OMTools Manual v1.4

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

March 10, 2018

Contents

1 Introduction

Optical mapping is a technique to capture specific enzyme sites on a long DNA molecule. The output data of this technology are the *optical map*, that is represented by a *tuple* (Figure 1).

Figure 1: An example of optical map. Green rectangle is the DNA backbone and the black columns are the enzyme site. The optical map contains 6 signals and 7 segments, and can be represented as a tuple: [1518; 15487; 8455; 1350; 25188; 17845; 4948]. The first and last segments are usually neglected in analysis because their sizes are inaccurate.

OMTools is a software package that provides efficient and intuitive data processing and visualization modules to handle optical mapping data.

2 Availability and Implementation

2.1 Availability

OMTools can be obtained from https://github.com/aldenleung/OMTools and released under a GPL license (See the software distribution for details).

2.2 Minimum requirements

OMTools is implemented in Java 1.8. The software has been tested on Ubuntu 14.10 and Microsoft Windows 7, 10.

2.3 Installation

All required libraries are placed in the folder lib/.

1. Compile the OMTools package in the OMTools folder:
   `javac -d bin -sourcepath src -cp "lib/*" @classes`
2. Build a runnable jar file for OMTools:
   `jar cvfm OMTools.jar manifest -C bin .`
3. Run OMTools:
   `java -jar OMTools.jar ModuleName`

3 Quick start

- `java -jar OMTools.jar FastaToOM --fastain Ecoli.fa --refmapout Ecoli.ref --enzyme BspQI`
- `java -jar OMTools.jar DataStatistics --optmapin Ecoli.ref --statout ReferenceStat.txt`
- `java -jar OMTools.jar OptMapDataGenerator --refmapin Ecoli.ref --optmapout EcoliExample.sdata --cov 100`
- `java -jar OMTools.jar DataStatistics --refmapin Ecoli.ref --optmapin EcoliExample.sdata --statout DataStat.txt`
- `java -jar OMTools.jar OMBlastMapper --refmapin Ecoli.ref --optmapin EcoliExample.sdata --optresout EcoliExample.omd --thread 2`
- `java -jar OMTools.jar ResultStatistics --refmapin Ecoli.ref --optmapin EcoliExample.sdata --optresin EcoliExample.omd --statout ResultStat.txt`
- `java -jar OMTools.jar OMView --viewrefin Ecoli.ref --viewresin EcoliExample.omd --viewregion 1:1-1000000`

4 Basic analysis procedures

Prior to the generation of optical mapping data, users are recommended to do simulation and determine the best enzyme to use. If users have sequence assembly files (e.g. short sequence contigs, long sequence scaffolds or existing reference sequences), they can perform an *in silico* digestion. Such conversion from *fasta* to *optical mapping* data can be done by using the **FastaToOM** module. Users obtain two pieces of information. First, by using the **DataStatistics** module, users can generate some basic statistics of the digested sequence. One of the most crucial number to look at is the density of enzyme sites (signal density). Data analysis in optical mapping becomes very difficult if the signal density is too high or too low (The range of "good" signal density depends on the platform used to generate optical mapping data, and users are suggested to consult the providers for more details).

Second, if the platform design is based on labeling with nicking enzyme, nicking site break can be a severe problem in continuity of optical mapping assembly. For each enzyme site, **FastaToOM** module provides the distance to the closest enzyme site. Users can predict the number of nicking site breaks by setting a distance cut-off (Again, the distance between two nicking sites that can lead to a break depends on the platform used to generate optical mapping data, and users are suggested to consult the providers for more details). In addition, if users want to complete genome assemblies using optical mapping data, looking at the location of predicted nicking site breaks at the contigs of interest is important. It is better to avoid nicking site break near the ends of contig.

After determining the enzyme used, users can proceed to the optical mapping experiment. When users receive raw high-throughput optical mapping data, it is recommended to generate statistics on the optical mapping data using the **DataStatistics** module. Usually several criteria are of great interest - data throughput, signal density, and average molecule length. Users can then use **DataTools** module to do data processing and filtering. Sometimes, the output data is duplicated when multiple raw optical mapping files are concatenated due to some unexpected human errors. In such case, users may want to do a quick scanning by using the **DuplicatedMoleculesDetection** module and remove any duplicated molecules by using the **DuplicatedMoleculesRemover** module.

Next, one of the upstream data analysis would be alignment. **OMTools** provides several alignment modules including the **OMBlastMapper**, **OMHAMapper** and **OMFMMapper** modules. A **PairwiseAlignment** module can also help to perform pairwise alignment across multiple data sets. Users can then process and filter the alignment results by using **ResultTools** module and generate

statistics using **ResultStatistics** module. Since several methods are recently published by other research groups that can perform alignment of optical mapping data, users can consider obtaining the union or intersection of alignment results from multiple methods using the **ResultMerger** module.

When it comes to development of new algorithms relating to optical mapping data, users may be interested in the simulation tools. These include **OptMapDataGenerator** module that generates optical mapping data given a reference optical map, and the **RandomReferenceGenerator** module that generates a random reference optical map. If users are developing the alignment algorithm, the **PrecisionRecallGraphDataGenerator** module generates a table that can be used for precision-recall graph generation.

Last but not least, users may want to further investigate the results and showcase some examples by visualizing the optical mapping data. **OMView** module serves as a multi-purpose visualizer on the optical mapping data for it.

Part I
Mapper

5 OMBlastMapper

Performs alignment of optical mapping data. OMBlast algorithm employs a seed-and-extend approach to align optical maps.

5.1 Common Mapper Options

`--minsig` Minimum signal of the query to align. [Default: 5]

`--minsize` Minimum size of the query to align. [Default: 50000]

`--exactmatch` Enable exact match of query to reference. Disable this option when performing self-alignment. [Default: true]

5.1.1 Overlapped Alignment Merging Module Options

`--overlapmergemode` Mode: 0: Disable merging step; 1: Merge same partial alignments; 2: Merge overlapping partial alignments [Default: 2]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--local` Enable local alignment [Default: true]

5.1.2 Result Filter Options

`--filtermode` Filter Mode. 0: No filter; 1: Filter by all the following options; 2: Filter by minimum score only [Default: 0]

`--minmatch` Minumum number of matches of a partial alignment [Default: 3]

`--maxfp` Maximum number of extra signals of a partial alignment [Default: 10000]

`--maxfn` Maximum number missing signals of a partial alignment [Default: 10000]

`--maxfpr` Maximum rate of extra signals of a partial alignment [Default: 1.0E-4]

`--maxfnr` Maximum rate of missing signals of a partial alignment [Default: 0.5]

`--minscore` Minimum score of a partial alignment [Default: 0.0]

`--minsubfragratio` Minimum subfragment ratio of a partial alignment [Default: 0.0]

`--minsigratio` Minimum aligned signal ratio of a partial alignment [Default: 0.0]

`--trimmode` Trim Mode. 0: Trim mode disabled; 1: Trim mode enabled [Default: 0]

`--maxtrim` Maximum number of trimming steps of a partial alignment [Default: 5]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

5.1.3 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 30]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.4]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

5.1.4 Result Reader Options

`--optresin` Input alignment result file for re-alignment

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

5.2 OMBlastMapper Options

`--local` Enable local alignment [Default: true]

`--allowequalrefquery` Allow equal reference and query in alignment. Disabling this options prohibits a query from aligning to itself as reference [Default: true]

`--allowdiffrefquery` Allow different reference and query in alignment. In contrast to allowequalrefquery, disabling this options prohibits a query from aligning to other queries as reference [Default: true]

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--falselimit` Maximum number of consecutive extra/missing signals [Default: 5]

`--maxseedno` Maximum similar seed number on query [Default: 10]

5.2.1 Seeding Options

`--seedingmode` Seeding mode: 1: Optimized for long k-mer (usually for k larger than 10); 2: Optimized for short k-mer (usually for k smaller than or equal to 10); -1: Auto-selection. [Default: -1]

`--k` Kmer length. [Default: 3]

`--maxnosignal` Maximum no signal region between signals for seeding. [Default: 10000000]

5.3 Multi-thread Options

`--thread` Number of threads [Default: 1]

5.4 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

5.5 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

5.6 Result Writer Options

`--optresout` Output alignment result file [Required]

`--optresoutformat` Result file format -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

`--writeunmap` Write discarded or unmapped molecules. [Default: true]

`--multiple` Write multiple maps for a molecule. [Default: true]

`--writeinfo` Write information of a molecule. [Default: true]

6 OMHAMapper

Performs alignment of optical mapping data. OMHA algorithm employs a heuristic approach to align optical maps.

6.1 Common Mapper Options

`--minsig` Minimum signal of the query to align. [Default: 5]

`--minsize` Minimum size of the query to align. [Default: 50000]

`--exactmatch` Enable exact match of query to reference. Disable this option when performing self-alignment. [Default: true]

6.1.1 Overlapped Alignment Merging Module Options

`--overlapmergemode` Mode: 0: Disable merging step; 1: Merge same partial alignments; 2: Merge overlapping partial alignments [Default: 2]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--local` Enable local alignment [Default: true]

6.1.2 Result Filter Options

`--filtermode` Filter Mode. 0: No filter; 1: Filter by all the following options; 2: Filter by minimum score only [Default: 0]

`--minmatch` Minumum number of matches of a partial alignment [Default: 3]

`--maxfp` Maximum number of extra signals of a partial alignment [Default: 10000]

`--maxfn` Maximum number missing signals of a partial alignment [Default: 10000]

`--maxfpr` Maximum rate of extra signals of a partial alignment [Default: 1.0E-4]

`--maxfnr` Maximum rate of missing signals of a partial alignment [Default: 0.5]

`--minscore` Minimum score of a partial alignment [Default: 0.0]

`--minsubfragratio` Minimum subfragment ratio of a partial alignment [Default: 0.0]

`--minsigratio` Minimum aligned signal ratio of a partial alignment [Default: 0.0]

`--trimmode` Trim Mode. 0: Trim mode disabled; 1: Trim mode enabled [Default: 0]

`--maxtrim` Maximum number of trimming steps of a partial alignment [Default: 5]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

6.1.3 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 30]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.4]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

6.1.4 Result Reader Options

`--optresin` Input alignment result file for re-alignment

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

6.2 OMHAMapper Options

`--local` Enable local alignment [Default: true]

`--localstart` Local start pos for alignment, 0: starts at every signal (exhausive), x: starts at first x signals, -x: starts without last x signals. [Default: 0]

`--scorefilter` Primary score filter during alginment [Default: 30]

`--deg` Degeneracy of close signals to handle resolution error. [Default: 1500]

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--falselimit` Max consecutive false signals [Default: 5]

6.3 Multi-thread Options

`--thread` Number of threads [Default: 1]

6.4 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

6.5 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

6.6 Result Writer Options

`--optresout` Output alignment result file [Required]

`--optresoutformat` Result file format -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

`--writeunmap` Write discarded or unmapped molecules. [Default: true]

`--multiple` Write multiple maps for a molecule. [Default: true]

`--writeinfo` Write information of a molecule. [Default: true]

7 OMFMMapper

Performs alignment of optical mapping data. OMFM algorithm employs an indexing approach to align optical maps.

