



SMAP File Format Specification Sheet

Document Number: 30041

Document Revision: G

Table of Contents

Legal Notice.....	3
Introduction.....	4
Format.....	4
Header Specifications.....	5
Header Specification Details.....	5
SV Types Definitions.....	8
Note on calculation of SVfreq.....	10
Technical Assistance.....	12

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 2020 Bionano Genomics, Inc. All rights reserved.

Introduction

The Bionano Genomics® SMAP file contains a list of structural variants (SV) detected between query maps and reference maps. Detailed information about each SV call is output in a tab-delimited, text-based format.

The SMAP file presents the information in two sections: 1) the SMAP information header, which describes the specific format of the data, and 2) the SV information block, which contains the data rows. This file format specification sheet provides descriptions, with examples, of the SMAP header and SV alignment information block format of the file.

When the data are imported into Bionano Access™, the SMAP file is automatically processed and ready for downstream analysis and visualization. SMAP files can also be opened in Excel for easy readability or in any tab-delimited, text-based editor.

Format

The SMAP file contains the following sections:

- SMAP header
 - # SMAP File Version:
 - # Reference Maps From:
 - # Query Maps From:
 - # XMAP Entries From:
 - #h
 - #f
- SV information block (each row as defined by the column headers in #h)
 - After the 4 IDs [SmappedEntryID, QueryContigID, RefContigID1, and RefContigID2] are the positions for query and reference of each SV [QueryStartPos, QueryEndPos, RefStartPos, RefEndPos].
 - Followed by the confidence scores of the SV calls and their corresponding SV type [Confidence, Type].
 - Then the fields XmapIDs provide the ID of xmap entries used to make the SV calls.
 - The next field LinkID references a SmappedEntryID when linked SMAP entries define a single SV call, especially for inversion breakpoints.
 - The Idxs [QueryStartIdx, QueryEndIdx, RefStartIdx, and RefEndIdx] are the label indices for query and reference labels for each SV call.
 - The final columns are Zygosity, Genotype, GenotypeGroup, RawConfidence, RawConfidenceLeft, RawConfidenceRight, RawConfidenceCenter, SVsize, SVfreq, and orientation.

Header Specifications

Header rows are prefixed by the pound sign (#).

Table 1. Required header lines

Header Line Tag	Header Line Description
# SMAP File Version:	Indicates the version of the SMAP file
# Reference Maps From:	A string denoting the path to the corresponding _r.cmap
# Query Maps From:	A string denoting the path to the corresponding _q.cmap
# XMAP Entries From:	A string denoting the path to the corresponding .xmap
#h	Defines the columns for each data row
#f	Defines the data type for each data column

Note: The above are required header line tags for Bionano Access to import SV data from an SMAP file. Required header line tags must be present and must precede the Alignment Information Block. Other header lines may contain auxiliary information and are optional.

Header Specification Details

The following tables provide the SMAP header's descriptions (including any specific formatting, limitations and requirements) and examples.

# SMAP File Version	
Header	# SMAP File Version:
Description	Indicates the version of the SMAP file.
Example	# SMAP File Version:<TAB>0.8

# Reference Maps From	
Header	# Reference Maps From:
Description	Denotes the path to the corresponding reference map, which contains the reference or anchor data.
Example	# Reference Maps From:<TAB>Example_r.cmap

# Query Maps From	
Header	# Query Maps From:
Description	Denotes the path to the corresponding query map, which contains the query data.
Example	# Query Maps From:<TAB>Example_q.cmap

# Xmap Entries From	
Header	# Xmap Entries From:
Description	A string denoting the path to the corresponding .xmap data, which contains information about the SV alignments.
Example	# Xmap Entries From:<TAB>Example_.xmap

#h and #f		
Header	#h	
Description	Description of the required tab-separated columns in #h:	
	SmapEntryID	A unique number for an entry in the SMAP file.
	QryContigID	Map ID of query map (Contig ID from .cmap). Both XmapID1 and XmapID2 contain alignments to this map.
	RefcontigID1	Reference contig ID (XmapID1). Map ID of the reference map from the .cmap reference file (the .cmap file may contain multiple reference maps). Note: RefContigIDs must be integers, but they need not be sequential.
	RefcontigID2	Reference contig ID (XmapID2). Map ID of the reference map from the .cmap reference file (the .cmap file may contain multiple reference maps). Note: These RefContigIDs are always the same for insertions, deletions, duplications, and inversion breakpoints.
	QryStartPos	Start of SV on the query.
	QryEndPos	End of SV on the query.
RefStartPos	Coordinate of reference contig ID1 aligned site which borders this SV. This site is always either a start or end of XmapID1 and it matches the site at the query start position (QryStartPos).	

