Bionano VCF File Format Specification Sheet

Document Number: 30459

Document Revision: A

Table of Contents

Legal Notice

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

Revision History

Revision	Notes
A	Initial release of document

Introduction

The Bionano Variant Call Format (VCF) file contains structural variation (SV) and copy number variation (CNV) calls in a VCF version 4.2 format. It is generated by a Python-based VCF converter from the SMAP and CNV output from the main SV analysis pipelines (PN 30041 SMAP File Format Specification Sheet, PN 30110 Solve Theory of Operation Structural Variant Calling) as well as the annotated versions of these files produced by the Variant Annotation Pipeline (PN 30168 Structural Variant Annotation Pipeline File Format Specification Sheet, PN 30461 Copy Number Variant Annotation Pipeline File Format Specification Sheet). Prior to the Bionano Solve 3.6 release, only SV calls were included in the Bionano VCF files. Bionano's VCF output has been validated by VCFtools' VCF validator (https://vcftools.github.io). For more information on how the SV and CNV calls are generated by the analysis pipelines, see Bionano Solve Theory of Operation Structural Variant Calling (PN 30110).

The VCF output is automatically generated at the end of each analysis pipeline and imported into Bionano Access. In Bionano Access, after the analysis results have been imported, users have the option to select variant calls of interest and output a VCF with the selected calls.

In addition, the VCF converter is a standalone tool that may be run on the command line. It is packaged as part of Bionano Solve. The tool requires Python 3.7 and Python libraries including pandas and numpy. The required input is the SMAP output from one of the SV analysis pipelines. The SMAP may be annotated by the Variant Annotation Pipeline (VAP). The command line tool optionally takes the output from the CNV analysis pipeline; the CNV data is included by default when the VCF is generated as part of the pipeline runs. See Appendix G in Bionano Solve Theory of Operation Structural Variant Calling (PN 30110).

Format

VCF is a text-based format; VCF files may be opened in Excel for easy readability or in any text-based editor. Each VCF file contains a meta-information section. Each line in this section starts with "##". It is followed by a single header line that starts with "#" and then the data lines, each containing information about a given variant call. Aspects of the representation of variant calls that are unique to Bionano are discussed. See the official VCF version 4.2 file specification for reference and additional information.

Meta-information

In the "meta-information" section, information lines are presented as key-value pairs. Fields tagged with "INFO"

such as SVTYPE, SVLEN and END provide basic information about each variant call. Variant calls are listed with SVTYPE and ALT alleles that follow VCF standard conventions. Original variant types from the Bionano SV and CNV callers are preserved in the BNGTYPE INFO field. Additional fields such as OVERLAPGENES, NEARGENE, and DGVOVERLAPS come from the variant annotation pipeline (VAP), which annotates variant calls based on, for example, genome and gene annotations. See Structural Variant Annotation Pipeline File Format Specification Sheet (PN 30168) and PN 30461 Copy Number Variant Annotation Pipeline File Format Specification Sheet for information about the annotations. VAP is run automatically for human and mouse datasets; those meta-information lines would be present regardless of whether the VAP fields are present in the input SMAP.

Header

The single header line has nine fixed fields: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, and FORMAT. The last field is variable, and it depends on the sample name provided. The default is “Sample1”.

Variant data

CHROM: the chromosome on which the variant is called. The set of possible chromosomes is indicated in the meta-information section. For human datasets, the chromosome IDs in the SMAP are automatically converted into the “chrN” format. Chromosomes “23” and “24” are converted to “chrX” and “chrY”, respectively. For translocations each breakpoint is listed as a separate entry where CHROM lists one of the chromosomes with the partner breakpoint referenced in the MATEID INFO field.

1. POS: the starting position of the variant interval is indicated. The ending position is indicated as END in the INFO field. When converting the SMAP into VCF, the VCF converter attempts to take into account of nearby, potentially unresolved labels and to estimate breakpoint uncertainty accordingly. Therefore, the coordinates in POS and END may not correspond to coordinates in the SMAP. For example, for insertion and deletion calls, the SV breakpoints are output as the midpoint between the last aligned label and the next label. The uncertainty of the breakpoints is indicated in the CIEND and CIPOS fields. See Appendix G in Bionano Solve Theory of Operation Structural Variant Calling (PN 30110) for information about the calculation.

