doi:10.1093/nar/gkr354

HapEdit: an accuracy assessment viewer for haplotype assembly using massively parallel DNA-sequencing technologies

Jong Hyun Kim[1,2], Woo-Cheol Kim[3], Lei M. Li[4,5] and Sanghyun Park[3,]*

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, [2]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, [3]Department of Computer Science, Yonsei University, 134 Shinchon-dong, Seoul, 120-749, Republic of Korea, [4]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1050 Childs way, Los Angeles, CA 90089, USA and [5]Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China

Received February 18, 2011; Revised April 10, 2011; Accepted April 26, 2011

ABSTRACT

The massively parallel sequencing technologies have recently flourished and dramatically cut the cost to sequence personal human genomes. Haplotype assembly from personal genomes sequenced using the massively parallel sequencing technologies is becoming a cost-effective and promising tool for human disease study. Computational assembly of haplotypes has been proved to be very accurate, but obviously contains errors. Here we present a tool, HapEdit, to assess the accuracy of assembled haplotypes and edit them manually. Using this tool, a user can break erroneous haplotype segments into smaller segments, or concatenate haplotype segments if the concatenated haplotype segments are sufficiently supported. A user can also edit bases with low-quality scores. HapEdit displays haplotype assemblies so that a user can easily navigate and pinpoint a region of interest. As inputs, HapEdit currently takes reads from the Polonator, Illumina, SOLiD, 454 and Sanger sequencing technologies.

INTRODUCTION

In transcriptome sequencing, epigenomics, targeted sequencing and whole-genome resequencing, the use of massively parallel sequencing technologies is widespread, some notable examples of which are sequencing-by-synthesis platform (Illumina) (1), sequencing-by-ligation platforms (Polonator; ABI SOLiD) (2), pyrosequencing platform (Roche 454) (3) and single-molecule sequencing platforms (Helicos Heliscope) (4) (Pacific Biosciences SMRT) (5). The massively parallel sequencing technologies continue to extend read length, increase throughput and shorten run time. Along with this, the massively parallel sequencing technologies are becoming