7.1 Common Mapper Options

`--minsig` Minimum signal of the query to align. [Default: 5]

`--minsize` Minimum size of the query to align. [Default: 50000]

`--exactmatch` Enable exact match of query to reference. Disable this option when performing self-alignment. [Default: true]

7.1.1 Overlapped Alignment Merging Module Options

`--overlapmergemode` Mode: 0: Disable merging step; 1: Merge same partial alignments; 2: Merge overlapping partial alignments [Default: 2]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--local` Enable local alignment [Default: true]

7.1.2 Result Filter Options

`--filtermode` Filter Mode. 0: No filter; 1: Filter by all the following options; 2: Filter by minimum score only [Default: 0]

`--minmatch` Minumum number of matches of a partial alignment [Default: 3]

`--maxfp` Maximum number of extra signals of a partial alignment [Default: 10000]

`--maxfn` Maximum number missing signals of a partial alignment [Default: 10000]

`--maxfpr` Maximum rate of extra signals of a partial alignment [Default: 1.0E-4]

`--maxfnr` Maximum rate of missing signals of a partial alignment [Default: 0.5]

`--minscore` Minimum score of a partial alignment [Default: 0.0]

`--minsubfragratio` Minimum subfragment ratio of a partial alignment [Default: 0.0]

`--minsigratio` Minimum aligned signal ratio of a partial alignment [Default: 0.0]

`--trimmode` Trim Mode. 0: Trim mode disabled; 1: Trim mode enabled [Default: 0]

`--maxtrim` Maximum number of trimming steps of a partial alignment [Default: 5]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

7.1.3 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 30]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.4]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

7.1.4 Result Reader Options

`--optresin` Input alignment result file for re-alignment

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

7.2 OMFMMapper Options

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--rfalselimit` Max consecutive false signals on reference [Default: 5]

`--qfalselimit` Max consecutive false signals on query [Default: 5]

`--cfalselimit` Max consecutive false signals on both reference and query [Default: 5]

`--minalignscore` Minimum score at alignment stage [Default: 20]

7.2.1 Seeding Options

`--seedingmode` Seeding mode: 1: Optimized for long k-mer (usually for k larger than 10); 2: Optimized for short k-mer (usually for k smaller than or equal to 10); -1: Auto-selection. [Default: -1]

`--k` Kmer length. [Default: 3]

`--maxnosignal` Maximum no signal region between signals for seeding. [Default: 10000000]

7.3 Multi-thread Options

`--thread` Number of threads [Default: 1]

7.4 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

7.5 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

7.6 Result Writer Options

`--optresout` Output alignment result file [Required]

`--optresoutformat` Result file format -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

`--writeunmap` Write discarded or unmapped molecules. [Default: true]

`--multiple` Write multiple maps for a molecule. [Default: true]

`--writeinfo` Write information of a molecule. [Default: true]

8 PairwiseAlignment

Performs pairwise alignment of data files based on OMBlastMapper. Input multiple data files for pair-wise alignment between each pair of them.

8.1 Common Mapper Options

`--minsig` Minimum signal of the query to align. [Default: 5]

`--minsize` Minimum size of the query to align. [Default: 50000]

`--exactmatch` Enable exact match of query to reference. Disable this option when performing self-alignment. [Default: true]

8.1.1 Overlapped Alignment Merging Module Options

`--overlapmergemode` Mode: 0: Disable merging step; 1: Merge same partial alignments; 2: Merge overlapping partial alignments [Default: 2]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--local` Enable local alignment [Default: true]

8.1.2 Result Filter Options

`--filtermode` Filter Mode. 0: No filter; 1: Filter by all the following options; 2: Filter by minimum score only [Default: 0]

`--minmatch` Minumum number of matches of a partial alignment [Default: 3]

`--maxfp` Maximum number of extra signals of a partial alignment [Default: 10000]

`--maxfn` Maximum number missing signals of a partial alignment [Default: 10000]

`--maxfpr` Maximum rate of extra signals of a partial alignment [Default: 1.0E-4]

`--maxfnr` Maximum rate of missing signals of a partial alignment [Default: 0.5]

`--minscore` Minimum score of a partial alignment [Default: 0.0]

`--minsubfragratio` Minimum subfragment ratio of a partial alignment [Default: 0.0]

`--minsigratio` Minimum aligned signal ratio of a partial alignment [Default: 0.0]

`--trimmode` Trim Mode. 0: Trim mode disabled; 1: Trim mode enabled [Default: 0]

`--maxtrim` Maximum number of trimming steps of a partial alignment [Default: 5]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

8.1.3 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 30]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.4]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

8.1.4 Result Reader Options

`--optresin` Input alignment result file for re-alignment

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

8.2 OMBlastMapper Options

`--local` Enable local alignment [Default: true]

`--allowequalrefquery` Allow equal reference and query in alignment. Disabling this options prohibits a query from aligning to itself as reference [Default: true]

`--allowdiffrefquery` Allow different reference and query in alignment. In contrast to allowequalrefquery, disabling this options prohibits a query from aligning to other queries as reference [Default: true]

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--falselimit` Maximum number of consecutive extra/missing signals [Default: 5]

`--maxseedno` Maximum similar seed number on query [Default: 10]

8.2.1 Seeding Options

`--seedingmode` Seeding mode: 1: Optimized for long k-mer (usually for k larger than 10); 2: Optimized for short k-mer (usually for k smaller than or equal to 10); -1: Auto-selection. [Default: -1]

`--k` Kmer length. [Default: 3]

`--maxnosignal` Maximum no signal region between signals for seeding. [Default: 10000000]

8.3 Multi-thread Options

`--thread` Number of threads [Default: 1]

8.4 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

8.5 Pairwise alignment options

`--output` output prefix [Required]

`--rerun` Rerun even if the result file exists [Default: false]

Part II
Simulation

9 OptMapDataGenerator

Generates simulated data from the reference.

9.1 Reference Reader Options

`--refmapin` Input reference map file

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

9.2 Multiple Reference Reader Options

`--refmaplistin` Input reference map file list with ratio

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

9.3 Data Generator Options

`--rsln` Resolution error [Default: 1200]

`--meas` Measurement error [Default: 500]

`--fsize` Average fragment size [Default: 200000]

`--fubound` Size upper boundary, inclusive [Default: 1000000]

`--flbound` Size lower boundary, inclusive [Default: 100000]

`--median` Median for scale [Default: 1.0]

`--scalesd` SD for scale [Default: 0.04]

`--subound` Scale upper boundary, inclusive [Default: 1.3]

`--slbound` Scale lower boundary, inclusive [Default: 0.7]

`--fpr` false positive rate [Default: 1.0E-5]

`--fnr` false negative rate [Default: 0.1]

`--seed` Random seed

`--indelsize` Random Insertion/Deletion size [Default: 0]

`--inversionmode` Inversion mode. 0: no inversion. 1: inversion of second half [Default: 0]

`--cov` Coverage of data output [Default: 10.0]

`--moleno` Number of molecules to be generated. Overriding coverage option if set to a positive number [Default: -1]

9.4 Data Writer Options

`--optmapout` Output optical map file [Required]

`--optmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

10 RandomReferenceGenerator

Generates random reference maps by shuffling the order of segments in the input reference maps.

10.1 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

10.2 Reference Writer Options

`--refmapout` Output reference map file [Required]

`--refmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

Part III
SV Detection

11 SVDetection

Provides a basic SV detection module for from optical mapping alignment

11.1 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

11.2 Data Reader Options

`--optmapin` Input optical map file

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

11.3 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

11.4 Standard SV Writer Options

`--svout` Output SV file

11.5 SV Detection

`--svmode` SV Mode. 1: split-map approach (experimental); 2: signal-based approach; 3: both split-map and signal based approach [Default: 3]

`--minsupport` Minimum molecule support for an SV [Default: 5]

`--flanksig` Minimum number of flanking signals for an alignment [Default: 5]

11.5.1 Split-map based detection options

`--closeref` The max distance (reference) between two results to be considered at same cluster. [Default: 250000]

`--closefrag` The max distance (fragment) between two results to be considered at same cluster. [Default: 250000]

`--closesv` Close SV (SVs are joined based on bp1bp2 comparison. [Default: 10000]

`--mergesv` Merge SV (SVs are joined based on simple region comparison). [Default: true]

`--flanksize` Minimum flanking size in alignment (Used in split-map detection only) [Default: 0]

`--minindelsize` Minimum indel size for an SV [Default: 1500]

`--maxindelsize` Maximum indel size for an SV [Default: 1000000]

`--maxsupport` Maximum molecule support for an SV [Default: 100]

`--minsvscore` Minimum sv score (ratio of support/opposition [Default: 0.5]

11.5.2 Signal-based detection options

`--mininvsig` Minimum signal involved in an inversion [Default: 4]

`--maxinvsize` Maximum inversion size [Default: 100000]

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--deg` Degeneracy of close signals to handle resolution error. [Default: 1500]

Part IV
Fasta Tools

12 FastaToOM

Performs *in silico* digestion on DNA sequence.

