1 The TWK Format Specification

1.1 TGZF specification

TGZF is a variation of BGZF used in the BAM and BCF formats (https://samtools.github.io/hts-specs/) that in turn are slightly modified gzip blocks concatenated together back-to-back to enable random access lookups. In short, TGZF lifts the size restriction in BGZF that uncompressed data cannot exceed 2^{16} bytes and instead asserts that the compressed TGZF data described the BSIZE field is smaller than 2^{32} bytes. When a block of uncompressed data is dispatched to the compressor and sent to disk is contextually determined and is available as a modifiable variable in many cases.

Notably, as the TGZF/BGZF format naturally extend the ZLIB library (http://zlib.net/), all multi-byte values are little endian (as required by the gzip specification). Endianness is not asserted in the uncompressed data.

Field	Description	Type	Value
List of compression blocks (until the end of the file)			
ID1	gzip IDentifier1	uint8_t	31
ID2	gzip IDentifier2	uint8_t	139
CM	gzip Compression Method	uint8_t	8
FLG	gzip FLaGs	uint8_t	4
MTIME	gzip Modification TIME	uint32_t	
XFL	gzip eXtra FLags	uint8_t	0
OS	gzip Operating System	uint8_t	255
XLEN	gzip eXtra LENgth	uint16_t	8
Extra subfield(s) (total size=XLEN)			
Additional RFC1952 extra subfields if present			
SI1	Subfield Identifier1	uint8_t	84
SI2	Subfield Identifier2	uint8_t	90
SLEN	Subfield LENgth	uint16_t	2
BSIZE	total Block SIZE	uint32_t	
Additional RFC1952 extra subfields if present			
CDATA	Compressed DATA by zlib::deflate()	uint8_t[BSIZE-XLEN-19]	
CRC32	CRC-32	uint32_t	
ISIZE	Input SIZE (length of uncompressed data)	uint32_t	

1.2 Template type data bounds

The Tomahawk data structure described below uses different integer types based on the number of samples in the imported file. This Template field is reinterpreted from the byte buffer as follows:

Field	Description
uint8_t	$0 < \text{n\_samples} < 2^4$
uint16_t	$2^4 < \text{n\_samples} < 2^{12}$
uint32_t	$2^{12} < \text{n\_samples} < 2^{28}$
uint64_t	$2^{28} < \text{n\_samples} < 2^{60}$

Using this approach, Tomahawk run-length encodes genotypes with the lowest 4 bits being the alleles and the remainder $\text{sizeof(Template)} * 8 - 4$ bits as the run-length.

1.3 TWK organization

A TWK file comprises of a binary (non-compressed) header followed by a series of TGZF-compressed binary blocks of TWK records and ended with a binary EOF marker.

The TWK format asserts that:

- Contig information is specified in the header: this data must minimally include a contig identifier (such as a unique string name in VCF or unique integer in BCF) and contig length in base pairs

- Entries must be bi-allelic SNVs
- Entries in the imported VCF / BCF file are sorted by contig identifier followed by genomic coordinates (base position of SNV)
- There is > 1 sample in the file
- If importing from the BCF format: is version 2.2 or later
- If importing from the VCF format: is version 4 or later

The fields INFO and FORMAT are dropped from imported VCF/BCF files as they are not used in the current implementation

1.4 TWI: Random access

Searching to the beginning of a specified TGZF block is aided by TWI entries specifying virtual offsets into the TWK file. Importantly, unlike BAM and BCF, none of TWK/TWI/TWO/TOI permits data to be split over multiple blocks. This intentional restriction guarantees that the uncompressed data is completely disjoint and this assertion renders parallel computing exceedingly more tractable.

Unlike BCI/BAI, TWI/TOI stores the virtual file offset to every single TGZF block. This permits complete random access to any part of a TWK/TWO file. This makes TWI/TOI indices larger than BCI/BAI indices but still remain very small. For example, the TWI file for HRC.v1-1 is < 2 MB.

Field	**Description**	**Type**	**Value**
MAGIC	Start of file identifier string	`char[10]`	`TOTEMPOLE\1`
version	Tomahawk major version	`float`	
samples	Number of samples	`uint64_t`	
controller	Currently unused	`uint8_t`	0
n_blocks	Number of TGZF blocks in Twk file	`uint32_t`	
n_largest	Size in bytes of largest uncompressed TGZF block	`uint32_t`	
header_offset	Relative disk offset until this position (start of data)	`uint32_t`	
header_offset_end	Relative disk offset until end of block (end of data)	`uint32_t`	
n_contigs	Number of contigs in header	`uint32_t`	
List of contig data (n_contigs)			
bp_contig	Length of contig in bases	`uint32_t`	
l_contig	Length of contig name	`uint32_t`	
contig_name	Contig name	`char[l_contig]`	
List of sample identifiers (n_samples)			
l_name	Length of sample name	`uint32_t`	
sample_name	Sample name	`char[l_name]`	
TGZF_block	Compressed DATA by zlib::deflate(). DATA keeping VCF header and any changes made to the file		
Totempole entries until end-of-file			
byte_offset	Virtual file offset to start to TGZF block	`uint64_t`	
contigID	All variants belong to this contig identifier	`int32_t`	
min_position	Smallest variant position	`uint32_t`	
max_position	Largest variant position	`uint32_t`	
n_variants	Number of variants	`uint16_t`	
uncompressed_size	Uncompressed size of data	`uint32_t`	
EOF_string	End-of-file marker	`char*`	