RefEndPos	Coordinate of reference contig ID2 aligned site which borders this SV. This site is always either a start or end of XmapID2 and it matches the site at the query end position (QryEndPos).
Confidence	Estimate of probability of being correct for insertions and deletions, and a quality metric for inversion and translocation breakpoints. Other SVs are given a placeholder value of '-1.00'. See Bionano Solve Theory of Operation: Structural Variant Calling (document 30110).
Type	Type of SV (See definitions in SV Types Definitions below).
XmapID1	XmapEntryID in the .xmap file of the first alignment from which this SV is derived.
XmapID2	XmapEntryID in the .xmap file of the second alignment from which this SV is derived.
LinkID	For some SV types, two SMAP entries may be linked using this field (e.g., inversion-partial, inversion-paired).
QryStartIdx	Index in query map of site at QryStartPos.
QryEndIdx	Index in query map of site at QryEndPos.
RefStartIdx	Index in reference map of site at RefStartPos.
RefEndIdx	Index in reference map of site at RefEndPos.
Zygoty	One of 'homozygous', 'heterozygous', or 'unknown' based on overlap with other SVs and alignments.
Genotype	'1' for homozygous SVs, '1' or '2' for heterozygous SVs, and
GenotypeGroup	Indels which overlap one another are assigned the same
RawConfidence	Minimum of next three columns for indels. '-1' for other SV
RawConfidenceLeft	Confidence of alignment to the left (on reference) of indel or
RawConfidenceRight	Confidence of alignment to the right (on reference) of indel or
RawConfidenceCenter	Indels only: outlier confidence.
SVsize	The estimated size of the SV (this is output for insertion, deletion, duplication, and inversion breakpoint calls.)
SVfreq	The estimated variant allele fraction. See Note.
orientation	This is computed only for translocation breakpoints and

Example	<pre>#h SmapEntryID<TAB>QryContigID<TAB>RefcontigID1 <TAB>RefcontigID2<TAB>QryStartPos<TAB>QryEndPos <TAB>RefStartPos<TAB>RefEndPos <TAB>Confidence <TAB>Type<TAB>XmapID1<TAB>XmapID2<TAB>LinkID <TAB>QryStartIdx<TAB>QryEndIdx<TAB>RefStartIdx<TAB>RefEndIdx <TAB>Zygotity<TAB>Genotype<TAB>GenotypeGroup<TAB>RawConfidence <TAB>SVsize<TAB>SVfreq<TAB>orientation</pre>
Header	#f
Description	Defines the numerical data type for each data column.
Example	<pre>#f int<TAB>int<TAB>int<TAB>int<TAB>float<TAB>float<TAB> float<TAB>float<TAB>float<TAB>string<TAB>int<TAB>int <TAB>int<TAB>int<TAB>int<TAB>int<TAB>string<TAB>int<TAB>int<TAB> float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>string</pre>

SV Types Definitions

Structural variants (SVs) are defined as any significant difference of between, typically, a *de novo* assembly of Bionano molecules and a reference. The assembly pipeline includes an SV detection stage. SVs are detected either as pairs of local alignments (MatchGroups) on the genome map or within a single alignment for indels. The following table provides an overview of the SV types currently included in the SMAP and describes the rules by which they are classified.

SV Types	Definition
insertion	Size difference which is larger on the query than on the reference. Query length - Reference length <= 5 Mbp.
insertion_nbase	Insertion with an N-base reference gap in between the insertion breakpoints. Must have a .bed file supplied to specify the gap.

deletion	Size difference which is larger on the reference than query.
deletion_nbase	Deletion with N-base reference gap in between the deletion breakpoints. Must have a .bed file supplied.
inversion	Two local alignments that have opposite orientation and no overlap. This is an inversion breakpoint, not a full inversion event.
inversion_paired	Two inversion events which are linked and form a full inversion. LinkID will point to other paired inversion.
inversion_partial	Extra information about inversion events. Not an independent event. LinkID will point to an inversion, inversion_nbase, or inversion_repeat event.
inversion_nbase	Inversion with N-base reference gap in between the inversion breakpoints. Must have bed file supplied.
inversion_repeat	Inversion call in which at least one matchgroup contains a simple repeat (adjacent regularly spaced label intervals) on the reference.
translocation_intrachr	Two local alignments which align to the same reference contig (chromosome) and are separated by more than 5 Mbp on the reference. Each local alignment is, on average, at least 100 kbp. They must also satisfy the translocation criteria described below.
translocation_interchr	Two local alignments which align to different reference contigs (chromosomes). They must also satisfy the translocation criteria described below.

trans_intrachr_common, trans_interchr_common	A translocation_intrachr / translocation_interchr with a breakpoint which overlaps a list of common translocation calls in euploid samples as specified in the .bed file argument to the Pipeline; presumed false positive call which is not displayed in Bionano Access by default.
trans_intrachr_segdupe, trans_interchr_segdupe	A translocation_intrachr / translocation_interchr with a breakpoint which overlaps annotated segmental duplication in the reference as specified in the .bed file argument to the Pipeline; presumed false positive call which is not displayed in Bionano Access by default.
duplication	A region of the reference which aligns to two places on a genome map.
duplication_inverted	A duplication with MatchGroups in opposite orientation.
duplication_split	A duplication inferred by the rearrangement of MatchGroups.
end	Unaligned region of at least 5 sites and 50 kbp at one end of the genome map.
complex	A pair of MatchGroups which do not satisfy any of the above criteria. For example, translocations which fail the criteria below.

