



CMAP File Format Specification Sheet

Document Number: 30039

Document Revision: H

Table of Contents

Legal Notice	3
Revision History	4
Introduction	4
Format	4
Header Specifications	4
Header Specification Details	5
Genome map quality scores	8
Genome map label attributes encoded in Mask column.....	8
Genome map information block specification.....	9
Example	9
Technical Assistance.....	10

Legal Notice

For Research Use Only. Not for use in diagnostic procedures.

This material is protected by United States Copyright Law and International Treaties. Unauthorized use of this material is prohibited. No part of the publication may be copied, reproduced, distributed, translated, reverse-engineered or transmitted in any form or by any media, or by any means, whether now known or unknown, without the express prior permission in writing from Bionano Genomics. Copying, under the law, includes translating into another language or format. The technical data contained herein is intended for ultimate destinations permitted by U.S. law. Diversion contrary to U. S. law prohibited. This publication represents the latest information available at the time of release. Due to continuous efforts to improve the product, technical changes may occur that are not reflected in this document. Bionano Genomics reserves the right to make changes in specifications and other information contained in this publication at any time and without prior notice. Please contact Bionano Genomics Customer Support for the latest information.

BIONANO GENOMICS DISCLAIMS ALL WARRANTIES WITH RESPECT TO THIS DOCUMENT, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO THOSE OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. TO THE FULLEST EXTENT ALLOWED BY LAW, IN NO EVENT SHALL BIONANO GENOMICS BE LIABLE, WHETHER IN CONTRACT, TORT, WARRANTY, OR UNDER ANY STATUTE OR ON ANY OTHER BASIS FOR SPECIAL, INCIDENTAL, INDIRECT, PUNITIVE, MULTIPLE OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH OR ARISING FROM THIS DOCUMENT, INCLUDING BUT NOT LIMITED TO THE USE THEREOF, WHETHER OR NOT FORESEEABLE AND WHETHER OR NOT BIONANO GENOMICS IS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Patents

Products of Bionano Genomics® may be covered by one or more U.S. or foreign patents.

Trademarks

The Bionano Genomics logo and names of Bionano Genomics products or services are registered trademarks or trademarks owned by Bionano Genomics in the United States and certain other countries.

Bionano Genomics®, Saphyr®, Saphyr Chip®, and Bionano Access® are trademarks of Bionano Genomics, Inc. All other trademarks are the sole property of their respective owners.

No license to use any trademarks of Bionano Genomics is given or implied. Users are not permitted to use these trademarks without the prior written consent of Bionano Genomics. The use of these trademarks or any other materials, except as permitted herein, is expressly prohibited and may be in violation of federal or other applicable laws.

© Copyright 2021 Bionano Genomics, Inc. All rights reserved.

Revision History

Revision	Notes
H	<ul style="list-style-type: none">• Added additional explanation of genome map quality scores• Added text on limitations of mosaicism simulation

Introduction

The Bionano Genomics® CMAP file is a data file which provides location information for label sites within a genome map or an *in silico* digestion of a reference or sequence data. The CMAP is a tab-delimited text-based file. Although the CMAP most commonly contains data from FASTA reference digestion and a *de novo* assembly, a BNX file (which typically contains raw molecule data) can also be converted to a CMAP.

A CMAP file contains two sections: 1) the CMAP information header, which describes the format of the data, and 2) the map information block, which contains the data values. This file format specification sheet provides descriptions, with examples, of the CMAP header and map information block format of the file.

CMAP files can be opened in Excel for easy readability or in any tab-delimited, text-based editor.

Format

The CMAP file contains the following sections:

- CMAP header
 - # CMAP File Version
 - # Label Channels
 - # Nickase Recognition Site
 - # Number of Consensus Maps
 - #h
 - #f
- Map information block
 - First label site in map
 - Next label site in map (repeated for all label sites)
 - Last label site is end of map

Header Specifications

Header rows are prefixed by the pound sign (#). “*” Denotes required header line tags.

