SMAP File Format Specification Sheet

Document Number: 30041

Document Revision: H

Table of Contents

Legal Notice

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

Revision History

Revision	Notes
H	• Added description of confidence score and variant allele fraction (VAF) columns • Added description of _overlap translocation types

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 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:

- **Header**. Contains metadata and description of the contents in the SV information block. There are both mandatory and optional lines.
  - **# SMAP File Version:**
  - **# Reference Maps From:**
  - **# Query Maps From:**
  - **# XMAP Entries From:**
  - **# Confidence scores:**
  - **# VAF:**
  - **#h**
  - **#f**
- **SV information block**. The content of each row is defined by the column headers in the header line #h. It is an open-ended, tab-delimited text format with no maximum number of columns defined, but there must be correspondence between the number of columns and the column names in #h. Columns:
  - After the 4 IDs [SmapEntryID, QryContigID, RefcontigID1, and RefcontigID2] are the positions for query and reference of each SV [QryStartPos, QryEndPos, 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 SmapEntryID when linked SMAP entries define a single SV call, especially for inversion breakpoints.
- The Idxs [QryStartIdx, QryEndIdx, RefStartIdx, and RefEndIdx] are the label indices for query and reference labels for each SV call.
- Other columns are pipeline and postprocessing dependent. For example, Zygosity, Genotype, GenotypeGroup, RawConfidence, RawConfidenceLeft, RawConfidenceRight, RawConfidenceCenter, SVsize, SVfreq, orientation, and VAF.

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 SV Information Block. Other header lines may contain auxiliary information and are optional.

Optional header lines

The confidence scores and VAF algorithms add lines to the header with details about versions and parameters used. To help with parsing, the values are stored in JSON format.

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 file of query maps, 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 file, which contains information about the map alignments.
Example	# Xmap Entries From:<TAB>Example_.xmap

# Confidence scores	
Header	# Confidence scores:
Description	A JSON entry recording details on confidence scoring model and parameters.

Example	# Confidence scores: {"translocations_score": {"pipeline_name": "denovo", "model_file": "/home/bionano/models/translocations_v0.5.2.joblib", "model_version": "0.5.2", "features": "LabelOccurrenceFrac,RegionCoverageFrac,RawConfidenceLeft,RawConfidenceRight,MeanCoverage,MeanOccurrence,MeanChimQuality,MeanSegDupL,MeanSegDupR,MeanFragileL,MeanFragileR,MeanOutlierFrac,InCentromere,InSegDup,StartRegion,EndRegion,Type"}, "inversions_score": {"pipeline_name": "denovo", "model_file": "/home/bionano/models/inversions_v0.5.0.joblib", "model_version": "0.5.0", "features": "MinAlignConfidence,MinLabelsAligned,BpPositionGap,RefLabelsUnaligned,LabelsInCMPR,MeanOutlierFrac,RegionCoverageFrac,LabelOccurrenceFrac,LabelsQryInvertedRegion,RefMatchGroupsOverlap,QryMatchGroupsOverlap,Conflict,BpLowerBoundInCentromere,BpLowerBoundInSegDup,BpUpperBoundInCentromere,BpUpperBoundInSegDup,BpLowerBoundRegion,BpUpperBoundRegion"}}

# VAF	
Header	# VAF:
Description	A JSON entry recording algorithm used for variant allele fraction
Example	# VAF: {"algorithm for non duplications": "BetaPriorBayesian", "version": "1.0","algorithm for duplications": "CoverageRatio", "pipeline": "de novo"}

#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 map. It is always the case for Indels (or anytime the 2 alignment MGs are not overlapped) that QryStartPos <= QryEndPos.
	QryEndPos	End of SV on the query map.
	RefStartPos	Coordinate of reference contig ID1 aligned position (typically a site, but can be between misresolved sites) which borders this SV. This position is always either a start or end of XmapID1 and it may correspond to either the query start position (QryStartPos) OR query end position (QryEndPos)
	RefEndPos	Coordinate of reference contig ID2 aligned position which borders this SV. This position is always either a start or end of XmapID2 and it matches the position of either the query end position (QryEndPos) or query start position (QryStartPos)
	Confidence	Estimate of probability of being correct for insertions, deletions over 500bp, inversions, and translocations. 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 nearest to QryStartPos.
	QryEndIdx	Index in query map of site nearest to QryEndPos.
	RefStartIdx	Index in reference map of site nearest to RefStartPos.
	RefEndIdx	Index in reference map of site nearest to at RefEndPos.
	Zygosity	One of 'homozygous', 'heterozygous', or 'unknown' based on overlap with other SVs and alignments.
	Genotype	'1' for homozygous SVs, typically '2' for heterozygous SVs (in general number of distinct SV clusters overlapping current SV), and '-1' for unknown zygosity and SVs which are not indels or translocations
	GenotypeGroup	Indels which overlap one another and belong to the same size cluster are assigned the same group.
	RawConfidence	Minimum of next three columns for indels. '-1' for other SV types.
	RawConfidenceLeft	Confidence of alignment to the left (on reference) of indel or translocation.
	RawConfidenceRight	Confidence of alignment to the right (on reference) of indel or translocation.
	RawConfidenceCenter	Indels only: outlier confidence.
	SVsize	The estimated size of the SV (this is output for insertion, deletion, duplication, and inversion breakpoint calls.)

