Nabsys

SV-Verify™ Pipeline Overview

Contents

Legal Notice

This document and its contents are proprietary to Nabsys 2.0 LLC and are intended for the proper use of the product(s) described herein by customers/collaborators. This document shall not be disclosed, distributed, reproduced or copied without written permission by Nabsys 2.0 LLC.

Instructions for use of the product(s) described herein should be strictly followed by appropriately trained personnel to ensure proper safety and use. Nabsys 2.0 LLC does not assume any liability for the improper use of product(s) described herein. The SV-Verify™ Pipeline Overview is subject to change and will be updated as necessary.

For research use only. Not for use in diagnostic procedures.

Copyright

Revision

| 755-00024-001 |
Rev B | 2026-08

Manufactured by
Nabsys
60 Clifford St.
Providence, RI 02903

Warranty

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Nabsys disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Nabsys shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Nabsys and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end-user customers. Nabsys provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Chapter 1: SV-Verify™ Overview

Introduction

SV-Verify™ is an application for secondary analysis deployed to Nabsys Navigator. It can be used to confirm hypothesized variants in electronic genome mapping (EGM) data from OhmX™ Analyzers via targeted alignment. It compares EGM molecules to a wild-type reference and a user-supplied candidate structural variant (SV) panel. Its primary objectives are to determine SV frequency (SVFreq), presence/absence, and zygosity. SV-Verify offers efficient high-confidence SV confirmation for EGM Assemblies, and may also be used as an orthogonal tool for cytogenetics and sequencing technologies.

This document provides an overview of the pipeline; for complete usage instructions, refer to the **Nabsys Navigator User Guide, 755-00026-001**.

SV-Verify Concepts Overview

Structural variant (SV) detection through whole-genome *de novo* assembly is a standard approach, especially when prior knowledge of SVs is limited. SV-Discover™ supports this by assembling molecules into contigs and aligning them to a reference genome for SV discovery, typically requiring high-coverage data.

When high coverage data is not available, SV-Verify offers an alternative method to verify candidate SVs without assembly. This is achieved through competitive remapping, where molecules are aligned to both the wild-type allele and the candidate SV allele (construct).

SV-Verify converts each user-defined candidate SV into a hypothesized haplotype (construct), then maps all molecules to a combined reference containing both the wild-type and construct sequences. Mapping behavior indicates SV presence:

- **Absent SV**—All molecules align to the wild type
- **Heterozygous SV**—Molecules split roughly evenly between wild type and construct
- **Homozygous SV**—All molecules align to the construct

Observed ratios may deviate from the truth due to factors such as random molecule sampling and misalignment. SV-Verify accounts for this by checking the uniqueness of a candidate SV and is described in **Chapter 3: Uniqueness Checking**.

SV Frequency, the ratio of molecules aligned to the construct versus all molecules aligned to either the construct or wild type, is used to infer the genotype of an SV event. When paired with the uniqueness evaluation of each candidate SV, a user can interpret the presence or absence of an SV from their EGM data with high accuracy.

For example, to verify a candidate deletion (DEL), SV-Verify constructs a hypothetical haplotype by joining the 250 kb regions flanking the breakpoints in the reference (**Figure 1**). The resulting construct has a single breakpoint, and molecules spanning this junction are considered informative. For construct designs for other SV types, see **Chapter 2: SV Panels**.

