WMAP

MAPPING NEXT GENERATION SEQUENCING (RNA-SEQ, CHIP-SEQ AND METHYLC-SEQ) DATA

USER MANUAL

TABLE OF CONTENTS

INTRODUCTION

WMap is a new sequence mapping software designed to map high-throughput sequencing reads as well as methylated-C reads, enabling usage of such sequencing data to various fields of biological research.

Note: Due to memory limitations, 32-bit WMap may not be able to handle large genomes and/or read files, please limit the size of both or use 64-bit WMap instead.

READ INPUT FILE FORMAT

FASTA format inputs are accepted by WMap (reads more than 76bp are preferable). Additionally, input read file will also be accepted if it meets with the following format criteria:

1. Each line contains only one read;
2. Each line contains exactly 197 bytes, and end in a return sign
3. The Read Sequence should begin at the 29th character of each line and continue for 76 base pairs and should not have any extraneous A, C, T, G, N within 20 characters of read sequence. The read must also be all capital letters.

Currently the first 5 characters of each read are truncated and this will be a user-controlled parameter in recent updates.

OUTPUT FILE FORMAT

The output file is compatible with SAM (Sequence Alignment/Map) Format.

FILE HEADER

The first several lines beginning with "@" are the header, they include some information for the genome or the entire mapping process and have no relationship with individual reads during the mapping.

@HD: This line shows the output file version, currently 1.0.

@SQ: This line shows the genome name the reads are mapped to (SN), and the length of the genome (LN).

MAPPING RESULT

After the header, every mapped read are represented by one line. The result is in tab-delimited format and every line consists of the following fields:

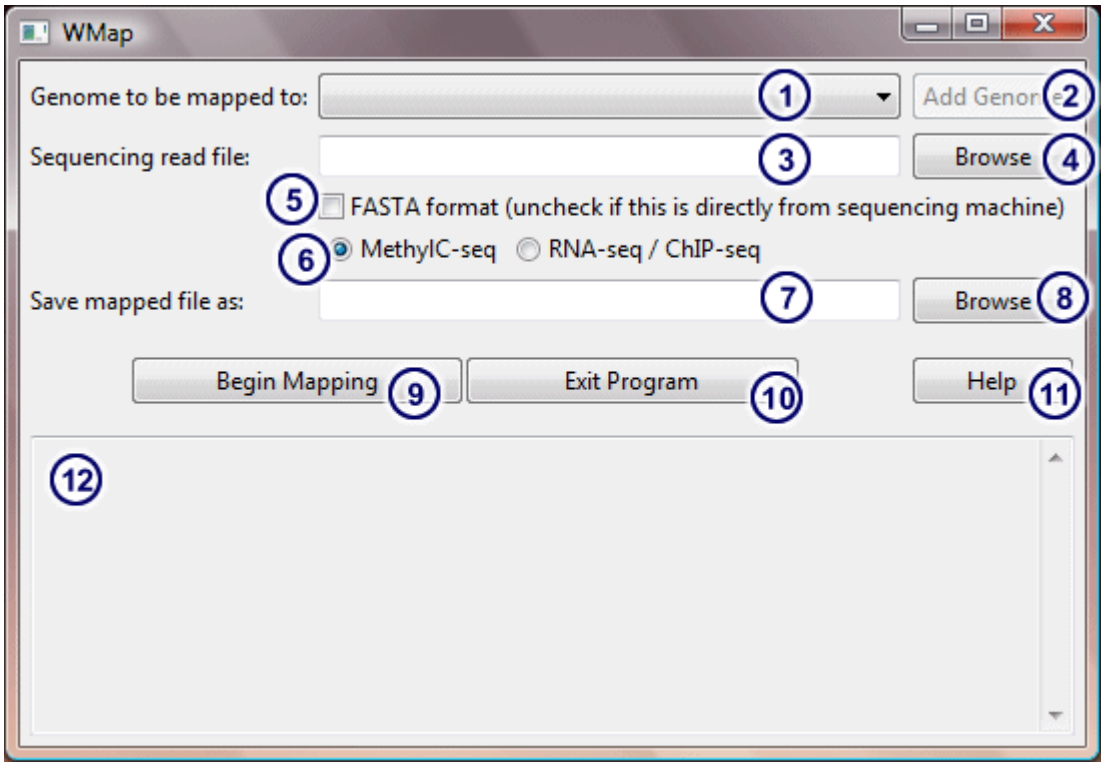
<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> \
<QUAL> NM:i:<NM_VALUE> XM:Z:<METH_INFO> XC:Z:<CHR_NAME> XP:i:<CHR_POS>

