BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 17 2011, pages 2426–2428
doi:10.1093/bioinformatics/btr398

Genome analysis

Advance Access publication July 19, 2011

Making whole genome multiple alignments usable for biologists

Daniel Blankenberg[1,†], James Taylor[2,*,†], Anton Nekrutenko[1,*,†] and The Galaxy Team[†]

[1]The Huck Institutes for the Life Sciences and Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA and [2]Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Here we describe a set of tools implemented within the Galaxy platform designed to make analysis of multiple genome alignments truly accessible for biologists. These tools are available through both a web-based graphical user interface and a command-line interface.

Availability and Implementation: This open-source toolset was implemented in Python and has been integrated into the online data analysis platform Galaxy (public web access: http://usegalaxy.org; download: http://getgalaxy.org). Additional help is available as a live supplement from http://usegalaxy.org/u/dan/p/maf.

Contact: james.taylor@emory.edu; anton@bx.psu.edu

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

Received on March 22, 2011; revised on May 18, 2011; accepted on June 27, 2011

1 INTRODUCTION

With the emergence and rapid proliferation of new sequencing technologies, data generation is no longer a major challenge in genomics. Unfortunately, the relative ease of genome sequencing does not automatically translate into the expanding of biological knowledge—it is still quite difficult to decipher the functional significance of genomic DNA. One reason for this is that the vast majority of functional studies have focused on annotating the genomes of human and model organisms. Whole genome alignments offer a solution to this challenge. By aligning newly sequenced genomes against well-annotated sequences, one can obtain a variety of functional, structural and evolutionary insights.

Yet there are still two formidable barriers preventing biomedical scientists, the ultimate 'consumers' of whole genome alignments, from effectively utilizing them in their research. First, whole genome alignments are very large. For example, the existing alignment sets of 28 mammalian species (Miller *et al*., 2007) and 46 vertebrate species (Fujita *et al*., 2011) occupy several hundred gigabytes of disk space and contain millions of alignment blocks. Handling data at this scale presents challenges even for researchers with extensive programming experience, while for most experimental biologists they are simply beyond reach. Second, the data exists in a specialized format, the Multiple Alignment Format (MAF; Supplementary Fig. S1). Although the MAF format is versatile and contains the information necessary for interpreting the alignments, it is currently not readily accepted or processed by downstream applications.

Here we describe a set of tools, available through both a web-based graphical user interface (GUI) and a command-line interface, designed to address challenges faced when working with these data. No downloads are required in order to use the GUI version of the tools, as they have been implemented into the web-based genome analysis platform Galaxy (Blankenberg *et al*., 2007, 2010; Goecks *et al*., 2010; Taylor *et al*., 2007). Galaxy aims to bridge the gap between the data and successful analysis.

Available freely as a public service (http://usegalaxy.org) and as an open source software project (http://getgalaxy.org), Galaxy can be deployed in individual labs and on Cloud resources (Afgan *et al*., 2010). Galaxy features a history system that tracks user inputs and parameter settings, ensuring that analyses can be reproduced precisely, as well as a seamless workflow system that allows reusable multiple tool pipelines to be created by extracting from an existing analysis history or through using an interactive drag and drop interface. Not only are researchers able to share their analysis histories and workflows with colleagues or the greater scientific community, but they can also compose complete analysis protocols (Pages) using a web-based word processor style (i.e. WYSIWYG) editor with built-in history and workflow embedding capabilities. Supplementary information for this manuscript is available as one of these Galaxy Pages (http://usegalaxy.org/u/dan/p/maf).

MAF format in brief: the MAF format has emerged as a *de facto* standard for storing and exchanging whole genome multiple alignments. Alignments stored in this format retain the sequence and genomic position information for the aligning sequence ranges. As a convention in Galaxy, sequences are named according to the source species genome build and sequence identifier within that build (generally a chromosome or contig); the genome build and sequence identifier are separated by a period. For example, the sequence of chromosome 21 from a March 2006 human genome reference assembly, known as hg18, would be named 'hg18.chr21'. Alignments are arranged in *blocks* separated by a blank line, where each block constitutes an individual set of sequence ranges (e.g. a single local alignment involving some set of species). These ranges need not be unique as a MAF set can contain overlapping blocks. In the MAF format, the genomic coordinates of alignments on the '-' strand are numbered relative to the reverse complement of the source sequence (unlike other common formats for genome annotation, such as GFF and BED). Though often an obstacle to biologists trying

[*]To whom correspondence should be addressed.