Header Line Tag	Header Line Description
-----------------	-------------------------

# CMAP File Version:	Version of CMAP*
# Label Channels:	The number of label channels (integer)*
# Nickase Recognition Site 1:	Comma separated list of label motif recognition sequences for channel 1 followed by semicolon and channel 1 color. There can be no spaces in this string. Color is optional. This can also refer to the label recognition sequence for a non-nicking enzyme (i.e. DLE-1).
# Number of Consensus Maps:	The total number of consensus genome maps in the CMAP file (integer)
#h	The columns for each data row
#f	The numerical data type for each data column

Header Specification Details

The following tables provide the CMAP header's descriptions (including any specific formatting, limitations and requirements) and examples. CMAP currently supports up to 2 label channels. Additional columns may be present but are not defined. Certain columns may be absent in earlier versions of the CMAP format.

# CMAP File Version	
Header	# CMAP File Version:
Description	Version of CMAP, auto-generated.
Example	# CMAP File Version:<TAB>0.2

# Label Channels	
Header	# Label Channels:
Description	The number of label channels (integer). Available values are: [1, 2].
Example	# Label Channels:<TAB>1

# Nickase Recognition Site 1	
Header	# Nickase Recognition Site 1:
Description	Comma separated list of label motif recognition sequences for channel 1 followed by semicolon and channel 1 color. There can be no spaces in this string. Color is optional. This can also refer to the label recognition sequence for a non-nicking enzyme (i.e. DLE-1).
Example	# Nickase Recognition Site 1:<TAB>gctctc,cctcagc;green_01

# Nickase Recognition Site 2 (optional)	
Header	# Nickase Recognition Site 2:

# Nickase Recognition Site 2 (optional)	
Description	Comma separated list of label motif recognition sequences for channel 2 followed by semicolon and channel 2 color. There can be no spaces in this string. Color is optional. This can also refer to the label recognition sequence for a non-nicking enzyme (i.e. DLE-1).
Example	# Nickase Recognition Site 2:<TAB>cctcagc;red_01

# Number of Consensus Maps	
Header	# Number of Consensus Maps:
Description	The total number of consensus genome maps in the CMAP file (integer).
Example	# Number of Consensus Maps:<TAB>81

