The Optical Map File Format Specification (v1.4)

2018/01/01

Alden Leung, Ting-Fung Chan
The Chinese University of Hong Kong

1. Optical Map Reference (REF) format

The REF format includes basic information for an optical map. Every optical map entry is represented in two lines.

Line 1:

Col	Field	Type	Brief Description
1	Fragment ID	String	Name of the optical map
2	Size	Long	Size of the optical map
3	TotalSignals	Integer	Total number of signals in the optical map

Line 2:

Position of the signals, separated by a white tab.

2. Optical Map Data (DATA) format

The DATA format includes basic information for an optical map. One line represents one optical map entry.

Col	Field	Type	Brief Description
1	Fragment ID	String	Name of the optical map
2	Size	Long	Size of the optical map
3	TotalSegments	Integer	Number of segments
4	SegmentDetail	String	Segment Information

1. **Fragment ID**: Optical map name.
2. **Size**: Size of the optical map. This is equal to number of segments – 1 + summation of size from all segments.
3. **TotalSegments**: Number of segments in the optical map. A query of n segments contains (n – 1) signals.
4. **SegmentDetail**: Optical map segment information. Contain length of all segments along the query, separated by semi-colon.

3. Optical Map Simulated Data (SDATA) format

The DATA format includes basic information for an optical map. One line represents one optical map entry.

Col	Field	Type	Brief Description
1	Fragment ID	String	Name of the optical map
2	Reference	String	Simulated reference name
3	Strand	String	Simulated strand
4	Start	Long	Simulated reference start position
5	Stop	Long	Simulated reference stop position
6	SimuInfoDetail	String	Detailed information of simulation
2	Size	Long	Size of the optical map
3	TotalSegments	Integer	Number of segments
4	SegmentDetail	String	Segment Information

1. **Fragment ID**: Optical map name.
2. **Reference**: The name of reference where the optical map comes from in simulation.
3. **Strand**: Strand in simulation. A strand could be in either “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] direction.
4. **Start**: The start coordinate on the reference where the optical map comes from.
5. **Stop:** The stop coordinate on the reference where the optical map comes from.
6. **SimuInfoDetail**: The source of the simulated signals, separated by semi-colon. Each simulated signal can come from multiple source signals. The source signals are separated by comma. The source signal is usually a reference signal or a false positive / extra signal. In the first case, the reference signal is represented in “R:xx” which can be interpreted as the signal at index “xx” on reference “R”. In the second case, a false positive signal is represented as “FP”. This field is optional.
7. **Size**: Size of the optical map. This is equal to number of segments – 1 + summation of size from all segments.
8. **TotalSegments**: Number of segments in the optical map. A query of n segments contains (n – 1) signals.
9. **SegmentDetail**: Optical map segment information. Contain length of all segments along the query, separated by semi-colon.

4. Optical Map Alignment (OMA) format

The OMA format includes basic alignment information and extra information on query. One line represents one partial alignment entry.

Col	Field	Type	Brief Description
1	QueryID	String	Query name
2	QuerySeg	Integer	Number of query segments
3	QuerySegInfo	String	Query segment Information
4	RefID	String	Reference name
5	Strand	String	Alignment strand
6	Score	Float	Alignment score
7	Confidence	Float	Alignment confidence
8	RefSegStart	Integer	Reference segment start
9	RefSegStop	Integer	Reference segment stop
10	QuerySegStart	Integer	Query segment start
11	QuerySegStop	Integer	Query segment stop
12	RefStartCoord	Long	Reference start coordinate
13	RefStopCoord	Long	Reference stop coordinate
14	Cigar	String	CIGAR String

1. **QueryID**: Query name. Entries with same **QueryID** are regarded as the alignment from the same query.
2. **QuerySeg**: Number of segments in the query. A query of n segments contains (n – 1) signals.
3. **QuerySegInfo**: Query segment information. Contain length of all segments along the query, separated by semi-colon.
4. **RefID**: The reference name. If reference name is “Unmapped” or “Discarded”, it indicates the query is not aligned. In this case later columns are not defined and should be left empty.
5. **Strand**: Alignment strand. A query is aligned in either “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] direction.
6. **Score**: Alignment score. This value reflects the quality of an alignment.
7. **Confidence**: Alignment confidence. The value ranges from 0 to 1, reflecting the specificity of the alignment.
8. **RefSegStart**: Reference segment start. The first reference segment in the alignment. **RefSegStart** is always smaller than or equal to **RefSegStop**

9. **RefSegStop**: Reference segment stop. The last reference segment in the alignment.
10. **QuerySegStart**: Query Segment Start. The first query segment aligned. If a query is forwardly/reversely aligned on the reference, **QuerySegStart** should be smaller/larger than or equal to **QuerySegStop**.
11. **QuerySegStop**: Query Segment Stop. The last query segment aligned.
12. **RefStartCoord**: Reference Start Coordinate. Refer to the position of first reference signal in alignment.
13. **RefStopCoord**: Reference Stop Coordinate. Refer to the position of last reference signal in alignment.
14. **Cigar**: CIGAR String. The Cigar String is defined as follows:

Character	Description
M	Signal match
I	Signal insertion to reference (Extra signal on query)
D	Signal deletion from reference (Missing signal on query)

"null" is output if Cigar string is not available

5. Optical Map Alignment with Details (OMD) format

The OMD format includes detailed alignment information and simulation information on query (if it is from simulated data). One line represents one partial alignment entry.