12.1 Stream Fasta Reader Options

`--fastain` fasta input file [Required]

12.2 Enzyme Input Options

`--enzyme` Built-in enzymes [BspQI, BbvCI, AlwI, BsmAI, BstNBI, BsmI, BsrDI, BssSI, BtsI] (Support multiple enzymes input)

`--enzymestring` Enzyme sequence (e.g. GCTCTTC)

12.3 Reference Writer Options

`--refmapout` Output reference map file [Required]

`--refmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

12.4 Nicking Site Break Prediction Options

`--nsbout` Potential nicking site breaks output (The prediction is useful for nicking enzyme-based data only)

Part V
Data Tools

13 DataTools

Provides basic functions for filtering and processing optical mapping data

13.1 Data Tools Options

`--idprefix` Add a prefix to all ids

`--idmodify` Convert all ids to x .. x + n - 1 (x: input value, n: number of optical maps in the data file). A negative value disables this function. [Default: -1]

`--idmodifylog` Log file containing the id conversions

`--staticid` Use a static id for all optical maps (Override all other id-related parameters

`--fix` Fix the data (negative signal-to-signal distance correction and etc.) [Default: true]

`--condense` Merge multiple signals closer than parameter into one signal [Default: 0]

`--removeseg` Remove segments smaller than the parameter [Default: -1]

`--minsize` Data with minimum size to retain [Default: 0]

`--minsig` Data with minimum signal to retain [Default: 0]

`--dataid` List of Data ID to be extracted

`--region` List of regions to be extracted.

`--shift` Shift forward (right) x bp (Assume circular) [Default: 0]

`--randdata` Number of random data to be extracted

`--seed` Seed used in random data extraction

`--concat` Concatenate all data entries into single entry. -1: not activated; Non-negative value: space (segment without any signal) between each data entry. Ignore any data modification functions [Default: -1]

13.1.1 ConcatInfo Reader Options

`--concatin` ConcatInfo file input.

13.1.2 ConcatInfo Writer Options

`--concatout` ConcatInfo file output.

13.1.3 Low complexity filtering

`--lowcom` Retain/Remove molecules with low complexity -1: Retain Low Complexity; 0: Do nothing; 1: Retain High Complexity [Default: 0]

`--maxdensity` Maximum density per 100kbp to filter [Default: 25.0]

`--maxseed` Maximum seed to filter [Default: 5]

13.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

13.3 Data Writer Options

`--optmapout` Output optical map file [Required]

`--optmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

14 DataQualityCheck

Performs basic data quality check, including distributions of molecule length, signal density and segment length

14.1 Reference Reader Options

`--refmapin` Input reference map file

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

14.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

14.3 Data Quality Check Options

`--name` Use data set name instead of file name.

`--prefix` Statistics output prefix [Default: output]

`--gradcolor` Use gradient color [Default: false]

`--imageformat` Formats of image to be saved. [svg; png; jpg;] [Default: png]

15 DataStatistics

Generates statistics of the data file.

15.1 DataStatistics Options

`--winsize` Window size for molecule size [Default: 25000]

`--maxsize` Maximum size. -1 for auto setup [Default: -1]

`--maxsignal` Maximum signal. -1 for auto setup [Default: -1]

`--statout` Statistics output [Required]

15.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

15.3 Reference Reader Options

`--refmapin` Input reference map file

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

16 DuplicatedMoleculesDetection

Detects duplicated molecules in an optical map data set. Duplicated molecules contain same number of total segment, and the difference between size of each segment is very small (usually smaller than 100 bp)

16.1 Multithread Seeding Options

`--meas` Measuring Errors. Usually it is much smaller than normal measuring errors in discovering duplicated molecules [Default: 100]

`--ear` Error acceptable range [Default: 0.0]

`--thread` Number of threads [Default: 1]

16.2 Seeding Options

`--seedingmode` Seeding mode: 1: Optimized for long k-mer (usually for k larger than 10); 2: Optimized for short k-mer (usually for k smaller than or equal to 10); -1: Auto-selection. [Default: -1]

`--k` Kmer length. [Default: 3]

`--maxnosignal` Maximum no signal region between signals for seeding. [Default: 10000000]

16.3 Duplicated Molecules Detecting Options

`--dupout` Files containing duplicated molecules [Required]

`--minseg` Minimum segments to be considered duplicated [Default: 15]

16.4 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

17 DuplicatedMoleculesRemover

Removes detected duplicated molecules from the data file

17.1 Duplicated Molecules Remover Options

`--dupin` Files containing duplicated molecules [Required]

17.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

17.3 Data Writer Options

`--optmapout` Output optical map file [Required]

`--optmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

18 FrequentKmerHighlight

Returns segment identifiers in kmers with more than minimum number of hit at the same query.

18.1 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

18.2 Segment Identifier Writer Options

`--siout` Output segment identifier file

18.3 Seeding Options

`--seedingmode` Seeding mode: 1: Optimized for long k-mer (usually for k larger than 10); 2: Optimized for short k-mer (usually for k smaller than or equal to 10); -1: Auto-selection. [Default: -1]

`--k` Kmer length. [Default: 3]

`--maxnosignal` Maximum no signal region between signals for seeding. [Default: 10000000]

Part VI
Alignment Tools

19 ResultTools

Provides basic functions for filtering and processing alignment results

19.1 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

19.2 Result Writer Options

`--optresout` Output alignment result file

`--optresoutformat` Result file format -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

`--writeunmap` Write discarded or unmapped molecules. [Default: true]

`--multiple` Write multiple maps for a molecule. [Default: true]

`--writeinfo` Write information of a molecule. [Default: true]

19.3 Reference Reader Options

`--refmapin` Input reference map file

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

19.4 Data Reader Options

`--optmapin` Input optical map file

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

19.5 Data Output Options

`--mapout` Mapped molecules output

`--unmapout` Unmapped molecules output

`--optmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

19.6 Result Tools Options

`--qprefix` Add prefix to query name.

`--rprefix` Add prefix to reference name.

`--disinvalid` Discard invalid results. [Default: true]

`--conf` Recalculating result confidence [Default: false]

`--dataid` List of Data ID to be extracted

`--region` Region in chrN:start-end or chrN:start format

`--refnamemodify` Modify the reference name according to the target file in format: src\tTarget

`--dataremoval` Remove result with query names in the file

`--joinresult` Represent partial alignments in one alignment (The gap is filled with extra and missing signals). Only works on partial alignments with indel relationship [Default: false]

19.6.1 Results Breaker Option

`--breakermode` Mode 0: Disable the breaking function; 1: Break the alignment at query/reference segment with size deviating too much, into multiple partial alignments [Default: 0]

`--meas` Measurement error [Default: 500]

`--ear` Error acceptable range (Scaling error tolerance) [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

19.6.2 Result Filter Options

`--filtermode` Filter Mode. 0: No filter; 1: Filter by all the following options; 2: Filter by minimum score only [Default: 0]

`--minmatch` Minumum number of matches of a partial alignment [Default: 3]

`--maxfp` Maximum number of extra signals of a partial alignment [Default: 10000]

`--maxfn` Maximum number missing signals of a partial alignment [Default: 10000]

`--maxfpr` Maximum rate of extra signals of a partial alignment [Default: 1.0E-4]

`--maxfnr` Maximum rate of missing signals of a partial alignment [Default: 0.5]

`--minscore` Minimum score of a partial alignment [Default: 0.0]

`--minsubfragratio` Minimum subfragment ratio of a partial alignment [Default: 0.0]

`--minsigratio` Minimum aligned signal ratio of a partial alignment [Default: 0.0]

`--trimmode` Trim Mode. 0: Trim mode disabled; 1: Trim mode enabled [Default: 0]

`--maxtrim` Maximum number of trimming steps of a partial alignment [Default: 5]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

19.6.3 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 0]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.0]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

19.6.4 Lift Over Options

`--liftoverin` Input liftOver file, Format: chromosome\t coordinate\t size\n

19.7 Simulated Results Analysis Options

`--rocout` Output a table for ROC curve plotting

20 ResultMerger

Merges alignment results from different alignment methods

20.1 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

20.2 Result Merger Options

`--resultkey` Keys (names) to represent the result files [Required]

`--gapallowed` Gaps allowed between results [Default: 0]

`--analyzeall` Analyze only if the query is present in all results [Default: false]

`--prefix` Output file prefix [Required]

`--outtype` Output file type [Default: .omd]

21 ResultStatistics

Generates statistics for alignment results

21.1 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

21.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

21.3 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

21.4 Result Stat Options

`--checkstrand` Checking strand for correctness [Default: true]

`--covout` Coverage output (Under construction)

`--statout` Statistics output

22 PrecisionRecallGraphDataGenerator

Generates a data table for precision recall graphs. This module assumes one alignment (it can contain multiple partial alignments) per one query. You need to use the same alignment joining module parameters if the alignment file is generated by OMTools mapper. If you are using other alignment tools, set alignmentjoinmode as 0.