#h																																			
Header	#h																																		
Description	<p>Defines the columns for each data row in #h rows:</p> <table border="1"> <tr> <td>CMapId</td> <td>Map ID, ordinal number</td> </tr> <tr> <td>ContigLength</td> <td>Map length in basepairs</td> </tr> <tr> <td>NumSites</td> <td>Total number of label sites in map</td> </tr> <tr> <td>SiteID</td> <td>Label ID, ordinal number</td> </tr> <tr> <td>LabelChannel</td> <td>Label channel of label sites The last LabelChannel field of each map is always 0.</td> </tr> <tr> <td>Position</td> <td>Position of label on map [0-based from map start] in basepairs</td> </tr> <tr> <td>StdDev</td> <td>Theoretical standard deviation in bases of label site interval between the current and next site. Value will be 0 for FASTA digestion of a reference.</td> </tr> <tr> <td>Coverage</td> <td> <p>Weighted coverage of aligned molecules across an interval. The values may be fractional. How much an alignment to a map contributes to the weighted coverage depends on whether the alignment is unique to that particular map. If a molecule aligns equally well to two maps, it would contribute 0.5 in coverage to each of the maps.</p> <p>Since Solve 3.5, coverage refers to the interval between the current and the next label. The header of the CMAP now includes a comment on whether coverage is based on the interval between labels.</p> </td> </tr> <tr> <td>Occurrence</td> <td>Number of molecules with a label aligned to a given label. This is also weighed. If a molecule spans an interval but its labels do not align to the label of interest, it would contribute to coverage but not occurrence. Generally, occurrence should be less than coverage. However, this may not be true in corner cases.</td> </tr> <tr> <td>ChimQuality</td> <td>Percent of molecules that align to both sides of the label out of all molecules that align on either side near this label.</td> </tr> <tr> <td>SegDupL</td> <td>See Note.</td> </tr> <tr> <td>SegDupR</td> <td>See Note.</td> </tr> <tr> <td>FragileL</td> <td>See Note.</td> </tr> <tr> <td>FragileR</td> <td>See Note.</td> </tr> <tr> <td>OutlierFrac</td> <td>Fraction of number of molecules with internal outlier which overlaps this site.</td> </tr> <tr> <td>ChimNorm</td> <td>This is the quantity (N1+N2+N3) described below.</td> </tr> <tr> <td>Mask</td> <td>64-bit hex value: each bit flags a possible attribute for each label. See below for currently used flags.</td> </tr> </table>	CMapId	Map ID, ordinal number	ContigLength	Map length in basepairs	NumSites	Total number of label sites in map	SiteID	Label ID, ordinal number	LabelChannel	Label channel of label sites The last LabelChannel field of each map is always 0.	Position	Position of label on map [0-based from map start] in basepairs	StdDev	Theoretical standard deviation in bases of label site interval between the current and next site. Value will be 0 for FASTA digestion of a reference.	Coverage	<p>Weighted coverage of aligned molecules across an interval. The values may be fractional. How much an alignment to a map contributes to the weighted coverage depends on whether the alignment is unique to that particular map. If a molecule aligns equally well to two maps, it would contribute 0.5 in coverage to each of the maps.</p> <p>Since Solve 3.5, coverage refers to the interval between the current and the next label. The header of the CMAP now includes a comment on whether coverage is based on the interval between labels.</p>	Occurrence	Number of molecules with a label aligned to a given label. This is also weighed. If a molecule spans an interval but its labels do not align to the label of interest, it would contribute to coverage but not occurrence. Generally, occurrence should be less than coverage. However, this may not be true in corner cases.	ChimQuality	Percent of molecules that align to both sides of the label out of all molecules that align on either side near this label.	SegDupL	See Note.	SegDupR	See Note.	FragileL	See Note.	FragileR	See Note.	OutlierFrac	Fraction of number of molecules with internal outlier which overlaps this site.	ChimNorm	This is the quantity (N1+N2+N3) described below.	Mask	64-bit hex value: each bit flags a possible attribute for each label. See below for currently used flags.
CMapId	Map ID, ordinal number																																		
ContigLength	Map length in basepairs																																		
NumSites	Total number of label sites in map																																		
SiteID	Label ID, ordinal number																																		
LabelChannel	Label channel of label sites The last LabelChannel field of each map is always 0.																																		
Position	Position of label on map [0-based from map start] in basepairs																																		
StdDev	Theoretical standard deviation in bases of label site interval between the current and next site. Value will be 0 for FASTA digestion of a reference.																																		
Coverage	<p>Weighted coverage of aligned molecules across an interval. The values may be fractional. How much an alignment to a map contributes to the weighted coverage depends on whether the alignment is unique to that particular map. If a molecule aligns equally well to two maps, it would contribute 0.5 in coverage to each of the maps.</p> <p>Since Solve 3.5, coverage refers to the interval between the current and the next label. The header of the CMAP now includes a comment on whether coverage is based on the interval between labels.</p>																																		
Occurrence	Number of molecules with a label aligned to a given label. This is also weighed. If a molecule spans an interval but its labels do not align to the label of interest, it would contribute to coverage but not occurrence. Generally, occurrence should be less than coverage. However, this may not be true in corner cases.																																		
ChimQuality	Percent of molecules that align to both sides of the label out of all molecules that align on either side near this label.																																		
SegDupL	See Note.																																		
SegDupR	See Note.																																		
FragileL	See Note.																																		
FragileR	See Note.																																		
OutlierFrac	Fraction of number of molecules with internal outlier which overlaps this site.																																		
ChimNorm	This is the quantity (N1+N2+N3) described below.																																		
Mask	64-bit hex value: each bit flags a possible attribute for each label. See below for currently used flags.																																		
Example	<pre>#h CmapId<TAB>ContigLength<TAB>NumSites<TAB>SiteID<TAB> LabelChannel<TAB>Position<TAB>StdDev<TAB>Coverage<TAB>Occurrence<TAB> ChimQuality<TAB>SegDupL<TAB>SegDupR<TAB>FragileL<TAB>FragileR<TAB>OutlierFrac<TAB>ChimNorm <TAB>Mask</pre>																																		

