bwttool: A tool for bigWig files

Andy Pohl1,2,* and Miguel Beato1,2
1Centre for Genomic Regulation (CRG), Barcelona, Spain.
2Universitat Pompeu Fabra, Barcelona, Spain.

ABSTRACT
BigWig files are a compressed, indexed, binary format for genome-wide signal data for calculations (e.g. GC percent) or experiments (e.g. ChIP-seq/RNA-seq read depth). bwttool is a tool designed to read bigWig files rapidly and efficiently, providing functionality for extracting data and summarizing it in several ways, globally or at specific regions. Additionally, the tool enables the conversion of the positions of signal data from one genome assembly to another, also known as “lifting”. We believe bwttool can be very useful for the analyst frequently working with bigWig data, which is becoming a standard format to represent functional signals along genomes.

1 INTRODUCTION
For many labs it has become an everyday task to generate or to analyze genome-wide data such as ChIP-seq read depth. To facilitate visualization of this data with tools such as the UCSC Genome Browser (Kent et al., 2002) or ENSEMBL (Flicek et al., 2013), or for further processing, it is common to use the wiggle (WIG) file format. This format is not without a few disadvantages, principally that the files can become quite large, particularly when care is not taken to store the data at a minimally-necessary decimal precision. Another disadvantage is that wiggles exist in three different forms, the choice of which depends on the sparseness of the data. Programs that expect WIG data do not always allow all three formats interchangeably.

The BigWig format (Kent et al., 2010) was created as a means for the UCSC Genome Browser to access real-valued signal data remotely hosted on HTTP/FTP servers worldwide. The format is binary, compressed, indexed, and allows random access to directly read bigWig read depth. bwttool is a tool designed to read bigWig files rapidly and efficiently, providing functionality for extracting data and summarizing it in several ways, globally or at specific regions. Additionally, the tool enables the conversion of the positions of signal data from one genome assembly to another, also known as “lifting”. We believe bwttool can be very useful for the analyst frequently working with bigWig data, which is becoming a standard format to represent functional signals along genomes.

2 DESCRIPTION
The bwttool program is designed to rapidly collect summary statistics and do common wiggle manipulations. The program is actually a collection of utilities (the names of which are in bold), which are taken to store the data at a minimally-necessary decimal precision. Another disadvantage is that wiggles exist in three different forms, the choice of which depends on the sparseness of the data. Programs that expect WIG data do not always allow all three formats interchangeably.

The BigWig format (Kent et al., 2010) was created as a means for the UCSC Genome Browser to access real-valued signal data remotely hosted on HTTP/FTP servers worldwide. The format is binary, compressed, indexed, and allows random access to directly query a subset of the larger dataset. In general, programs designed to read bigWig files should treat remote URLs of bigWigs the same as if they were local to that computer. BigWig uses an indexing strategy similar to other binary/indexed formats such as bigBed (Kent et al., 2010), BAM (Li et al., 2009), and tabix-based formats (Li, 2011), but unlike BAM or tabix-based formats, BigWig is specific to numerical data. WIG and BAM are both common data formats and are utilized by many applications, e.g. MACS (Zhang et al., 2008) and MISO (Katz et al., 2010) respectively, but to date there are not many applications that accommodate bigWig data.

We have created command-line software under the UNIX operating system called bwttool in a similar spirit to bedtools (Quinlan and Hall, 2010) or samtools (Li et al., 2009) that offers the possibility to carry out a number of diverse operations on bigWigs in a convenient way. Until now, the common procedure to access the data within bigWig files has been to use the tools available from UCSC: bigWigToWig, bigWigSummary, bigWigAverageOverBed, bigWigMerge, bigWigCorrelate, or bigWigInfo. These offer some basic usability for bigWigs. bigWigInfo provides instant information about a bigWig file and is useful for glancing at the overall mean and standard deviation as well as seeing how many bases are covered by the signal. bigWigToWig is indispensable as it is occasionally necessary to convert a bigWig into the original WIG to utilize legacy software. Beyond those two, bwttool provides additional features and flexibility not found in other software.

1 INTRODUCTION
For many labs it has become an everyday task to generate or to analyze genome-wide data such as ChIP-seq read depth. To facilitate visualization of this data with tools such as the UCSC Genome Browser (Kent et al., 2002) or ENSEMBL (Flicek et al., 2013), or for further processing, it is common to use the wiggle (WIG) file format. This format is not without a few disadvantages, principally that the files can become quite large, particularly when care is not taken to store the data at a minimally-necessary decimal precision. Another disadvantage is that wiggles exist in three different forms, the choice of which depends on the sparseness of the data. Programs that expect WIG data do not always allow all three formats interchangeably.