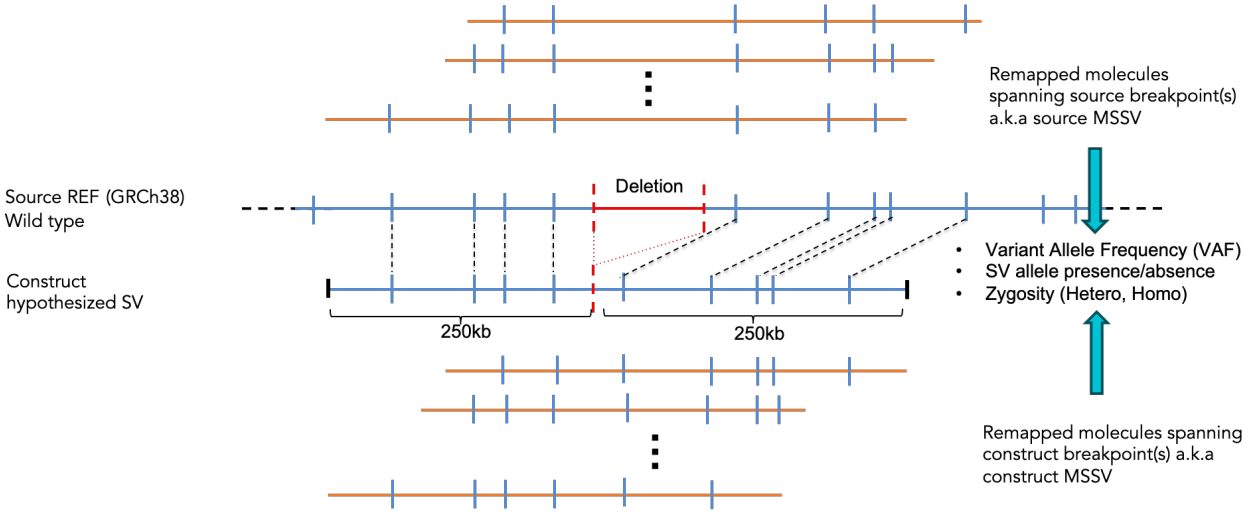

Figure 1. Example of a molecule alignment supporting a heterozygous deletion. Blue vertical lines represent labels in both the reference and construct.

Chapter 2: SV Panels

Putative SV List

SV-Verify starts with a putative SV list, or a list of SV hypotheses based on orthogonal techniques or anticipated events in a sample. SV hypotheses are relative to a specified wild-type reference, such as GRCh38. The putative SV list must be appropriately formatted and contain the type of event, the start position, and the size of the SV, at a minimum. The types of events that may be hypothesized are:

- Deletions (**DEL**)
- Insertions with unknown sequence or label pattern (**INS_N**)
- Insertions with known sequence or label pattern (**INS_wSeq**)
- Inversions (**INV**)
- Tandem Duplications (**TandemDUP**)

The putative SV list is built as a tab-delimited text file. The first three lines in the file are denoted as metadata using two slashes (//) and must contain the following:

```
```
//FileType: GeneralizedPutativeSVDefinition
//FileFormatRevision: 1
//ReferenceFile: <PATH_TO_REFERENCE_GENOME_FASTA_FILE> OR <None>
```

Below the metadata, there will be candidate SV information described in 12 columns. All SV types require values for columns 1–5. Additional data is required in columns 6–12 based on SV type. For BPF (breakpoint fusion) SVs, values in columns 6–9 are also required. For `INS_wSeq`, a string in column 10 is required.
```

Table 1. Putative SV list file format guide.

Column #	Column Name	Units	Variable Type	Description
1	`SVIndex`	N/A	Int	Unique # ID of putative SV; **required** for all SV types.
2	`SVType`	N/A	String	SV type; acceptable values include `DEL`, `INS_wSeq`, `INS_N`, `INV`, `BPF`, `TandemDUP`; **required** for all SV types.
3	`ChrA`	N/A	Int	Chromosome index/ordinal in the base reference FASTA file; for `BPF`, this is the chromosome of the first breakpoint; **required** for all SV types.
4	`StartA`	bp	Int	SV starting position; for `BPF`, this is the position of the first breakpoint; **required** for all SV types.
5	`SizeA`	bp	Int	SV size; acceptable values include `None` for `BPF` and positive integers for `DEL`, `INS`, `INV`, `TandemDUP`; **required** for all SV types.
6	`SideA`	N/A	String	Side (left or right) of the flanking segment of the first breakpoint that joins with the second breakpoint for `BPF` acceptable values include `L` or `R` for `BPF` and required to read `None` for other SV types.
7	`ChrB`	N/A	Int	Chromosome index/ordinal in the base reference FASTA file of the second breakpoint; acceptable values include positive integers for `BPF` and required to read `None` for other SV types.
8	`StartB`	bp	Int	Starting position of the second breakpoint; acceptable values include positive integers for `BPF` and required to read `None` for other SV types.
9	`SideB`	N/A	String	Side (left or right) of the flanking segment of the second breakpoint that joins with the first breakpoint for `BPF`; acceptable values include `L` and `R` for `BPF` and required to read `None` for other SV types.
10	`Seq`	N/A	String	Known inserted sequence for `INS_wSeq` type SV; required to read `None` for other SV types. Note that for `INS_N`, N-bases of length equal to `SizeA` is used as the inserted sequence.
11	`SVZygosity`	N/A	String	Zygosity of the putative SV, acceptable values include `HOMO`, `HET`, and `Unknown`. If `HOMO`, `SVHaplotype` should be `HAP12`; if `HET`, `SVHaplotype` should be `HAP1`, `HAP2`, or `Unknown`; if `Unknown`, `SVHaplotype` should be `Unknown`.
12	`SVHaplotype`	N/A	String	Haplotype of the putative SV, acceptable values include `HAP1`, `HAP2`, `HAP12`, and `Unknown`. If `HAP1` or `HAP2`, `SVZygosity` should be `HET`; if `HAP12`, `SVZygosity` should be `HOMO`; if `Unknown`, `SVZygosity` should be `HET` or `Unknown`.