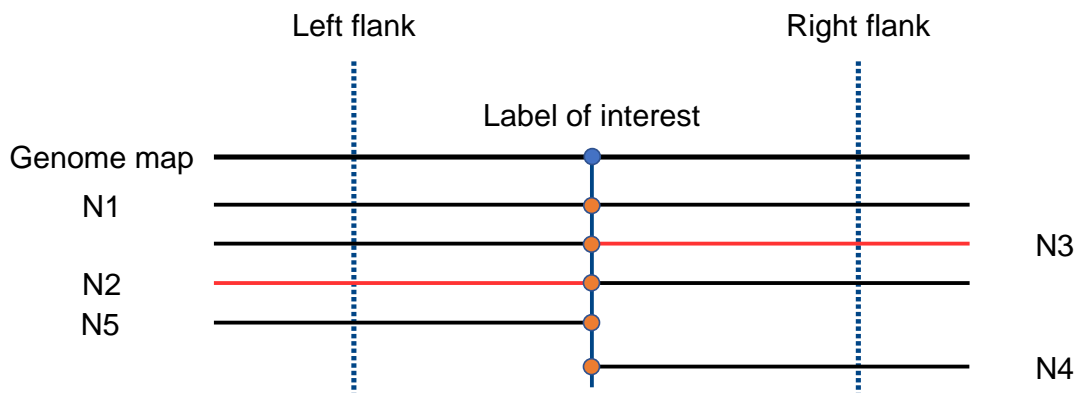
#f	
Header	#f
Description	Defines the numerical data type for each data column.
Example	<pre>#f<TAB>int<TAB>float<TAB>int<TAB>int<TAB>int<TAB>float<TAB>float<TAB>float <TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>Hex</pre>

For Research Use Only.
Not for use in diagnostic procedures.

Genome map quality scores

Based on the molecule-to-genome map alignment, we compute the following genome map quality scores for each label of the genome map. In the example below, N1-N5 are representative molecules which align to the genome map. They all contain the label of interest, for which the score is computed. The numbers may be fractional, since coverage is typically weighted (a molecule that aligns to 2 regions of the genome gets a weight of 0.5 for each location).

- First, the following quantities are computed for each label in the genome map. For N2 through N5, up to two missing and one extra label are allowed next to the label for which the score is computed.
- N1: the number of molecules which align over both left and right flanks. Each flank is 36 kbp (see CovTrimLen in refineFinal section of optArguments.xml)
- N2/N3: number of molecules which align on one flank, but have an endoutlier (unaligned portion, shown in red below) which spans the second flank.
- N4/N5: same as N2/N3 but no endoutlier is present
- The genome map quality scores are defined by the following (they are expressed as fractions):
 - $\text{ChimQuality} = N1/(N1+N2+N3)$
 - $\text{SegDupL} = N2/(N1+N2+N3)$
 - $\text{SegDupR} = N3/(N1+N2+N3)$
 - $\text{FragileL} = N4/(\text{coverage})$
 - $\text{FragileR} = N5/(\text{coverage})$



Genome map label attributes encoded in Mask column

The following bits are currently used to flag attributes of labels in the genome map (the default bit value is 0):

