

Genome analysis

HAPLOWSER: a whole-genome haplotype browser for personal genome and metagenomeJong Hyun Kim¹, Woo-Cheol Kim², Michael S. Waterman³, Sanghyun Park^{2,*} and Lei M. Li^{3,*}¹Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA,²Department of Computer Science, Yonsei University, 134 Shinchon-dong, Seoul, 120-749, Republic of Korea and³Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs way, Los Angeles, CA 90089, USA

Received on January 26, 2009; revised on June 9, 2009; accepted on June 25, 2009

Advance Access publication June 27, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Haplotype assembly is becoming a very important tool in genome sequencing of human and other organisms. Although haplotypes were previously inferred from genome assemblies, there has never been a comparative haplotype browser that depicts a global picture of whole-genome alignments among haplotypes of different organisms. We introduce a whole-genome HAPLotype brOWSER (HAPLOWSER), providing evolutionary perspectives from multiple aligned haplotypes and functional annotations. Haplowser enables the comparison of haplotypes from metagenomes, and associates conserved regions or the bases at the conserved regions with functional annotations and custom tracks. The associations are quantified for further analysis and presented as pie charts. Functional annotations and custom tracks that are projected onto haplotypes are saved as multiple files in FASTA format. Haplowser provides a user-friendly interface, and can display alignments of haplotypes with functional annotations at any resolution.

Availability: Haplowser, written in Java, supports multiple platforms including Windows and Linux. Haplowser is publicly available at <http://embio.yonsei.ac.kr/haplowser>

Contact: sanghyun@cs.yonsei.ac.kr; lilei@usc.edu

Supplementary information: Supplementary data are available at <http://embio.yonsei.ac.kr/haplowser>

1 INTRODUCTION

Comparative genomics and comparative genome browsers have focused on comparing haploid sequences of multiple organisms and visualizing haploid sequences. Although haplotypes were inferred in the highly polymorphic genomes of *Ciona intestinalis* and *Ciona savignyi* by using computational methods (Kim *et al.*, 2007; Vinson *et al.*, 2005), an effective way of comparing haplotypes and its visualization has yet to be explored. While haplotypes inferred from the genome assembly of an individual human showed promising information (Levy *et al.*, 2007; Wang *et al.*, 2008), haplotype comparisons of personal genomes will play significant roles in the forthcoming era of 'Personal Genomics' (Church, 2005; Wheeler

et al., 2008). In metagenomics, whole-genome shotgun sequencing of environmental samples and application of genomic techniques to those samples are routinely conducted to unravel population structures and evolutionary features. If haplotypes are inferred from metagenome assemblies of multiple environmental samples, the resulting haplotype information can be compared to reveal the distinct characteristics of microbial communities.

Although web servers or stand-alone programs are widely used to visualize comparative data (Engles *et al.*, 2006; Frazer *et al.*, 2004; Karolchik *et al.*, 2008), none of them display haplotypes. Haplowser is a comparative browser for comparing and visualizing haplotypes from individual genomes and metagenomes. In addition to aligned haplotypes, Haplowser is capable of displaying functional annotations and user-defined custom tracks that are cross-linked to alignments.

2 IMPLEMENTATION

The main view of Haplowser consists of five windows [Alignment Plot (AP) Window, Multiple Alignment (MA) Window, Vertical Annotation (VA) Window, Horizontal Annotation (HA) Window and Property Window]; see Figure 1A. Haplowser imports alignment data in MUMmer (Kurtz *et al.*, 2004) or AVID format (Bray *et al.*, 2003), displaying them in the AP Window (Figure 1A). Alignments run as colored diagonal lines in the AP Window, where the vertical axis indicates the genomic coordinates of the reference sequence and the horizontal axis indicates the genomic coordinates of the aligned sequence. After multiple haploid sequences are aligned to a single haploid reference sequence, haplotypes can be easily aligned to the haploid reference sequence and haploid aligned sequences because haplotypes are of the same length as their haploid sequence (Kim *et al.*, 2007). Whole-genome alignments of a haploid reference sequence and haploid aligned sequences are first generated by using either MUMmer or AVID. As alignments of a haploid reference sequence and haploid aligned sequences are imported, haplotypes are imported and aligned to the haploid reference sequence and haploid aligned sequences, both of which are imported as well. The AP Window can be zoomed in or out by using the mouse operation (scroll wheel) or the scale controller. The resolution of

*To whom correspondence should be addressed.