2. ID: these are output based on the SMAP entry IDs (“SMAP” followed by SMAP ID) and the CNV entry IDs (“CNV” followed by CNV ID). Breakpoints for inter-chromosomal translocations with be listed with a ‘bnd_’ prefix and a unique numeric suffix. For example, a translocation called with SMAP ID 4122 will have two breakpoint entries with IDs bnd_SMAP4122_1 and bnd_SMAP4122_2.

3. REF: because the Bionano optical mapping platform does not provide single-base level resolution, the precise base for a given variant is not relevant. “N” is output for all variants.

4. ALT: the variant type (defined in the meta-information section) is output.

5. QUAL: the variant confidence scores from the SMAP and CNV output are converted into Phred scale. The maximum confidence is capped at 20.

6. FILTER: “Masked” is output for masked calls. “LowConfidence” is output for variants that do not meet the minimum recommended confidence score. “PASS” is output for all other calls. Masking is performed by default during the SV and CNV calling steps using separate masks. The SV mask include regions where false positive translocation calls were made in control samples with no known translocations. These regions are often segmental duplication loci and cannot be aligned uniquely. The CNV mask include regions with elevated coverage noise, defined based on control samples with no known large CNV events. False positive CNV calls are more common in high coverage noise regions. SV and CNV calls overlapping with the masks are masked and of lower confidence. See Bionano Solve Theory of Operation Structural Variant Calling (PN 30110) for information about the masks and the masking procedure as well as recommended minimum confidence scores for each variant type.

7. INFO: the fields are defined in the meta-information section.

8. FORMAT: only “GT” is currently output for structural variant calls. CNV calls will have “CN” and “CNF” fields that record the rounded copy number for the gain or loss as an integer and the fractional copy number as a floating point.

9. Genotype field: this can be hemizygous (“1”, inferred based on copy number data), heterozygous (“0/1”), homozygous (“1/1”), or unknown (“./.”). If zygosity is not present in the input SMAP, “./.” is output. Hemizygous chromosomes are defined to be the ones where the average chromosome copy number is between 0.9 to 1.1.

Notes:

- Both intra-chromosomal fusions and inter-chromosomal translocations are represented by paired breakend (BND) entries linked by MATEID. The orientations of the breakends are converted based on the orientation column in the input SMAP and output according to the VCF specification.

- Unlike the uncertainty for other SV types, the uncertainty for CNV calls is fixed at 30 kbp and set based on empirical data. It is used for both CIPOS and CIEND. It is subject to change, as new methods for estimating the breakpoint uncertainty become available.

- The two-entry inversion calls in the input SMAP are converted into single-entry VCF lines. The smallest

coordinate in the two entries is taken to be POS, and the largest coordinate is taken to be END. POS and END then represent the outer bounds of where the inversion of interest might be.