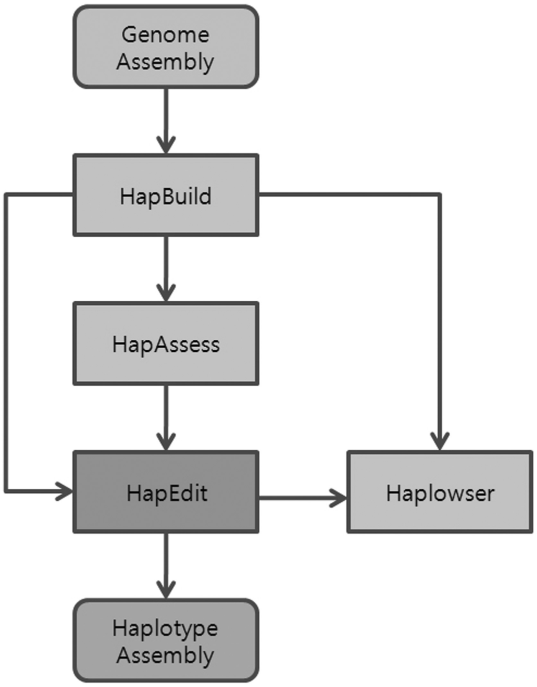
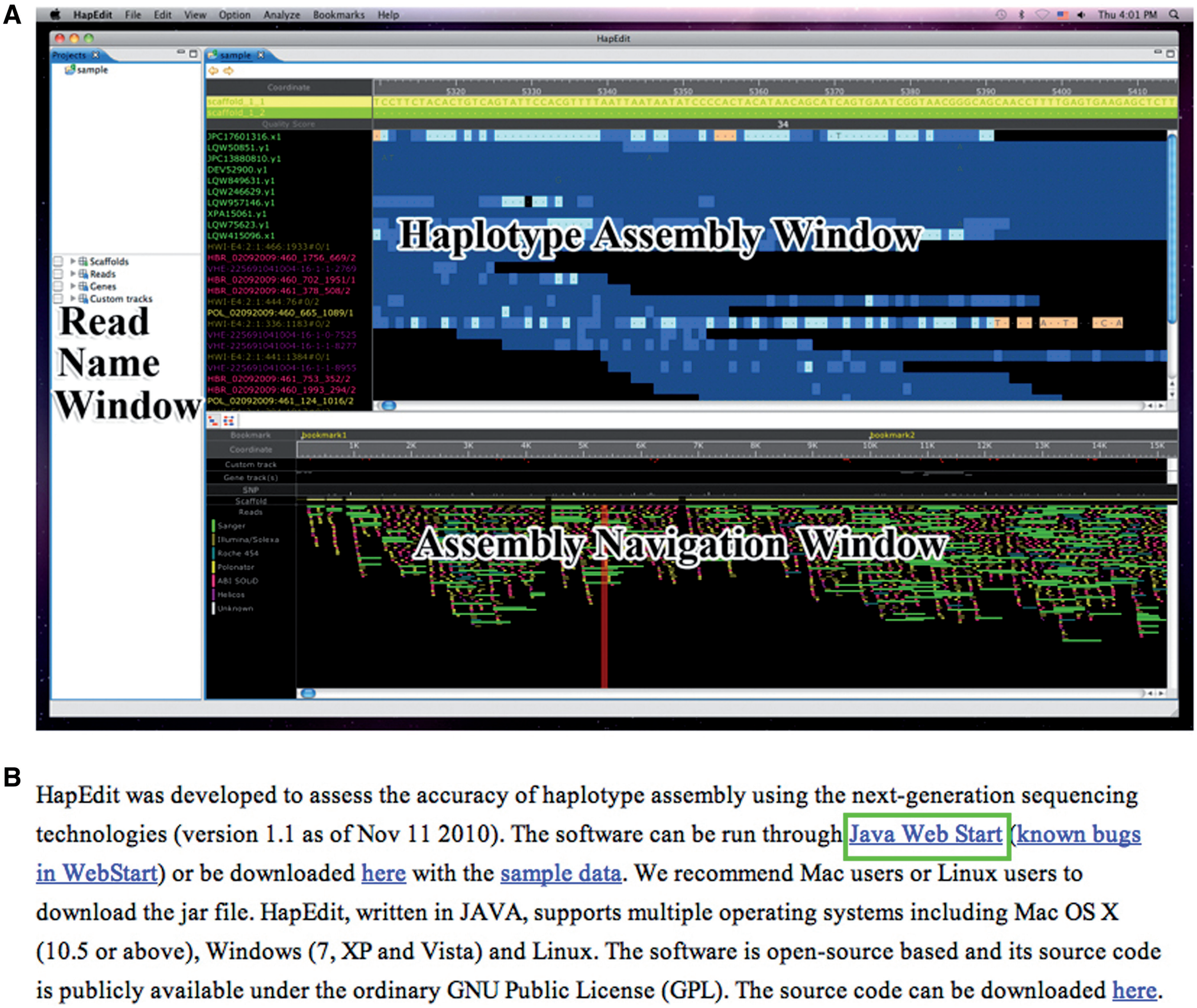


Figure 1. Workflow of haplotype assembly. The input is a sequence assembly, taken by HapBuid as an input. The final output is a haplotype assembly. For the description of each component software, see the main text.

*To whom correspondence should be addressed. Email: sanghyun@cs.yonsei.ac.kr



HapEdit was developed to assess the accuracy of haplotype assembly using the next-generation sequencing technologies (version 1.1 as of Nov 11 2010). The software can be run through Java Web Start (known bugs in WebStart) or be downloaded here with the sample data. We recommend Mac users or Linux users to download the jar file. HapEdit, written in JAVA, supports multiple operating systems including Mac OS X (10.5 or above), Windows (7, XP and Vista) and Linux. The software is open-source based and its source code is publicly available under the ordinary GNU Public License (GPL). The source code can be downloaded here.