Note: Translocation criteria: If the two local alignments (MatchGroup) do not overlap, they must be no further than 500 kbp apart on the query. When they do overlap, they must not overlap by more than 30% (of the minimum MatchGroup size) and by no more than 200 kbp.

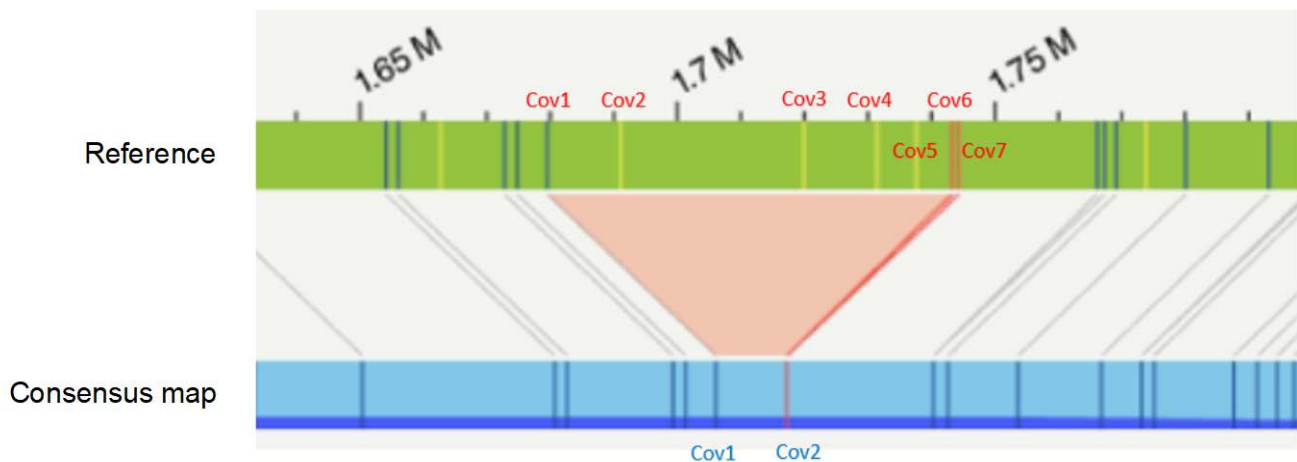
Note on calculation of SVfreq

SVfreq provides information about the prevalence of an allele is in a sample relative to other alleles. This is most relevant for *de novo* assembly pipeline data. Conceptually, SVfreq reflects the ratio between the number of molecules that are unique to a given allele map and the number of molecules that align to a particular reference region.

SVfreq is calculated based on the weighed molecule coverage during the final refinement stage (refineFinal1; output/contigs/exp_refineFinal1/EXP_REFINEFINAL1.cmap) of the assembly. The molecule coverage data are saved in the “coverage” column on the consensus genome map CMAP. If a molecule could align to two maps, the coverage it contributes would be halved accordingly. Currently, the molecule alignment counts towards coverage of the CMAP from the first to last aligned label, but the coverage is recorded for label intervals, so it would correspond to the 1st through the 2nd last aligned label in CMAP. SVfreq is then computed during SV calling and output in the SMAP.

The number of molecules that align to a particular reference region (overall coverage on the reference map regardless of the alleles) is computed by averaging the coverage of all consensus maps that align to the reference region. For each SV, the coverage of the allele-specific consensus map that called the SV (averaged for the map region where the SV is called) is divided by the coverage of the reference. See below for an example diagram.

Because coverage is weighed, if the same SV is called by two or more consensus maps, SVfreq across the maps need to be summed to get the overall variant allele frequency. In the following example, if a homozygous SV event is observed and called in two maps (only one map is shown), each SV call is expected to have an SVfreq of 0.5. In the figure below, Cov2 to Cov6 on the reference are expected to be zero. The sum of Cov1 and Cov2 on



$$\text{Ref coverage} = \text{Average} (\text{Cov1} + \text{Cov2} + \text{Cov3} + \text{Cov4} + \text{Cov5} + \text{Cov6} + \text{Cov7})$$

$$\text{SV coverage} = \text{Average} (\text{Cov1} + \text{Cov2})$$

$$\text{SV frequency} = \text{SV coverage} / \text{Ref coverage}$$

the map is expected to be roughly half of the sum of Cov1 and Cov7 on the reference, because molecules would align to both maps containing the deletion. The “SV coverage” (weighed coverage of the labels on the consensus maps) would then be roughly half of the “Ref coverage”

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