SVMerge Output File Format Specification Sheet

Document Number: 30165

Document Revision: F

Table of Contents

Legal Notice

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

Introduction

The Bionano SVMerge tool can merge structural variant (SV) calls from two single-enzyme genome map assemblies of the same sample into a single integrated set of calls (*PN#30110, Bionano Solve Theory of Operation: Structural Variant Calling*). SVMerge is currently recommended for NLRS data only. It merges insertion, deletion, inversion breakpoint, translocation breakpoint, and duplication calls. The merged calls are output to a text file with the suffix *_mergedSV.txt*.

The *_mergedSV.txt* file reports merged SV calls as well as SV calls detected only in one of the single-enzyme assemblies. Each data line contains the merged SV start and end coordinates and their locations in each individual enzyme as shown in each SMAP file using a tab-delimited, text-based file.

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

When imported into Bionano Access™ along with proper SMAP and XMAP files, the *_mergedSV.txt* file is automatically processed and ready for visualization. The *_mergedSV.txt* files can be opened in Excel for easy readability or in any tab-delimited, text-based editor.

Format

The *_mergedSV.txt* file header contains the following sections:

- **The *_mergedSV.txt* file header:**
  - **# SVMergeVersion:**
  - **# SMAP of Enzyme 1:**
  - ***# <headers copied from SMAP file of Enzyme 1>* (*See 30041, SMAP File Format Specification Sheet*)**
  - **# SMAP of Enzyme 2:**
  - ***# <headers copied from SMAP file of Enzyme 2>* (*See 30041, SMAP File Format Specification Sheet*)**
  - ***# "data column names"***

- **The *_mergedSV.txt* file information block (each row as defined by the *data column names*). The information columns can be grouped in three categories:**
  - **Column 1-12 and 37, 38: [SVIndex, Type, RefcontigID1, RefcontigID2, RefStartPos, RefEndPos, Confidence, RawConfidence, Size, Zygosity, E1Id, E2Id, Orientation1, Orientation2] - information for merged call**
  - **Columns 13-24: [Type1, Confidence1, RawConfidence1, QryContigID1, QryStartPos1, QryEndPos1, QryStartIdx1, QryEndIdx1, RefStartPos1, RefEndPos1, RefStartIdx1, RefEndIdx1] – SMAP entry information of the first enzyme used for SV merging. Copied from the original SMAP entry.**
  - **Columns 25-36: [Type2, Confidence2, RawConfidence2, QryContigID2, QryStartPos2, QryEndPos2, QryStartIdx2, QryEndIdx2, RefStartPos2, RefEndPos2, RefStartIdx2, RefEndIdx2] – SMAP entry information of the second enzyme used for SV merging. Copied from the original SMAP entry.**

Header Specifications

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

Table 1: Header Order

Header Line Tag	Header Line Description
# SVMergeVersion:	Indicates the version of the _mergedSV.txt file*
# SMAP of Enzyme 1:	A string denoting the path to the first SMAP file used for SVMerge*
# <headers copied from SMAP file of Enzyme 1>	Header information copied from the first SMAP file excluding the column header (#h) and data type (#f) lines (see SMAP file format for detail)
# SMAP of Enzyme 2:	A string denoting the path to the second SMAP file used for SVMerge*
# <headers copied from SMAP file of Enzyme 2>	Header information copied from the second SMAP file excluding the column header (#h) and data type (#f) lines (see SMAP file format for detail)
# SVIndex	Defines the columns for each data row*

Note: *Denotes the required header line tags for _mergedSV.txt file. Required header line tags must be present and must precede the SV Information Block to read _mergedSV.txt file. Other header lines are optional and may be omitted.