22.1 Reference Reader Options

`--refmapin` Input reference map file [Required]

`--refmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

22.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

22.3 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

22.4 Alignment Joining Options

`--alignmentjoinmode` Mode. 0: No joining. 1: Standard indel joining. 2: Standard indel-inv joining. 3. Standard transloc joining [Default: 0]

`--closeref` The maximum distance (reference) between two partial alignments to be joined [Default: 250000]

`--closefrag` The maximum distance (query) between two partial alignments to be joined [Default: 250000]

`--minmatch` Minimum matching signals to be considered as a valid partial alginment. [Default: 3]

`--maxtrim` Maximum trimming steps for a partial alignment [Default: 5]

`--trimear` Scaling error tolerance during trimming [Default: 0.1]

`--match` Score for matching signal [Default: 5]

`--fpp` Penalty for extra signal [Default: 2]

`--fnp` Penalty for missing signal [Default: 2]

`--indelp` Penalty for joining partial alignments with indel relationship [Default: 10]

`--invp` Penalty for joining partial alignments with inversion relationship [Default: 30]

`--transp` Penalty for joining partial alignments with translocation relationship [Default: 50]

`--localpenalty` Enable local-alignment penalty for the final alignment (treated as global alignment) [Default: false]

`--minjoinscore` Minimum score of the joined final alignment [Default: 30]

`--minconf` Minimum confidence (uniqueness) of the final alignment [Default: 0.4]

`--minjoinedfragratio` Minimum ratio of aligned length against query length [Default: -1.0]

`--minjoinedsigratio` Minimum ratio of number of aligned signals against total number of query signals [Default: -1.0]

`--overlapalign` Allow overlapping final alignments to be output [Default: true]

`--maxalignitem` Maximum number of final alignments output. -1: No limit on the number of final alignments [Default: 1]

22.5 Precision Recall Graph Options

`--prgout` Precision recall graph table output [Required]

`--checkstrand` Checking strand for correctness [Default: true]

`--sortstrat` Sort by "score" or "confidence" [Default: score]

23 QueryReverse

Reverses the query according to the alignment results so that they have the same strand as the reference. Only one reference and one strand is allowed in the alignment results for each query (or an exception will be thrown). Inversions are not supported.

23.1 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

23.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

23.3 Data Writer Options

`--optmapout` Output optical map file [Required]

`--optmapoutformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

24 AlignmentHighlight

Highlights segments used in alignments as segment identifiers.

24.1 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

24.2 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

24.3 Segment Identifier Writer Options

`--siout` Output segment identifier file

24.4 BED Writer Options

`--bedout` Output BED file

24.5 Alignment Highlight Options

`--mode` Highlight segments in - 1: Reference only; 2: Query only; 3: Both query and reference [Default: 3]

`--overlap` Require overlapping query and reference (with at least one signal) [Default: false]

`--strict` Highlight segments only in consecutive signal match (without extra or missing signals in-between) [Default: false]

Part VII
Multiple Alignment

25 MultipleAlignment

Performs multiple alignment taking multiple optical maps as queries

25.1 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

25.2 Multiple Alignment Options

`--maconfig` Multiple alignment configuration file

`--minlinksize` Minimum segment size in linking process [Default: 1000]

`--maref` References for multiple alignment

`--maorderby` 1: First appearance of block; 2: similarity clustering [Default: 1]

25.3 Collinear Block Writer Options

`--cblout` Multiple alignment collinear blocks output.

25.4 Collinear block order writer

`--cboout` Multiple alignment collinear blocks order output

25.5 Multiple alignment configuration

The whole multiple alignment process can be broken down into three major sub-modules - link, build and merge/mergeproximity.

The first sub-module "link" processes the input alignment, clustering and multiple alignment files to create potential "segment links" among segments from different queries. These links are essential to the grouping strategies of segments into collinear blocks. Links created for this sub-module should be highly confident because they determine the global structure of the multiple alignment.

The second sub-module "build" compiles the segment links created from previous sub-module into collinear blocks.

The third sub-module "merge" and "mergeproximity" combines similar collinear blocks into a single collinear block without disrupting the global structure of multiple alignment. Since conservative strategy is usually taken to trade sensitivity for higher specificity in the "link" step, some segments that belong to the same collinear block may be put in several collinear blocks. The merging is based on (1) less confident links ("merge") or (2) length of proximal blocks ("mergeproximity").

An alternative sub-module "mask" labels and prevents segments that lead to non-specific multiple alignment from linking and merging in above sub-modules. This module is only used when the queries consist of many repetitive signaling patterns or severe segmental duplications that disrupt the multiple alignment process.

25.5.1 Multiple alignment configuration file

In the configuration file, each line starts with a command corresponding to a module in multiple alignment procedures, followed by parameters and input files.

mask (Optional) *mask [Input file]*
Prohibit the segments used in the alignment file from being processed in the "link" and "merge" steps.

link *link [Input file] [k=?(clustering);maxnosignal=?(clustering)]*
Create segment links from alignment, clustering and multiple alignment files. Two parameters k and *maxnosignal* are required for clustering file input.

build *build [allowrearrangement=true]*
Compile these segment links into collinear blocks.

mergeproximity (Optional) *mergeproximity* [meas=?(measurement error);ear=?(scaling error)]
Merge blocks in proximity if their block lengths are sufficiently close subjected to the level of errors tolerated.

mergeproximitysimple (Optional) *mergeproximitysimple* [meas=?(measurement error);ear=?(scaling error)]
Merge blocks in proximity if their block lengths are sufficiently close subjected to the level of errors tolerated. This command employs a simplified version of mergeproximity for reduced processing time and memory but only merge proximate blocks with same and unique parent block.

merge (Optional) *merge [Input file] [mergechain=true;k=?(clustering);maxnosignal=?(clustering)]*
Merge collinear blocks using segment links. Input files and parameters are similar to those in *link* command.

A configuration file must contain at least one "link" command followed by one "build" command. The "mask", "mergeproximity", "mergeproximitysimple" and "merge" commands are optional.

Example configuration file

mask SegmentMaskFile.si

link AlignmentFile1.oma AlignmentFile2.oma

link ClusteringFile1.kc k=7;maxnosignal=1000000

link MultipleAlignmentFile1.cbl

build allowrearrangement=true

mergeproximitysimple meas=500;ear=0.1

mergeproximity meas=500;ear=0.1

merge AlignmentFile3.oma mergechain=true

25.5.2 Optimization strategy for configuration

Since the parameter optimization is largely dependent on the nature of queries, it is difficult to have one single configuration that suits all cases. In this section, the optimization strategy for configuration is generalized and introduced. Impatient users could skip the details here and read sections below for configuration templates on specific cases.

The major optimization efforts are on what files should be provided in "link" and "merge" steps, where users have to decide what to provide as clues for segment links. Other steps are less complicated and do not require much optimization.

mask Segments that result in non-specific alignment but with high score should be masked because they lead to incorrect segment links for initial collinear block construction and disrupt the multiple alignment structure. Usually these segments come from repetitive signaling patterns or segmental duplications. Users can locate these segments by self-alignment of each query:

Mask

OMBlastMapper –refmapin query.data –optmapin query.data –alignmentjoinmode 0 –filtermode 1 –minconf 0 –allowdiffrefquery false –exactmatch false –maxalignitem -1 –minjoinscore 50 –fpp 10 –fnp 10
AlignmentHighlight –mode 3 –optmapin query.data –optresin queryselfalign.oma –siout query-mask.si

Users could decrease the parameters minjoinscore, fpp and fnp for less stringent alignment so that more potential segments involved in segmental duplications are excluded. This step is not necessary if segmental duplication is absent in the queries.

link The "link" step is the most crucial part for optimization. The segment links should be built using highly confident evidence because every segment link determines the final global multiple alignment structure. Users should be conservative and use stringent parameters for alignment. The main objective of optimization is to link as many queries as possible while keep the structure correct.

The optimization can be simplified as the process of selecting the right files. First, users can import the alignment files from high to low confidence. For ambiguous overall multiple alignment structure users should import less alignment files with low confidence. For multiple alignment with a lot of similar patterns not combined into same collinear blocks, instead, users should import more alignment files with low confidence.

Multiple alignment file can also be used for linking. This usually occurs in multiple alignment of large eukaryotic genomes where queries were partially aligned in sliding windows.

Note that during the optimization of the "link" step, users may disable "merge" and "mergeproximity" command.

build Users do not need to optimize the "build" step.

merge Based on the segment links, the "merge" step only combines two collinear blocks if the merge does not disrupt the global multiple alignment structure (i.e. create rearrangement). The optimization process of "merge" is similar to that of "link", except the files for segment links can be of lower confidence in general. This step is highly NOT recommended to use this in a region with copy number variations. Notice that this step is very time- and memory-consuming and is not applicable to eukaryotic genomes.

mergeproximity Users need to consider only one factor in optimizing "mergeproximity" step - error tolerance (measurement error and scaling error). Basically, higher error tolerance should be applied on multiple alignment of raw optical mapping molecules. On the contrary, a lower error tolerance can be used in multiple alignment of assembled optical mapping contigs or *in-silico* digested sequences. Users will find "mergeproximity" step useful in almost all circumstances. The step is time- and memory-consuming and a simplified but less powerful version "mergeproximitysimple" can be used in larger eukaryotic genomes.

mergeproximitysimple The "mergeproximitysimple" is a simplified version of "mergeproximity". It restricts to merge proximate blocks only when they have one single parent block.

25.5.3 Configuration Template

Prokaryotic genomes

Alignment

OMBlastMapper –refmapin query.data –optmapin query.data –optresout query1.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 15 –fnp 15 –ear 0.05 –meas 500 –minjoinscore 50
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query2.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 10 –fnp 10 –ear 0.05 –meas 500 –minjoinscore 50
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query3.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 5 –fnp 5 –ear 0.05 –meas 500 –minjoinscore 50
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query4.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 50 –fnp 50 –ear 0.05 –meas 500 –minjoinscore 30
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query5.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 50 –fnp 50 –ear 0.05 –meas 500 –minjoinscore 20
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query6.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 2 –fnp 2 –ear 0.05 –meas 500 –minjoinscore 50

Multiple alignment configuration file

link query1.oma
link query2.oma
link query3.oma
build allowrearrangement=true
mergeproximity meas=500;ear=0.05
merge query1.oma
merge query2.oma
merge query3.oma
merge query4.oma
merge query5.oma
merge query6.oma
mergeproximity meas=500;ear=0.05

Copy number variations

Alignment

OMBlastMapper –refmapin query.data –optmapin query.data –optresout query1.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 10 –fnp 10 –ear 0.05 –meas 500 –minjoinscore 100
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query1.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 5 –fnp 5 –ear 0.05 –meas 500 –minjoinscore 50
OMBlastMapper –refmapin query.data –optmapin query.data –optresout query1.oma –filtermode 1 –alignmentjoinmode 0 –maxalignitem -1 –minconf 0 –thread 2 –allowequalrefquery false –exactmatch false –writeinfo false –writeunmap false –fpp 50 –fnp 50 –ear 0.05 –meas 500 –minjoinscore 30

Multiple alignment configuration

link query1.oma
link query2.oma
link query3.oma
build allowrearrangement=true
mergeproximity meas=500;ear=0.05

Part VIII
Multiple Alignment Tools

26 CBLTools

Provides basic functions for filtering and processing multiple alignment results

26.1 Data Reader Options

`--optmapin` Input optical map file

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

26.2 Collinear Block Reader Options

`--cblin` Multiple alignment collinear blocks input.

