Supplementary Figures

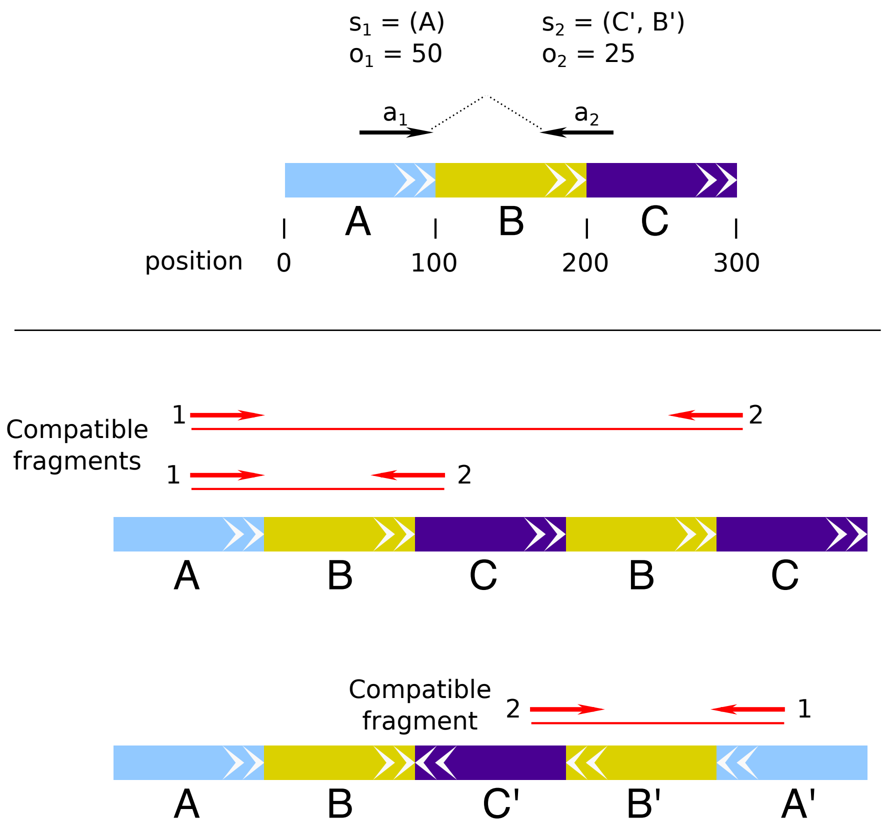


Figure 3: Illustration of fragment model (top) and compatible fragments concept (bottom). The paired alignment (a_1, a_2) has two compatible fragments under $\theta_1 = ABCBC$ and one compatible fragment under $\theta_2 = ABC'B'A'$

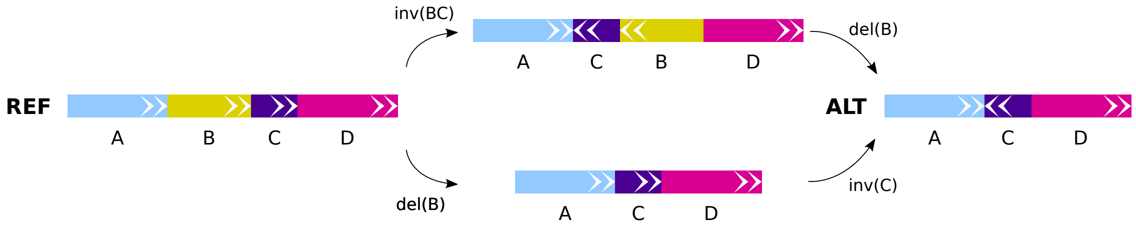


Figure 4: In general, there may not be a unique sequence of interval operations to transform the reference sequence (left) into a complex SV (right).

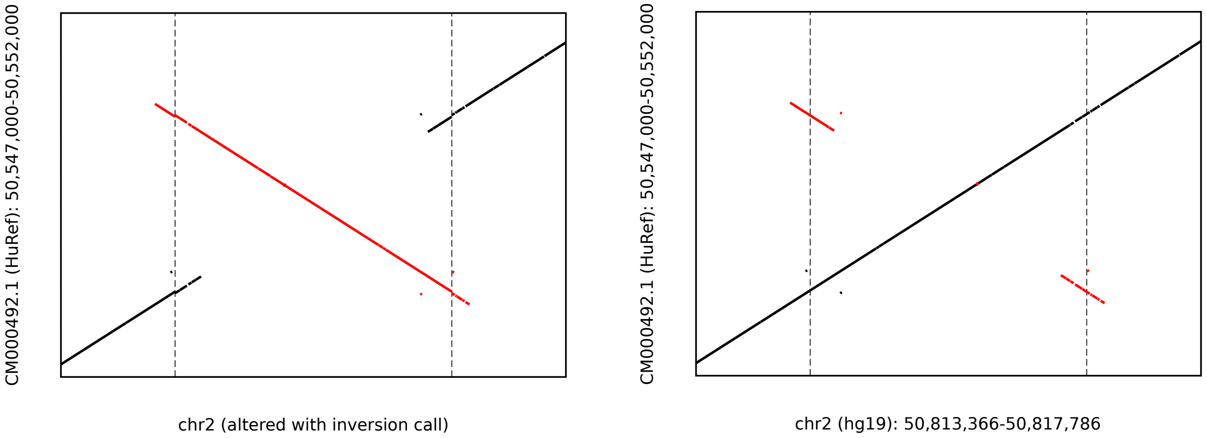


Figure 5: An inversion call (left, indicated by dotted lines) that does not match the HuRef assembly, and thus should not be considered validated according to our criteria. The alignment to the reverse strand (red) is high quality, with 160 bp of flanking sequence on each side. However, direct comparison between the hg19 reference and HuRef contig sequence shows there is no inversion present (right).