Header Specification Details

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

# SVMergeVersion:	
Header	# SVMergeVersion:
Description	Indicates the version of the _mergedSV.txt file.
Example	# SVMergeVersion:<TAB>0.9.5

# SMAP of Enzyme 1:	
Header	# SMAP of Enzyme 1:
Description	A string denoting the path to the first SMAP file used for SV Merge
Example	# SMAP of Enzyme 1:<TAB>output/contigs/exp_refineFinal1_sv/merged_smaps/exp_refineFinal1_merged_filter_inversions.smap

# SMAP of Enzyme 2:	
Header	# SMAP of Enzyme 2:
Description	A string denoting the path to the second SMAP file used for SV Merge
Example	# SMAP of Enzyme 2:<TAB>output/contigs/exp_refineFinal1_sv/merged_smaps/exp_refineFinal1_merged_filter_inversions.smap

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	SVIndex	A unique line number for the data lines in the SMAP file.
	Type	Type of SV (insertion, deletion, inversion, translocation. See definitions in 30041, SMAP File Format Specification Sheet).
	RefcontigID1	Reference contig ID. **Note:** RefcontigIDs must be integers, but they need not be sequential.
	RefcontigID2	Reference contig ID.
	RefStartPos	SV breakpoint coordinate on reference map ID1
	RefEndPos	SV breakpoint coordinate on reference map ID2
	Confidence	Probability of an insertion or deletion call being correct, and a quality metric for translocation breakpoints. '-1.00' for other SV types.
	RawConfidence	Only applies to indels. '-1' for other SV types.
	Size	Size for insertion,deletion and duplication calls. For merged SV calls, the size is the average sizes of the two single-enzyme calls. '-1' for other SV types.
	Zygosity	One of 'homozygous', 'heterozygous', or 'unknown'. For merged calls, if one of the single-enzyme calls is 'heterozygous', the merged call is 'heterozygous'.
	LinkID	For some SV types, two SMAP entries may be linked using this field (e.g., inversion-partial).
	E1Id	The first SV SMAP entry id participated in the merging (For file name see header "**# SMAP of Enzyme 1:**"). '-1' if the SV is only detected in the second single-enzyme assembly.
	E2Id	The second SV SMAP entry id participated in the merging (For file name see header "**# SMAP of Enzyme 2:**"). '-1' if the SV is only detected in the first single-enzyme assembly.

	[Type1, Confidence1, RawConfidence1, QryContigID1, QryStartPos1, QryEndPos1, QryStartIdx1, QryEndIdx1, RefStartPos1, RefEndPos1, RefStartIdx1, RefEndIdx1, LinkID1]	The first SV SMAP entry information participating in SV merging. Copied from the original SMAP file as indicated in the header - "**# SMAP of Enzyme 1:**". (See 30041, SMAP File Format Specification Sheet for detail.) All values would be -1 except Type1 (NA) if the SV is only detected in the second single-enzyme assembly.
	[Type2, Confidence2, RawConfidence2, QryContigID2, QryStartPos2, QryEndPos2, QryStartIdx2, QryEndIdx2, RefStartPos2, RefEndPos2, RefStartIdx2, RefEndIdx2, LinkID2]	The second SV SMAP entry information participating in SV merging. Copied from the original SMAP file as indicated in the header - "**# SMAP of Enzyme 2:**". (See 30041, SMAP File Format Specification Sheet for detail.) All values would be -1 except Type2 (NA) if the SV is only detected in the first single-enzyme assembly.
	Orientation1	Only applies to translocation and inversion breakpoints; -1 for other SV types. For translocation and inversion orientation, see Figure 1 and Figure 2, respectively.
	Orientation2	Only applies to translocation and inversion breakpoints; -1 for other SV types. For translocation and inversion orientation, see Figure 1 and Figure 2, respectively.