Figure 2. (**A**) Screenshot of HapEdit main window. At the top of the HA Window, haplotype sequences with the genomic coordinates are displayed, where SNPs are highlighted in red. Quality scores for assembled haplotype sequences are displayed below the haplotype sequences. A haplotype assembly is located below quality scores, where read names are colored according to the sequencing technologies used. (**B**) HapEdit web site. HapEdit can be run simply by pressing the 'Java Web Start' link.

indispensable in genomic variation detection and clinical diagnosis.

Haplotype assembly is a useful tool for genome analysis. One example is to characterize the causal relationship between cis variation and gene expression. As genome-wide association studies have progressed, it is now essential to understand how cis variations are correlated with phenotypes. To advance this study, haplotype assembly is necessary to determine the phases between *cis*-regulatory regions and coding regions. The Sanger sequencing (6) or Illumina sequencing technology has been used to assemble personal human haplotypes (7,8). It is anticipated that whole-genome resequencing using the massively parallel sequencing technologies will become routine as the sequencing cost for a personal genome drops under $1000 within several years. If personal genomes can be sequenced at that low cost, haplotypes will be more frequently assembled for clinical use.

It has been a common practice to infer a haploid consensus sequence from a genome assembly even when reads were generated from two haplotypes in eukaryotes. In haploid assembly, inferring a haploid consensus sequence only requires a simple statistical method (9). To computationally assemble haplotypes from sequenced reads, however, it is necessary to disentangle reads from two haplotypes and infer two consensus sequences. The complexity of haplotype assembly is known to be NP-hard (10). Several computational methods have been developed to assemble haplotypes, which are based on Markov chain Monte Carlo approaches (11,12), heuristic approaches (7,13), and a combinatorial approach (14).

