The Journal of Open Source Software

Visualizing genome synteny with xmatchview

Rene L Warren[1]

1 BC Cancer Agency, Genome Sciences Centre, Vancouver, BC, Canada

DOI: 10.21105/joss.00491

Software

- Review
- Repository
- Archive

Submitted: 07 December 2017
Published: 08 December 2017

Summary

In genomics research, the visual representation of DNA sequences is of prime importance. When displayed with additional information, or tracks, like the position of annotated genes, alignments of sequence of interest, etc. these displays facilitate our understanding of genome and gene structure, and become powerful tools to assess the relationship between various sequence data. They can be used for troubleshooting, in-depth sequence analysis, and eventually find their way in publications and oral presentations as they often translate complex and abundant data succinctly, with esthetically pleasing images. In bioinformatics, daily use of ENSEMBL (https://www.ensembl.org) (Hubbard et al. 2002), the UCSC genome browser (https://genome.ucsc.edu) (Karolchik, Hinrichs, and Kent 2002), and IGV (Robinson et al. 2011) for visualizing such data is common. The former allows for an easy-to-use visual navigation of the ENSEMBL genome databases. The latter two are customizable and flexible tools that can be used to situate sequence [read] alignments within a genome reference or draft assembly context, either online (UCSC) or as a stand-alone tool (IGV). These tools are also useful to bioinformatics software development and debugging code as well as in *de novo* genome sequencing projects as they are incredibly effective for troubleshooting sequence assemblies. Circos, a highly cited stand-alone visualization tool represents data as concentric circles, allowing for abundant data (eg. human genome scale) to be represented succinctly in full, within a computer screen window (M. Krzywinski et al. 2009). The success of circos has been in part due to its flexibility, versatility and customization in representing complicated relationships between data of all sorts, not just genomics. As attractive and convenient as circles are for displaying relationships between data, linear representations of synteny blocks between two DNA sequences remain more intuitive. Here, I introduce xmatchview (https://github.com/warrenlr/xmatchview), a tool for visualizing DNA sequence alignments produced by cross_match (unpublished, http://www.phrap.org/), a robust implementation of the sensitive Smith-Waterman algorithm for DNA alignments. The software requires python and the python imaging library (PIL) to produce publication-ready images in a variety of formats (PNG, BMP, JPEG, PS and TIFF) and cross_match for performing the DNA alignments. With xmatchview, users can compare any two DNA sequences, including but not limited to gene reconstructions, genome assemblies, cDNA, nanopore reads, etc and visually 1) identify collinear blocks, 2) assess the relationship between them, 3) analyze the sequence identity between repeated segments, and 4) view their frequency at given coordinates.

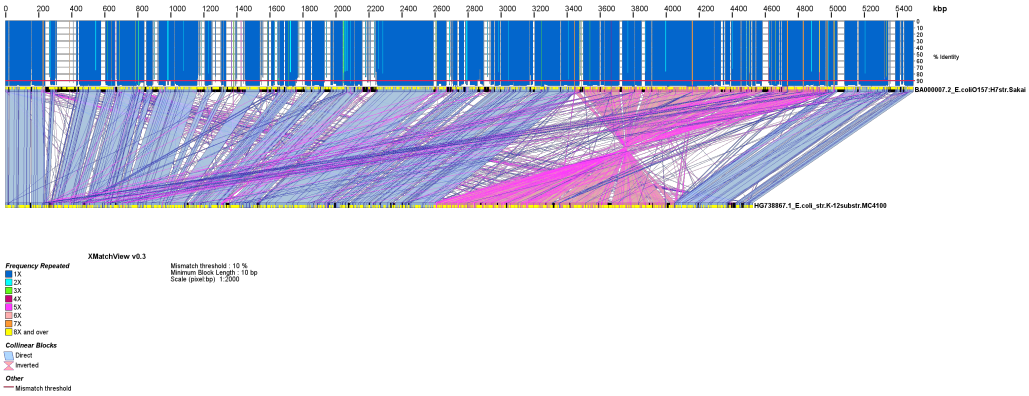

Figure 1. Genome sequence synteny between *E. coli* strains O157:H7 str. Sakai and K12 str. MC4100 (Genbank accessions BA000007.2 and HG738867.1). A large inversion is seen in one genome relative to the other. Open reading frames in both genomes are displayed in yellow.