1. Bit 0 (Value 1) is set for end labels to mark a broken end when a genome map is broken at an ambiguous CMPR (complex multi-path region). See Bionano Solve Theory of Operation: Structural Variant Calling (PN# 30110) for detail.
2. Bit 1 (Value 2) is set for end labels to mark the end of an alternate allele map (similar to assembly graph bubbles). Typically, such a map consists of the alternate region plus 300 kbp at either end of the shared homozygous region. They are generated when haplotype-aware assembly is performed. For a haplotype-aware assembly, most of these alternate maps are assigned to one of the two allelic maps, but any alternate maps that could not be assigned to either of the two dominant alleles will have their ends marked with this Bit 1.
3. Bit 2 (Value 4) is set for all labels in a region that is a suspected CMPR (complex multi-path region): these are genome map regions that closely resemble regions in other genome maps (other than the matching allelic map pair) and could be mediated by segmental duplications. By default, such regions over 140 kbp are likely to be broken with both pieces sharing the CMPR region and the broken ends marked with Bit 0 (see above). We also provide the option to not break them. Currently, CMPR regions under 140 kbp are NOT broken but marked with Bit 2.
4. Bit 3 (Value 8) is used in Hybrid Scaffold to mark ends derived from a Bionano genome map.
5. Bit 4 (Value 16 OR 0x10) is used in Hybrid Scaffold to mark ends derived from an NGS sequence (or sequence scaffold).

Note that a hybrid scaffold can have one end derived from a Bionano genome map *and* the other end derived from an NGS sequence. During Hybrid Scaffold, Mask bits 3 and 4 are used to prevent merging scaffold ends that are both derived from NGS sequence.

Genome map information block specification

The data is grouped per each genome map represented in the CMAP file. Each group starts with the first label site, followed by each label site in the map, and ends with the map length. Each group follows this convention:

- Genome map information block
 - First label site in map
 - Next label site in map [repeated for all label sites]
 - End location of genome map. This position encodes the final coordinates of the map.

Example

```
# CMAP File Version: 0.2
# Label Channels: 1
# Nickase Recognition Site 1: cttaag:green_01
# Number of Consensus Maps: 459
# Values corresponding to intervals (StdDev, HapDelta) refer to the interval between current site and next site
#h CMapId ContigLength NumSites SiteID LabelChannel Position StdDev Coverage Occurrence ChimQuality SegDupL
  SegDupR FragileL FragileR OutlierFrac ChimNorm Mask
#f int float int int int float float float float float float float float float float float Hex
182 58474736.7 10235 1 1 58820.9 35.4 13.5 13.5 -1.00 -1.00 -1.00 3.63 0.00 0.00 -1.00 0
182 58474736.7 10235 2 1 70333.1 36.5 13.6 13.6 -1.00 -1.00 -1.00 0.00 0.00 0.00 -1.00 0
182 58474736.7 10235 3 1 84845.3 30.7 14.6 13.7 -1.00 -1.00 -1.00 0.31 0.00 0.00 -1.00 0
182 58474736.7 10235 4 1 87470.9 36.7 14.6 14.6 -1.00 -1.00 -1.00 0.31 0.00 0.04 -1.00 0
182 58474736.7 10235 5 1 106152.6 34.9 14.6 14.5 -1.00 -1.00 -1.00 0.00 0.00 0.10 -1.00 0
182 58474736.7 10235 6 1 119659.3 30.7 14.6 13.2 100.00 0.00 0.00 0.03 0.00 0.00 13.64 0
182 58474736.7 10235 7 1 122330.5 29.9 15.1 14.1 96.66 3.34 0.00 5.33 0.00 0.00 14.99 0
```

Technical Assistance

For technical assistance, contact Bionano Genomics Technical Support.

You can retrieve documentation on Bionano products, SDS's, certificates of analysis, frequently asked questions, and other related documents from the Support website or by request through e-mail and telephone.

Type	Contact
Email	support@bionanogenomics.com
Phone	Hours of Operation: Monday through Friday, 9:00 a.m. to 5:00 p.m., PST US: +1 (858) 888-7600
Website	www.bionanogenomics.com/support