26.3 Collinear Block Writer Options

`--cblout` Multiple alignment collinear blocks output.

26.4 CBL Tools Options

`--cboout` Collinear block order output

26.4.1 Filtering Options

`--allblock` Extract queries that contains all blocks in this list

`--reservequery` Reserve the queries even they do not match other criteria

26.4.2 Sorting Options

`--sortblock` Sort the queries by block existence

`--sortdesflanklen` Sort the queries by length between two flanking blocks (FB1a FB1b) in descending order

27 MultipleAlignmentPerformanceAnalysis

Analyzes performance of multiple OM alignment based on the multiple sequence alignment.

27.1 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

27.2 MAF Reader

`--mafin` Multiple alignment format input.

27.3 Collinear Block Reader Options

`--cblin` Multiple alignment collinear blocks input.

27.4 multiple alignment options

`--statout` Statistics output

28 BlockConnectionGraphGeneration

Generates a dot graph file representing the progression of collinear blocks.

28.1 Collinear Block Reader Options

`--cblin` Multiple alignment collinear blocks input.

28.2 Block connection graph generation options

`--maref` References for multiple alignment

`--reverse` Reverse the direction of progression of collinear blocks [Default: false]

`--minedgeweight` Filter with minimum edge weight [Default: 0]

`--displayall` Output all blocks (Set false to stop the program from outputing blocks without any linkage to other block) [Default: true]

`--dotout` Dot file output [Required]

Part IX
Phylogenetics

29 UPGMATreeConstruction

Reconstructs phylogenetic tree based on multiple alignment results using the UPGMA approach

29.1 UPGMATreeConstruction options

`--matrixout` Output the distance matrix [Required]

`--treeout` Output trees in newick format [Required]

`--startblock` Start block (Experimental parameters)

`--stopblock` Stop block (Experimental parameters)

29.2 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

29.3 Collinear Block Reader Options

`--cblin` Multiple alignment collinear blocks input.

Part X
Visualization

30 OMView

Visualizes optical mapping data. OMView provides a GUI to visualize optical mapping data for different purposes.

30.1 Data Loading

`--viewrefin` Load references

`--viewmapin` Load molecules

`--viewresin` Load alignment results

`--viewcblin` Load collinear blocks

`--viewcboin` Load collinear blocks (order)

`--viewcbcin` Load collinear blocks (color)

`--viewannoin` Load annotations

`--viewin` Automatic file input

30.2 View Opening

`--viewregion` Show a specific region on a regional view

`--viewanchor` Show a specific anchor on an anchor view

`--viewanchorregion` Specify the region for the anchor on an anchor view

`--viewalignment` Show a specific alignment

`--viewma` Automatically open multiple alignment view [Default: false]

`--viewmab` Automatically open multiple alignment block view [Default: false]

`--viewmolecule` Automatically open molecule view [Default: false]

`--viewsave` Save views to specific location instead of starting OMView

`--viewsaveformat` Formats of image to be saved. [svg; png; jpg;] [Default: png]

30.3 View Settings

`--dnaratio` Default DNA ratio [Default: 400.0]

`--zoom` Default zoom level [Default: 1.0]

`--viewbreakresult` Enable Result Breaker [Default: false]

`--viewunmap` Enable Unmapped Portion [Default: false]

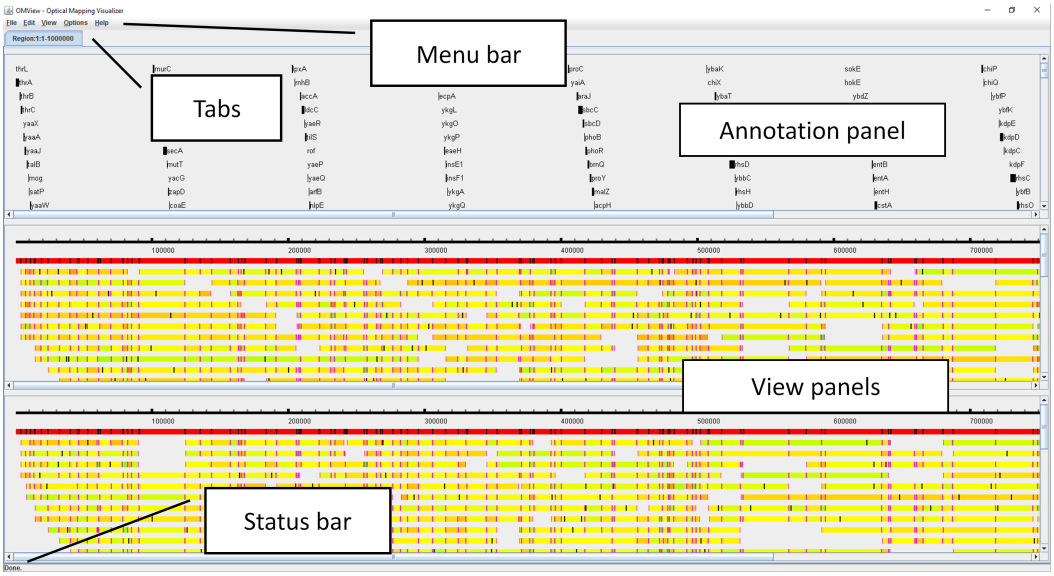
`--viewsettingin` The OMView setting file input

30.4 Help

`--help` Display help menu

30.5 Visualization Procedures

30.5.1 Layout of OMView

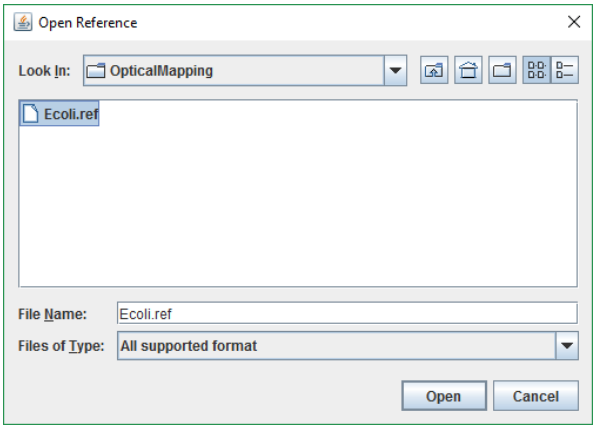


30.5.2 Load required data

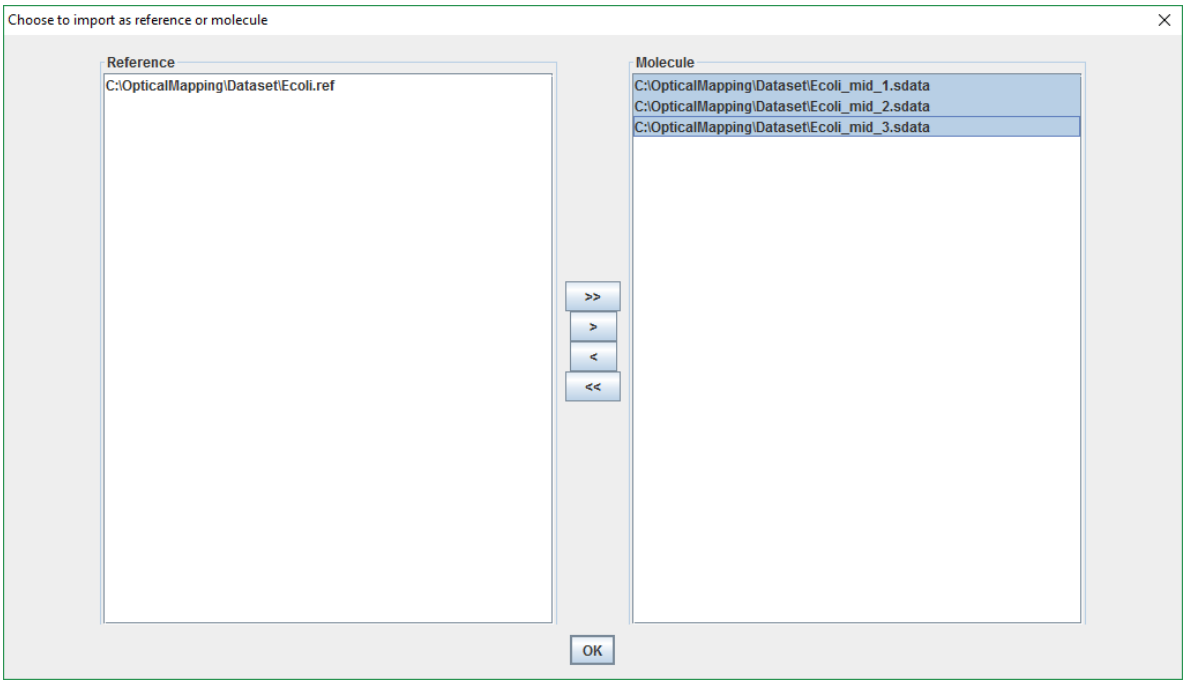
There are two ways to load data into OMView: (1) Load datasets from the menu option and (2) Drag and drop data sets into the program.

