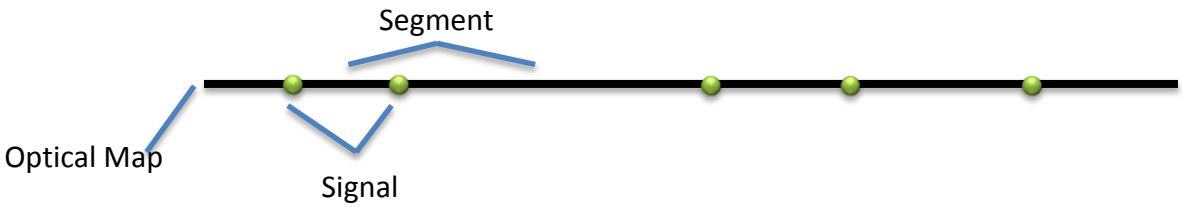
[Year]

The Optical Map File Format Specification

(v1.1) 2015/12/01

Alden Leung

1. Introduction



Optical Map: A DNA molecule marked with fluorescent signals, represented by as a tuple

Signal: Fluorescent signal. Fluorescent signals are marked in the DNA molecule with a specific pattern.

Segment: The regions divided by the signals are called “segment”. An optical map is divided by (n – 1) signals into n segments.

Alignment: The alignment of query optical map to the reference optical map

Definition of CIGAR String:

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

“null” is output if Cigar string is not available

2. Optical Map Alignment (OMA) format

The OMA format includes basic alignment information and extra information about the query.

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 the same **QueryID** are regarded as alignments 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. Contains the length of all segments along the query, separated by semi-colons.
4. **RefID**: The reference name. If the reference name is “Unmapped” or “Discarded”, it indicates the query is not aligned. In this case later columns are undefined and can be left empty.
5. **Strand**: Alignment strand. A query is aligned in either the “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 aligned in the forward/reverse direction 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. Refers to the position of the first reference signal in the alignment.
13. **RefStopCoord**: Reference Stop Coordinate. Refers to the position of the last reference signal in the alignment.
14. **Cigar**: CIGAR String.

3. Optical Map Alignment with Details (OMD) format

This an extended format to OMA, with more information about the query and statistics about the alignments

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 the query
7	QuerySeg	Integer	Number of query segments
8	QuerySegInfo	String	Query segment information
9	RefID	String	Reference name
10	Strand	String	Alignment strand
11	RefSegStart	Integer	Reference segment start
12	RefSegStop	Integer	Reference segment stop
13	QuerySegStart	Integer	Query segment start
14	QuerySegStop	Integer	Query segment stop
15	RefStartCoord	Long	Reference start coordinate
16	RefStopCoord	Long	Reference stop coordinate
17	AlignedSegRatio	Float	Aligned length of query divided by the query size
18	Score	Float	Alignment score
19	Cigar	String	CIGAR string
20	Confidence	Float	Alignment confidence
21	FP	Integer	Number of false positives / extra signals
22	FN	Integer	Number of false negatives / missing signals
23	Scale	Float	The scaling of the query with respect to the reference
24	FPRate	Float	Ratio of false positive signals to length of alignment
25	FNRate	Float	Ratio of false negative signals to total reference signals
26	simuCorrectlyMapped	Boolean	Correctness

1. **QueryID**: Query name. Entries with the same **QueryID** are regarded as alignments from the same query.
2. **simuRefID**: The name of the reference where the query comes from in simulation
3. **simuStrand**: Strand in the simulation. A strand could be in either the “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] directions.
4. **simuStart**: The start coordinate on the reference where the query comes from
5. **simuStop:** The stop coordinate on the reference where the query comes from
6. **simuSize:** The length of the query in bp
7. **QuerySeg**: Number of segments in the query. A query of n segments contains (n – 1) signals.
8. **QuerySegInfo**: Query segment information. Contains the length of all segments along the query, separated by semi-colons.
9. **RefID**: The reference name. If the reference name is “Unmapped” or “Discarded”, it indicates the query is not aligned. In this case later columns are undefined and can be left empty.
10. **Strand**: Alignment strand. A query is aligned in either the “forward” [Also: “1”, “+”], or “reverse” [Also: “-1”, “-“] directions.
11. **RefSegStart**: Reference segment start. The first reference segment in the alignment. **RefSegStart** is always smaller than or equal to **RefSegStop**
12. **RefSegStop**: Reference segment stop. The last reference segment in the alignment.
13. **QuerySegStart**: Query Segment Start. The first query segment aligned. If a query is aligned in the forward/reverse direction on the reference, **QuerySegStart** should be smaller/larger than or equal to **QuerySegStop**.
14. **QuerySegStop**: Query Segment Stop. The last query segment aligned.
15. **RefStartCoord**: Reference Start Coordinate. Refers to the position of the first reference signal in the alignment.
16. **RefStopCoord**: Reference Stop Coordinate. Refers to the position of the last reference signal in the alignment.
17. **AlignedSegRatio**: Ratio of the length of the query aligned in this alignment entry to the length of all but the first and last segments.
18. **Score**: Alignment score. This value reflects the quality of an alignment.
19. **Cigar**: CIGAR String.
20. **Confidence**: Alignment confidence. The value ranges from 0 to 1, reflecting the specificity of the alignment.

21. **FP**: Number of false positives / extra signals in the alignment
22. **FN**: Number of false negatives / missing signals in the alignment
23. **Scale**: The scaling factor of the query with respect to the reference. This is calculated as the length of the aligned region of the query divided by the length of the reference.
24. **FPRate**: The false positive rate, which equals to **FP** divided by the aligned length of the query
25. **FNRate**: The false negative rate, which equals to **FN** divided by the total signals present in the aligned region on the query
26. **simuCorrectlyMapped**: Correctness of the alignment according to the simulation