Sample Header

Following is a sample header that includes definitions for all INFO and FILTER entries

```
##fileformat=VCFv4.2
##fileDate=2021-08-23
##source=Bionano Solve 3.7 variant annotation pipeline
##command=bionano_vcf_converter.py <args>
##sample=<sex=male>
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=BNGTYPE,Number=.,Type=String,Description="Original BNG variant type from
SMAP or CNV">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant
described in this record">
##INFO=<ID=MATEID,Number=.,Type=String,Description="ID of mate breakends">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between
REF and ALT alleles">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Breakpoint uncertainty for
start position POS">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Breakpoint uncertainty for end
position END">
##INFO=<ID=CT,Number=1,Type=String,Description="Breakpoint connection type">
##INFO=<ID=EXPERIMENT,Number=1,Type=String,Description="experiment_id from dbVar
submission of the experiment that generated this call">
##INFO=<ID=SAMPLE,Number=1,Type=String,Description="sample_id from dbVar
submission. Each call must have only one of either SAMPLE or SAMPLESET">
##INFO=<ID=SAMPLESET,Number=1,Type=Integer,Description="sampleset_id from dbVar
submission. Each call must have only one of either SAMPLE or SAMPLESET">
##INFO=<ID=OVERLAPGENES,Number=.,Type=String,Description="Set of genes overlapped
by structural variant">
##INFO=<ID=NEARGENE,Number=1,Type=String,Description="Nearest non-overlapping
gene">
##INFO=<ID=NEARGENEDIST,Number=1,Type=Integer,Description="Distance to nearest non-
overlapping gene">
##INFO=<ID=DGVOVERLAPS,Number=1,Type=Integer,Description="Number of overlapped
variants in DGV database">
##INFO=<ID=INPARENTS,Number=1,Type=String,Description="Found in parents' datasets">
##INFO=<ID=ISCN,Number=.,Type=String,Description="ISCN annotation">
##INFO=<ID=UCSC1,Number=1,Type=String,Description="UCSC web link 1">
##INFO=<ID=UCSC2,Number=1,Type=String,Description="UCSC web link 2">
##INFO=<ID=SAMPLETYPE,Number=1,Type=String,Description="SV sample type in VAP
pipeline">
##INFO=<ID=ALG,Number=1,Type=String,Description="Algorithm used in VAP">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural
variation">
##INFO=<ID=PCNTBNG,Number=1,Type=Float,Description="Percent of BNG control samples
with SV">
##INFO=<ID=PCNTBNGENZ,Number=1,Type=Float,Description="Percent of BNG control
samples with the same enzyme with SV">
##INFO=<ID=PCNTBNGHOM,Number=1,Type=Float,Description="Percent of BNG control
samples with homozygous SV">
##INFO=<ID=PCNTBNGHET,Number=1,Type=Float,Description="Percent of BNG control
samples with heterozygous SV">
##INFO=<ID=PCNTBNGSVAFR,Number=1,Type=Float,Description="Percent of AFR BNG control
samples with SV">
##INFO=<ID=PCNTBNGSVAMR,Number=1,Type=Float,Description="Percent of AMR BNG control
samples with SV">
```

```
##INFO=<ID=PCNTBNGSVEUR,Number=1,Type=Float,Description="Percent of EUR BNG control
samples with SV">
##INFO=<ID=PCNTBNGSVEAS,Number=1,Type=Float,Description="Percent of EAS BNG control
samples with SV">
##INFO=<ID=PCNTBNGSVSAS,Number=1,Type=Float,Description="Percent of SAS BNG control
samples with SV">
##INFO=<ID=PCNTBNGSVUNK,Number=1,Type=Float,Description="Percent of unknown BNG
control samples with SV">
##INFO=<ID=FAILCHIM,Number=1,Type=String,Description="Fail assembly chimeric
score">
##INFO=<ID=GENFUS,Number=0,Type=Flag,Description="Putative gene fusion">
##INFO=<ID=INCTRLASSM,Number=0,Type=Flag,Description="Found in control sample
assembly">
##INFO=<ID=INPAIRCTRL,Number=0,Type=Flag,Description="Found in paired control
sample CNVs">
##INFO=<ID=INPARASSM,Number=1,Type=String,Description="Found in parents'
assemblies">
##INFO=<ID=ZYG,Number=1,Type=String,Description="Zygosity of SV">
##INFO=<ID=ZYGPAIRASSM,Number=.,Type=String,Description="Zygosity in paired control
sample assembly">
##INFO=<ID=ZYGMASSM,Number=.,Type=String,Description="Zygosity in mother assembly">
##INFO=<ID=ZYGFASSM,Number=.,Type=String,Description="Zygosity in father assembly">
##INFO=<ID=SELFMOL,Number=0,Type=Flag,Description="Found in self molecules">
##INFO=<ID=INCTRLMOL,Number=0,Type=Flag,Description="Found in paired control sample
molecules">
##INFO=<ID=PARMOL,Number=1,Type=String,Description="Found in parents' molecules">
##INFO=<ID=SELFMOLCNT,Number=1,Type=String,Description="Self molecule count">
##INFO=<ID=CTRLMOLCNT,Number=1,Type=String,Description="Paired control sample
molecule count">
##INFO=<ID=MMOLCNT,Number=1,Type=String,Description="Mother molecule count">
##INFO=<ID=FMOLCNT,Number=1,Type=String,Description="Father molecule count">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=INS,Description="Insertion">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=BND,Description="Breakend">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for
imprecise events">
##FORMAT=<ID=CNF,Number=1,Type=Float,Description="Fractional copy number for
imprecise events">
##FILTER=<ID=LowConfidence,Description="Does not meet minimum recommended
confidence score for variant type">
##FILTER=<ID=Masked,Description="Masked due to low quality in problematic calling
regions">
```

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