Chapter 3: Uniqueness Checking

The uniqueness of each allele at an SV locus relative to other alleles and the broader genome directly impacts alignment accuracy, which in turn affects the calculated SV frequency and the reliability of SV detection and zygosity inference.

To address this, SV panel generation includes a uniqueness checking step. SV-Verify uses empirical error models to generate *in-silico* (ISG) molecules from both reference and variant alleles. These ISG molecules are then remapped to both the reference and variant constructs to measure how distinctly each allele can be identified. This enables the calculation of well-defined uniqueness metrics and expected SV frequencies for each possible genotype.

Two key metrics are computed per SV locus and genotype:

1. **F1 Score [0, 1]**—Measures alignment accuracy by balancing false positives and false negatives; higher values indicate greater uniqueness and lower risk of misalignment
2. **SVFreq Expectation**—Predicts the SV frequency for a given genotype based on ISG remapping; a highly unique SV will yield an expected SVFreq close to the true value

SV-Verify currently supports **constitutional SVs only**, so simulated genotypes are limited to VAF values of 0 (0/0), 0.5 (0/1), and 1 (1/1).

The table below illustrates this with two autosomal SVs. **SV ID 55** has F1 scores near 1 across all genotypes, suggesting high alignment accuracy and SVFreq expectations close to true values. In contrast, **SV ID 66** shows lower F1 scores and greater deviation in SVFreq expectations, indicating reduced uniqueness and higher likelihood of misalignment.

Table 2. Examples of unique SV hypotheses (SV ID 55) and non-unique SV hypotheses (SV ID 66) which may impact the detection of certain variants.

	genotype = 0/0 true SVFreq = 0		genotype = 0/1 true SVFreq = 0.5		genotype = 1/1 true SVFreq = 1	
SV ID	**F1**	**Expected SVFreq**	**F1**	**Expected SVFreq**	**F1**	**Expected SVFreq**
55	0.98	0.06	0.97	0.52	0.96	0.97
66	0.46	0.35	0.53	0.62	0.63	0.73

Chapter 4: Alignment-Based SV Verification

Deletions and Insertions

Deletions (DEL) and Insertions (INS) are the simplest SV types and are well-suited for competitive remapping using SV-Verify. For both types, the construct is built by merging the two flanking regions. DEL events have two informative breakpoints on the reference genome and one on the construct; INS events have one on the reference genome and two on the construct. Molecules bridging these breakpoints are considered supporting evidence for the candidate SV.

For candidate insertions, SV-Verify supports two types: those with a provided sequence (INS_wSeq) and those without (INS_N). For INS_wSeq, constructs are built by inserting the provided sequence where labels are placed when the sequence contains label recognition sites (**Figure 2**).