As seen on Fig. 1, the xmatchview display consists of three main components: 1) The sequence objects, represented as black rectangles. Additional features such as exons, coding sequences (CDS), mRNA, ORFs, etc are provided to xmatchview via a simple tab-separated file enumerating each start and end positions and are plotted as yellow rectangles. Stretches of Ns in the reference and query sequences, when applicable, are shown as red rectangles on top of black rectangles. 2) Relationships between co-linear block of sequences are represented by blue and pink polygons between the two black rectangles, depending on their direct or inverted associations, respectively. 3) A histogram on top of the reference sequence (upper most) black rectangle shows the sequence identity (top to bottom, from 0 to 100%) with the query sequence (lower rectangle). When a sequence is repeated, the color of the histogram changes to reflect its frequency. Visualizing repeat frequency is a feature unique to xmatchview that can be used to readily assess sequence complexity between (Fig. 1) or within (Fig. 2) sequences. When the same sequence is given as input to xmatchview, internal repeats within that sequence are shown instead, representing only the reference sequence and the relationships between repeated blocks as arcs, instead of polygons (Fig. 2). Users can control whether to show the position of exons (CDS, or other features) on the reference and query (*-e* and *-y* options), show co-linear blocks of a certain length (*-r* option) when their mismatch rates are below a threshold (*-m* option). The histogram is generated by moving a sliding window with a step length (*-l* recommended between 10-50). The color space in xmatchview is RGBA and the alpha channel is used for visualizing the relationship between co-linear blocks (*-a* option, transparent to solid, 0 to 255). A shell script that pipelines cross_match and xmatchview is included with the distribution (runCompareTwoGenomesColinear.sh). Typically, sequences <10 Mbp in length are compared with cross_match and displayed in less than a few minutes using this pipeline, depending on your system. Images from xmatchview have been used in a number of peer-reviewed publications to showcase co-linearity and/or highlight differences between genome sequences (Bakkeren et al. 2006) (D'Souza et al. 2011) (R. L. Warren et al. 2013) (Coombe et al. 2016). A modified version developed specifically for comparing conifer DNA sequences with an evergreen tree representation, xmatchview-conifer, is co-released with xmatchview. The conifer tree representation differs from that of xmatchview. In the former, the sequence identity is shown within the synteny block relationships instead of a histogram (Fig. 3). Both xmatchview and xmatchview-conifer are implemented in python and released under GPLv3. As seen on Fig. 1, the xmatchview display consists of three main components: 1) The sequence objects, represented as black rectangles. Additional features such as exons, coding sequences (CDS), mRNA, ORFs, etc are provided to xmatchview via a simple tab-separated file enumerating each start and end positions and are plotted as yellow rectangles. Stretches of Ns in the reference and query sequences, when applicable, are shown as red rectangles

on top of black rectangles. 2) Relationships between co-linear block of sequences are represented by blue and pink polygons between the two black rectangles, depending on their direct or inverted associations, respectively. 3) A histogram on top of the reference sequence (upper most) black rectangle shows the sequence identity (top to bottom, from 0 to 100%) with the query sequence (lower rectangle). When a sequence is repeated, the color of the histogram changes to reflect its frequency. Visualizing repeat frequency is a feature unique to xmatchview that can be used to readily assess sequence complexity between (Fig. 1) or within (Fig. 2) sequences. When the same sequence is given as input to xmatchview, internal repeats within that sequence are shown instead, representing only the reference sequence and the relationships between repeated blocks as arcs, instead of polygons (Fig. 2). Users can control whether to show the position of exons (CDS, or other features) on the reference and query (*-e* and *-y* options), show co-linear blocks of a certain length (*-r* option) when their mismatch rates are below a threshold (*-m* option). The histogram is generated by moving a sliding window with a step length (*-l* recommended between 10-50). The color space in xmatchview is RGBA and the alpha channel is used for visualizing the relationship between co-linear blocks (*-a* option, transparent to solid, 0 to 255). A shell script that pipelines cross_match and xmatchview is included with the distribution (runCompareTwoGenomesColinear.sh). Typically, sequences <10 Mbp in length are compared with cross_match and displayed in less than a few minutes using this pipeline, depending on your system. Images from xmatchview have been used in a number of peer-reviewed publications to showcase co-linearity and/or highlight differences between genome sequences (Bakkeren et al. 2006) (D'Souza et al. 2011) (R. L. Warren et al. 2013) (Coombe et al. 2016). A modified version developed specifically for comparing conifer DNA sequences with an evergreen tree representation, xmatchview-conifer, is co-released with xmatchview. The conifer tree representation differs from that of xmatchview. In the former, the sequence identity is shown within the synteny block relationships instead of a histogram (Fig. 3). Both xmatchview and xmatchview-conifer are implemented in python and released under GPLv3.