[†]http://usegalaxy.org.

to work with these files, this important difference in coordinate systems is resolved internally within this toolset and requires no additional effort or consideration on the part of users.

2 A SUITE OF TOOLS FOR MULTIPLE ALIGNMENT ANALYSIS

2.1 Alignment extractors

Commonly, using all of an alignment of entire genomes is neither practical nor desired. A frequent first step in an analysis utilizing a whole genome alignment is to extract a specific subset of interest, such as those corresponding to genes or other genomic elements. Thus, it is critical to be able to quickly and efficiently identify a set of alignment blocks that overlap a given set of genomic intervals. This presents the need for the first set of tools, known as alignment extractors (Supplementary Fig. S2). Within the public web server implementation, users can extract from a collection of alignments locally cached on the Galaxy server, from alignments provided via uploading from a computer or by copy and pasting a URL, or from alignments acquired directly from an external data source. Blocks that have start and end positions exceeding the requested region boundaries are trimmed. Additionally, blocks are automatically output in the strand-orientation of the provided genomic intervals.

2.2 Format converters

2.2.1 MAF to FASTA Most current alignment analysis programs are unable to recognize MAF alignment files, which presents the need for the second set of tools: format converters. Two approaches of MAF to FASTA conversion (Supplementary Fig. S3) are provided: one that creates a block-by-block multiple alignment FASTA file and another that creates a single alignment block. The tool that creates a multiple alignment FASTA file additionally allows the user to exclude blocks that are missing a requested species. The latter tool concatenates the sequences for all blocks, resulting in one large alignment block, where each species has exactly one sequence and blocks that lack a particular species will have the unaligned regions filled with gaps.

2.2.2 MAF to interval The MAF to interval converter allows users to extract genomic interval information from an alignment. This tool creates a tabular file containing the aligned sequence data and converts the included genomic region information into a zero-based half-open (BED-like) format.

2.3 Stitchers

In addition to the alignment extractors and format converters, there is a type of tool that acts as a hybrid of these two concepts. These tools, known as MAF ‘stitchers’ (Supplementary Fig. S4), produce exactly one alignment for each user-provided interval by resolving overlaps and ‘stitching’ adjacent blocks together. Two forms of stitching are provided here: one that works on standard genomic intervals and another that works on genomic intervals having additional fields that define protein-coding exons (i.e. BED12). BED12 files can be retrieved, for example, directly from the UCSC Table browser (Karolchik *et al.*, 2004) by selecting a gene track (e.g. RefSeq Genes) and selecting the option to create one BED record per ‘Whole gene’.

For each genomic interval that is specified, a single FASTA alignment block is created. This alignment block contains only the genomic positions that appear within the genome to which the genomic intervals belong; insertions in aligning species, relative to this reference species, are discarded. Overlapping blocks, allowed within MAF alignments, must be taken into account. Conceptually, these blocks are split at the boundaries of the overlap, and the original score for the alignment block is used to determine which aligning sequences are used on a per species basis; if a sequence is present for a species in a lower scoring alignment block but not in any higher-scoring block, then the sequence for that species is taken from the lower scoring alignment block. When performing this operation on a gene basis, only the positions, which are part of the protein-coding sequence for the reference species are included in the output; individual coding exons are processed as separate genomic intervals, as above, and these exons are concatenated together for each requested gene.

2.4 Tools for MAF manipulation

2.4.1 Species restriction The number of species included in publicly available pre-computed multiple alignments continues to increase, which makes these alignments more useful, but also more difficult to analyze, as more species results in more fragmented alignments. Extra gaps, such as those left over from an insertion found only in the genome of a removed species, should be eliminated. The limiting of an alignment to only desired species (Supplementary Fig. S5) is available as a separate tool or can be accomplished directly within the Extract MAF blocks tool interface.

2.4.2 Block Joiner A variant of the filter species tool is the Join MAF blocks by Species tool (Supplementary Fig. S6), which not only removes undesired species, but also concatenates adjacent MAF blocks. When the species that caused a block to be divided are removed, the split blocks can be joined together to create a single MAF alignment block. This requires the genomic positions of each species in the blocks-to-be-joined to have genomic positions that start and end directly adjacent to each other. When joining, the strand of each sequence is important, as positions listed in opposing orientations are considered different even if they are equivalent locations.