Note: Users must load the reference and molecule files first before loading other files (Alignment, annotations and multiple alignments files)

Select data files from menu After choosing the data file to loaded (**File→Load**) select the target file and click **Open**.



Drag and drop datasets Multiple files can be dragged and dropped into the program at the same time.

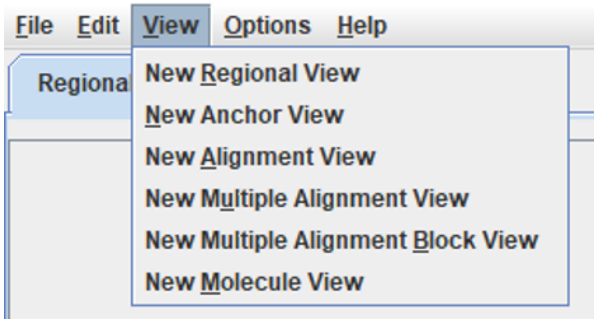


Since the formats for reference and molecule files are the same, users need to specify whether the files are loaded as reference or molecules.

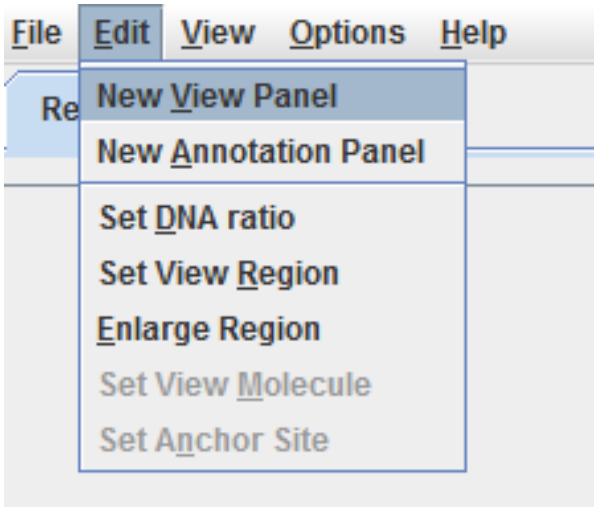
File dependency

- Reference and molecules: No requirements on other data files. Note that the reference ID should not be duplicated among all reference files. Similarly, molecule ID should not be duplicated.
- Alignments: Require references and molecules (If alignment file contains molecule information, the loading of molecule file is not needed)
- Annotations: Require reference
- Multiple alignments: Require molecules

30.5.3 Starting a view



Users can open new view tabs under the menu **View**. By default, OMView will initialize with a new blank regional view.

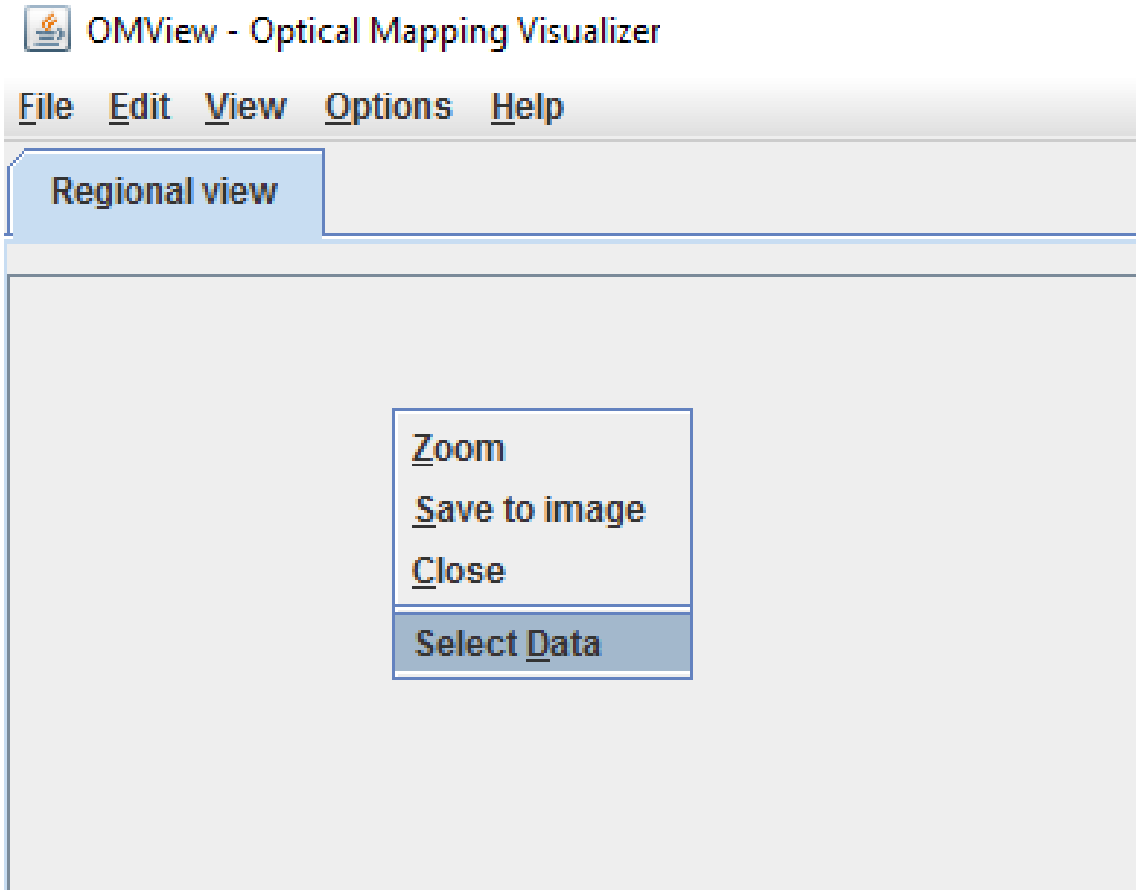


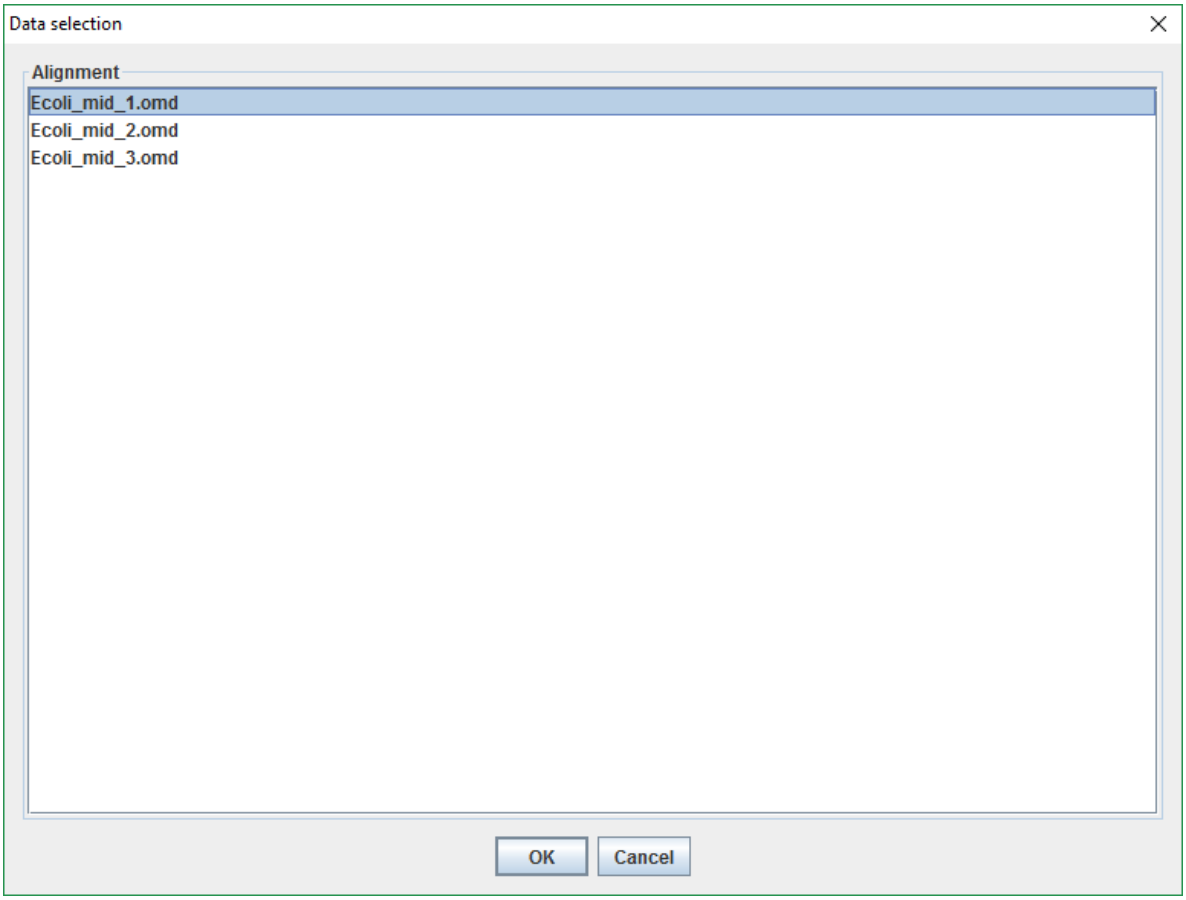
Each tab contains a view panel by default. Some views support visualization of more than one view panels (regional view, anchor view and molecule view) and annotation panels (regional view and anchor view) in the same tab. Users can insert new panels under the menu **Edit**.

30.5.4 Select data in the view panel

From the right click menu for each view panel, choose **Select Data** to select the data to be displayed.