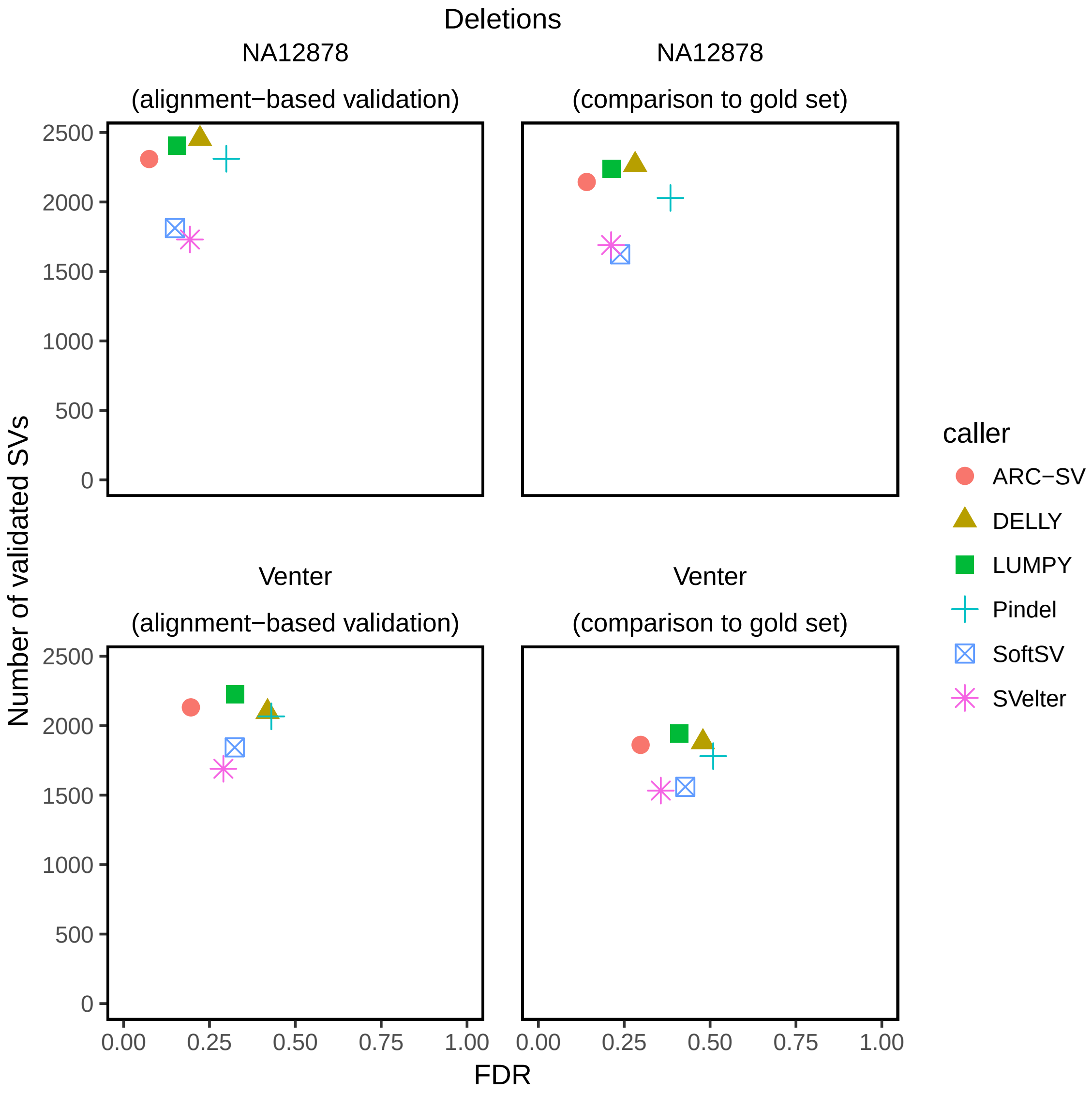


Figure 6: False discovery rate and sensitivity of deletion detections in Venter and NA12878, as evaluated by both validation methods. Results are prior to filtering by random forests.

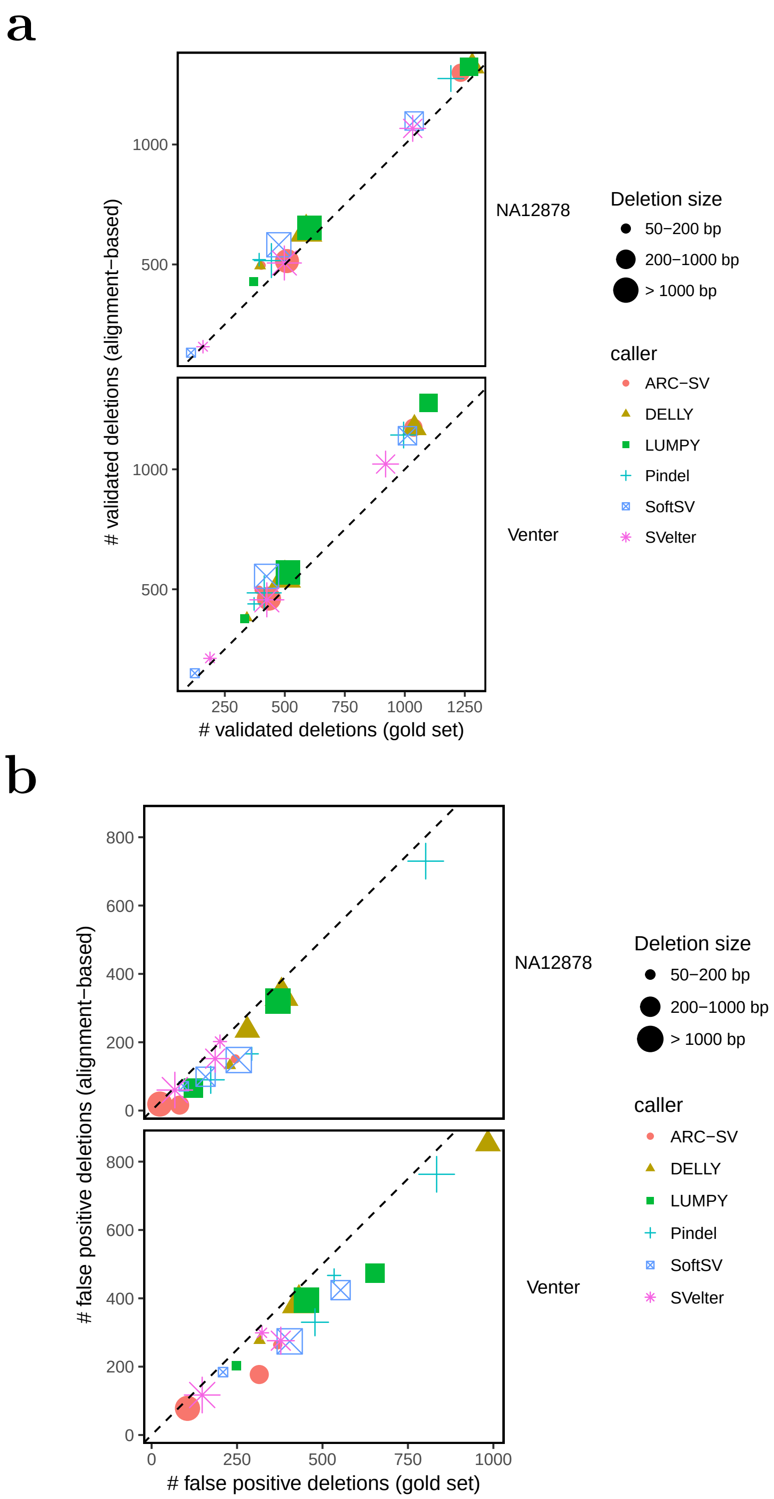


Figure 7: Comparison of (A) true positive and (B) false positive deletion counts based on our alignment-based validation and gold set overlap. Results are stratified by SV caller and SV size range. Results are prior to filtering by random forests.