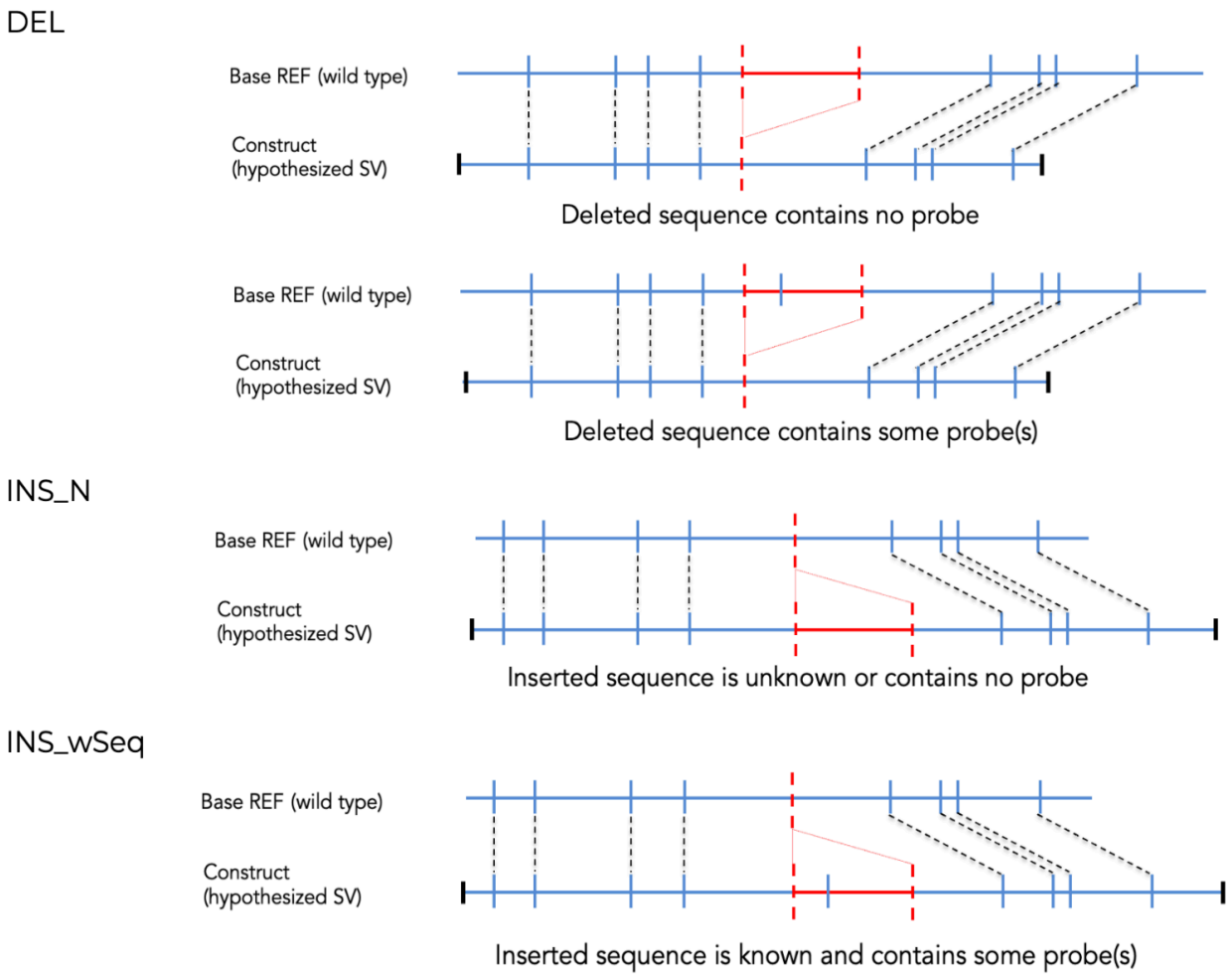

Figure 2. Visual representation of a deletion and two Insertion types as constructs.

A known limitation is that as the N-base region length increases in INS_N candidate SVs, the number of supportive molecules on the construct decreases, leading to underestimation of SV allele coverage and a bias in SVFreq toward 0 for 0/1 and 1/1 genotypes (**Figure 3**). For large insertions (> 100 kb), it is recommended to define the hypothesis as an **INS_wSeq** SV type. Alternatively to sequencing data, a genome mapping contig may be used if appropriately converted into the required format.