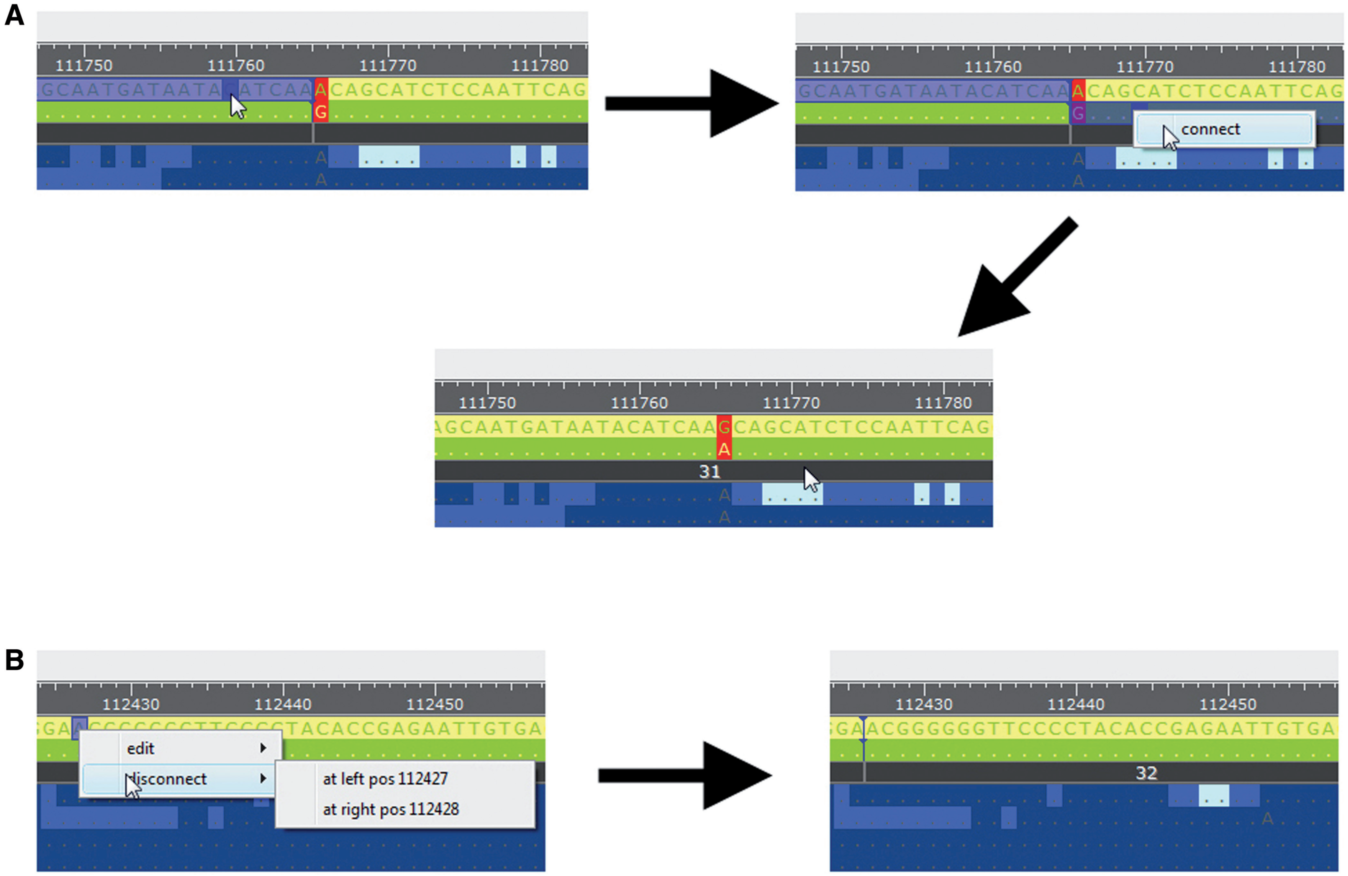

A
111750
111760
111770
111780
GCAATGATAATA ATCAA A CAGCATCTCCAATTCAG
G
GCAATGATAATACATCAA A CAGCATCTCCAATTCAG
connect
AGCAATGATAATACATCAA G CAGCATCTCCAATTCAG
A
31
B
112430
112440
112450
GA
TACACCGAGAATTGTGA
edit
disconnect
at left pos 112427
at right pos 112428
GA ACGGGGGGTTCCCCTACACCGAGAATTGTGA
32

Figure 3. (**A**) Using a combination of a mouse and key operation, a user can connect haplotype segments. Haplotype segments to be connected are selected by pressing the left-mouse button while holding the shift key. Then, the connection menu pops up with a mouse right-click (control key + a mouse click on MacOSX). Haplotype segments are connected by clicking the 'connect' menu. (**B**) A user can choose any region of haplotypes to be disconnected by pressing the right-mouse button. Haplotype segments are disconnected by selecting the position to be disconnected.

The assembly viewer Consed was originally developed to assess and edit haploid genome assemblies from reads obtained by Sanger sequencing, but now also supports reads obtained from massively parallel sequencing methods. (15). Recently, EagleView was developed to view genome assemblies by massively parallel sequencing technologies (16). However, none of these was designed to view, assess, and edit haplotype assemblies.

HapEdit was designed to assess assembled haplotypes and edit misassembled haplotypes, supporting reads sequenced by the five massively parallel sequencing technology platforms (Illumina, Polonator, ABI SOLiD, Roche 454, and Helicos) and the Sanger sequencing technology.

WORKFLOW

Software package

Figure 1 shows the flowchart of haplotype assembly. HapEdit imports a haplotype assembly from HapBuild (11). Optionally, HapEdit can import and display quality scores for assembled haplotypes, which are calculated by HapAssess (17). A user can compare haplotypes from different individuals, using a comparative browser, Haplowser (18). HapEdit is provided as a component in a software package for haplotype assembly.

Web start and standalone program

A user can download the binary files compiled for the three operating systems (MacOSX, Windows and Linux). Alternatively, a user can run HapEdit directly on the web site through Java web start (Figure 2B).

