metadamage(?) formats

tsk

June 8, 2022

This document describe some of the internal formats used by the metadamage software. These are at the current time.

- .bdamage.gz
- .lca
- .stat

1 bdamage format

bdamage files are files that contain counts of mismatchs conditional on strand and cycle (position within read). These are generated with metadamage lca or metadamage getdamage. The first 8 bytes magic number determines which bdamage version. If no magic number is present then version0 is assumed.

1.1 version 0

First version of the bdamage file is a single bgzf compressed file. MAXLENGTH occurs once in the beginning of the file, followed by succesive blocks of data[1-8]. Block[3-5] indicates the actual mismatch counts for the forward which we have MAXLENGTH times. Block[6-8] indicates the actual mismatch counts for the reverse strand which will also occur MAXLENGTH times.

Col	Field	Type	Brief description
0	MAXLENGTH	int	Number of cycles
1	ID	int	Id for mismatch type[1]
2	NREADS	size_t	Number of reads used supporting the mismatch matrix
3	1	int[16]	mismatch rate for first cycle from the 5prime
4	*i*	int[16]	mismatch rate for the *i*'th cycle from the 5prime
5	MAXLENGTH	int[16]	mismatch rate for the last cycle from the 5prime
6	1	int[16]	mismatch rate for first cycle from the 3prime
7	*i*	int[16]	mismatch rate for the *i*'th cycle from the 3prime
8	MAXLENGTH	int[16]	mismatch rate for the last cycle from the 3prime

Table 1: Content of bdamage.gz file. Note[1] This is either the taxidID or the referenceID relative to the SAM/BAM header for single species resequencing projects or it is the *taxid* if output has been generated with metadamage lca 3) Order is given by AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT, with first base indicatting reference nucleotide and second base indicating observed nucleotide
.

1.2 version 1

First version of the bdamage file is a single bgzf compressed file. MAXLENGTH occurs once in the beginning of the file, followed by succesive blocks of data[1-8]. Block[3-5] indicates the actual mismatch counts for the forward which we have MAXLENGTH times. Block[6-8] indicates the actual mismatch counts for the reverse strand which will also occur MAXLENGTH times.

Col	Field	Type	Brief description
0	MAXLENGTH	int	Number of cycles
1	ID	int	Id for mismatch type[1]
2	NREADS	size_t	Number of reads used supporting the mismatch matrix
3	NREADS2	size_t	Number of unique reads
4	1	int[16]	mismatch rate for first cycle from the 5prime
5	*i*	int[16]	mismatch rate for the *i*'th cycle from the 5prime
6	MAXLENGTH	int[16]	mismatch rate for the last cycle from the 5prime
7	1	int[16]	mismatch rate for first cycle from the 3prime
8	*i*	int[16]	mismatch rate for the *i*'th cycle from the 3prime
9	MAXLENGTH	int[16]	mismatch rate for the last cycle from the 3prime

Table 2: Content of bdamage.gz file. Note[1] This is either the taxidID or the referenceID relative to the SAM/BAM header for single species resequencing projects or it is the *taxid* if output has been generated with metadamage lca 3) Order is given by AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT, with first base indicatting reference nucleotide and second base indicating observed nucleotide
.

2 .lca

This section describes the test output generated by a metadamage lca subfunctionality and contains information at the readlevel regarding both taxonomic information and statistics pertaining usefull readinformation.

First line of the file begins with a hashtag followed by the actual command used for generating the file. Last entry of the line is again a hashtag followed by the git commit id which will serve as a primitive versioncontrol.

Each line consists of a number of items seperated by tabspace. First entry contains readID together with other information seperated by colon. After the first entry succesive blocks of the type taxid:name:”taxlevel” from the lca toward the root. The complete specification is seen in table below.

Col	Brief description	
1	readID	readID, this might contains colon
2	seq	The actual sequence
3	length(seq)	The length of the sequence
4	nAlignments	The number of alignments used for inferring the lca
5	gc-content	The GC content for the sequence
6	taxid	the taxomic id (integer)
7	taxid	the taxonomic name(string)
8	“taxlevel”	the taxonomic level

Table 3: Content of a .lca file. Note that 1) seperate between field[1-5] is colon, but the readID might also contain colon. 2) the quotes around field[8] is intentional since taxlevels might contain spaces. 3) Not that the number of tab seperated entries is different between reads since this is the number of nodes needed to traverse from lca to root

3 .stat

Very simple tabseperated flatfile

1. taxid
2. Number of supporting reads
3. Mean lengths of supporting reads
4. Variance of the lengths of the supporting reads
5. Mean gccontent of supporting reads
6. Variance of the gccontent of supporting reads
7. name of lca in quotes (if relevant, otherwise NA)
8. name of taxomic level of lca in quotes (if relevant, otherwise NA)