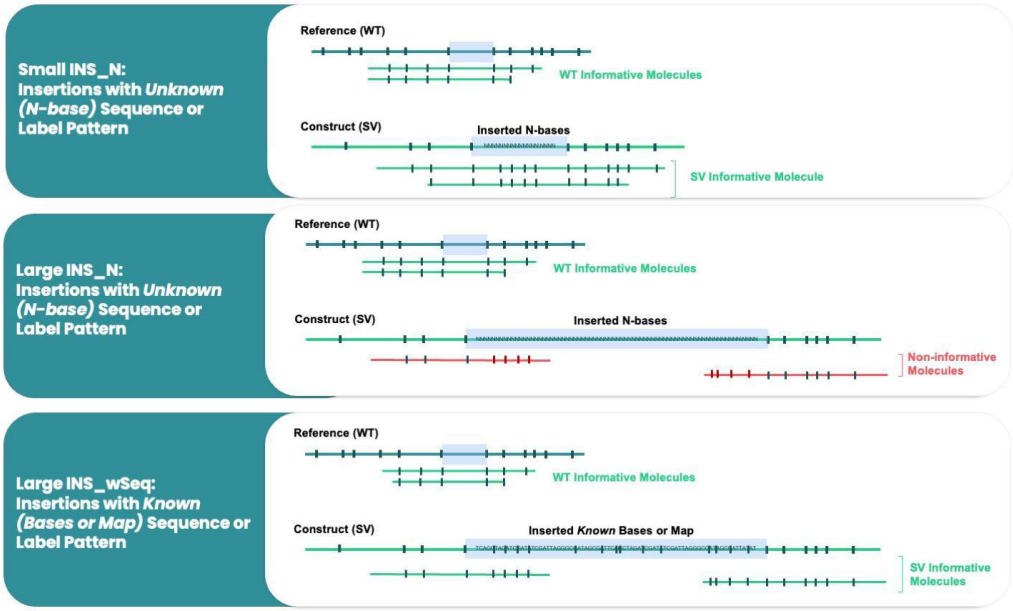

Figure 3. Guide for choosing the best Insertion SV type to ensure informative molecules can be found.

Inversions

Inversions (INV) are also suitable SV types for competitive remapping using SV-Verify. Given the two breakpoints on the reference sequence, the reverse complementary sequence of the inverted region will be joined with the two flanking regions on the reference to form the construct, resulting in two construct breakpoints. At least one label should be contained in the inverted region so that the built construct would be different from the reference at the map level. Informative molecules, referred to as Molecules Spanning SV Breakpoint (MSSV), must bridge the breakpoints (dotted red lines) where the inversion occurs, but are not required to stretch the entire length of the event (**Figure 4**).

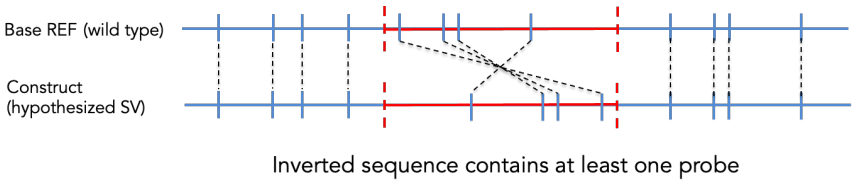

Figure 4. Visual representation of Inversion SVs as constructs. Breakpoints are depicted by red dotted lines and labels are depicted as blue vertical lines.

Tandem Duplications

SV-Verify supports verification of tandem duplications, where a sequence is duplicated and inserted immediately after its original location—regardless of label presence within the duplicated region (**Figure 5**).

