources for biomedical applications as well as basic molecular biological studies. For example,

in genome-wide association studies, there are often cases where no genes exist in the linkage disequilibrium (LD) block that contains statistically significant markers. If there are some conserved non-coding regions in the LD block, these can be candidates for disease susceptibility (Ward and Kellis, 2012). Similarly, during exome analysis, single nucleotide variants (SNVs) that are specifically identified in patients may be implicated in pathogenesis even if they are in synonymous sites because they have functions other than coding such as exonic splicing regulators (Chamary *et al.*, 2006). Evaluating the conservation of the positions of synonymous SNVs in multiple genome sequence alignments may facilitate the inference of their functional relevance to pathogenicity.

Funding

Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan [26550089 to T.S.; 21510215 and 22132005 to M.S.] and Kyushu University Interdisciplinary Programs in Education and Projects in Research Development (P&P).

Conflict of Interest: none declared.

References

Blankenberg,D. *et al.* (2011) Making whole genome multiple alignments usable for biologists. *Bioinformatics*, **27**, 2426–2428.

Chamary,J.V. *et al.* (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.*, **7**, 98–108.

Dutheil,J.Y. *et al.* (2014) MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, **15**, 53.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Håndstad,T. *et al.* (2011) A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites. *PLoS One* **6**, e18430.

Karolchik,D. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.

Kuhn,R.M. *et al.* (2013) The UCSC Genome Browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.