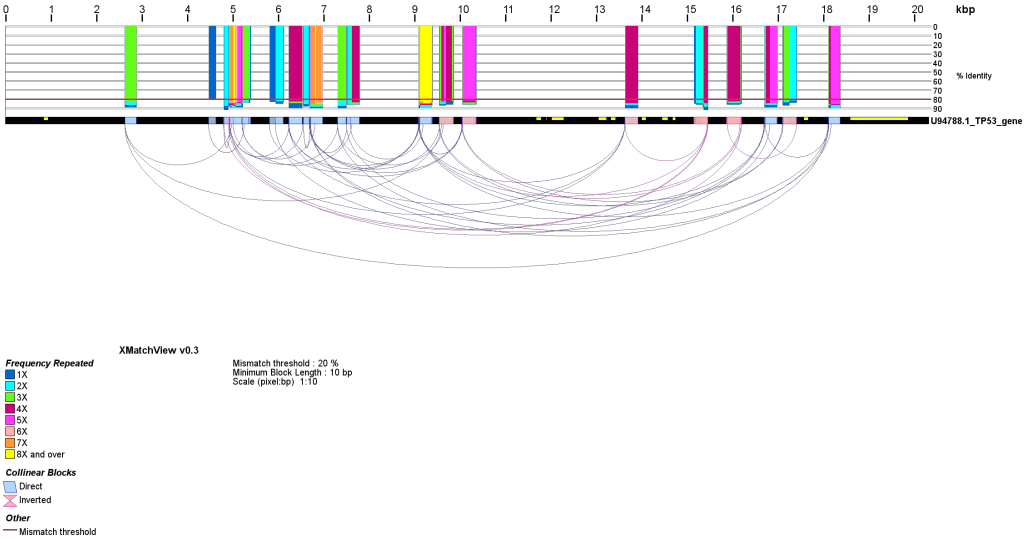

Figure 2. Sequence repeats within the human TP53 gene (Genbank accession U94788.1). TP53 mRNA sequence coordinates within the gene are shown by yellow rectangles.