2.4.3 MAF Filter Many times there is no need to modify the contents of alignment blocks, but instead it is desirable to apply a set of filters that remove blocks that do not match a set of conditions. The Filter MAF by specified attributes tool allows users to build complex, multiple step filters that are applied to each alignment block. Examples of this include removing blocks, which lack species, removing blocks which have aligned species occurring between non-syntenic chromosomes or strands, removing blocks which are missing desired species and removing blocks which fall outside of a desired size range (Supplementary Fig. S7).

2.5 Alignment coverage

After modifying or filtering an alignment (or without modification), the MAF Coverage Stats tool allows viewing coverage information about the remaining blocks in reference to a particular set of intervals. There are two different types of output: one that provides information on a per interval basis and another that provides a summary over all intervals provided. Only positions that exist in the genome of the supplied genomic intervals are included in the output.

3 CONCLUSIONS

Due to the size of multiple species whole genome alignments, searching through the entirety of their contents to locate desired blocks is not practical; this has led to the utilization of a compression-capable indexing implementation that is a variation of the positional binning approach (Kent *et al*., 2002; Miller *et al*., 2007) stored on disk. The indexes for MAFs appearing in a user's history are generated during history item creation; when an index is not available, the command-line tools will create temporary index files on the fly. For larger locally cached alignments, the source MAF files are compressed and an associated lookup table is created to allow the interoperability of the indexes with the compressed data.

It is worth mentioning that each of the alignment sets locally cached by the public Galaxy server are actually composed of several individual MAF files. These files tend to be split by and named for the chromosomes of the reference (projected) genome of the alignment. For example, the 28-way alignment is divided according to the human chromosome found within each alignment block. This results in 49 individual compressed MAF files, indexes and lookup tables; this number is larger than the number of human chromosomes due to the 'random' chromosomal regions, several chromosomal haplotypes and the mitochondrial genome. It is not required that MAF sets be divided in this (or any) fashion, as the indices indicate which blocks are found in a particular MAF file, but this is a common release practice of the research groups creating the alignments and can allow greater flexibility with hardware and system concerns.

While all of the tools are designed to work directly out-of-the-box for personal Galaxy installations, additional steps are required to provide a collection of pre-cached source alignments to the extraction tools. These steps include obtaining source alignments, generating indexes and compressing the source MAF files (when desired); the steps required to perform these actions are outlined at the Galaxy wiki, with direct links provided in the Supplementary Material. Setting up these locally cached alignment sources is not required, as users are able to directly upload and use their own alignment files in any of the tools.

The tools described here are implemented in Python, allowing seamless cross-platform compatibility, and utilize the bx-python package (https://bitbucket.org/james_taylor/bx-python/). The GUI version of this toolset has been made available through the public Galaxy server (http://usegalaxy.org) allowing users to access not only the tools detailed above, but also additional genome analysis tools and data sources, all within one unified interface. The command-line tools and the graphical configuration files are distributed as part of the standard Galaxy distribution (http://getgalaxy.org). These tools and the entire Galaxy framework are released as open-source under the academic free license—allowing developers to modify and redistribute the applications with few restrictions.

ACKNOWLEDGEMENTS

We would like to thank UCSC Genome Biology group (and especially Jim Kent and David Haussler) for generating, maintaining and providing access to alignments. Efforts of the Galaxy Team (Enis Afgan, Dannon Baker, Dan Blankenberg, Nate Coraor, Jeremy Goecks, Greg Von Kuster, Ross Lazarus, Kanwei Li, Kelly Vincent) were instrumental for making this work happen.

Funding: Beckman Foundation Young Investigator Award (to A.N.); National Science Foundation grant (DBI 0850103); National Institutes of Health grant (HG004909, to A.N. and J.T.); Penn State University; Huck Institutes for the Life Sciences (to A.N.); Emory University (to J.T.); Pennsylvania Department of Health using Tobacco Settlement Funds (in part).

Conflict of Interest: none declared.

REFERENCES

Afgan,E. *et al*. (2010) Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics*, **11** (Suppl. 12), S4.

Blankenberg,D. *et al*.(2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol*., **Chapter 19**, Unit 19.10.1–19.10.21.

Blankenberg,D. *et al*. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res*., **17**, 960–964.

Fujita,P.A. *et al*. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*., **39**, D876–D882.

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.

Karolchik,D. *et al*. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*., **32**, D493–D496.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Miller,W. *et al*. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*., **17**, 1797–1808.

Taylor,J. *et al*. (2007) Using galaxy to perform large-scale interactive data analysis. *Curr. Protoc. Bioinform*., **19**, 10.5.1–10.5.25.