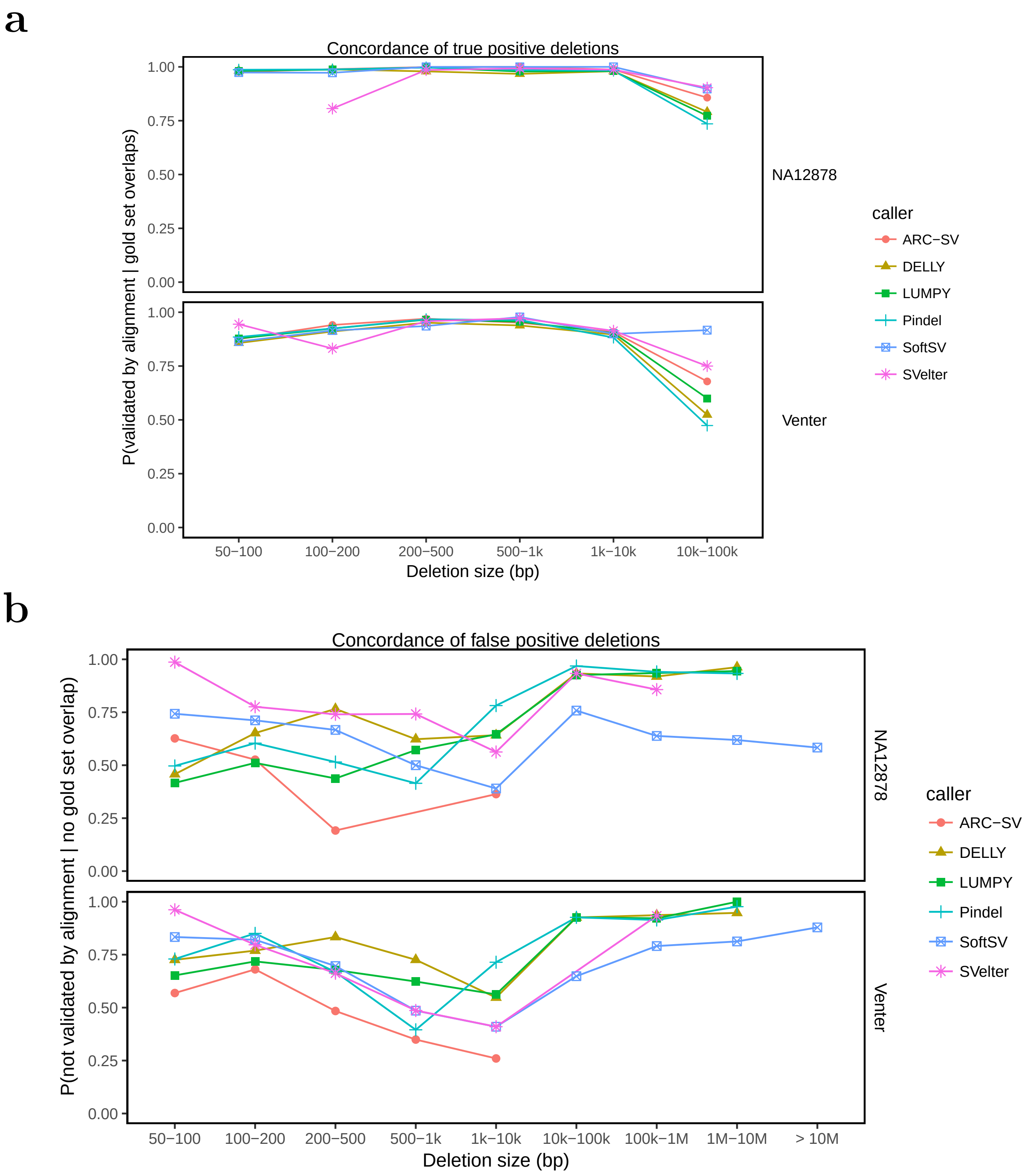


Figure 8: Concordance of (A) true positive and (B) false positive deletions between our alignment-based validation and gold set overlap (higher is better). Results are stratified by caller and SV size range, and only bins with more than 10 deletions are displayed. Results are prior to filtering by random forests.

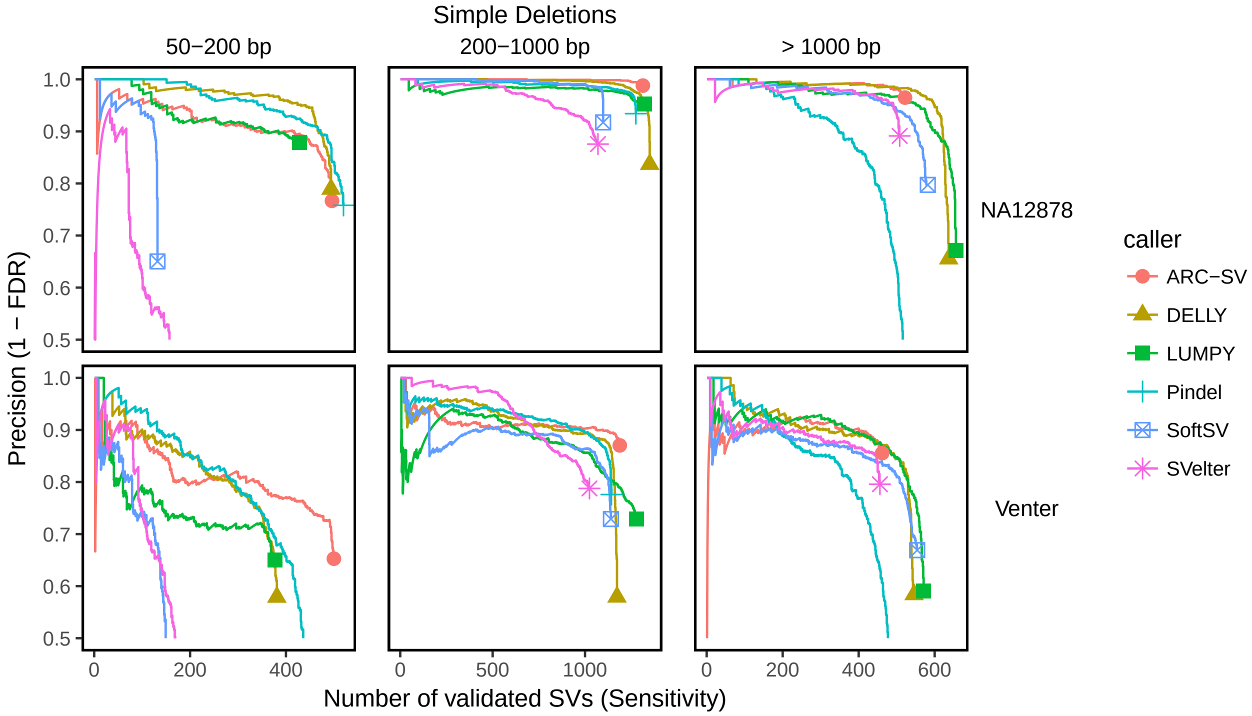


Figure 9: Sensitivity and precision of deletion calls across each size range. The results without random forest filtering (points) are the same as those given in Supp. Table 1. Less reliable calls are removed as we increase the stringency of the filter.

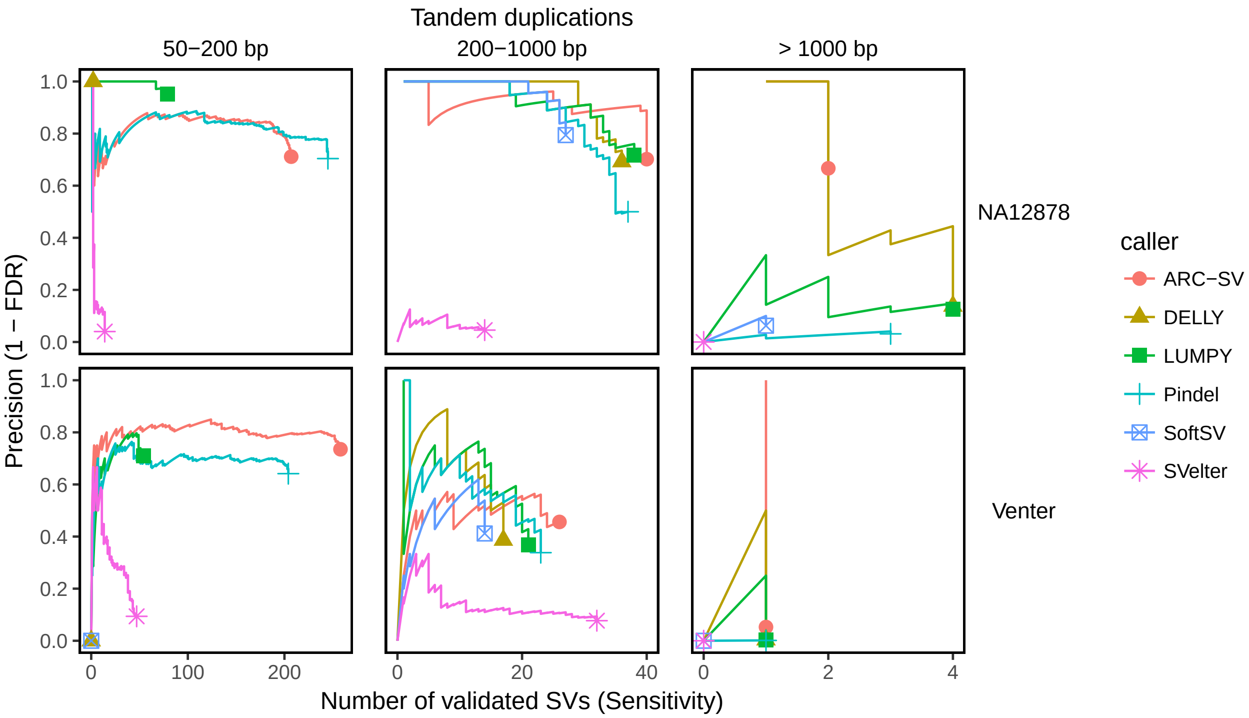


Figure 10: Sensitivity and precision of tandem duplication calls across each size range. The results without random forest filtering (points) are the same as those given in Supp. Table 1. Less reliable calls are removed as we increase the stringency of the filter. (Note the lack of data in the "> 1000 bp" category — the most sensitive callers achieve only 5 true detections across both samples.)