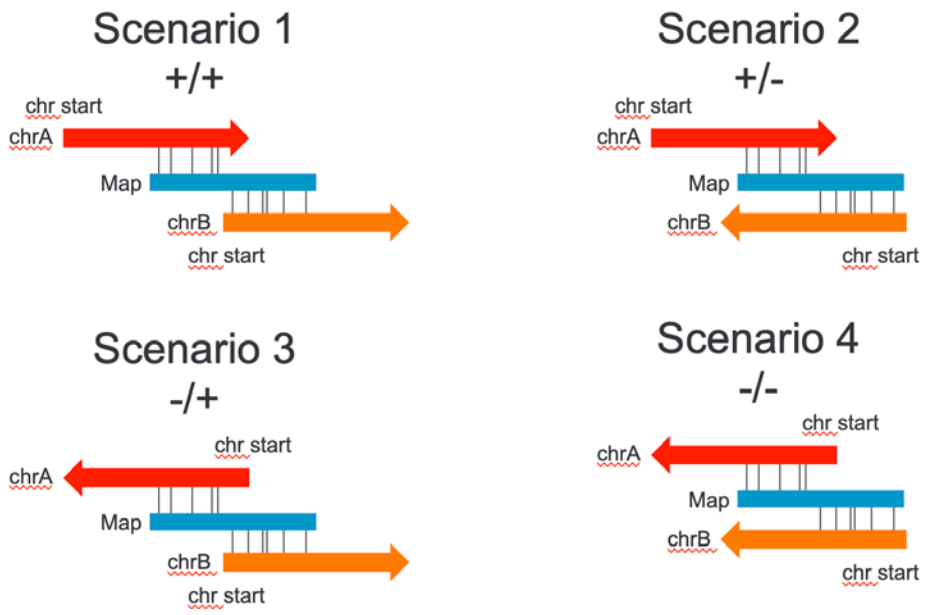

Figure 1. Determination of translocation orientation.

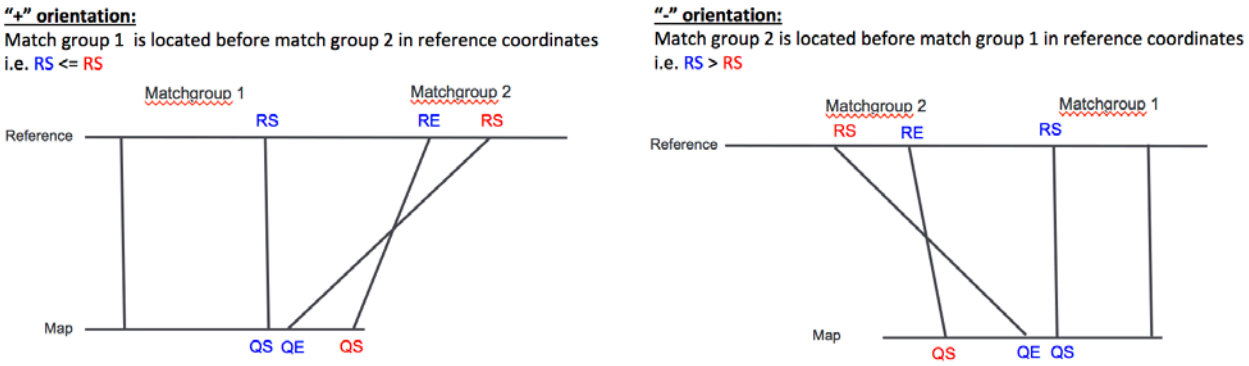

Figure 2. Determination of inversion orientation.

Merged SV Information Block Specification

The data is grouped such that each data row represents one structural variant – merged or detected in one of the single-enzyme assemblies.

Merging SV with variant annotation

Variant annotation of SV is performed for each enzyme, and SVMerge combines the annotation of the two sets with minor edits. The output file name, when run using command line, has the suffix **_mergedSV_genes.txt*.

Note: when one downloads the results from Bionano Access, the output file name's suffix is **_mergedSV.txt.* The following describes the addition of variant annotation information to the basic SVMerge output, and it assumes that the two enzymes used were Nt.BspQI and Nb.BssSI. Please also

refer to 30190 Bionano Solve Theory of Operation, Variant Annotation and the single enzyme 30168 Structural Variant Annotation Pipeline File Format Specification Sheet.

There are two additional header lines denoting the sample name given to the two single enzyme experiments.

