ACSESS proceedings
2010

Using a custom Ciona intestinalis genome browser to visualize genomic conservation with Ciona savignyi

Jerry S. Chen and Robert W. Zeller

AP10-08

CSRC an interdisciplinary gateway to the future
SAN DIEGO STATE UNIVERSITY

Using a custom Ciona intestinalis genome browser to visualize genomic conservation with Ciona savignyi

Jerry S Chen[*,1], *Robert W Zeller* [1]
[1] *Computational Science Research Center and Department of Biology*
San Diego State University, 5500 Campanile Drive, San Diego, CA 92182
[*] *Corresponding author: jchen@alumni.caltech.edu*

Abstract: We have successfully designed a custom genome browser for visualization of the genome of the invertebrate chordate Ciona intestinalis. We describe the use of the genome browser to visualize the conservation between Ciona intestinalis and a related species, Ciona savignyi. Visualizing genomic conservation between these two organisms is important for determining genes and regulatory regions that differ between these closely related species. We also describe the use of our genome browser in verifying 3UTR predictions for gene targets of an important regulatory molecule called microRNA-124.

Keywords: genome browser, Ciona intestinalis, Ciona savignyi, genome conservation

1 Introduction

The GBrowse Genome Browser
GBrowse is a web-based application for visualizing genomic annotations and other features for a model organism [1]. The base components required for genome visualization are a set of sequence scaffolds or chromosomes collectively comprising the raw genomic sequence, and a gene model file containing feature information for each gene such as whether the gene is transcribed from the forward or reverse DNA strand, the location of the transcription start site and the location of exons and introns. With this base alone, one can visualize many aspects of the model organism genome such as relative locations of genes within each scaffold or chromosome, exon and intron nucleotide composition and relative location within a particular gene, differences between alternative transcripts of a gene, and GC nucleotide content within genomic regions of interest.

GBrowse also allows the input of additional genome annotation tracks for comparative analysis. For example, if our model organism is the fruit fly, Drosophila melanogaster, adding an annotation track containing a plot showing conservation with the human genome can allow one to see which genes are conserved between fruit fly and human. Or similarly, tracks containing raw nucleotide or protein data from other species can be used to visualize multi-species alignments and find regions of conservation. Indeed, the UCSC genome browser [2] was used recently to visualize alignment and conservation among 28 vertebrate species, revealing insights into evolutionary insertions and deletions within protein coding regions and conservation of genomic elements such as start and stop codons [3]. More recently, DNA-protein interaction data is being produced in high volumes using ChIP-SEQ [4, 5]. With ChIP-SEQ, millions of short sequence reads are produced which correspond to the locations where a studied transcription factor protein binds to the genome. By creating an annotation track containing information on the genomic location of these sequence reads, DNA-binding sites can be visualized. The research questions that can then be explored are numerous. For example, in which genomic features do these sites tend to reside? Do binding sites tend to cluster together, or are they spread apart?

2 Methods

The Ciona intestinalis JGI scaffold and gene models were obtained from the Ciona JGI v.1.0 ANISEED databse (http://crfb.univ-mrs.fr/aniseed/). Source code files for the web-based Generic Genome Browser (GBrowse) software were downloaded from the GMOD GBrowse website (http://gmod.org/wiki/GBrowse). The latest production release (version 1.69) was downloaded. Initial installation of GBrowse

from source code was performed with the help of a Perl installation script. Apache (version 2.2.11), MySQL (version 14.12) and Perl (version 5.10.0) were all updated and configured for use with GBrowse version 1.69. All software, data and necessary backend modules and programs were stored on a local server running Fedora Core (RedHat) release 10 powered by dual quad-core 2.3 Ghz AMD Opteron processors.

3 Results

For visualization of the Ciona intestinalis genome, we have successfully implemented a genome browser based on the Generic Genome Browser (GBrowse) (Figure 1). The implementation of our C.intestinalis browser is described in a report that is being published concurrently (Chen and Zeller, submitted). We are currently using the browser to verify gene target sites of an important regulatory molecule called microRNA (miR)-124. Specifically, we are using the browser to verify target site sequences we have predicted previously [7], which are found at the end of gene transcripts in regions called 3'UTRs (Figure 2).

On top of this browser base, we have successfully implemented a conservation track showing genomic conservation between Ciona intestinalis and the closely related Ciona savignyi. The conservation track displays data representing raw genome conservation between C.intestinalis and C.savignyi. The display has been configured so that peaks occur when there is at least 30% conservation. The higher the peak, the greater the conservation (Figure 3). We have also successfully incorporated the conservation track into an SQL database. We then integrated this database with the genome browser so that, when a query is input into the browser, the browser on the backend does not have to scan through an entire file but rather only has to process an appropriate SQL query.