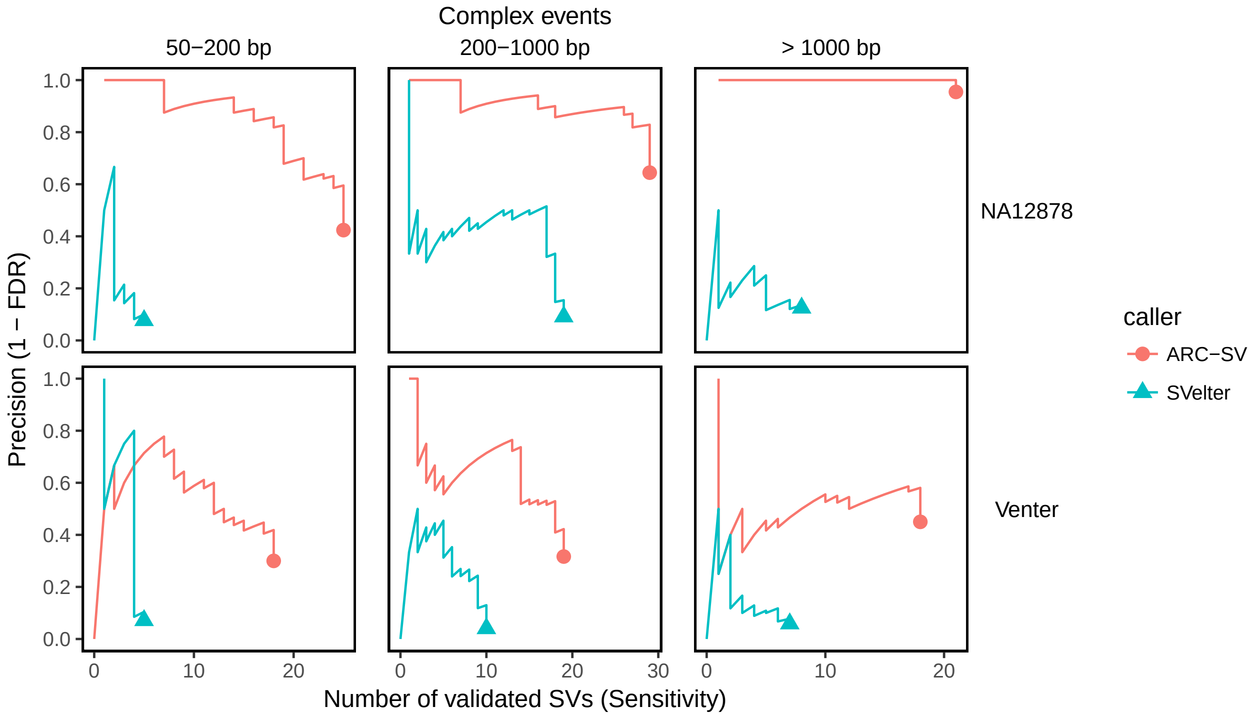


Figure 11: Sensitivity and precision of complex SV calls across each size range. The results without random forest filtering (points) are the same as those given in Supp. Table 1. Less reliable calls are removed as we increase the stringency of the filter.

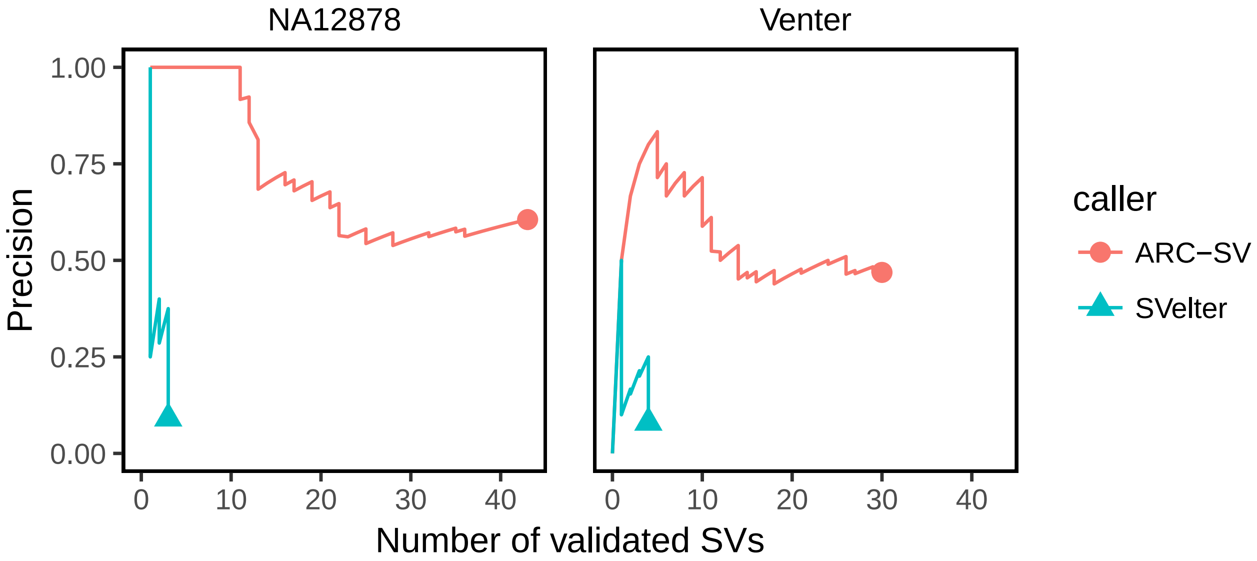


Figure 12: Sensitivity and precision of compound SV calls. The results without random forest filtering (points) are the same as those given in Supp. Table 1. Less reliable calls are removed as we increase the stringency of the filter.