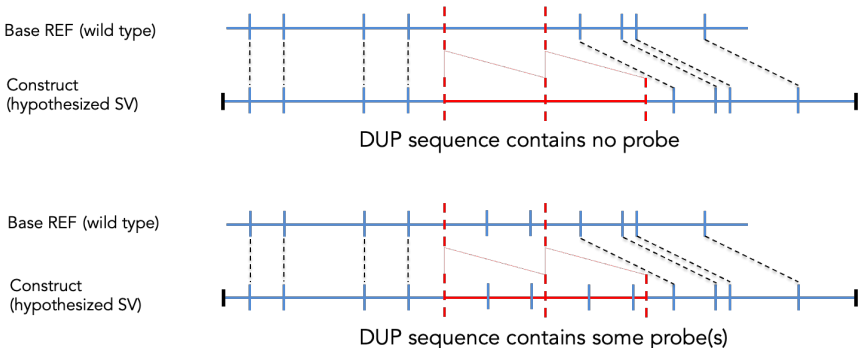

Figure 5. Visual representation of tandem duplications as constructs with (top) or without (bottom) duplicated labels.

A key distinction from other SV types is that **not all** MSSVs are informative in the case of TandemDUPs. For the **wild-type allele**, only MSSVs with molecules spanning **both outer breakpoints** on the reference are considered informative (**Figure 6**). For the **SV allele**, only MSSVs with molecules spanning the **middle breakpoint** on the construct–where the original and duplicated segments join–are informative. As the candidate duplication (TandemDUP) length increases, fewer supportive molecules are found on the reference, leading to an underestimation of wild-type allele coverage and causing the SVFreq to be biased toward 1 for both 0/0 and 0/1 genotypes.

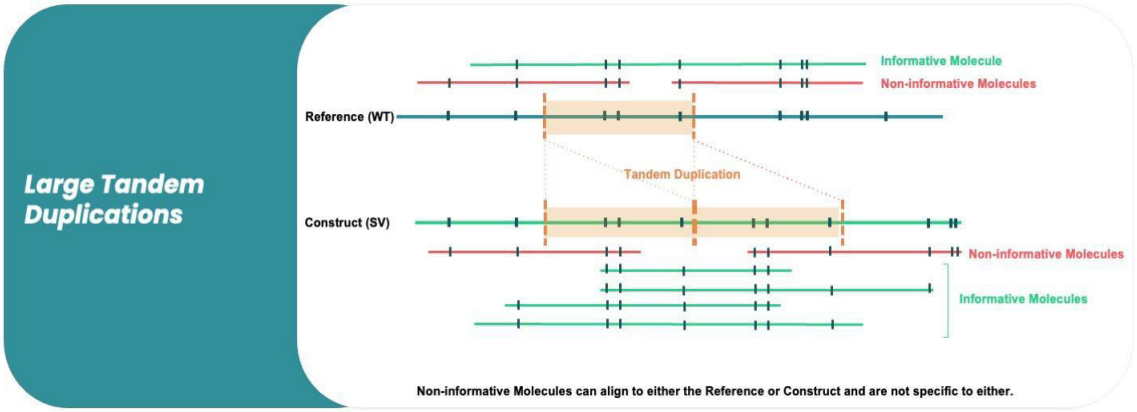

Figure 6. Depiction of informative and non-informative molecules for the reference and construct in a tandem duplication candidate SV.

Breakpoint Fusions

A breakpoint fusion (BPF) SV type refers to the fusion of two disjointed sequences in the source reference that are joined in the construct. There are four BPF subtypes based on the orientation of the joined sequences (**Figure 7**). **SideA** and **SideB** indicate whether the **left** or **right** side of each breakpoint is joined; these are distinguished given the final fusion orientation shown in the figure.

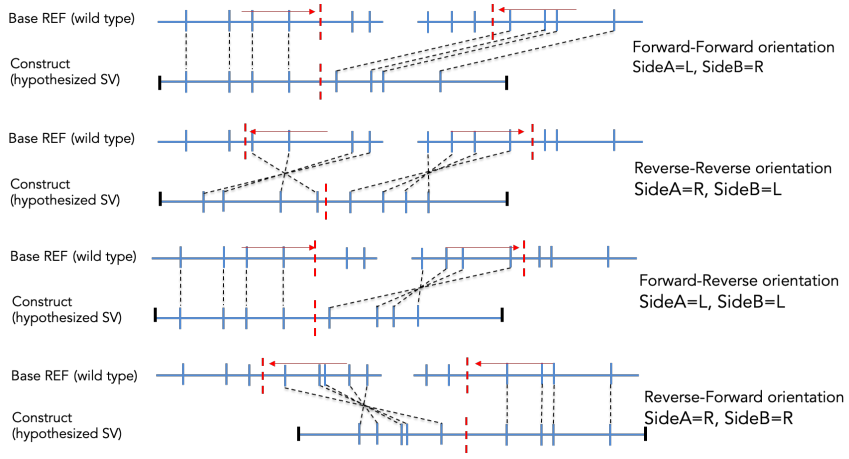

Figure 7. Example orientations of various BPF SVs.