The implemented conservation track will faciliate future conservation studies between C.intestinalis and C.savignyi. We are in the process of updating the conservation track for the latest Ciona intestinalis genome assembly [6]. These conservation studies are important for finding important similarities and differences between organisms of the same genus. In general, coding regions of genes are significantly conserved relative to non-coding regions (Figure 3). However, exceptions may occur, depending on the gene. Some genes may have very conserved non-coding regions as well. Or perhaps they may have good conservation with the exception of a single region; these may be sites where the regulatory mechanisms differ between the species.

4 Acknowledgments

The author would like to thank Pierre Khoueiry at EMBL for kindly providing the conservation data and for his help with implementing the data in GBrowse. This work was supported by NSF grants IBN-0347937 (through 28 Feb 2010) and IOS-0951347 (since 01 Mar 2010).

5 References

1. Stein, L.D., et al., The generic genome browser: a building block for a model organism system database. Genome Res, 2002. 12(10): p. 1599-610.
2. Karolchik, D., et al., The UCSC Genome Browser Database. Nucleic Acids Res, 2003. 31(1): p. 51-4.
3. Miller, W., et al., 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res, 2007. 17(12): p. 1797-808.
4. Jothi, R., et al., Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res, 2008. 36(16): p. 5221-31.
5. Johnson, D.S., et al., Genome-wide mapping of in vivo protein-DNA interactions. Science, 2007. 316(5830): p. 1497-502.
6. Satou, Y., et al., Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. Genome Biol, 2008. 9(10): p. R152.
7. Chen, J., RW Zeller, Regulation of gene expression by the microRNA miR-124 in the developing nervous system of C.intestinalis. ACSESS Proceedings, 2009.

6 Figures

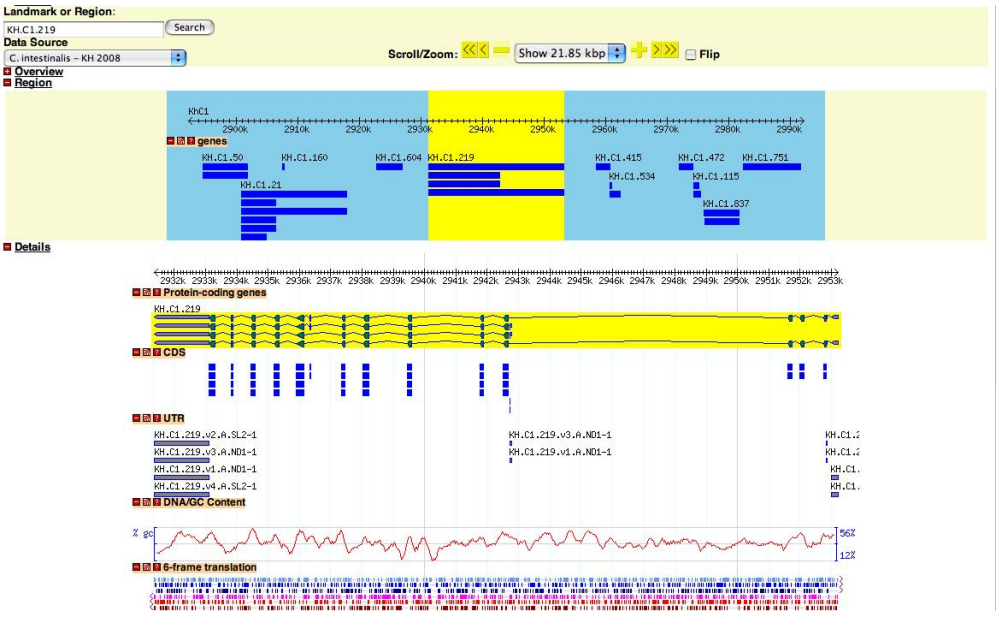

Fig. 1: A representative image of the Ciona intestinalis custom genome browser. By default, the browser shows an overview panel with the scaffold coordinates, a region panel showing a landscape of the genes around and within the field of view, and five annotation tracks specific to the highlighted region of interest: a protein-coding genes track showing exon (green arrows), intron (jagged lines), and untranslated region (gray arrows) for genes; a CDS track showing gene coding regions (blue boxes); a UTR track showing untranslated regions (gray boxes); a DNA/GC content track showing the percentage of guanine and cytosine nucleotides across the region; and a 6-frame translation track showing, when zoomed in, the protein translation of the underlying DNA sequence for all six reading frames.

