Supplementary material

Additional file 1:

Table S1: A brief summary of the detected SVs by the two calling methods.

Variation	Data 1 by T_1		Data 2 by T_2	
	Raw data1(%)	Filtered data1(%)	Raw data2(%)	Filtered data2(%)
All	109378	1115	76431	990
INS: length of base pairs				
<50bp	18706(53.54)	253(61.71)	12834(53.84)	204(56.51)
>50bp	16231(46.46)	157(38.29)	11004(46.16)	157(43.49)
Total	34937(21.94)	410(36.77)	23838(31.19)	361(36.46)
DEL: length of base pairs				
<50bp	31638(62.29)	322(90.45)	23461(63.09)	326(92.61)
>50bp	19153(37.71)	34(9.55)	13724(36.91)	26(7.39)
Total	50791(46.44)	356(31.93)	37185(48.65)	352(35.56)
DUP: length of base pairs				
<50bp	12214(61.39)	149(75.25)	8511(63.11)	163(81.91)
>50bp	7682(38.61)	49(24.75)	4975(36.89)	36(18.09)
Total	19896(18.19)	198(17.76)	13486(17.64)	199(20.10)
INV: length of base pairs				
200bp-1Mb	86(17.48)	—	39(21.97)	—
>1Mb	406(82.52)	—	174(78.03)	—
Total	492(0.45)	—	223(0.29)	—
CNV: length of base pairs				
1kb-1Mb	191(86.04)	120(79.47)	95(82.61)	58(74.36)
>1Mb	31(13.96)	31(20.53)	20(17.39)	20(25.64)
Total	222(0.20)	151(13.54)	115(0.15)	78(7.88)
TRA: number of translocation (unit: times)				
2	1520	—	792	—
Total	3040(2.78)	—	1584(2.07)	—

Note: Mb,millone base; kb,kilobase; bp, base pair; 1 Mb =1,000 kb= 1,000,000 bp.

Additional file 2: The posterior estimates if the two calling methods are conditionally independent

In this section, we retained the notations and assumptions used in the main text, and assume T_1 and T_2 were conditionally independent, that is, the covariance $Covp^{SV}$ and $Covn^{SV}$ were