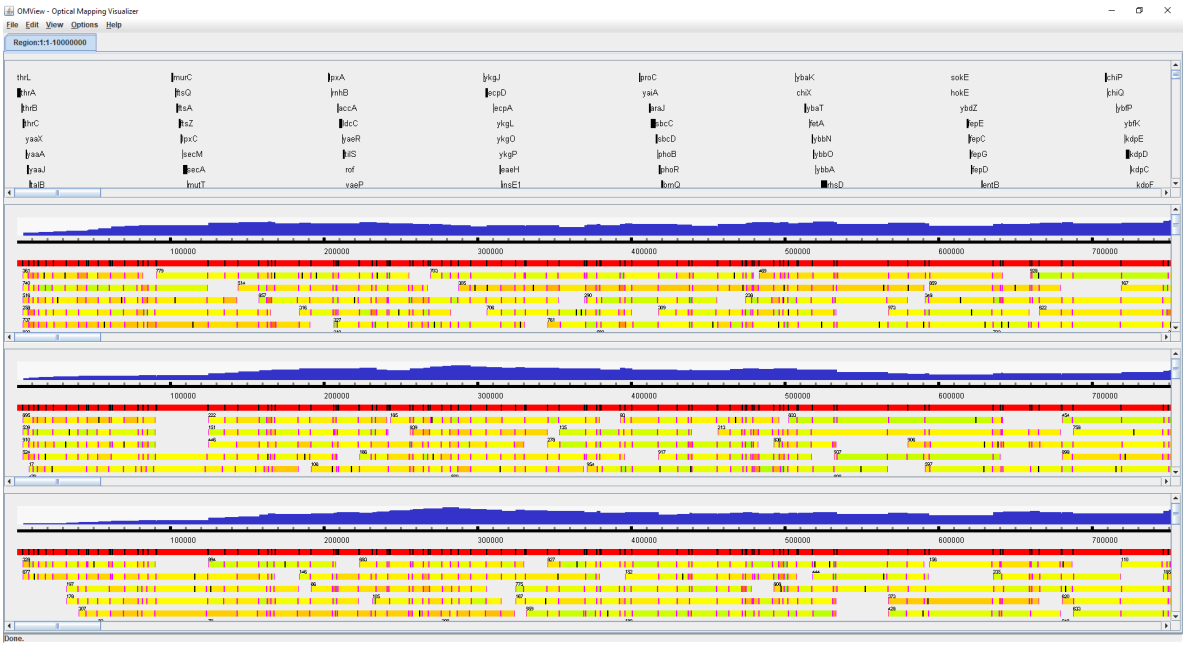
A dialog will open for users to choose the data. Note that if more than one set of alignment results are loaded in the same view panel, make sure their respective molecule IDs are distinct.


OMView - Optical Mapping Visualizer
File Edit View Options Help
Regional view
Zoom
Save to image
Close
Select Data


Data selection
Alignment
Ecoli_mid_1.omd
Ecoli_mid_2.omd
Ecoli_mid_3.omd
OK
Cancel

30.5.5 Procedures on specific views

Regional view Regional view displays alignments as an overview at a selected region.



After opening a regional view tab, set the target region as X:NNNNNN-NNNNNN (**Edit→Set View Region**). A reference (red rectangle) with signals (black vertical bars) should appear.

Note: The reference name should be consistent with that in the dataset. A common mistake in setting the region is messing up the reference name "chr1" or "1".

After selecting the alignments and annotations, the results will be displayed in the view panels. Aligned portion of molecules are shown in a color spectrum (from green to red depending on the scaling factor from 0.5 to 1.5, where yellow implies scaling factor 1) with pink and black signals indicating mapped or unmapped signals.

Move the cursor on the signal to display its information. To display details of alignment of a certain molecule, simply click on the target molecule and a new alignment view tab will be created. To view all molecules aligned to a specific reference signal, click on that signal to create a new anchor view.

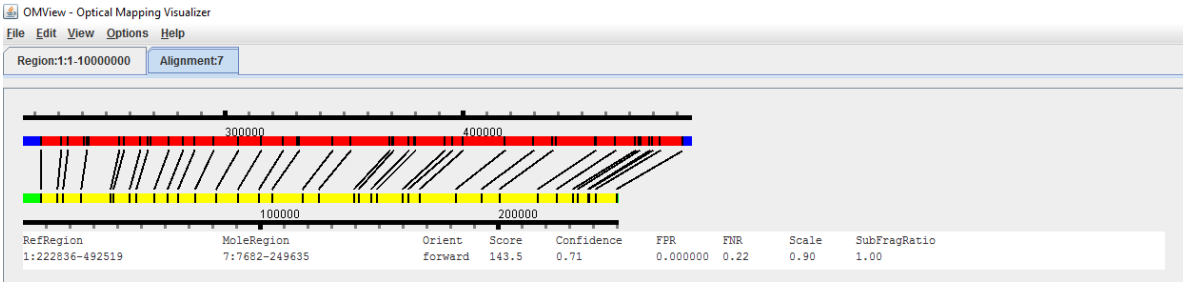
Example: (Reference) Ecoli.ref, Ecoli_mid_1.omd, Ecoli_mid_2.omd, Ecoli_mid_3.omd, Ecoli.gff

Anchor view Anchor view displays alignments that match selected signals to validate structural variations

The procedure of setting is similar to that in regional view. After opening the anchor view tab, set the anchor site as X:NNNNNN-NNNNNN (**Edit→Set Anchor Site**). Note that the anchor site must represent the position of one or two signals. Users can set the region as X:NNNNNN-NNNNNN (**Edit→Set Region**). By default region is set to 200 kbp away from the anchor sites.

Example: (Reference) Ecoli.ref, Ecoli_mid_1.omd, Ecoli_mid_2.omd, Ecoli_mid_3.omd, Ecoli.gff

Alignment view Alignment view displays alignment detail of a single molecule.

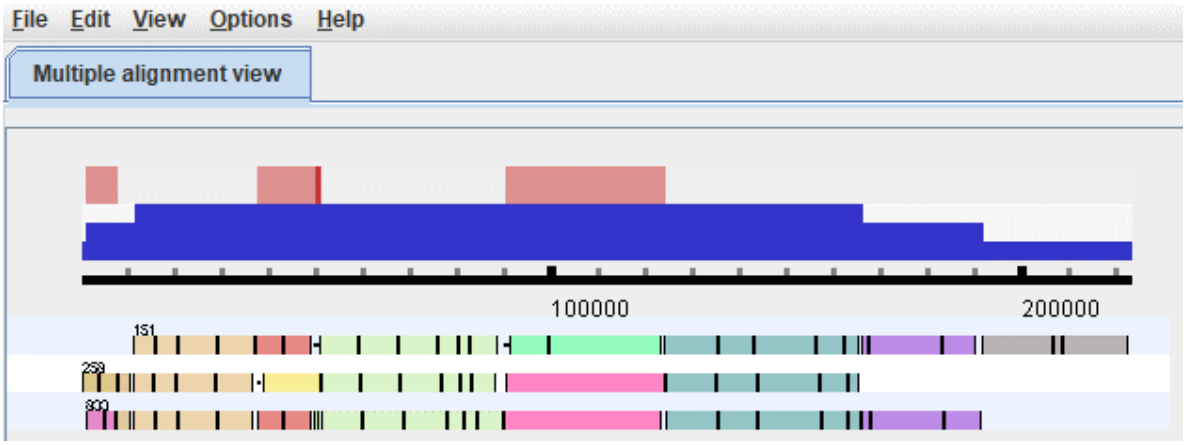


In alignment view, set the molecule ID to view the details of an alignment (**Edit→Set View Molecule**). Users will receive a warning message if an alignment does not exist for the selected molecule.

The top and the bottom rectangles represent the reference and the molecule respectively. Two sets of rulers indicate the coordinates relative to the start of the reference chromosome and the start of the optical molecule. Note that molecule is not scaled and remains in the forward orientation, while reference can be scaled and reversed for better visualization. Users can look at the text box below to obtain further details of the alignment.

Example: (Reference) Ecoli.ref, Ecoli_mid_1.omd

Multiple alignment view Multiple alignment view displays the multiple alignments of all queries for genomic comparison.