	SVfreq	See *Calculation of SVfreq* below.
	Orientation	This is computed only for translocation breakpoints and indicates the orientation of the translocation events. It can be "+/+", "+/-", "-/+", or "-/-". See Theory of Operation: Structural Variant Calling (PN 30110) for detail.
	VAF	Variant allele fraction as calculated in Bionano Solve 3.7. See Theory of Operation: Structural Variant Calling (PN 30110) for detail.

Example	**#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>Zygosity<TAB>Genotype<TAB>GenotypeGroup<TAB>RawConfidence** **<TAB>SVsize<TAB>SVfreq<TAB>orientation**
Header	#f
Description	Defines the numerical data type for each data column.
Example	#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>int<TAB>string<TAB>int<TAB>int<TAB> float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>float<TAB>string

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 covering at least 40% of the reference breakpoint interval. Must have a .bed file supplied to specify the gap.
insertion_tiny	Insertion smaller than 5% of smaller of reference or query range AND both reference and query ranges include at least 2 misaligned labels: possibly a balanced indel, like a small inversion.
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.
deletion_tiny	Deletion smaller than 5% of smaller of reference or query range AND both reference and query ranges include at least 2 misaligned labels: possibly a balanced indel, like a small inversion.
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 primarily consists of a simple repeat (adjacent regularly spaced label intervals) on the reference, that may occur in multiple locations in the Genome.
translocation_intrachr	Two local alignments which align to the same reference contig (chromosome) and are typically (but not always) separated by more than 5 Mbp on the reference. The minimum confidence (and size) of

	each local alignment is around 1/3 lower than for other SVs. They must also satisfy the translocation criteria described below.
translocation_interchr	Two local alignments which align to different reference contigs (chromosomes). The minimum confidence (and size) of each local alignment is around 1/3 lower than for other SVs. 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_overlap, trans_interchr_overlap	A translocation_intrchr / translocation_interchr with a breakpoint that overlaps another distinct SV call for the same Map: Indicates a possible incorrect call.
trans_intrachr_segdupe, trans_interchr_segdupe	A translocation_intrachr / translocation_interchr with a breakpoint which overlaps an annotated segmental duplication (50kb or larger) 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, but the Map does not include complete copies of both duplicates.
end	Unaligned region of at least 5 sites and 50 kbp at one end of the genome map : possibly an incomplete insertion or translocation region.
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 (or have an intermediate MatchGroup 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 140 kbp.

Calculation of SVfreq

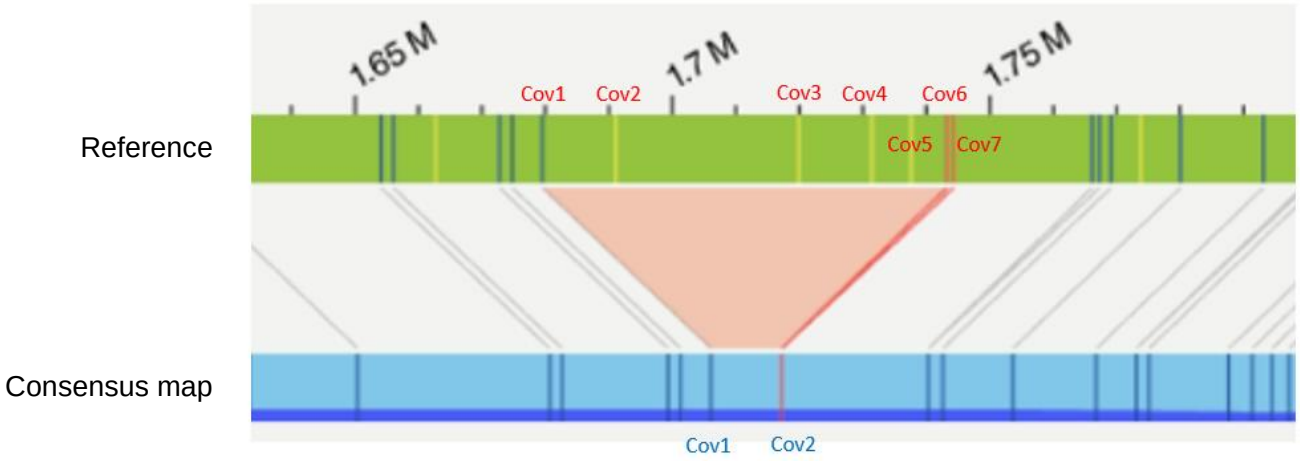
***Note**: Bionano Solve 3.7 provides a new way of estimating the allele fraction with results stored in the VAF column. See Bionano Solve Theory of Operation Structural Variant Calling (PN 30110) for information about the VAF calculation. The column SVfreq is kept for backwards compatibility. Details on the original algorithm are presented here for historical purposes but will be removed in future releases*

SVfreq provides information about the prevalence of an allele 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 weighted, 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 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”.



Ref coverage = Average (Cov1 + Cov2 + Cov3 + Cov4 + Cov5 + Cov6 + Cov7)
SV coverage = Average (Cov1 + Cov2)
SV frequency = SV coverage / 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**

Bionano Genomics, Inc.
9540 Towne Centre Drive, Suite 100
San Diego, CA 92121