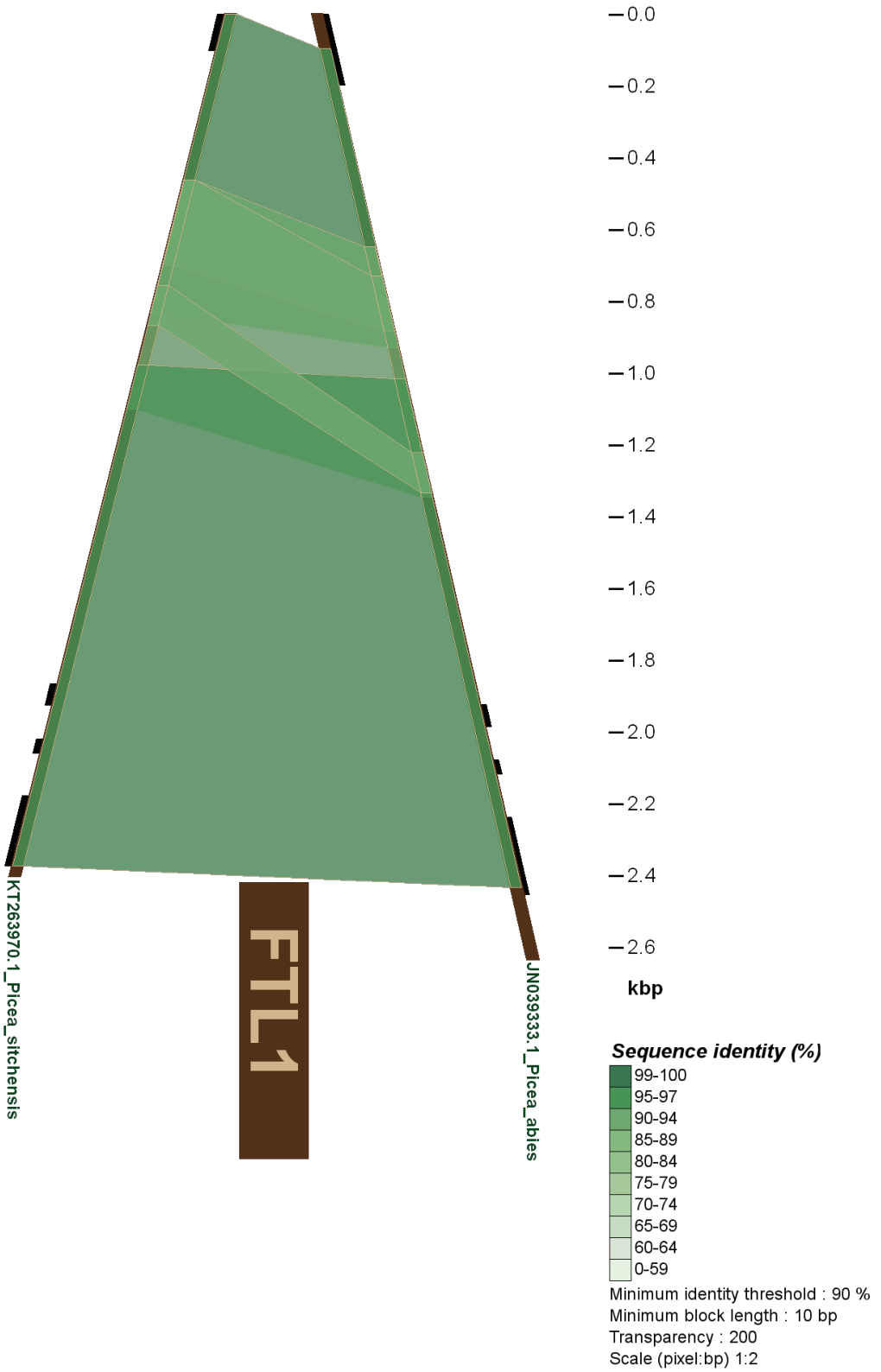

Figure 3. Sequence comparisons of the flowering locus gene FTL1 in Norway and Sitka spruce genes, *P. abies* and *P. sitchensis* with xmatchview-conifer (Genbank accessions JN039333.1 and KT263970.1). The position of exons is indicated by the black rectangles outside on the outer edge of the tree representation.

Funding

This work has been partly supported by the National Human Genome Research Institute of the National Institutes of Health (under award number R01HG007182). Additional funds were received through Genome Canada, Genome Quebec, Genome British Columbia and Genome Alberta for the Spruce-Up (243FOR) project (www.spruce-up.ca). The content reported here is solely the responsibility of the author, and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

References

Bakkeren, Guus, Guoqiao Jiang, René L. Warren, Yaron Butterfield, Heesun Shin, Readman Chiu, Rob Linning, et al. 2006. "Mating Factor Linkage and Genome Evolution in Basidiomycetous Pathogens of Cereals." *Fungal Genetics and Biology* 43 (9): 655–66. doi:https://doi.org/10.1016/j.fgb.2006.04.002.

Coombe, Lauren, René L. Warren, Shaun D. Jackman, Chen Yang, Benjamin P. Vandervalk, Richard A. Moore, Stephen Pleasance, et al. 2016. "Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10x Genomics' Gemcode Sequencing Data." *PLOS ONE* 11 (9). Public Library of Science: 1–13. doi:10.1371/journal.pone.0163059.

D'Souza, C. A., J. W. Kronstad, G. Taylor, R. Warren, M. Yuen, G. Hu, W. H. Jung, et al. 2011. "Genome Variation in Cryptococcus Gattii, an Emerging Pathogen of Immunocompetent Hosts." *mBio* 2 (1). doi:10.1128/mBio.00342-10.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, et al. 2002. "The Ensembl Genome Database Project." *Nucleic Acids Research* 30 (1): 38–41. doi:10.1093/nar/30.1.38.

Karolchik, Donna, Angie S. Hinrichs, and W. James Kent. 2002. "The Ucsc Genome Browser." In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. doi:10.1002/0471250953.bi0104s40.

Krzywinski, Martin, Jacqueline Schein, İnanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19 (9): 1639–45. doi:10.1101/gr.092759.109.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29: 24–26. doi:10.1038/nbt.1754.

Warren, René L., Douglas J. Freeman, Stephen Pleasance, Peter Watson, Richard A. Moore, Kyla Cochrane, Emma Allen-Vercoe, and Robert A. Holt. 2013. "Co-Occurrence of Anaerobic Bacteria in Colorectal Carcinomas." *Microbiome* 1 (1): 16. doi:10.1186/2049-2618-1-16.