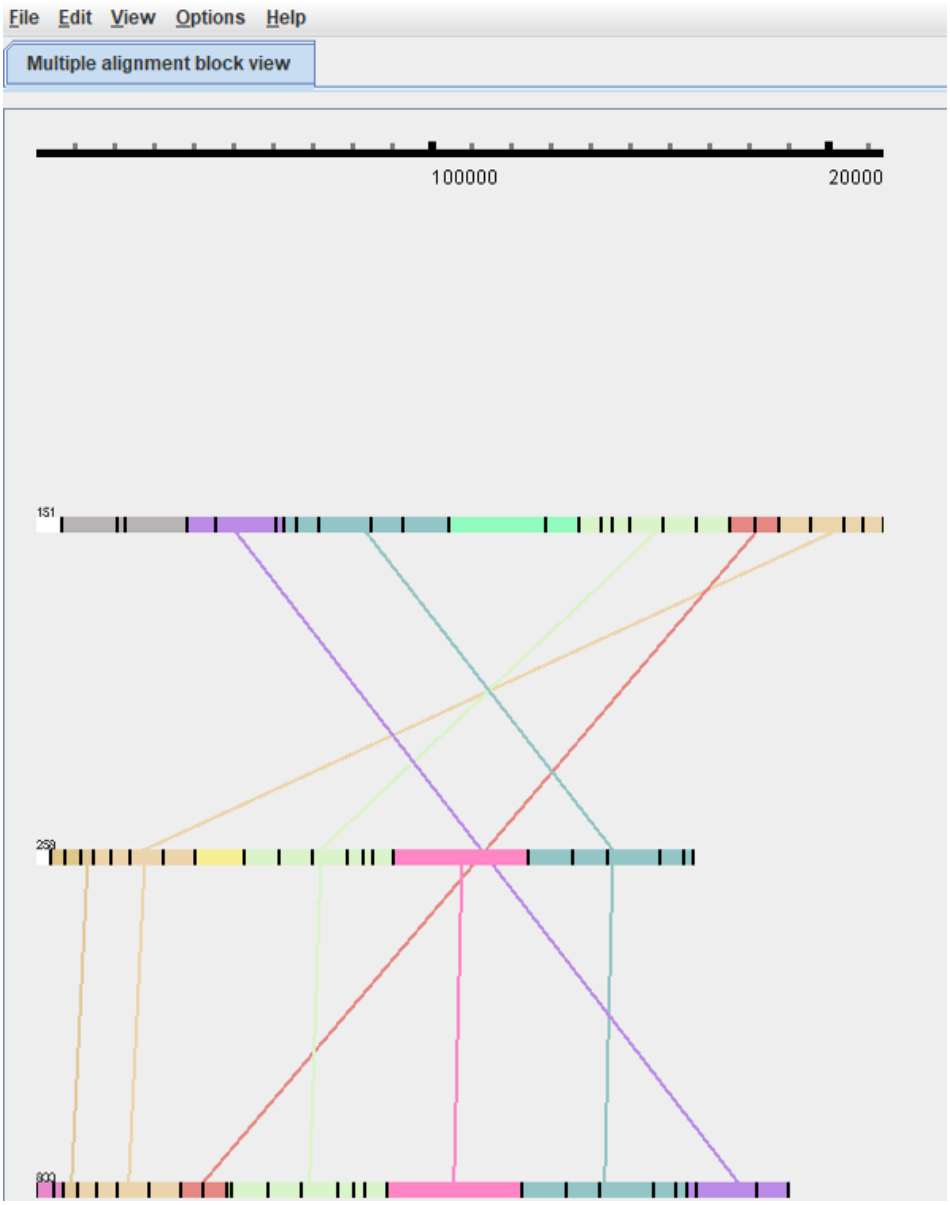


Multiple alignment view displays automatically once you set the data. If the color file (.cbc) is not provided, random color will be assigned to the collinear blocks.

Right click to sort and manipulate the multiple alignments. Users could save the current multiple alignment after the manipulation. A counting function is also available and the statistics is output in the console.

Example: (Molecule) Ecoli_MA.sdata, Ecoli_MA.cbl, Ecoli_MA.cbo, Ecoli_MA.cbc

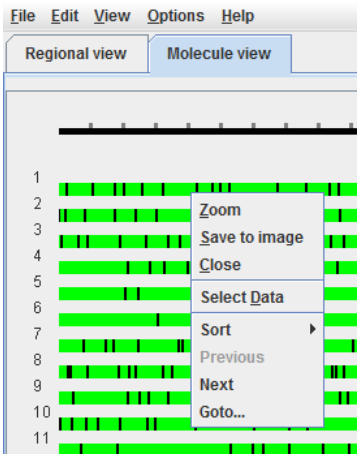
Multiple alignment block view Multiple alignment block view displays an alternative visualization of multiple alignment. Queries remain at the raw form without modifications. Segments from the same collinear block are connected by lines with the same color.



Multiple alignment block view displays automatically once you set the data. If the color file (.cbc) is not provided, random color will be assigned to the collinear blocks.

Example: (Molecule) Ecoli_MA.sdata, Ecoli_MA.cbl, Ecoli_MA.cbo, Ecoli_MA.cbc

Molecule view In molecule view, a panel displays a page that contains 100 molecules. Go to another page using the right click menu items Previous, Next, and Goto. A sorting function is also available to sort the molecules by molecule size, number of signals in the molecules, or molecule name.

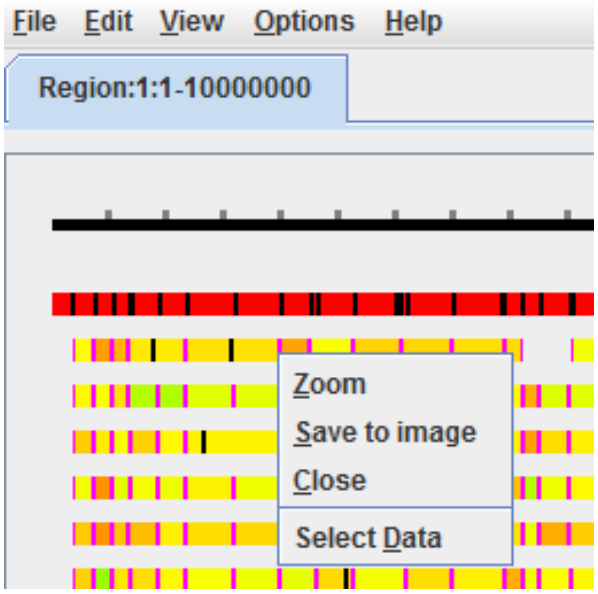


Right click to sort molecules.
Example: (Molecule) Ecoli_mid_1.sdata

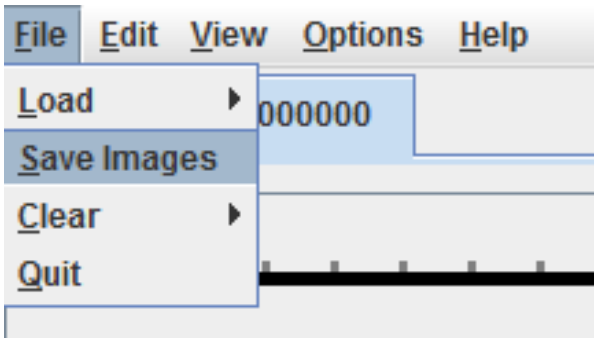
30.5.6 Export image

Images could be exported in individual panels or multiple panels in SVG, PNG and JPG formats.

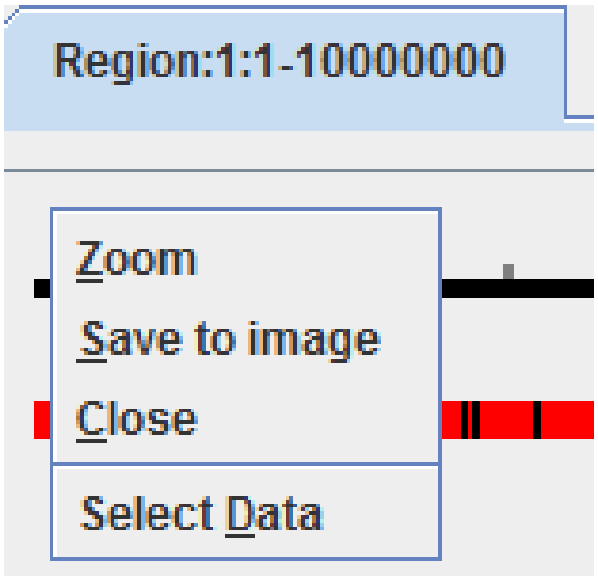
Individual panel From the right click menu, users could select Save to image to save the current view panel.



All panels under the same tab Users could select the menu **File→Save images** to save all view panels in the current tab to a single image.



30.5.7 Closing tabs and panels



To close the tab, click the tab with middle mouse button. To close the view panel, right click on the panel and select close.

30.6 Frequently Asked Questions (FAQs)

Q: Is there a batch mode to generate images?

A: Yes. You can use the `--viewsave` option for batch image generation.

Q: Can I choose multiple items at the same time to import reference or select data?

A: Yes. Hold the control key when you choose multiple items.

Q: I loaded all the files including references, molecules and alignment results but don't see anything. Whats wrong?

A: To visualize the reference and the alignment results, (1) you have to select the data in the view panel (**Select Data** from the right click menu); and (2) a view region must be set by **Edit→Set View Region**.

Q: I set view region or anchor site to visualize my data, but get the error of Reference not found. Whats wrong? A: The most common cause for the error is an incorrect input of a reference name. Note for the difference between "chr1" "Chr1", "CHR1", and "1".

Q: Why is the file loading speed very slow after loading half of my data?

A: Ensure that you have enough memory to store all the data. Use the parameter Xmx to allocate more memory to Java machine. Dont load too many data into an OMView instance if your machine does not have enough memory.

Q: The loading speed of regional view is slow.

A: Try to limit the range of region. Usually a region larger than 1 Mbp takes some time to completely load.

Q: What formats does OMView accept?

A: Reference and molecule file formats: REF, FA01, SPOTS, DATA, SDATA, BNX, CMAP, OPT, SILICO, and OpGen XML
Alignment result formats: OMA, OMD, XMAP, Valouev et al., SOMA v2 Unique Match, and Twin PSL
Annotation formats: BED, GVF, and OSV, AGP

Part XI
Other Scripts

31 TWINResultRepeatRemover

Removes repeat alignment results from TWIN.

31.1 Result Reader Options

`--optresin` Input alignment result file [Required]

`--optresinformat` -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

31.2 Result Writer Options

`--optresout` Output alignment result file [Required]

`--optresoutformat` Result file format -1: Auto-detected from file extension; 0: OM Alignment Format (OMA); 1: OM Detailed Alignment Format (OMD); 2: XMAP format (XMAP); 3: Valouev et al. format; 4: SOMA v2 Unique Match Format; 5: Twin PSL Format; 6: Maligner ALN Format; [Default: -1]

`--writeunmap` Write discarded or unmapped molecules. [Default: true]

`--multiple` Write multiple maps for a molecule. [Default: true]

`--writeinfo` Write information of a molecule. [Default: true]

32 SeparateBNXScan

Separates a bnx file into multiple bnx files according to the global scan number

32.1 Data Reader Options

`--optmapin` Input optical map file [Required]

`--optmapinformat` -1: Auto-detected from file extension; 0: Reference Standard Format (REF) (Equivalent to SILICO format); 1: fasta-01 format (FA01); 2: Spots File Format (SPOTS); 3: Molecule Standard Format (DATA); 4: Molecule Simulation Format (SDATA); 5: BNX File Format (BNX); 6: CMAP File Format (CMAP); 7: SOMA opt format (OPT); 8: SOMA silico format (SILICO) (Equivalent to REF format); 9: OpGen XML Format; 10: Valouev data format; 11: Maligner maps format; [Default: -1]

`--bnxsnr` BNX SNR filter value [Default: 3.0]

32.2 SeparateBNXScan Options

`--prefix` Output prefix [Default: Scan]