

MapView Format

MVF (MapView Format) is a novel file format designed for fast and memory efficiency visualization of huge amount of short reads alignment.

The maximum length of reference sequence: 1 - 4294967295 ($2^{32}-1$).

The maximum length of short read: 1 - 255 (2^8-1).

All the sequences are stored by two bits unit:

A: 00

C: 01

G: 10

T: 11

Characters other than "ACGT" are recorded as unidentified base and displayed as "N".

MVF file use 40 bit addressing and support file size up to 1024GB.

Main File (.MVF)

File head (150 bytes)

0 ... 35	MVF_04_2009-03-02 17:15_MapView 2.7
36 ... 39	The offset address of compressed reference sequence *
40 ... 43	The length of reference sequence
44 ... 47	The number of base "A" in reference sequence
48 ... 51	The number of base "T" in reference sequence
52 ... 55	The number of base "G" in reference sequence
56 ... 59	The number of base "C" in reference sequence
60 ... 63	The number of base "N" in reference sequence
64 ... 67	The number of base that has an item in base-statistic-information section
68 ... 71	The number of short reads
72 ... 75	The number of short reads in negative strand
76 ... 79	The number of short reads that have mismatch
80 ... 83	The total number of mismatched bases in short reads
84 ... 87	The number of short reads that have unidentified base
88 ... 91	The number of unidentified base in short reads
92 ... 95	The number of short reads index

96 ... 99	NULL
100	0 0 0 0 0 0 pair-end quality score
101 ... 105	The offset address of unidentified base on reference sequence §
106 ... 110	The offset address of short reads index ※
111 ... 115	The offset address of short reads item ☆
113 ... 149	NULL

Data

150	The name of reference file \t The name of mapping file \t documents *Compressed reference sequence The position of unidentified base on reference sequence ☆The information of each short read item: The starting position on reference sequence The starting position of paired read on reference sequence (when paired-end = 1) Compressed short read sequences (theReadSeqLength/4 + 0.9 bytes) Quality score of this read (theReadSeqLength bytes, when Quality = 1) The number of mismatch in this read The position of the first mismatch to the last one on this read The number of unidentified base in this read The position of the first unidentified base to the last one on this read The length of read name (low 7 bits, highest bit: 0, forward strand; 1, reverse strand) The name of this read
-----	---

Index table

※The address of short reads index table. For each index:
The starting position on reference sequence
The offset address of the file where the short reads mapping on this position

File end

The length of MVF file: FL

FL-13 ... FL-16	Default: 12 34 AB CD
FL-9 ... FL-12	Default: 12 34 AB CD
FL-1 ... FL-8	MVF__END

Additional file (.sta)

18 bytes for each reference site

- 1 ... 4 The position on reference sequence
- 5 ... 6 The number of short reads covered on this position
- 7 ... 8 The number of "A" on this position
- 9 ... 10 The number of "T" on this position
- 11 ... 12 The number of "G" on this position
- 13 ... 14 The number of "C" on this position
- 15 ... 16 The number of "N" on this position

The highest two bits of byte 15 store the base on reference sequence in this position

- 17 ... 18 The variant frequency without consider quality score

The highest two bits of byte 17 store the variant base

3 bytes for each quality score

- 1 quality score character
- 2 ... 3 Percentage of the quality score.