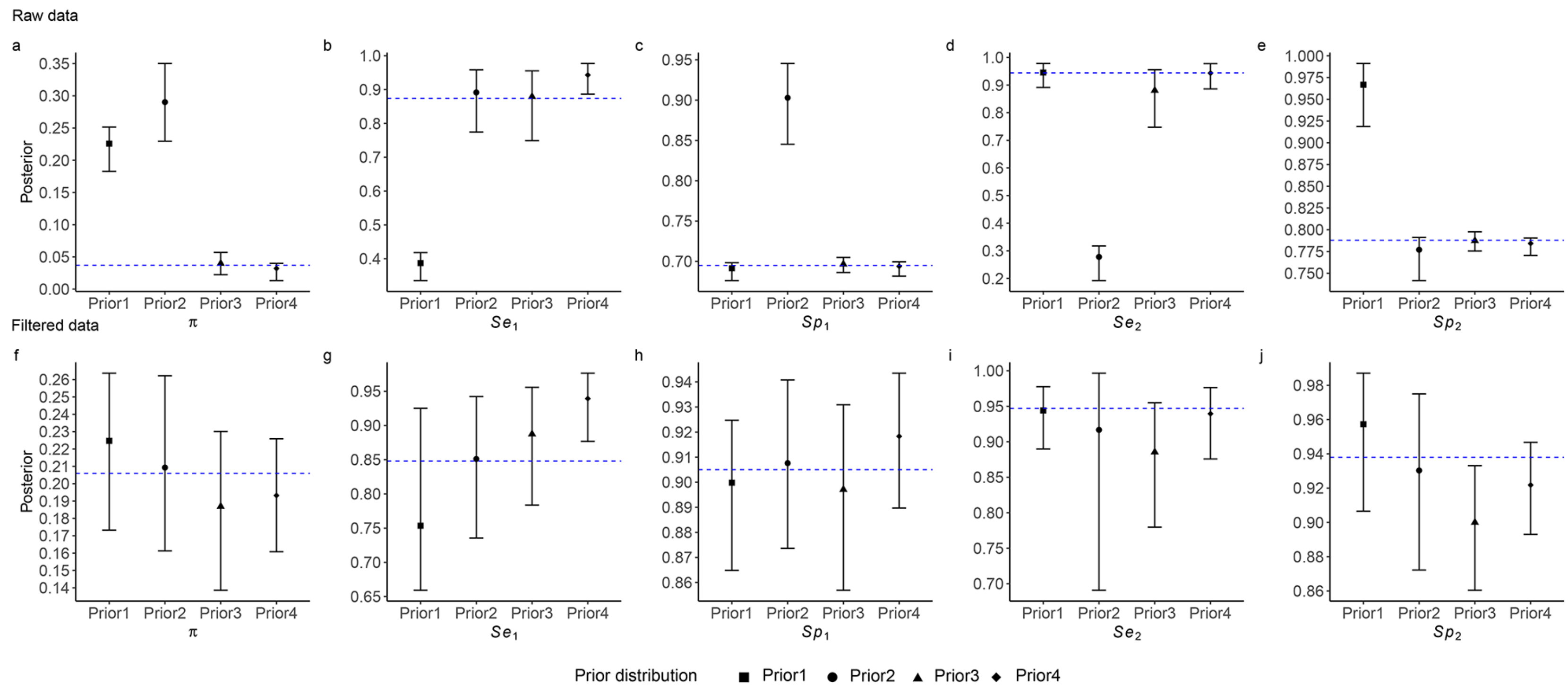

Figure S1: Posterior estimates and 95% CIs of DEL under four prior distributions. The dashed lines in each subplot correspond to the posterior estimates in Table 3 of the main text.