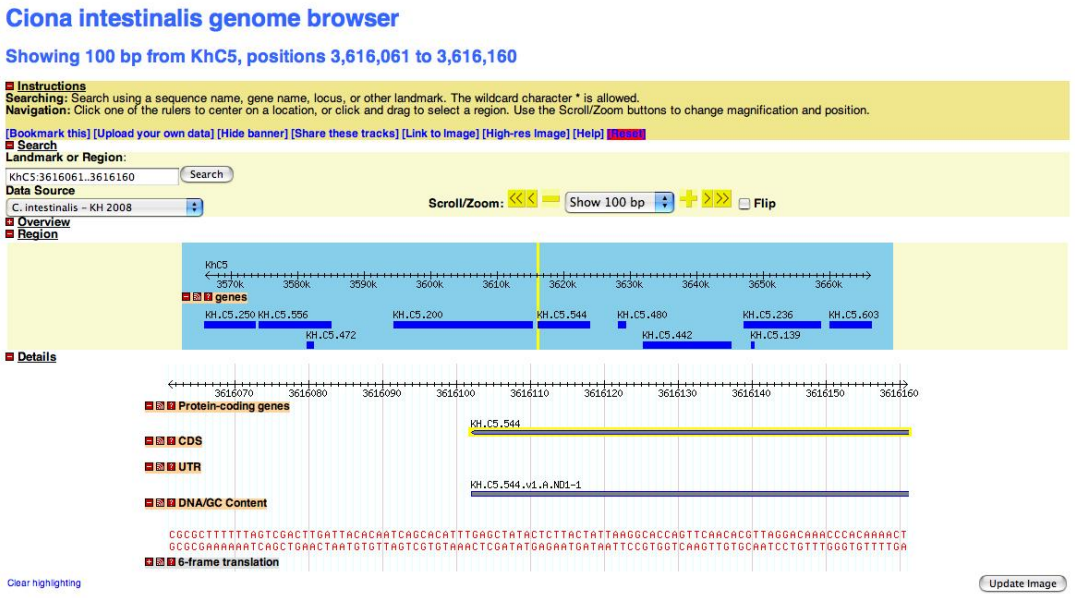

Fig. 2: The Ciona intestinalis genome browser can be used to visualize raw genomic DNA inline with gene features such as exons and introns. Here a representative three prime untranslated region (3'UTR) is shown, which is found at the tail end of the respective gene. We are currently using the genome browser to verify 3'UTR predictions that were done in our lab previously [7], highlighting a useful application of our browser.

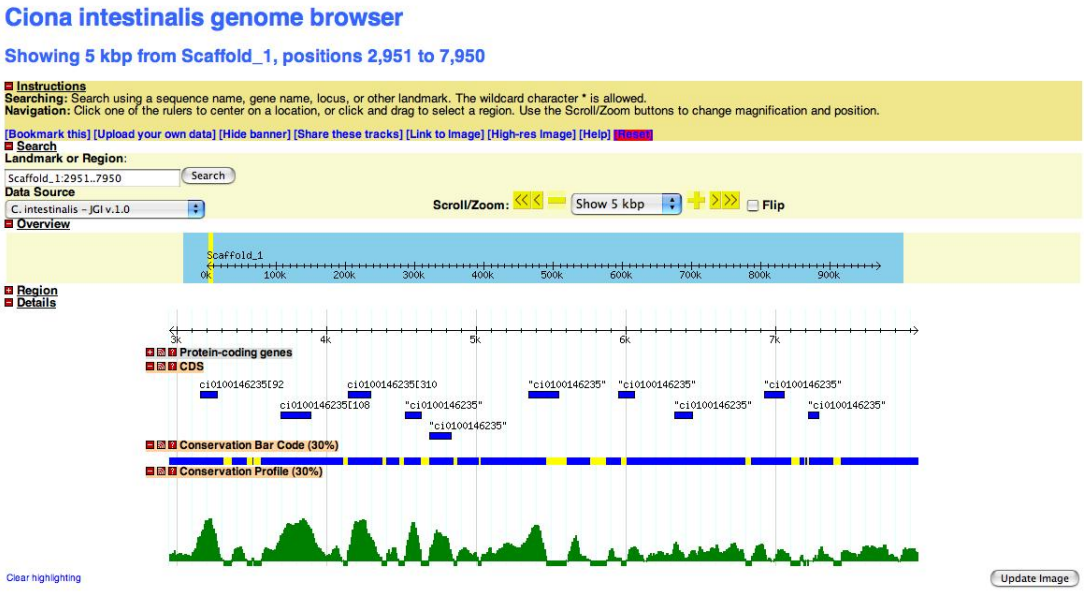

Fig. 3: Panel showing the conservation profile of a representative gene. The blue bars under the CDS feature highlight the coding regions of the gene that will eventually be translated into the corresponding protein. The blue regions within the conservation bar code refer to regions that are at least 30% conserved with with the sister species Ciona savignyi, while yellow regions are less than 30% conserved. The green peak plot is the conservation profile track, showing the amount of conservation of a particular region with the corresponding region in Ciona savignyi. Notice in this example that conserved peaks tend to occur preferentially in coding regions.