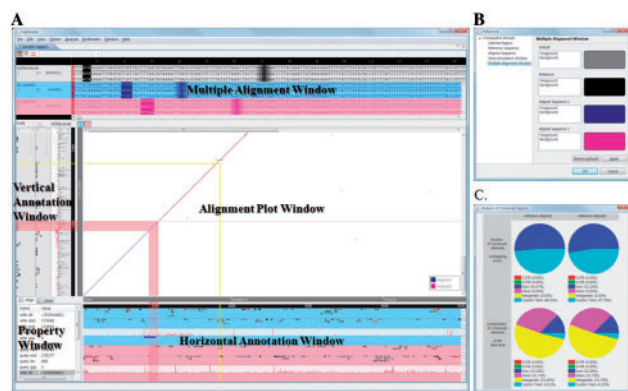


Fig. 1. Main view of Haplowser and analysis window. (A) In the AP Window, the vertical axis corresponds to the reference haplotypes, while the horizontal axis corresponds to the aligned haplotypes. The haplotypes from multiple genomes or metagenomes can be aligned to the single reference haplotypes. (B) The colors of the objects (eg. genes, alignments, custom tracks, etc.) shown in (A) can be configured by a user. (C) The conserved regions or alignments overlapping each category of annotation (e.g. exons, introns, 5'UTR, 3'UTR and intergenic regions) are analyzed and displayed in a pop-up window. For details, see text.

the AP Window varies from the entire haplotype level up to the base level.

The region selected by a mouse click in the AP Window is synchronously displayed in the MA Window, where the multiple alignments of haplotypes are shown. The VA and HA Windows are also synchronized with the AP Window, consisting of 4 panes (Forward-directed Gene Pane, Reverse-directed Gene Pane, SNP Pane and Custom Track Pane); the VA Window represents the annotations of the reference haplotypes and the HA Window represents the annotations of the aligned haplotypes. When a user selects a specific region in the AP Window, the range that corresponds to the selected region is denoted by lines in the VA and HA Windows. Similarly, when a user selects an alignment, the ranges spanned by the alignment are highlighted in the VA and HA Windows.

The VA and HA Windows display genes, SNPs, and custom tracks. Gene annotation can be imported in GFF, UCSC or SGF (Simple Gene Format) format. Genes with different annotation types (e.g. tRNA, rRNA and protein coding genes) are represented in different colors (Figure 1B). When a gene is selected, the gene and the orthologous genes are simultaneously highlighted. The Property Window displays the information related to the selected gene (type, direction, starting point, etc.). Because imported gene sequences are internally mapped to haplotypes, a user can save haplotype-mapped gene sequences in FASTA format. In this case, the number of saved files is equivalent to the number of imported haplotypes.

A user can add custom tracks in a dialog box, or load text files in other gene annotation formats (UCSC or SGF format). Alignment data in BLAST (Altschul *et al.*, 1997) or BLAT (Kent 2002) format can also be displayed in the Custom Track Pane. Similar to gene sequences, a user can save haplotype-mapped custom tracks in FASTA format files because an added custom track

is internally mapped to haplotypes as well. For example, a user can save haplotype-mapped enhancer sequences, importing the enhancer sequences as a custom track.

Haplowser enables users to annotate alignments or conserved regions by calculating the fractions of alignments or conserved regions associated with exons, introns, 5'UTR, 3'UTR and intergenic regions (Kim *et al.*, 2007). First, the fractions of conserved regions (or alignments) overlapping exons, falling within introns, and falling within intergenic regions are calculated and represented as a pie chart (Figure 1C). Second, the base composition of conserved regions (or alignments) associated with exons, introns, 5'UTR, 3'UTR and intergenic regions is calculated and represented as a pie chart (Figure 1C). If a user adds custom tracks (e.g. enhancer sequences), the fractions of conserved regions associated with the custom tracks are also included in the pie charts.

3 CONCLUSION

Haplowser is a comparative browser to compare haplotypes inferred from genome assemblies or metagenome assemblies. It offers a convenient way to navigate haplotype sequences and functional annotations, both of which operate synchronously. Along with zooming, a user can navigate any region of haplotypes and functional annotations at any resolution.

ACKNOWLEDGEMENTS

We thank Daehyun Kim for helpful comments.

Funding: The Korea Science and Engineering Foundation, Ministry of Sciences & Technology (2007-03965 to J.H.K., W.K. and S.P.); The National Institutes of Health, USA (HG002790 to M. S. W. and L. M. L.).

Conflict of Interest: none declared.

REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search program. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bray,N. *et al.* (2003) AVID: a global alignment program. *Genome Res.*, **13**, 97–102.
- Church,G.M. (2005) The personal genome project. *Mol. Syst. Biol.*, **1**, 2005.0030.
- Engels,R. *et al.* (2006) Combo: a whole genome comparative browser. *Bioinformatics*, **22**, 1782–1783.
- Frazer,K.A. *et al.* (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.
- Karolchik,D. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
- Kent,J.W. (2002) BLAT-The BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
- Kim,J.H. *et al.* (2007) Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.*, **17**, 1101–1110.
- Kurtz,S. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
- Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
- Vinson,J. *et al.* (2005) Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Res.*, **15**, 1127–1135.
- Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
- Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.