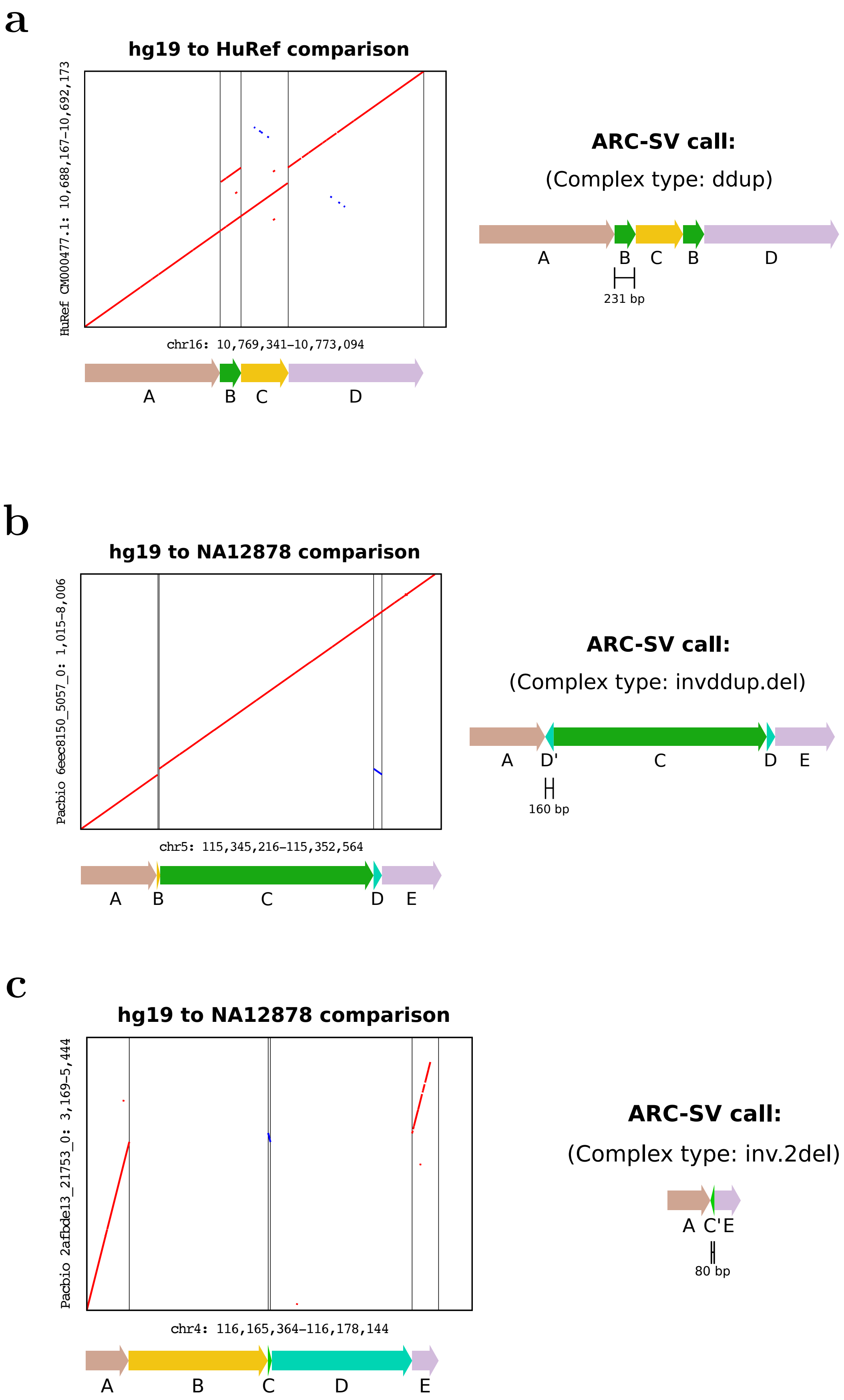


Figure 13: Validated complex ARC-SV calls: (A) interspersed duplication; (B) inverted interspersed duplication with deletion; (C) inversion with two flanking deletions

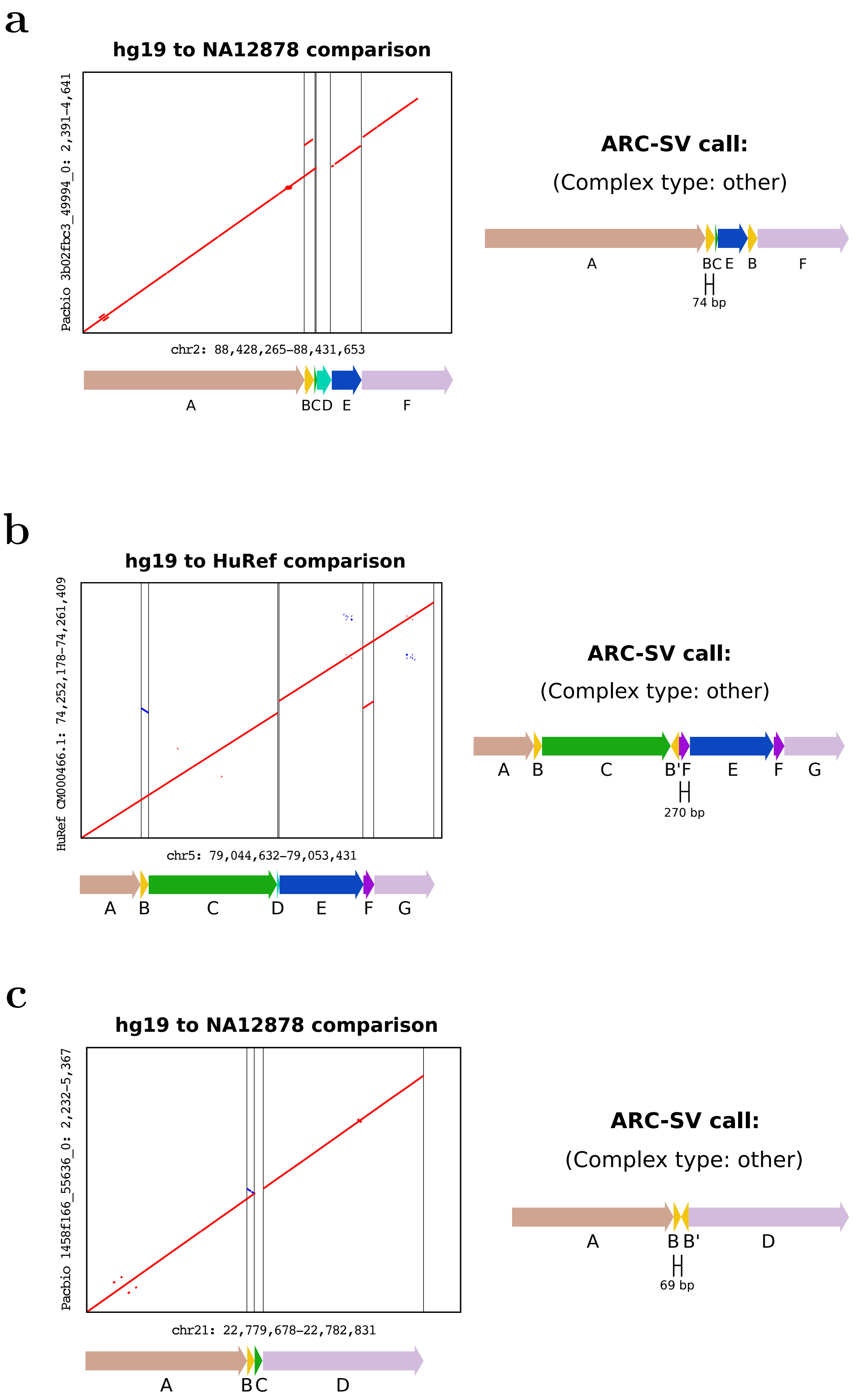

a
hg19 to NA12878 comparison
Pacbio 3b02fbc3_49994_0: 2,391–4,641
chr2: 88,428,265–88,431,653
A
BCD E
F
ARC-SV call:
(Complex type: other)
A
BC E
B
F
74 bp
b
hg19 to HuRef comparison
HuRef CM000466.1: 74,252,178–74,261,409
chr5: 79,044,632–79,053,431
A B C D E F G
ARC-SV call:
(Complex type: other)
A B C B'F E F G
270 bp
c
hg19 to NA12878 comparison
Pacbio 1458f166_55636_0: 2,232–5,367
chr21: 22,779,678–22,782,831
A BC D
ARC-SV call:
(Complex type: other)
A B B' D
69 bp

Figure 14: Examples of validated complex ARC-SV calls not fitting into our named categories.

