metadamage(?) formats

tsk

October 13, 2023

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

- .bdamage.gz
- .lca
- .stat
- .dfit.txt.gz
- .dfit.stat.txt.gz

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

.

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	lca taxid	the taxomic id (integer)
7	lca taxid	the taxonomic name(string)
8	lca "taxlevel"	the taxonomic level
9	taxid	the taxomic id (integer)
10	taxid	the taxonomic name(string)
11	"taxlevel"	the taxonomic level

Table 2: Content of a .lca file. Note that 1) seperate between fields[1-6,8-9] is tab, readID might also contain colon. 2) the quotes around field[7,8] is intentional since taxlevels and names might contain spaces andor colons. 3) Number of tab seperated entries is consistent across different reads, since seperator between block 9-11 is semicolon. Seperator between 6-9,9-11 is colon

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)

4 .dfit.txt.gz

This output format was added 8d28d337 Sun Jul 23. Depending on runmode and parameters supplied to the program. It will contain the per file, the per reference or the per species estimate of damage.

Col	Brief description	
1	id	identifier see paragraph for details
2	A	Dfit statistic. Damage at position one, taking into account offset
3	q	per cycle decrease
4	c	background substitution rate or noise baseline
4	ϕ	
5	llh	Likelihood for our MLE
6	nopt	Number of function calls used for obtaining our MLE
7	Zfit	Z value
8	Zconf	significance
9	K_i	Number of CT og AG observations
10	N_i	Count of either CA,CC,CG,CT or AA,AG,AC,AT observations
11	Dx_i	Fitted value of cycle specific damage
13	$Dconf_i$	Confidence interval for the cycle specfic damage

Table 3: Content of a .dfit.txt.gz file. Note that entry nine to 13 is repeated for each cycle first of the 5ánd then from the 3ź With the total number of times repeated is given by 1.1.

Depending on which parameters that was supplied to both *getdamage* and *dfit*, the content of the id will be different. If a bamfile is supplied (with -bam) then each line will be the information associated with the different refids in the bam file and the id will be the referenceids from the bam file. The case scenario for this would be either obtaining perchromosome estimates of damage or per reference damage which could be relevant for metagenomic studies. If user are computing the damagesignal in the context of the lca. Then the id column will contain the taxid. If -names has been supplied to the dfit program, then the id column will the taxid:*scientific name* . If -nodes has not been defined the dfit.txt.gz will only contain information for the observed references. If -nodes has been defined the program will aggregate the summary statitics for the internal nodes.

5 .dfit.stat.txt.gz

This output format was added 8d28d337 Sun Jul 23. Depending on runmode and parameters supplied to the program. It will contain the per file, the per reference or the per species estimate of the statistic. This format extends 3 to include the internal nodes.