IMPLEMENTATION

HapEdit provides different views of a haplotype assembly through three windows [Read Name Window (RN), Haplotype Assembly Window (HA) and Assembly Navigation Window (AN)]; see Figure 2A. In the Haplotype Assembly Window (HA), a detailed view of a haplotype assembly is displayed with zooming function, where haplotype sequences and alignments with quality scores are also shown. The name of each read in the alignments is differentially colored based on the sequencing technology used. Similarly, each base-call is colored based on its quality score. In this manner, a user can easily identify low-quality bases and the sequencing technology used for the bases. At the top of the HA window, a user can manually edit haplotype sequences in three ways. First, erroneous bases can be fixed by directly modifying the bases. Second, a user can connect haplotype segments if the connection is judged to be significantly supported by any read (Figure 3A). Third, a user can consider the quality scores for assembled haplotype segments, and break haplotype segments potentially containing phasing errors into pieces (Figure 3B).

The AN Window is synchronized with the HA window to depict a global view of a haplotype assembly. In the AN window, the sequencing technology used for a read is indicated by the color of the read. A user can navigate any region of a haplotype assembly in a mouse click. The region selected by the mouse click in the AN Window is synchronously displayed in the HA Window. Conversely, The region shown in the HA window is traced and marked by a red bar in the AN window. The gene annotation [in GFF, UCSC or SG (Simple Gene) format] and custom track information (in BLAT or BLAST format) can be imported, and displayed in two optional panes (Gene Pane and Custom Track Pane) of the AN window. The SNP information obtained from haplotype sequences is also displayed in an optional pane (SNP Pane).

The RN Window enumerates the names and genomic coordinates of all the reads included in the haplotype assembly in the HA window. A user can move to the starting point (or ending point) of a read of interest by right-clicking the name of the read and selecting the pop-up menu. The names of gene names and custom tracks shown in the AN window are also enumerated in the RN window.

Integrating different sequencing technologies to detect structural variants in a cost-effective way has been recently explored through a simulation study (19). However, finding the optimal composition rate of each sequencing technology between cost and N50 haplotype length in assembling haplotypes is yet to be explored. To fit the composition rates into the ideal composition rates (or the rates that a user initially planned), HapEdit facilitates a user to assess the deviation from those; the composition rates of reads and clone among the entire reads and clones are summarized in pie charts (Figure 4A). Similarly, the sequence coverage and clone coverage of each sequencing technology are analyzed and summarized in bar charts (Figure 4B).

CONCLUSION

HapEdit is an accuracy assessment tool to view haplotype assemblies by massively parallel sequencing technologies and edit misassembled haplotypes. It offers a graphical user interface to navigate haplotype assemblies and helps a user to fit the composition rates of the reads sequenced by the (up to) six different sequencing technologies to the ideal composition rates.

ACKNOWLEDGEMENTS

We thank Michael Sismour and John Aach for helpful comments.