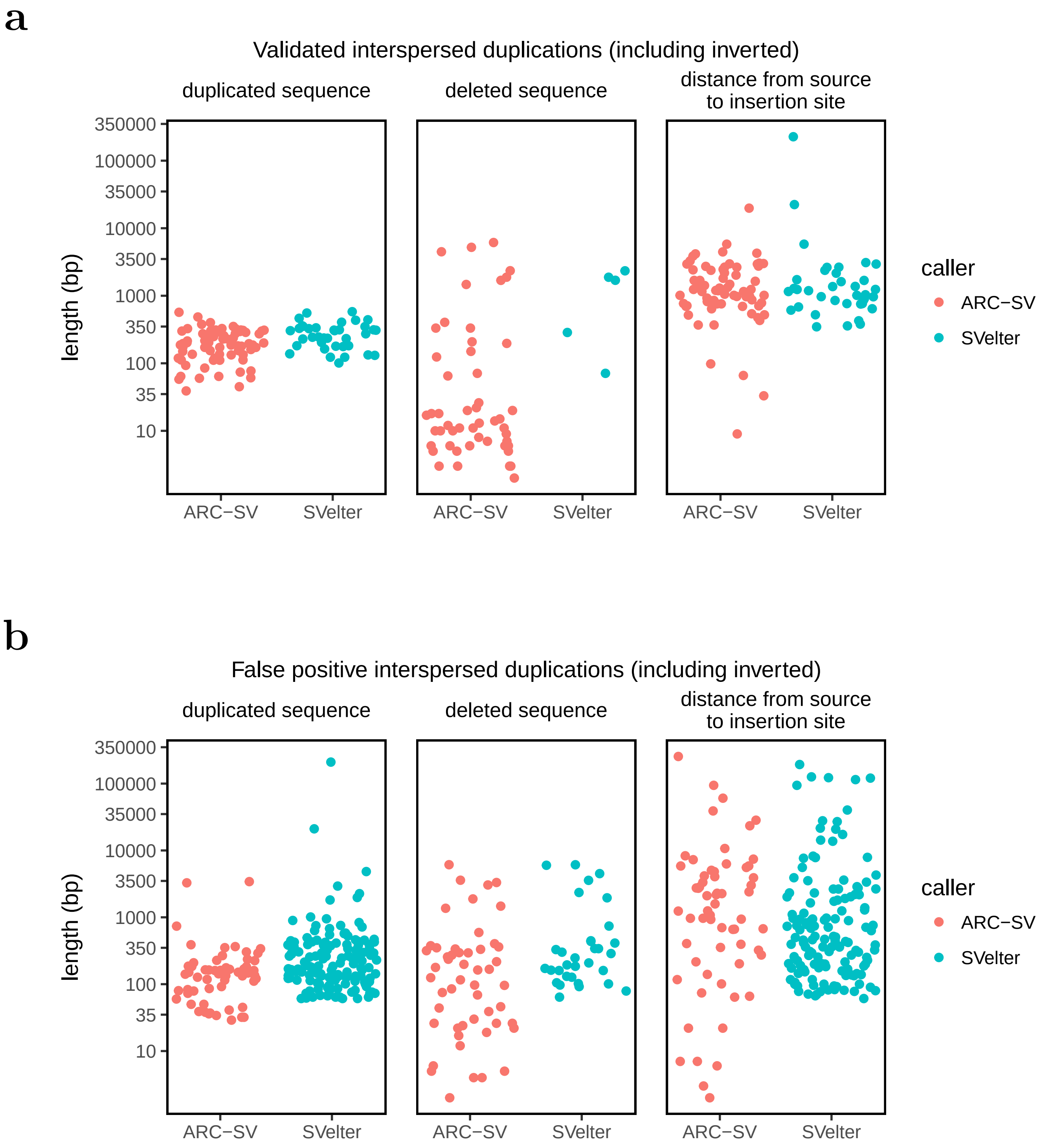


Figure 15: Sizes of interspersed (non-tandem) duplications. ARC-SV has 71 validated detections and 61 false positives; SVelter has 37 validated and 140 false positives. Limiting to interspersed duplications with deleted sequence (middle panels), ARC-SV has 50 validated detections and 51 false positives; SVelter has 5 validated and 32 false positives.

		ARC-SV		SVelter		DELLY		LUMPY		Pindel		SoftSV	
		TP	Prec										
	Complex	**75**	**59%**	32	9%								
	Compound	**43**	**60%**	3	9%								
NA12878	Deletion	2331	**92%**	1735	81%	**2481**	77%	2407	84%	2311	70%	1812	85%
	Inversion	6	**50%**	9	26%	**14**	3%	11	15%	9	1%	10	22%
	Tandem dup	249	71%	28	4%	42	51%	121	**72%**	**285**	55%	28	56%
	Complex	**55**	**34%**	22	5%								
	Compound	**30**	**46%**	4	8%								
Venter	Deletion	2150	**80%**	1692	71%	2100	58%	**2226**	67%	2068	57%	1844	68%
	Inversion	5	**71%**	6	11%	**11**	3%	10	15%	8	1%	8	8%
	Tandem dup	**285**	**66%**	79	8%	18	6%	76	20%	228	20%	14	6%

Table 1: Overall SV calling results. TP = # validated SVs; Prec = precision = 1 - FDR. Results are before random forest filtering, and excluding simple and tandem repeat regions. Blank cells indicate that no calls were made.

		ARC-SV		SVelter		DELLY		LUMPY		Pindel		SoftSV	
		TP	Prec										
	Complex	**22**	**13%**	13	5%								
	Compound	**9**	**15%**	3	9%								
NA12878	Deletion	741	**61%**	488	54%	529	46%	407	42%	**1822**	29%	242	37%
	Inversion	0	0%	0	0%	0	0%	**4**	**13%**	4	2%	0	0%
	Tandem dup	246	**40%**	116	22%	105	29%	112	19%	**915**	17%	53	24%
	Complex	21	**17%**	**33**	9%								
	Compound	**28**	**20%**	5	8%								
Venter	Deletion	944	**57%**	637	54%	548	40%	721	48%	**1834**	34%	347	33%
	Inversion	1	**50%**	0	0%	2	3%	3	7%	3	1%	**5**	5%
	Tandem dup	304	**40%**	288	30%	29	7%	53	7%	**671**	14%	87	11%

Table 2: Results for SV calls within the excluded tandem and simple repeat regions.

		ARC-SV			SVelter		
		TP	N	Prec	TP	N	Prec
NA12878	ddup	**8**	9	**88%**	**8**	50	16%
	ddup.del	**14**	23	**60%**	1	11	9%
	dup.del	**3**	7	**42%**	0	14	0%
	dup.dup				0	32	0%
	inv.2del	**14**	14	**100%**	1	5	20%
	inv.del	3	6	**50%**	**7**	21	33%
	invddup	3	4	**75%**	**12**	23	52%
	invddup.del	**18**	23	**78%**	1	4	25%
	invddup.del.inv	**4**	4	**100%**	0	1	0%
	invddup.inv						
	invdup				**1**	39	**2%**
	other	**8**	36	**22%**	1	130	1%
	trans				0	6	0%
	trans.inv				0	1	0%
Venter	ddup	**6**	10	**60%**	4	48	8%
	ddup.del	**7**	29	**24%**	2	14	14%
	dup.del	**1**	5	**20%**	0	28	0%
	dup.dup	**1**	1	**100%**	0	56	0%
	inv.2del	**7**	14	**50%**	0	1	0%
	inv.del	**8**	11	**72%**	5	28	18%
	invddup	4	8	**50%**	**8**	19	42%
	invddup.del	**11**	26	**42%**	1	8	12%
	invddup.del.inv	0	6	0%	0	2	0%
	invddup.inv	0	3	0%	0	1	0%
	invdup				0	27	0%
	other	**10**	47	**21%**	2	190	1%
	trans				0	11	0%
	trans.inv				0	3	0%

Table 3: Complex SV calling results. TP = # validated SVs; N = # calls; Prec = precision = 1 - FDR. Results are before random forest filtering, and excluding tandem repeat regions. Blank rows indicate that no SV calls were made. See Table 4 for examples of each complex SV type.

SV type	**abbreviation**	**structure**
reference		$ABCDE$
deletion		$ACDE$
tandem duplication		$ABBCDE$
INV		$AB'CDE$
interspersed duplication	(ddup)	$ABCBDE$
interspersed duplication + deletion	(ddup.del)	$ABCBE$
tandem duplication + deletion	(dup.del)	$ABBDE$
double tandem duplication	(dup.dup)	$ABBCCDE$
inversion + 2 flanking deletion	(inv.2del)	$AC'E$
inversion + flanking deletion	(inv.del)	$AC'DE$
interspersed inverted duplication	(invddup)	$ABCB'DE$
invddup + intermediate inversion	(invddup.inv)†	$ABC'B'DE$
invddup + deletion	(invddup.del)	$ABCB'E$
invddup + deletion + intermediate inversion	(invddup.del.inv)	$ABC'B'E$
inverted tandem duplication	(invdup)	$ABB'CDE$
translocation	(trans)†	$ACBDE$
inverted translocation	(trans.inv)†	$ACB'DE$
other complex SV		$AEBDE$
		$AD'B'E$
		$ABC'BCD$
		$\vdots$
compound SV		$ACDDE$
		$ABBCD'E$
		$ACDE$
		$\vdots$

Table 4: Example genomic structures for different simple, complex, and compound SVs.
† No validated calls of this SV type.