Here are the description of the fields:

Field	Description
<QNAME>[1]	The sequence read file name
<FLAG>[1]	Pair-mapping flag
<RNAME>[1]	Genome name
<POS>[1]	The mapped position of the read in the entire genome
<MAPQ>[1]	Mapping quality
<CIGAR>[1]	CIGAR format for mapping information
<MRNM>[1]	Mate reference sequence
<MPOS>[1]	Mate position
<ISIZE>[1]	Inferred insert size
<SEQ>[1]	The sequence of the read
<QUAL>[1]	Query quality
<NM_VALUE>	Number of mismatches in mapping the read
<METH_INFO>	Methylated C information: every letter represent the status of one nucleotide: M means normal match (if the nucleotide is a C, it's not methylated), X means methylated-C, N means mismatch
<CHR_NAME>	The chromosome name that the read is mapped to
<CHR_POS>	The position on the chromosome that the read is mapped at

[1] These fields are part of SAM format specification. For more details about these fields or about SAM format specification not covered in this manual, please refer to SAM format specification at http://samtools.sourceforge.net/SAM1.pdf

GUI FOR WINDOWS (32 BIT) / LINUX (64 BIT) PLATFORM



The GUI interface of WMap will appear like the above image. (*Depending on the OS, the exact appearance of the GUI will be slightly different from the image. However, such difference should not interfere with the function.*) Here is an introduction of every part in the GUI:

1. **Pre-computed genome selection.** The GUI will search under its folder for pre-computed genomes (must be extracted). If you have downloaded a pre-computed genome but it doesn't appear in the list, exit GUI, extract the pre-computed genome to the same folder that executable / GUI is in and restart GUI.
2. **Add new pre-computed genome. (Not available under 32-bit platform due to memory usage limitation)** Pre-compute a new genome (pure sequence or FASTA) for mapping. **All genomes must be first pre-computed before short sequences can be mapped to them.**
3. **File containing the short sequence reads.** Input the file name manually or use the button to browse.
4. **Browse for file containing the short sequence reads.** The file name will be automatically updated after browse.
5. **FASTA format.** Check this if the read file is in FASTA format, otherwise leave it blank. **The non-FASTA format should meet the requirement mentioned above.**
6. **Sequencing type.** Select "Bisulfite sequencing" if your reads come from a bisulfite sequencing (to detect cytosine methylation). Select "RNA / ChIP sequencing" if your reads have not been modified in sequence.
7. **Output file.** To specify where you want to save the output file.
8. **Browse output file.** The file name will be automatically updated after browse.
9. **Begin Mapping.** After you provided all needed information, click this to begin the mapping process.
10. **Exit Program.** Exit the GUI.
11. **Help.** You can open the user manual from here.
12. **Log for W-Map.** The detailed mapping log will appear here.

TUTORIAL

To get W-Map up and running on a 64-bit Linux / 32-bit Windows machine the following steps will be needed:

1. Download WMap from http://biocomp.bioen.uiuc.edu/wmap/download.html.
2. Extract all files in the package to the directory you want to install WMap at.
3. (Required for 32-bit Windows platform) Download pre-computed genome from http://biocomp.bioen.uiuc.edu/wmap/download.html, extract and put all pre-computed genome in the same directory.
4. Execute "./WMap.jar" from command line (Linux) / "WMap.jar" (Windows) to start WMap GUI.
5. Select read file, format, sequence type and output file name to begin mapping;
   or (64-bit Linux only) pre-compute new genome
6. Read the output file for results
7. Sample reads file can be also be downloaded at
   http://biocomp.bioen.uiuc.edu/wmap/download.html

FAQ

1. *What is Bisulfite Sequence Mapping?*
   Currently a huge amount of bisulfite treated reads have been generated in the attempt to understand which specific base pairs in a given genome are methylated. Bisulfite sequencer will take these millions of reads and map them onto a reference genome and will take into account conversions from cytosine to thymine as well as allowing for mismatches. Currently we allow for 2 mismatches.
2. *Will WMap miss detection of some reads?*
   Currently yes, we will miss detection of certain types of input read sequences. Because of the algorithm we use, if one third of the sequence is of unusually low complexity after converting thymine to cytosine, i.e. a long string of c's, then these read sequences will be thrown out.