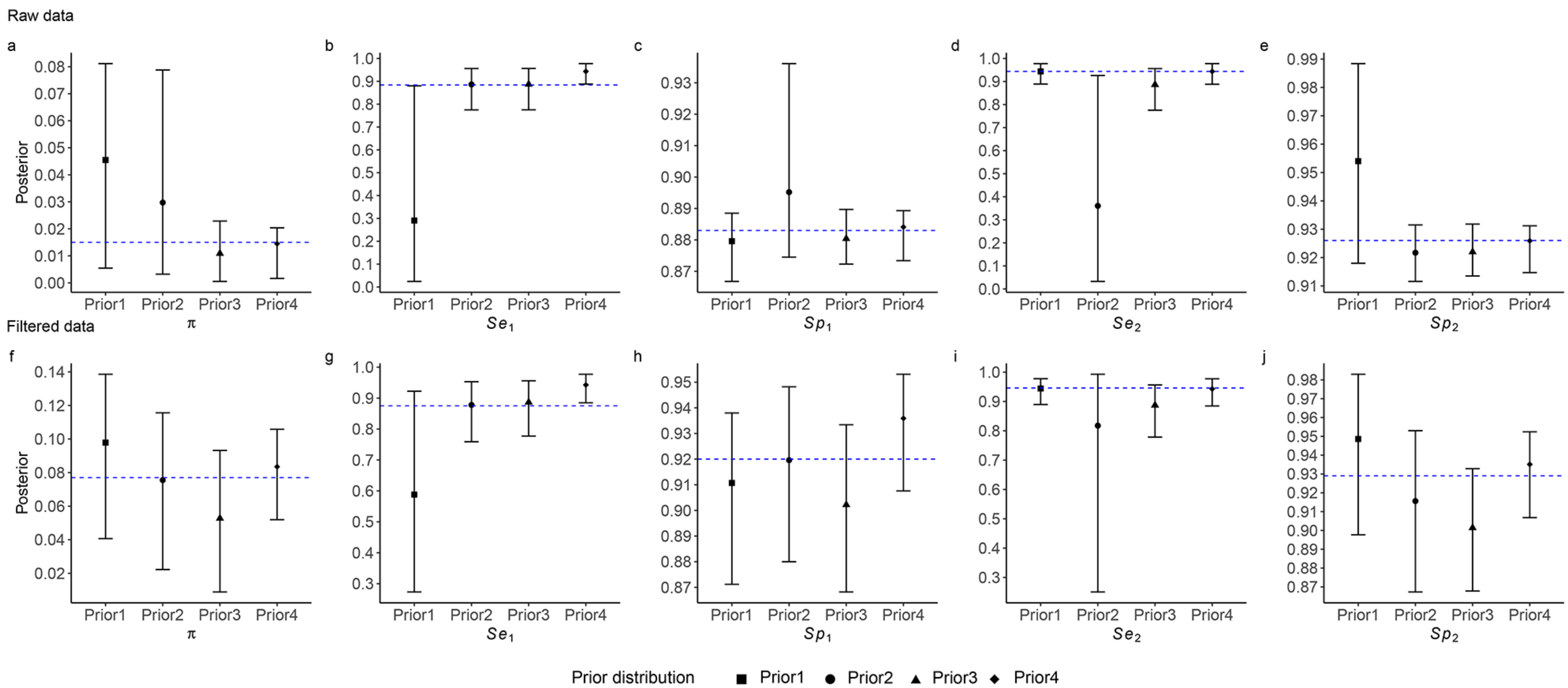

Figure S2: Posterior estimates and 95% CIs of DUP under four prior distributions. The dashed lines in each subplot correspond to the posterior estimates in Table 3 of the main text.

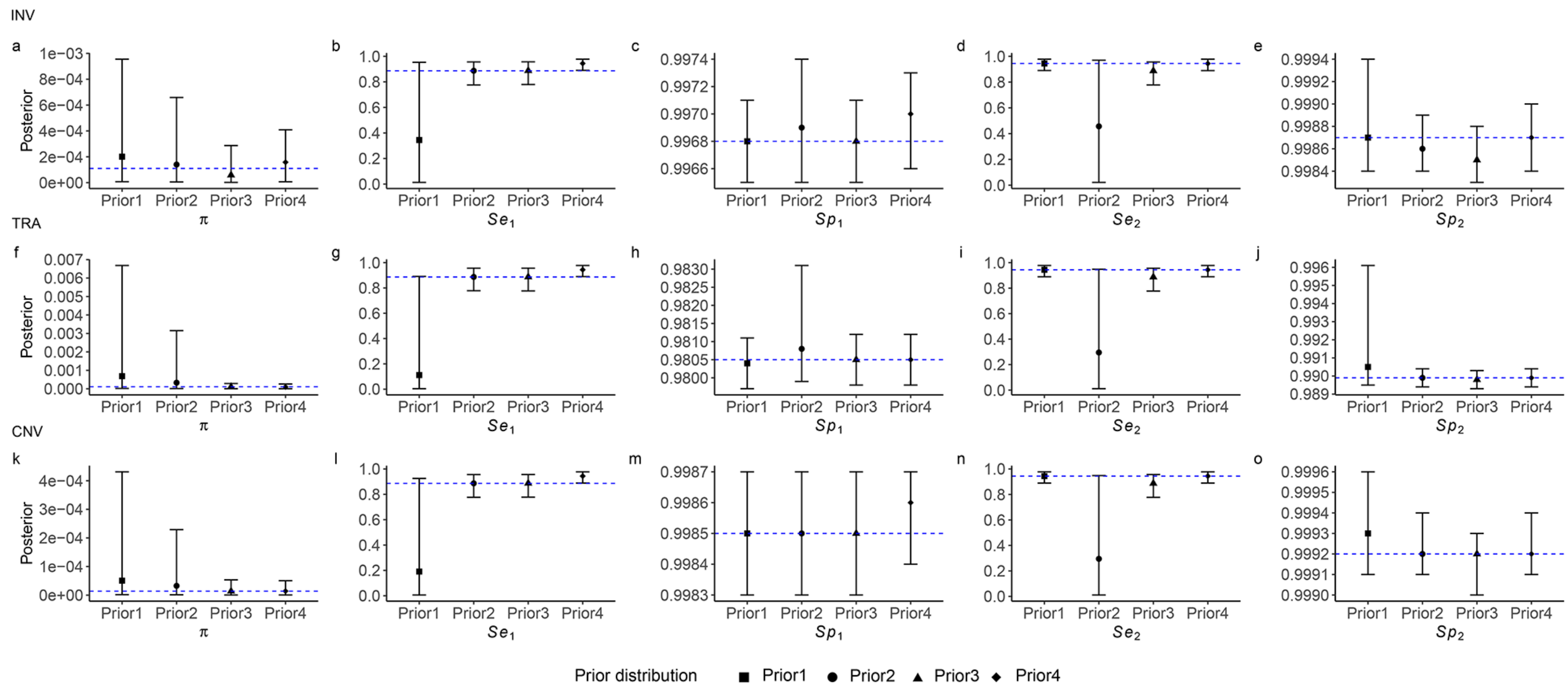

Figure S3: Posterior estimates and 95% CIs of INV, TRA, and CNV under four prior distributions in raw data. Panels (a)-(e), (f)-(j), (k)-(o) are the posteriors corresponding to INV, TRA, and CNV, respectively. The dashed lines in each subplot correspond to the posterior estimates in Table 3 of the main text.