ARC-SV (simple SV)	SV size; genotype; SR support; PE support; SV score; runner-up score; difference between best and runner-up scores; # paths (haplotypes) considered; # candidate breakpoints not used in call[†]; total breakpoint uncertainty
ARC-SV (complex SV)	in addition to the above: # affected base pairs; size of largest affected block; lowest split support across breakpoints; lowest paired-end support across breakpoints; number of breakpoints in SV
SVelter (simple SV)	SV size; genotype; # candidate breakpoints not used in call[†]; SV score
SVelter (complex SV)	in addition to the above: # affected base pairs; size of largest affected block; # breakpoints in SV
LUMPY	SV size; SR support; PE support; start/end position 95% confidence interval lengths; VCF tags INV_PLUS, INV_MINUS
DELLY	SV size; SR support; PE support; genotype; start/end position confidence interval lengths; VCF tags MAPQ, SRQ, CE, CT, FT, GQ, RC, RCR, RCL, CN, DR, DV, RR, RV
Pindel	SV size; SR support; genotype; length of microhomology (HOMLEN); Number of bases inserted in place of deleted code (NTLEN)
SoftSV	SV size; SR support; PE support

Table 5: Features used in random forest classification.

[†] For example, a deletion of B reported as $ACDE$ has 2 unused candidate breakpoints: between C and D; and between D and E

Supplementary Note 1: Uniform distribution inference

Let $X_1, \dots, X_n$ be drawn iid from a Uniform distribution on $[a, b) \subset \mathbb{R}$. We construct a confidence interval for a having the form

$$[X_{(1)} - c(X_{(n)} - X_{(1)}), X_{(1)}],$$

where $c > 0$ depends on n and the confidence level $1 - \alpha$.

To solve for c we reduce the problem to one based on Uniform$(0, 1)$ order statistics $U_{(1)}, \dots, U_{(n)}$. Defining

$$Z_j = \frac{\sum_{i=1}^{j} E_i}{\sum_{i=1}^{n+1} E_i},$$

where E_i are independent Exponential(1) random variables, it is known that $(U_{(1)}, \dots, U_{(n)})$ and $(Z_1, \dots, Z_n)$ have the same joint distribution [1]. Now we have

$$\begin{aligned}
1 - \alpha =& P\left(a \geq X_{(1)} - c(X_{(n)} - X_{(1)})\right) \\
=& P\left(\frac{a - X_{(1)}}{X_{(n)} - X_{(1)}} \geq c\right) \\
=& P\left(\frac{-U_{(1)}}{U_{(n)} - U_{(1)}} \geq c\right) \\
=& P\left(\frac{-E_1}{\sum_{i=2}^{n} E_i} \geq c\right).
\end{aligned}$$

The final term concerns Exp(1) divided by an independent Gamma$(n - 1,\ 1)$, which by definition follows a scaled F-distribution: $\frac{2}{2n-2} F_{2,2n-2}$. The confidence interval for b is given by the same argument.

References

1. DasGupta, A. Finite sample theory of order statistics and extremes. In *Probability for Statistics and Machine Learning*, pages 221-248. Springer, 2011.

Supplementary Note 2: Affected sequence calculation

A simple SV is defined by a single position and length in the reference genome. Complex SVs, however, may span large regions but make only small modifications to the sequence. In our analysis of complex SV calls, we keep the usual requirement that SVs must affect at least 50 bp of sequence. We propose a simple method that defines which reference segments are "affected" by an SV.

Suppose the reference region (separated into segments by the breakpoints) is given by $B_1^0 B_2^0 \ldots B_n^0$, and the rearranged version is $B_{r_1}^{o_1} B_{r_2}^{o_2} \ldots B_{r_m}^{o_m}$, where o_i is 1 if the segment is in reverse orientation. We find a pairwise alignment between these two strings of genomic segments (not the nucleotide sequences), treating blocks $B_i^{o_i}$ and $B_j^{o_j}$ as equal if $i = j$ and $o_i = o_j$. Alignments were computed using a very large score (1000) for matches and equally small penalties (-1) for mismatches and gap extensions. After finding an optimal alignment, blocks that are mismatched or opposite gaps in either sequence are considered "affected." The affected length is then computed as the sum of affected reference block lengths. Note that used the `Biostrings::pairwiseAlignment` function in R [1], which returns only a single optimal alignment. To partially mitigate the problem of non-unique optima, we computed a second alignment by switching the subject and query sequences and adding any additional affected blocks.

References

1. Pags H, Aboyoun P, Gentleman R and DebRoy S (2017). Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.44.2.

Supplementary Note 3: Materials and Methods for Venter SV Validation

Genomic DNA

Genomic DNA sample of Craig Venter (NS12911) was purchased from the Coriell Institute. Two male genomic DNA were purchased from EMD Millipore (Cat. 70572) and Promega (Cat. G1471). DNA were aliquoted and stored at -80^{o}C for long-term usage.

Experimental verification of variants using Oxford Nanopore MinION™ Sequencing

PCR verification

At least two sets of PCR primers were designed for the control region and the target variant using Primer 3 (**http://bioinfo.ut.ee/primer3-0.4.0/primer3/**; **Table 1**). The length of each PCR product is about 3-5 kb, and each target locus was amplified using 200 ng of Venter DNA and two control male genomic DNA, separately. The amplification was performed using Q5 high-fidelity DNA polymerase (NEB, Ipswich, MA) under the following conditions: initial denature at 98^{o}C for 30 sec, total 30 cycles of: 98^{o}C for 10 sec, 65^{o}C for 20 sec, and 72^{o}C for 2 min (or 4 min for longer PCR product), and final extension at 72^{o}C for 10 min. The PCR products were then analyzed by agarose gel electrophoresis.

Library Preparation

For those PCR with correct size prediction, we designed tailing PCR primer for Oxford Nanopore carrying the universal sequences: 5'-TTTCTGTTGGTGCTGATATTGC-[project-specific forward primer sequence]-3' and 5'-ACTTGCCTGTCGCTCTATCTTC-[project-specific reverse primer sequence]-3'. To construct the library, a first PCR was performed using the tailing PCR primer, 200 ng of Venter DNA, and Q5 High-Fidelity DNA polymerase, under the same PCR condition mentioned above. The PCR products were then analyzed by agarose gel electrophoresis. For those PCR with only one single band, the PCR products were purified using 1.0 x by volume AMPure XP beads (Beckman Coulter, Brea, CA) to remove proteins, salts, dNTPs and primers. For those PCR with two bands, the PCR products were gel purified using NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel, Bethlehem, PA). DNA concentration was measured using Qubit dsDNA HS Assay kit (Invitrogen, Carlsbad, CA). Then a second PCR was performed to incorporate the Oxford Nanopore barcode sequence into each amplicon following the instruction of the PCR 96 barcoding kit (R9 version, Oxford Nanopore, Oxford, UK). Multiple bands generated from the same PCR were labelled with the same barcode and a total of 11 barcodes was used in this study. Barcoded PCR libraries were quantified with Qubit dsDNA HS assay kit, normalized, and pooled to a final amount of 1 μg. For sequencing, the libraries were end-repaired and dA-tailed using NEBNext End repair/dA-tailing Module (NEB, Ipswich, MA), followed by the ligation of hairpin and Oxford Nanopore-specific leader adapters using NEB Blunt/TA Ligase Master mix (NEB, Ipswich, MA). Next, a HP tether was bound to both the leader and hairpin adapter, serving as a motor protein pulling each molecule through the nanopore one base at a time during sequencing. Finally, the adapted and tethered DNA library was enriched with MyOne C1 Streptavidin beads (Life Technologies, Carlsbad, CA) and eluted in 25 μl of elution buffer prior to loading into the MinION^{TM} flow cell.

Oxford Nanopore MinION^{TM} sequencing

To prime the MinION^{TM} flow cell, 500 μl of the priming mix was loaded into the sample loading port. After 10 minutes, this priming process was repeated a second time. To load the library, 12 μl of DNA library was mixed with 75 μl of RBF1 and 63 μl of nuclease-free water, and a 48 hour sequencing protocol was initiated using the MinKNOW^{TM} software (version 1.1.20k). The flow cell was ‘topped-up’ with a freshly diluted library 6 hours after the run. Read event data were base-called by the software Metrichor^{TM} agent (version 2.42.2) using Barcoding Plus 2D Basecalling RNN for SQK-NSK007.