The BigWig format (Kent et al., 2010) was created as a means for the UCSC Genome Browser to access real-valued signal data remotely hosted on HTTP/FTP servers worldwide. The format is binary, compressed, indexed, and allows random access to directly query a subset of the larger dataset. In general, programs designed to read bigWig files should treat remote URLs of bigWigs the same as if they were local to that computer. BigWig uses an indexing strategy similar to other binary/indexed formats such as bigBed (Kent et al., 2010), BAM (Li et al., 2009), and tabix-based formats (Li, 2011), but unlike BAM or tabix-based formats, BigWig is specific to numerical data. WIG and BAM are both common data formats and are utilized by many applications, e.g. MACS (Zhang et al., 2008) and MISO (Katz et al., 2010) respectively, but to date there are not many applications that accommodate bigWig data.

We have created command-line software under the UNIX operating system called bwttool in a similar spirit to bedtools (Quinlan and Hall, 2010) or samtools (Li et al., 2009) that offers the possibility to carry out a number of diverse operations on bigWigs in a convenient way. Until now, the common procedure to access the data within bigWig files has been to use the tools available from UCSC: bigWigToWig, bigWigSummary, bigWigAverageOverBed, bigWigMerge, bigWigCorrelate, or bigWigInfo. These offer some basic usability for bigWigs. bigWigInfo provides instant information about a bigWig file and is useful for glancing at the overall mean and standard deviation as well as seeing how many bases are covered by the signal. bigWigToWig is indispensable as it is occasionally necessary to convert a bigWig into the original WIG to utilize legacy software. Beyond those two, bwttool provides additional features and flexibility not found in other software.
to exclude specific regions in the genome. Unequally sized intervals can be also extracted with the `extract` utility.

- Another way to extract data from multiple bigWigs is to use the `paste` utility. This outputs tab-delimited data from a set of bigWigs, one base per line. Pasting bigWigs together makes it possible to perform many complex calculations with small auxiliary scripts. In this way, the functionality of bwtool can easily expand upon the functionality of bigWigMerge and bigWigCorrelate from UCSC.

- Discretize the real-valued signal into letters, using the SAX algorithm (Shieh and Keogh, 2009).

- Removing data based on thresholds and specific regions if desired. Conversely, regions missing data in a bigWig can be replaced with a constant using the `fill` utility.

- Summarize data at specific regions. This functionality is similar to the combined programs of bigWigSummary and bigWigAverageOverBed, with the addition of median and optional quartile information in the output.

Common options to many of the features include the ability to specify the decimal precision, to fill missing bases with a given value, or to provide a bed file specifying specific regions of the bigWig to read.

3 USAGE AND AVAILABILITY

bwtool is command-line software for UNIX, a common platform for bioinformatics researchers to conduct analysis. Running the bwtool command without additional parameters displays a description of the various utilities and some general options. Combined with a utility name, bwtool will display specific information about how to perform an operation using that utility. A detailed guide has been created on bwtool's web page (http://cromatina.crg.eu/bwtool) to provide thorough examples of using the program.

bwtool is written in C. The source code for the program is available on its GitHub web page. Distributed (with permission) with bwtool is the basic C library from Jim Kent that is needed for routines specific to bigWig data, as well as other algorithmic code. He and the University of California hold the copyright to this specific library, but the remaining code is covered by the GNU Public License v3. bwtools makes use of GNU autotools to simplify the installation process to the standard “make config”, “make”, “make install” procedure most UNIX users will be familiar with. To verify the accuracy of the software, tests may be run with “make check”: bwtool does not require additional libraries that are not typically found in common UNIX environments, but if the GNU Scientific Library is installed, it will make use of that for the random utility.

ACKNOWLEDGEMENTS

Thanks to Daniel Soronellas, João Curado, Alessandra Breschi, Roderic Guigó, Jakob Skou Pedersen, Jim Kent, and Brian Raney for testing the program and providing feedback and advice prior to release.

**Fig. 1.** Example of aggregated plots of different histone modification ChIP sequence read-depth signals from MCF7 cells from ENCODE aligned at each of the 20,330 protein-coding gene transcription start sites in GENCODE release v17 (Harrow et al., 2012). See supplement for instructions on how to reproduce this plot. The raw signals in this example are not normalized, so specific values cannot be compared between signals, however the morphological differences in averaged profiles are nevertheless useful in characterizing the patterns of each histone mark.

**REFERENCES**


**Funding:** Supported by by grants to M.B. from the Departament d’Innovació Universitat i Empresa (DIIUE), Ministerio de Educación y Ciencia (MEC) BFU2010-01531 and Consolider (CSD2006-00049).