Furthermore, complex SVs, such as translocations and distal duplications, could be decomposed into a set of verifiable BPF events. By combining the verification result of such BPFs, the associated complex SV could be verified. However, this is not an automated detection feature of SV-Verify at this time. Please contact Nabsys Support for assistance.

Unsupported SV Types

Translocations

Translocations involve the relocation of a sequence to a different genomic location. In practice, a translocation can be represented as a set of related **breakpoint fusion (BPF)** events. For example, a simple non-reciprocal translocation may be decomposed into three BPF events, illustrated by the three colored arrow pairs in **Figure 8**. Other translocation subtypes can be similarly broken down into sets of BPFs.

Verifying these BPF events collectively can provide meaningful insight into the underlying translocation.

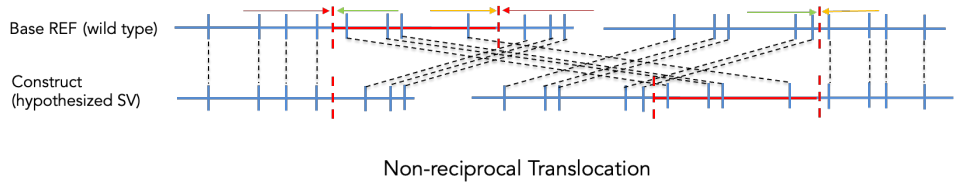

Figure 8. Example breakpoint fusions associated with a non-reciprocal translocation that might be detected by SV-Verify to provide evidence of the translocation SV. Each pair of arrowed lines is assigned a distinct color to denote the corresponding BPF: red, green, and orange.

CNV Losses and Gains

Copy number losses can be verified as **DEL** SV types in SV-Verify. Copy number gains, however, are more complex. While they are often represented as **tandem duplications (TandemDUP)**, other configurations are possible depending on the genomic context. Therefore, accuracy of verifying copy number gains in SV-Verify is limited by how much is known about the candidate SV.

SV Frequency Computation

Molecules supporting either the reference (WT) or construct (SV) are computed. SV Frequency is calculated as the proportion of molecules aligning to the construct out of the total number of molecules aligning to either the reference or the construct.

Chapter 5: Interpretation

Genotyping

The genotyping mode used in the sample analysis is output in the `Results.zip` for each sample. It is found on line 9 in the `<sample_name>_cr_combined_result_report.txt` file header. Earlier models for genotyping in SV-Verify v1.0.1 and earlier employed Genotyping Mode 1, solely based on observed SV Frequency. For more information on SV-Verify v1.0.1, please contact Nabsys Support.

Genotyping Mode 2 uses a Bayesian inference model to determine the most likely genotype for each structural variant (SV) detected in a sample. The method integrates multiple sources of evidence from the run to estimate the probability of each possible genotype and reports both a genotype call and a confidence score.

The model evaluates several factors when assessing each SV:

- Coverage of the SV allele and the reference (wild-type) allele
- Observed SV Frequency
- Expected SVFreq, adjusted based on how uniquely the SV can be detected in the genome
- A genotype prior, which reflects expected genotype frequencies

Using this information, the Bayesian model calculates the probability of each possible genotype and reports the genotype with the highest probability.

Autosomal Structural Variants

For variants located on autosomes, the model evaluates the following possible genotypes:

- 0/0 – No copies of the variant (reference genotype)
- 0/1 – One copy of the variant (heterozygous)
- 1/1 – Two copies of the variant (homozygous)

Structural Variants on Sex Chromosomes

For variants located in haploid regions of sex chromosomes, only two genotypes are considered:

- 0/0 – Variant absent
- 1/1 – Variant present

The genotype with the highest probability is reported as the final genotype call, along with a confidence score indicating the strength of the supporting evidence.