1.5 TWK format

Field	Description	Type	Value
MAGIC	Start of file identifier string	char[9]	TOMAHAWK\1
version	Tomahawk major version	float	
samples	Number of samples	uint64_t	
TGZF blocks until end-of-file			
For n_variants (described in TWI)			
pos_plus	pos<<30|phased<<1|missing; Genomic coordinate; flag if all data is phased; flag if any data is missing	uint32_t	
ref_alt	REF<<4|ALT	uint8_t	
MAF	Minor allele frequency	float	
HWE_P	Hardy-Weinberg P-value (Fisher's exact test)	float	
n_runs	Number of runs for this variant	Template	
Until end of TGZF block			
RLE	Run-length encoded data	Template*	
EOF_string	End-of-file marker	char*	

1.6 TWO format

TWO entry

Field	Description	Type	Value
MAGIC	Start of file identifier string	char[17]	TOMAHAWK~OUTPUT\1
version	Tomahawk major version	float	
samples	Number of samples	uint64_t	
n_contigs	Number of contigs in header	uint32_t	
List of contig data (n_contigs)			
bp_contig	Length of contig in bases	uint32_t	
l_contig	Length of contig name	uint32_t	
contig_name	Contig name	char[l_contig]	
TGZF_block	Compressed DATA by zlib::deflate(). DATA keeping VCF header and any changes made to the file		
TGZF blocks of TWO entries until end-of-file			

TWO entries

Field	Description	Type	Value
FLAG	Bit-wise flags	uint16_t	
AcontigID	Variant A contig map identifier, $0 \leq$ AcontigID $<$ n_ref	uint32_t	
Aposition	pos<<30|phased<<1|missing	uint32_t	
BcontigID	Variant B contig map identifier, $0 \leq$ BcontigID $<$ n_ref	uint32_t	
Bposition	pos<<30|phased<<1|missing	uint32_t	
p1	Haplotype counts for A1B1 (Ref-Ref). Is estimated if FLAG bit 1 is unset	float	
p2	Haplotype counts for A1B2 (Ref-Alt). Is estimated if FLAG bit 1 is unset	float	
q1	Haplotype counts for A2B1 (Alt-Ref). Is estimated if FLAG bit 1 is unset	float	
q2	Haplotype counts for A2B2 (Alt-Alt). Is estimated if FLAG bit 1 is unset	float	
D	Coefficient of linkage disequilibrium (D)	float	
Dprime	Normalised D value	float	
R2	Correlation coefficient squared	float	
P	Fisher's exact test P-value	double	
chiSqFisher	Exact Fisher's test or Chi-squared test (see FLAG bit) for the 2x2 haplotype contingency table	double	
chiSqModel	Chi-squared critical value for the 3x3 genotype contingency table	double	

1.7 TWO FLAG field

The TWO FLAG

Bit	Description
1	Both variant lines were phased OR equations used for phased genotypes was used
2	Either variant has missing values
3	A field is incomplete (A1B1, A1B2, A2B1, or A2B2 has 0 observations)
4	There are multiple possible biological solutions (valid roots in cubic equation)
5	Both variants are on the same contig
6	There is $>$ 1 million base pairs between the variants
7	Variant A failed Hardy-Weinberg test ($P < 10^{-6}$)
8	Variant B failed Hardy-Weinberg test ($P < 10^{-6}$)
9	Variant A has a low minor allele frequency ($<$ 1%)
10	Variant B has a low minor allele frequency ($<$ 1%)
11	Currently unused
12	Currently unused
13	Currently unused
14	Currently unused
15	Currently unused
16	Currently unused

1.8 TOI format

Field	Description	Type	Value
MAGIC	Start of file identifier string	`char[9]`	`TOMAHAWK~OUTPUT~INDEX\1`
version	Tomahawk major version	`float`	
n_samples	Number of samples	`uint64_t`	
n_entries	Number of `TOI entries`	`uint32_t`	
controller	sorted<<7|expanded<<6|partial_sort<<5|unused	`uint8_t`	
List of `TOI` *entries until* `n_entries`			
byte_offset	Virtual data offset into `TWO`	`uint64_t`	
byte_offset_end	Virtual data offset end `TWO`	`uint64_t`	
n_entries	Number of `TWO` entries	`uint32_t`	
uncompressed_size	`TGZF DATA` uncompressed size	`uint32_t`	
Extra subfield(s) if `controller` *is sorted and expanded*			
List of TOI sorted entries for `n_contigs`			
from_block	BlockID start offset in `TWO`	`int32_t`	
fromBlock_entries_offset	Start position in block at `TWO` position	`uint32_t`	
to_block	BlockID end offset in `TWO`	`int32_t`	
toBlock_entries_offset	End position in block at `TWO` position	`uint32_t`	
List of TOI sorted entry bins for `n_contigs`			
List of TOI sorted entries for 1024 iterations			
from_block	BlockID start offset in `TWO`	`int32_t`	
fromBlock_entries_offset	Start position in block at `TWO` position	`uint32_t`	
to_block	BlockID end offset in `TWO`	`int32_t`	
toBlock_entries_offset	End position in block at `TWO` position	`uint32_t`	