Col	Field	Type	Brief Description
1	QueryID	String	Query name
2	simuRefID	String	Simulated reference name
3	simuStrand	String	Simulated strand
4	simuStart	Long	Simulated reference start position
5	simuStop	Long	Simulated reference stop position
6	QuerySize	Long	The size of query
7	QuerySeg	Integer	Number of query segments
8	QuerySegInfo	String	Query segment information
9	RefID	String	Reference name
10	RefStartCoord	Long	Reference start coordinate
11	RefStopCoord	Long	Reference stop coordinate
12	Strand	String	Alignment strand
13	RefSegStart	Integer	Reference segment start
14	RefSegStop	Integer	Reference segment stop
15	QuerySegStart	Integer	Query segment start
16	QuerySegStop	Integer	Query segment stop
17	AlignedSegRatio	Float	Aligned length of query divided by query size
18	Score	Float	Alignment score
19	Cigar	String	CIGAR string
20	Confidence	Float	Alignment confidence
21	FP	Integer	Number of false positive / extra signals
22	FN	Integer	Number of false negative / missing signals
23	Scale	Float	The scaling of query with respect to the reference
24	FPRate	Float	Ratio of false positive signals
25	FNRate	Float	Ratio of false negative signals

26	simuCorrectlyMapped	Boolean	Correctness

1. **QueryID**: Query name. Entries with same **QueryID** are regarded as the alignment from the same query.
2. **simuRefID**: The name of reference where the query comes from in simulation[1].
3. **simuStrand**: Strand in simulation. A strand could be in either “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] direction[1].
4. **simuStart**: The start coordinate on the reference where the query comes from[1].
5. **simuStop:** The stop coordinate on the reference where the query comes from[1].
6. **simuSize:** The length of query in bp[1].
7. **QuerySeg**: Number of segments in the query. A query of n segments contains (n – 1) signals.
8. **QuerySegInfo**: Query segment information. Contain length of all segments along the query, separated by semi-colon.
9. **RefID**: The reference name. If reference name is “Unmapped” or “Discarded”, it indicates the query is not aligned. In this case later columns are not defined and should be left empty.
10. **RefStartCoord**: Reference Start Coordinate. Refer to the position of first reference signal in alignment.
11. **RefStopCoord**: Reference Stop Coordinate. Refer to the position of last reference signal in alignment.
12. **Strand**: Alignment strand. A query is aligned in either “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] direction.
13. **RefSegStart**: Reference segment start. The first reference segment in the alignment. **RefSegStart** is always smaller than or equal to **RefSegStop**
14. **RefSegStop**: Reference segment stop. The last reference segment in the alignment.
15. **QuerySegStart**: Query Segment Start. The first query segment aligned. If a query is forwardly/reversely aligned on the reference, **QuerySegStart** should be smaller/larger than or equal to **QuerySegStop**.
16. **QuerySegStop**: Query Segment Stop. The last query segment aligned.
17. **AlignedSegRatio**: Ratio of length of query aligned in this alignment entry
18. **Score**: Alignment score. This value reflects the quality of an alignment.
19. **Cigar**: CIGAR String. The Cigar String is defined as follows:

[1] Blank or -1 if the simulation information does not exist

Character	Description
M	Signal match
I	Signal insertion to reference (Extra signal on query)
D	Signal deletion from reference (Missing signal on query)

"null" is output if Cigar string is not available

20. **Confidence**: Alignment confidence. The value ranges from 0 to 1, reflecting the specificity of the alignment.
21. **FP**: Number of false positive / extra signals in the alignment
22. **FN**: Number of false negative / missing signals in the alignment
23. **Scale**: The scaling factor of query with respect to the reference. This is calculated as length of aligned region of query divided by that of reference.
24. **FPRate**: The false positive rate, which equals to **FP** divided by aligned length of the query
25. **FNRate**: The false negative rate, which equals to **FN** divided by total signals present in the aligned region on the query
26. **simuCorrectlyMapped**: Correctness of alignment according to the simulation

6. Segment Identifier (SI) format

The SI format includes basic segment identifiers of queries. One line represents one segment.

Col	Field	Type	Brief Description
1	QueryID	String	Query name
2	SegmentIndex	Integer	Index of segment in the query

1. **QueryID**: Query name.
2. **SegmentIndex**: The index of segment in the query. The first segment of a query has an index 0.

7. CBL format

The CBL format stores the multiple alignment as a list of blocks. One line represents one block.

Col	Field	Type	Brief Description
1	BlockName	String	Block name
2	BlockDetails	String	Details of blocks from queries

1. **BlockName**: Block name.
2. **BlockDetails**: A semi-colon separated query fragments included in the blocks. Each fragment is represented as “Q:S-ED”, where Q, S, E, and D represent query name, start signal, end signal and direction respectively.

8. CBO format

The CBO format includes the order of queries in multiple alignment displayed in OMView. One line represents one query.

Col	Field	Type	Brief Description
1	GroupName	String	Group name
2	QueryID	String	Query name

1. **GroupName**: Group name of the query. Use the query name if the query is not assigned to any group.
2. **QueryID**: Query name.

9. CBC format

The CBC format includes the color of multiple alignment to be visualized in OMView. One line represents one block.

Col	Field	Type	Brief Description
1	GroupName	String	Group name
2	QueryID	String	Query name

3. **GroupName**: Group name of the queries. Use the query name if queries are not assigned to a group.
4. **QueryID**: Query name.