zero. Then, the likelihood function can be written as

$$\begin{aligned}
L(\hat{\boldsymbol{\theta}}) \propto & \left[\pi^{SV} Se_1^{SV} Se_2^{SV}\right]^{n_{11}} \times \left[\pi^{SV} Se_1^{SV}(1 - Se_2^{SV})\right]^{n_{10}} \\
& \times \left[\pi^{SV}(1 - Se_1^{SV}) Se_2^{SV}\right]^{n_{01}} \times \left[\pi^{SV}(1 - Se_1^{SV})(1 - Se_2^{SV})\right]^{n_{00}} \\
& \times \left[(1 - \pi^{SV})(1 - Sp_1^{SV})(1 - Sp_2^{SV})\right]^{N_{11} - n_{11}} \\
& \times \left[(1 - \pi^{SV})(1 - Sp_1^{SV}) Sp_2^{SV}\right]^{N_{10} - n_{10}} \\
& \times \left[(1 - \pi^{SV}) Sp_1^{SV}(1 - Sp_2^{SV})\right]^{N_{01} - n_{01}} \\
& \times \left[(1 - \pi^{SV})(Sp_1^{SV} Sp_2^{SV})\right]^{N_{00} - n_{00}}.
\end{aligned}$$

The posterior distribution can be obtained by combining the likelihood function with prior distributions in Table 2. The above functions can be defined similarly for INS, DEL, DUP, INV, TRA, and CNV, respectively.

Table S2 indicated that T_2 had higher sensitivity and specificity both in raw data and filtered data when the two calling results were conditionally independent. Compared with Table 3 in the main text, the majority posterior estimates without considering the covariance were slightly larger with narrower confidence intervals. While the posterior estimates were close to each other under the condition that T_1 and T_2 had the same prior distributions (Table S3), we pointed out that there were situations where the sensitivity of T_2 did not exceed that of T_1. But T_2 still performed better because of its higher summation of sensitivity and specificity, except for DUP in filtered data. The analysis again showed that the posterior estimates were sensitive to the prior distributions, which was consistent with that in the main body.

Additional file 3: Detailed data processing of data 2

In this section, we present the software and its settings used for alignment and SVs calling.

Alignment

Tool: *pbmm2*, Version 1.3.0.
Genome Reference: hg38 (From National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/).

Usage: Align CCS fastq input and sort output.

```
pbmm2 index ../hg38.fasta hg38.mmi --preset CCS
pbmm2 align hg38.mmi input.fastq output.bam --preset CCS --sort
--rg '@RG\tID:example'
```

Details and default parameters can be found at
`https://github.com/PacificBiosciences/pbmm2`.

Structural variant calling

Tool: *pbsv*, Version 2.6.2.
Usage: Discover signatures via –tandem-repeats from GRCh38
(`https://github.com/PacificBiosciences/pbsv/blob/master/annotations/human_GRCh38_no_alt_analysis_set.trf.bed`) and call structural variation.

```
pbsv discover output.bam output.svsig.gz --tandem-repeats
../human_GRCh38_no_alt_analysis_set.trf.bed
pbsv call --ccs ../hg38.fasta output.svsig.gz output.vcf
```

Details and default parameters can be found at
`https://github.com/PacificBiosciences/pbsv`.

Table S2: Posterior estimates and 95%CIs for the detection of INS, DEL, DUP, INV, TRA, and CNV