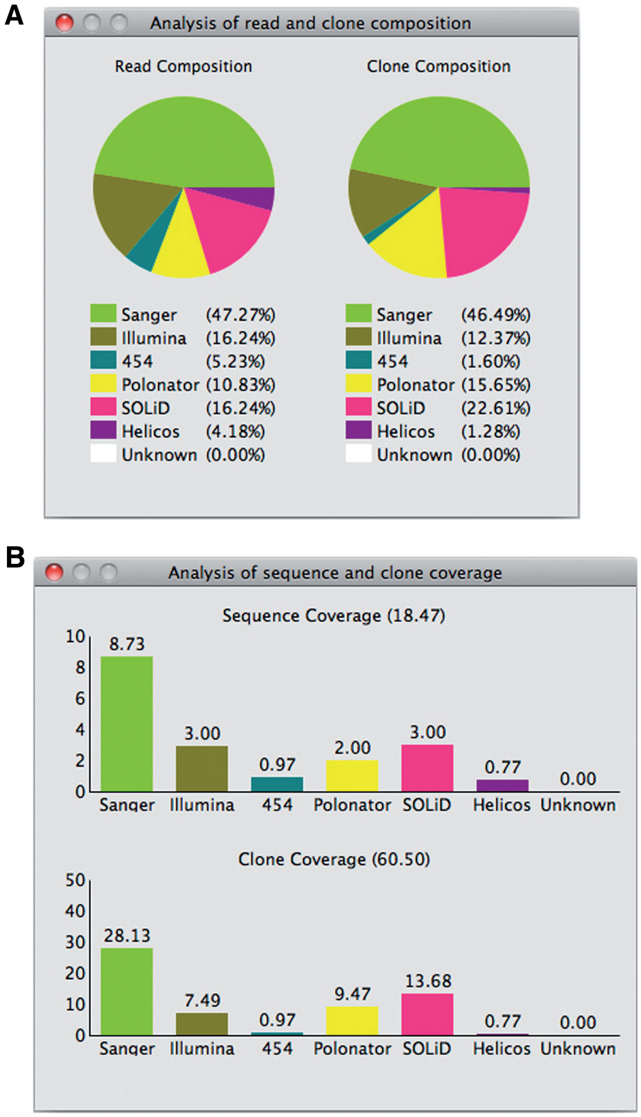


Figure 4. (**A**) The composition rates of reads (left) and clones (right) sequenced by the different technologies are displayed in pie charts. The different colors represent the different sequencing technologies. (**B**) The sequence coverage (upper) and clone coverage (lower) are calculated and presented in a form of bar chart. Each bar indicates the coverage by a specific sequencing technology.

FUNDING

The National Research Foundation of Korea (2009-0083311 to J.H.K., W.K. and S.P.; 2011-0004382 to S.P.); The National Institutes of Health, USA (HG002790 to L.M.L.). Funding for open access charge: The National Research Foundation of Korea [2009-0083311].

Conflict of interest statement. None declared.

REFERENCES

1. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome

sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
2. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, **309**, 1728–1732.
3. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braveman,M.S., Chen,Y.J. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
4. Pushkarev,D., Neff,N.F. and Quake,S.R. (2009) Single-molecule sequencing of an individual human genome. *Nature Biotech.*, **17**, 847–850.
5. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time sequencing from single polymerase molecules. *Science*, **323**, 133–138.
6. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
7. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The Diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
8. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
9. Churchill,G.A. and Waterman,M.S. (1992) The accuracy of DNA sequence: Estimating sequence quality. *Genomics*, **14**, 89–98.
10. Lippert,R., Schwartz,R., Lancia,G. and Istrail,S. (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinform.*, **3**, 23–31.
11. Kim,J.H., Waterman,M.S. and Li,L.M. (2007) Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.*, **17**, 1101–1110.
12. Bansal,V., Halpern,A.L., Axelrod,N. and Bafna,V. (2008) An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.*, **18**, 1336–1346.
13. Long,Q., MacArthur,D., Ning,Z. and Tyler-Smith,C. (2009) HI: haplotype improver using paired-end short reads. *Bioinformatics*, **25**, 2436–2437.
14. Bansal,V. and Bafna,V. (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, i153–i159.
15. Gordon,D., Abajian,C. and Green,P. (1998) Consed: A graphical tool for sequencing finishing. *Genome Res.*, **8**, 195–202.
16. Huang,W. and Marth,G. (2008) EagleView: A genome assembly viewer for next-generation sequencing technologies. *Genome Res.*, **18**, 1538–1543.
17. Kim,J.H., Waterman,M.S. and Li,L.M. (2007) Accuracy assessment of diploid consensus sequences. *IEEE Trans. Comput. Biol. and Bioinfo.*, **4**, 88–97.
18. Kim,J.H., Kim,W.C., Waterman,M.S., Park,S. and Li,L.M. (2009) HAPLOWSER: whole-genome haplotype browser for personal genome and metagenome. *Bioinformatics*, **25**, 2430–2431.
19. Du,J., Bjornson,R.D., Zhang,Z.D., Kong,Y., Snyder,M. and Gerstein,M.B. (2009) Integrating sequencing technologies in personal genomics: optimal low cost reconstruction of structural variants. *PLoS Comput. Biol.*, **5**, e1000432.