Confidence Score

For each structural variant (SV), the model evaluates all possible genotype states and calculates the probability of each one. The confidence score is the probability of the genotype with the highest probability.

For autosomal SVs, this corresponds to the maximum posterior probability across 0/0, 0/1, and 1/1. For sex chromosome SVs (haploid regions), this corresponds to the maximum posterior probability across 0/0 and 1/1.

Higher confidence reflects stronger statistical support for the inferred genotype relative to alternative genotype states.

Chapter 6: Performance

Methodology

To establish performance baselines for key SV types, we simulated variants and molecule data using an *in silico* diploid genome model based on GRCh38. The study included five SV types: insertions with novel sequence (INS_wSeq), deletions (DEL), inversions (INV), tandem duplications (TandemDUP), and breakpoint fusions (BPF) representing non-reciprocal translocations.

Each SV type was simulated in three genotypes—negative (absent), heterozygous (HET), and homozygous (HOMO)—in a 2:1:1 ratio to ensure balanced representation across presence/absence and zygosity. HET variants were placed on one autosomal chromosome copy, while HOMO variants were placed on both. All positive SVs were randomly positioned on autosomes, with a minimum spacing of 500 kb between events to prevent regional interference. SVs were excluded from N-base gaps and placed at least 250 kb away from large low-complexity regions.

A total of 900 DEL and 900 INS_wSeq were simulated, ranging from 300 bp to 1 Mb, at both 50X and 150X coverage. An equal number of true negative hypotheses were also tested. INV and TandemDUP were simulated across a size range of 20 kb to 1 Mb, with 800 variants of each type and an equal set of true negatives. For BPF, 600 true positive non-reciprocal translocations were generated, including an even mix of inter- and intra-chromosomal, as well as inverted and non-inverted events. An equivalent number of true-negative events were also tested. Since each translocation yields three BPFs, this resulted in a total of 3,600 BPFs.

To create a realistic sample context, a diploid reference genome was assembled with autosomes, one X chromosome, and one Y chromosome, with HET and HOMO SVs embedded accordingly. ***In silico*-generated (ISG) molecule data** were then produced for each coverage level. This simulated dataset enables robust assessment of SV detection performance across SV type, size, genotype, and coverage.

Genotyping Mode 1 is used to assess current performance, although updated performance metrics will soon be available for Genotyping Mode 2. We do not anticipate significant deviation from the presented performance standard based on initial evaluations.

ISG Data

There are several advantages to using *in silico*-generated (ISG) data for performance evaluation. Although ISG data cannot fully replicate the complexity of real biological samples—including the presence of polymorphisms, missing or extra labels, or other variability—it provides a controlled sample in which a large and diverse set of SVs can be systematically tested. This enables high-throughput benchmarking across a wide range of variant types, sizes, and zygosities, increasing statistical power and supporting robust performance characterization. Additionally, because ISG data is derived from a known reference genome, every variant has a clearly defined ground truth. This allows for precise evaluation of sensitivity, specificity, and zygosity classification across all variant types.

Currently, all performance metrics are based on ISG data alone. However, future updates to this document may include concordance results from real biological samples to further support validation in applied settings.

Performance Summary

- At 150X coverage, SV-Verify can detect the following heterozygous or homozygous variants:
- Indels > 500 bp with > 90% accuracy
- Inversions > 20kb with > 95% accuracy
- Tandem duplications 20–100 kb with > 95% accuracy
- Breakpoint fusions with > 95% accuracy

Additionally, many SVs can be detected with high accuracy with as little as 50X coverage and smaller indels (300–500 bp) may be detected with marginally lower accuracy. Expanded tandem duplication performance is in development. For more information on confirmation of SVs of varying size, see **Chapter 4: Alignment-Based SV Verification**, or contact Nabsys Support.

Technical Support

Nabsys
60 Clifford Street
Providence, Rhode Island
02903 USA

support@nabsys.com

For more information, visit nabsys.com.