Data	Para.	INS		DEL		DUP	
		Median	95% CI	Median	95% CI	Median	95% CI
Raw data	π	0.0692	(0.0612, 0.0858)	0.0445	(0.0380, 0.0580)	0.0223	(0.0199, 0.0267)
	Se_1	0.8747	(0.7447, 0.9513)	0.8724	(0.7447, 0.9502)	0.8829	(0.7641, 0.9545)
	Sp_1	0.8236	(0.8205, 0.8278)	0.6990	(0.6960, 0.7024)	0.8893	(0.8875, 0.8913)
	Se_2	0.9411	(0.8820, 0.9761)	0.9411	(0.8828, 0.9765)	0.9438	(0.8884, 0.9773)
	Sp_2	0.9050	(0.8985, 0.9197)	0.7936	(0.7886, 0.8038)	0.9328	(0.9305, 0.9366)
Data	Para.	INV		TRA		CNV	
		Median	95% CI	Median	95% CI	Median	95% CI
Raw data	π	4.3E-4	(4.4E-6, 4.0E-4)	2.2E-4	(5.1E-6, 2.8E-4)	2.7E-5	(6.4E-7, 5.3E-5)
	Se_1	0.8887	(0.7804, 0.9567)	0.8863	(0.7753, 0.9562)	0.8866	(0.7763, 0.9560)
	Sp_1	0.9971	(0.9968, 0.9974)	0.9806	(0.9799, 0.9813)	0.9985	(0.9983, 0.9987)
	Se_2	0.9449	(0.8911, 0.9777)	0.9438	(0.8890, 0.9773)	0.9441	(0.8891, 0.9774)
	Sp_2	0.9989	(0.9988, 0.9991)	0.9900	(0.9895, 0.9905)	0.9993	(0.9991, 0.9994)
Data	Para.	INS		DEL		DUP	
		Median	95% CI	Median	95% CI	Median	95% CI
Filtered data	π	0.2343	(0.2034, 0.2691)	0.2171	(0.1861, 0.2527)	0.1003	(0.0810, 0.1247)
	Se_1	0.8881	(0.8179, 0.9517)	0.8762	(0.7915, 0.9486)	0.8813	(0.7720, 0.9532)
	Sp_1	0.8978	(0.8770, 0.9189)	0.9214	(0.9031, 0.9392)	0.9390	(0.9248, 0.9519)
	Se_2	0.9434	(0.8907, 0.9771)	0.9490	(0.9021, 0.9793)	0.9486	(0.8995, 0.9793)
	Sp_2	0.9589	(0.9336, 0.9842)	0.9497	(0.9234, 0.9788)	0.9516	(0.9349, 0.9702)

Table S3: Posterior estimates and 95%CIs under the same prior distribution

Data	Para.	INS		DEL		DUP	
		Median	95% CI	Median	95% CI	Median	95% CI
Raw data	π	0.0630	(0.0586, 0.0697)	0.0396	(0.0358, 0.0449)	0.0207	(0.0191, 0.0228)
	Se_1	0.9415	(0.8841, 0.9765)	0.9414	(0.8840, 0.9764)	0.9435	(0.8888, 0.9772)
	Sp_1	0.8235	(0.8205, 0.8277)	0.6989	(0.6959, 0.7023)	0.8893	(0.8875, 0.8913)
	Se_2	0.9413	(0.8842, 0.9767)	0.9418	(0.8856, 0.9765)	0.9430	(0.8886, 0.9769)
	Sp_2	0.8996	(0.8965, 0.9044)	0.7900	(0.7868, 0.7900)	0.9314	(0.9298, 0.9332)
Data	Para.	INV		TRA		CNV	
		Median	95% CI	Median	95% CI	Median	95% CI
Raw data	π	0.0004	(0.0003, 0.0005)	0.0002	(0.0001, 0.0003)	2.5E-5	(6.7E-6, 6.3E-5)
	Se_1	0.9443	(0.8905, 0.9776)	0.9439	(0.8892, 0.9775)	0.9440	(0.8891, 0.9774)
	Sp_1	0.9972	(0.9969, 0.9974)	0.9806	(0.9799, 0.9813)	0.9986	(0.9984, 0.9988)
	Se_2	0.9440	(0.8891, 0.9774)	0.9439	(0.8895, 0.9775)	0.9438	(0.8887, 0.9774)
	Sp_2	0.9989	(0.9987, 0.9991)	0.9900	(0.9895, 0.9905)	0.9993	(0.9991, 0.9994)
Data	Para.	INS		DEL		DUP	
		Median	95% CI	Median	95% CI	Median	95% CI
Filtered data	π	0.2242	(0.1959, 0.2571)	0.2057	(0.1791, 0.2366)	0.0948	(0.0779, 0.1142)
	Se_1	0.9393	(0.8844, 0.9752)	0.9384	(0.8803, 0.9749)	0.9421	(0.8869, 0.9766)
	Sp_1	0.9051	(0.8827, 0.9301)	0.9307	(0.9109, 0.9519)	0.9460	(0.9314, 0.9593)
	Se_2	0.9351	(0.8730, 0.9739)	0.9382	(0.8803, 0.9749)	0.9423	(0.8866, 0.9768)
	Sp_2	0.9445	(0.9250, 0.9662)	0.9339	(0.9144, 0.9550)	0.9454	(0.9309, 0.9587)