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	Present_in_%_of_BNG_control_samples	The percentage of samples in the Bionano control SV database that also carry that SV. SVMerge takes the maximum number between the two enzymes' variant
	Present_in_%_of_BNG_control_samples_with_ the_same_enzyme	Same as above, but the percentage is calculated only based on those database samples having the same enzyme as the sample being annotated. SVMerge takes the maximum number between the two enzymes' variant annotation results.
	Algorithm_BspQI Algorithm_BssSI	The calls are based on comparing the *de novo* assembly of the sample with the reference, and so the algorithm is called "assembly comparison".
	Fail_BspQI_assembly_chimeric_score Fail_BssSI_assembly_chimeric_score	A flag used to denote whether a potential chimeric join occurred during *de novo* assembly at the variant locus. This denotes whether a minimal chimeric quality score of 35 and coverage of 10 have been achieved around each SV breakpoint. A value of 'pass' means that the two criteria have been met; a 'fail' denotes the criteria not met; a 'not_applicable' value denotes that the check has not been performed. Notice that this check is performed only for inversion and translocation calls. **Note:** a chimeric quality score of a label on a genome map is the percent of molecules that align to both sides of the label out of all molecules that align on either side near this label.

The next sets of columns vary depending on whether trio, dual or single analyses have been selected upon execution of variant annotation pipeline for each enzyme.

Trio analysis

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	Found_in_parents_BspQI_assemblies Found_in_parents_BssSI_assemblies	Whether the SV call is also identified in the father's or mother's assembly. The possible values are 'mother', 'father', 'both' and 'none'. Since inversion partial calls are not annotated, a value of '-' is shown for any inversion partial call.
	Found_in_parents_BspQI_molecules Found_in_parents_BssSI_molecules	These columns show whether there is a sufficient number of parents' molecules supporting the proband's genome map at the SV breakpoints. The possible values are 'mother', 'father', 'both' and 'none'. Since inversion partial calls are not annotated, a value of '-' is shown for inversion partial calls. **Note** that the minimum numbers of molecules required are defined as parameters by the users upon running the variant annotation pipeline.
	Found_in_self_BspQI_molecules Found_in_self_BssSI_molecules	These columns denote whether there is a sufficient number of proband's molecules supporting the proband's genome map at the SV breakpoints. Since inversion partial calls are not annotated, a value of '-' is shown for inversion partial calls. **Note** that the minimum number of molecules required is defined as a parameter by the users upon running the variant annotation pipeline.

Dual analysis

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	Found_in_self_BspQI_molecules Found_in_self_BssSI_molecules	These columns denote whether there is a sufficient number of case sample's molecules supporting the case sample's genome map at the SV breakpoints. The possible values are 'yes' and 'no'. Since inversion partial calls are not annotated, a value of '-' is shown for inversion partial calls. **Note** that the minimum number of molecules required is defined as a parameter by the users upon running
	Found_in_control_sample_BspQI_assembly Found_in_control_sample_BssSI_assembly	Whether the SV call is also identified in the control sample's assembly. The possible values are 'yes' or 'no'. Since inversion partial calls are not annotated, a value of '-' is shown for any inversion partial call.
	Found_in_control_sample_bspqi_molecules Found_in_control_sample_bspqi_molecules	These columns show whether there is a sufficient number of control sample's molecules supporting the case sample's genome map at the SV breakpoints. The possible values are 'yes' and 'no'. Since inversion partial calls are not annotated, a value of '-' is shown for inversion partial calls. **Note** that the minimum number of molecules required is defined as a parameter by the users upon running the variant annotation pipeline.

Single analysis

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	Found_in_self_BspQI_molecules Found_in_self_BssSI_molecules	These columns denote whether there is a sufficient number of case sample's molecules supporting the case sample's genome map at the SV breakpoints. The possible values are 'yes' and 'no'. Since inversion partial calls are not annotated, a value of '-' is shown for inversion partial calls. **Note** that the minimum number of molecules required is defined as a parameter by the user upon running the

Gene overlap

The variant breakpoints are refined during SVMerge, so the gene-overlap is recomputed using the refined positions. The last few columns show the genes overlapping or closest to merged variants.

# SVIndex		
Header	# SVIndex	
Description	Description of the required tab-separated columns in # SVIndex	
	OverlapGenes	A semi-colon separated list indicating which genes overlap with the SV.
	NearestNonOverlapGene	The next closest gene to the SV.
	NearestNonOverlapGeneDistance	The distance between the SV and the next closest gene.

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