Experimental verification of variants using Sanger sequencing

PCR primers were designed on sequence around the target variant (**Table 2**). The length of each PCR product is about 1 kb. Each target locus was amplified using 200 ng of Venter DNA, and the amplification was performed using Q5 High-Fidelity DNA polymerase (NEB, Ipswich, MA). PCR conditions were as follows: initial denature at 98^{o}C for 30 sec, total 32 cycles of: 98^{o}C for 10 sec, 65^{o}C for 20 sec, and 72^{o}C for 30 sec, and final extension at 72^{o}C for 10 min. The PCR products were then analyzed by agarose gel electrophoresis, and each band was cut out and purified using NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel, Bethlehem, PA). PCR products were then sequenced using ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA) with 5 pmol of each sequencing primer list on **Table 2**.

Acknowledgement

The authors would like to thank Stanford Pan Facility Alberto Lovell for Sanger sequencing. This work was supported by NIH grant (?).

Table 1. Primers used in PCR analysis*

	Ref (bp)	**Venter (bp)**
Control-1:		
5’-gctaaacgagcaccctgttc-3’	3196	3196
5’-gaaacacgaggtggaggaaa-3’		
Control-2:		
5’-agaccctgtccaccaaaatg-3’	3297	3297
5’-agagcttgcaggagaagctg-3’		
Deletion-1:		
5’-tgggtattgtgatggtgtgg-3’	4199	3181
5’-ccaacaggattgcttgacag-3’		
Deletion-2:		
5’-ggaggggaacatcacacact-3’	3153	2823
5’-gtaggaggagacaggcagca-3’		
Duplication-1:		
5’-cacctgtcctgtgggaaagt-3’	3000	3000
5’-aagctcccttccttgagagc-3’		4546
Duplication-2:		
5’-ctgaggtgggtggatcatct-3’	2698	2698
5’-taagcaggaaggatggcaag-3’		2909
Complex-1:		
5’-gcccatatgaaacctgaagc-3’	3050	3050
5’-tgtgggcatcttgggtaaat-3’		3288
Complex-2:		
5’-catgacaccaaaagcacagg-3’	3098	3098
5’-gcctctgcaccatattgaca-3’		3437
Complex-3:		
5’-acggaggatgtcagtgtggt-3’	3489	3462
5’-cagagaggagaaggcaatcg-3’		

Complex-4:		
5’-gttgccttccatttcagcat-3’	4897	4897
5’-tcaggggagcagcattattc-3’		5341
Complex-5:		
5’-cttcagggagcagcaaattc-3’	2849	2849
5’-ggctggctttctgtgtaagg-3’		3118

*The first PCR amplification in Oxford Nanopore library preparation requires tailed primers. Here is the PCR primer design.
5’-TTTCTGTTGGTGCTGATATTGC-[locus-specific forward primer sequence]-3’
5’-ACTTGCCTGTCGCTCTATCTTC-[locus-specific reverse primer sequence]-3’

Table 2. PCR primer and sequencing primer used in Sanger sequencing

PCR primer:

Duplication-2:

5’-tctgggtaggtggtgtgtta-3’
5’-ggccaacatggtgaaactct-3’

Complex-2:

5’-gcggtgtattcagtcaatgg-3’
5’-gatcacctgaggtcagaagt-3’

Complex-4:

5’-taagggttcgaatccagcc-3’
5’-cctcctaaatgaacagtggc-3’

Complex-5:

ABC
5’-tgcacagcgtagcttctgtt-3’
5’-caatcctgctttcaggaagg-3’

AB’DE
5’-agcatctgtgtgctaatccc-3’
5’-aacactccttctccaccaca-3’

EBC
5’-accgagcaggctcactaaat-3’
5’-agccttggtagtgcctacaa-3’

Sequencing primer:

Duplication-2:
5’-gagaccaggagtttgagagt-3’

Complex-2:
5’-gctaagtgaatgaaccaagac-3’

Complex-4:
5’-ctgggcaacatagtaagacc-3’

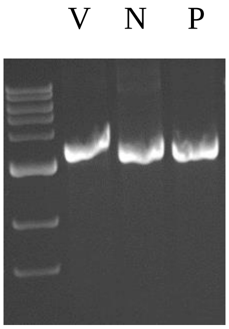
Complex-5:
ABC:
5’-tcttatggcagggctctgaa-3’

AB’DE:
5’-tcttatggcagggctctgaa-3’

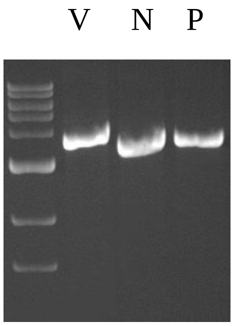
EBC:
5’-tgtggtggagaaggagtgt-3’

Figure 1: PCR verification for Oxford Nanopore Sequencing. V: Venter DNA from Coriell Institute; N: male genomic DNA from EMD Millipore; P: male genomic DNA from Promega. 1kb DNA Ladder from NEB was used in the gel electrophoresis.

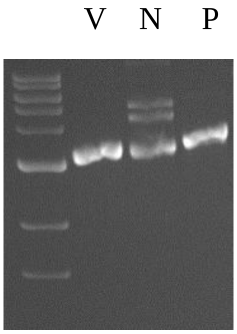
Control-1:



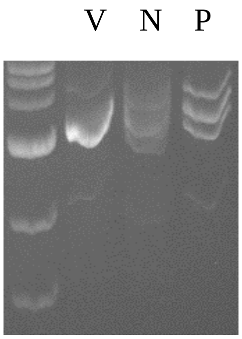
Control-2:



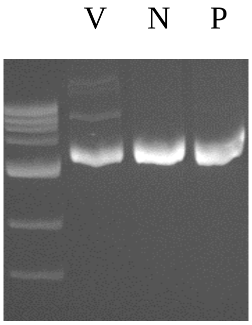
Deletion-1:



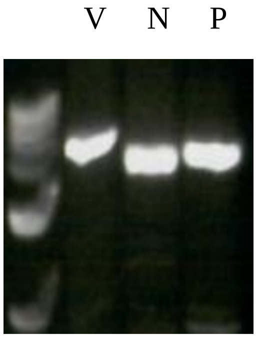
Deletion-2:



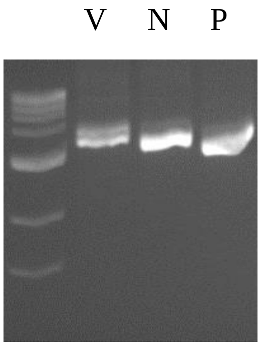
Duplication-1:



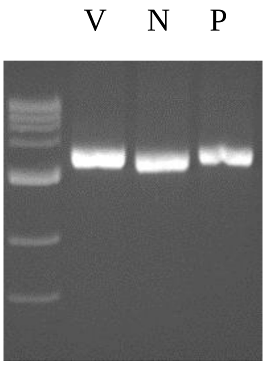
Duplication-2:



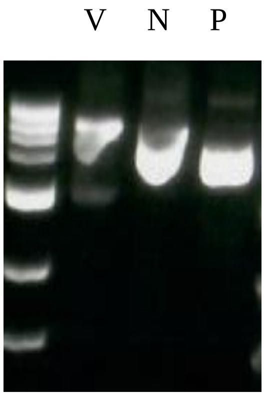
Complex-1:



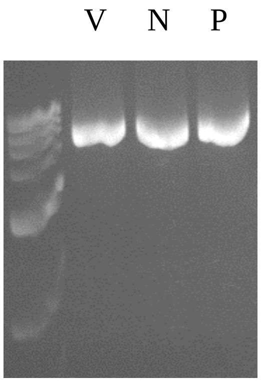
Complex-2:



Complex-3:



Complex-4:



Complex-5: