Large indel detection in region-based phased diploid assemblies from linked-reads

Can Luo, Brock A. Peters, and Xin Maizie Zhou

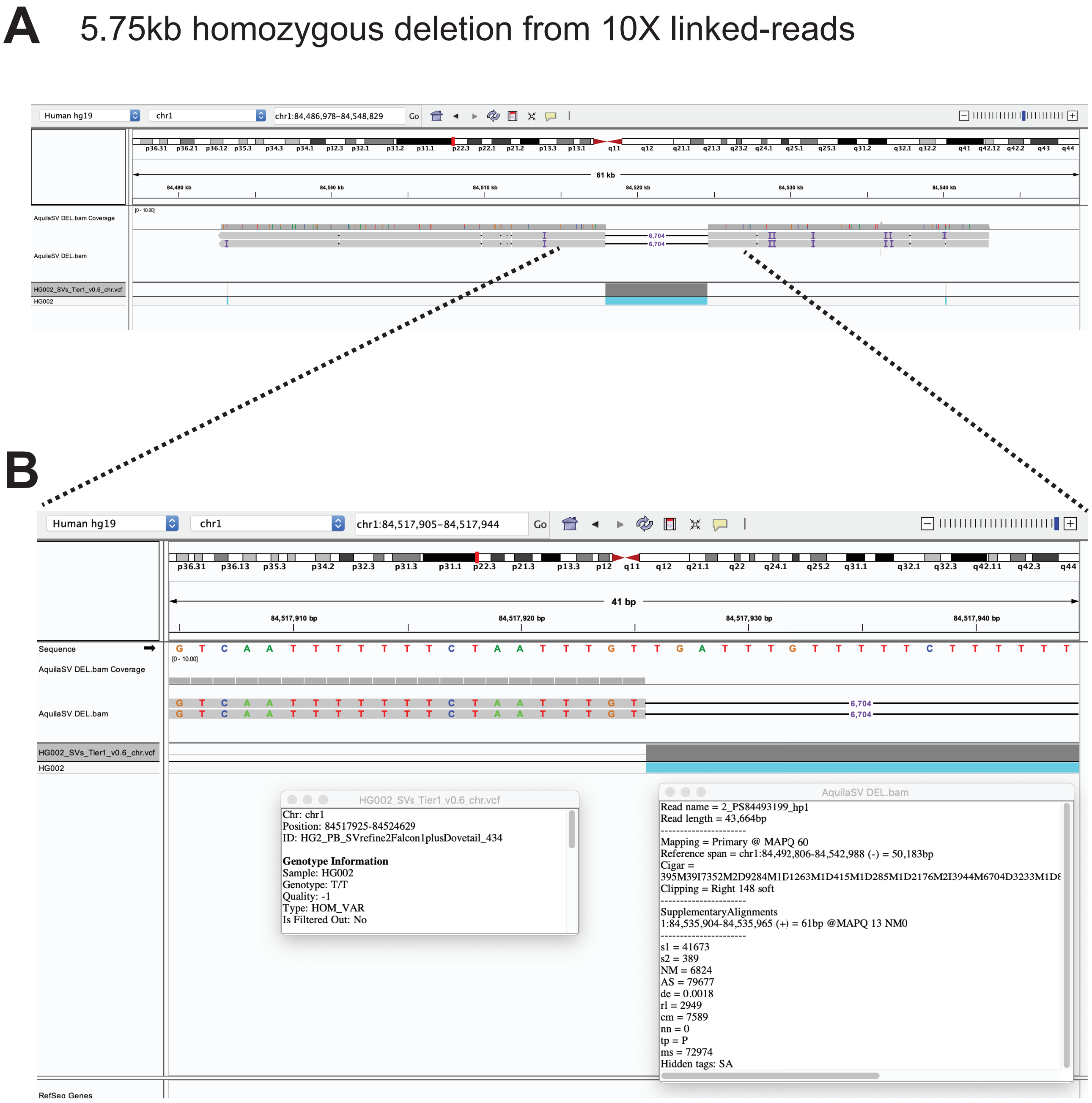
Supplementary Information

Deletions		flanking region (10kb)	flanking region (25kb)	flanking region (50kb)	flanking region (100kb)
50-1k	benchmark	58			
	TP	35	48	48	49
	TP_GT	28	40	41	39
	FP	28	53	76	123
	FN	23	10	10	9
	Recall	60.3%	82.8%	82.8%	84.5%
	Precision	55.6%	47.5%	38.7%	28.5%
	F1	57.9%	60.4%	52.7%	42.6%
	GT_accuracy	80.0%	83.3%	85.4%	79.6%
1k-10k	benchmark	7			
	TP	3	6	6	5
	TP_GT	3	6	6	5
	FP	0	1	7	6
	FN	4	1	1	2
	Recall	42.9%	85.7%	85.7%	71.4%
	Precision	100.0%	85.7%	46.2%	45.5%
	F1	60.0%	85.7%	60.0%	55.6%
	GT_accuracy	100.0%	100.0%	100.0%	100.0%

Table S1: Evaluation of deletions (>=50bp) on chromosome 21 in HG002 from stLFR linked-reads by different lengths of flanking regions in RegionIndel. Abbreviations: true positives (TP), false positives (FP), false negatives (FN), and genotype (GT).

Insertions		flanking region (10kb)	flanking region (25kb)	flanking region (50kb)	flanking region (100kb)
50-1k	benchmark	95			
	TP	14	19	23	22
	TP_GT	8	13	12	10
	FP	1	3	5	8
	FN	81	76	72	73
	Recall	14.7%	20.0%	24.2%	23.2%
	Precision	93.3%	86.4%	82.1%	73.3%
	F1	25.5%	32.5%	37.4%	35.2%
	GT_accuracy	57.1%	68.4%	52.2%	45.4%
1k-10k	benchmark	16			
	TP	1	1	1	1
	TP_GT	0	0	0	0
	FP	0	0	0	0
	FN	15	15	15	15
	Recall	6.2%	6.2%	6.2%	6.2%
	Precision	100.0%	100.0%	100.0%	100.0%
	F1	11.8%	11.8%	11.8%	11.8%
	GT_accuracy	0.0%	0.0%	0.0%	0.0%

Table S2: Evaluation of insertions (>=50bp) on chromosome 21 in HG002 from stLFR linked-reads by different lengths of flanking regions in RegionIndel. Abbreviations: true positives (TP), false positives (FP), false negatives (FN), and genotype (GT).



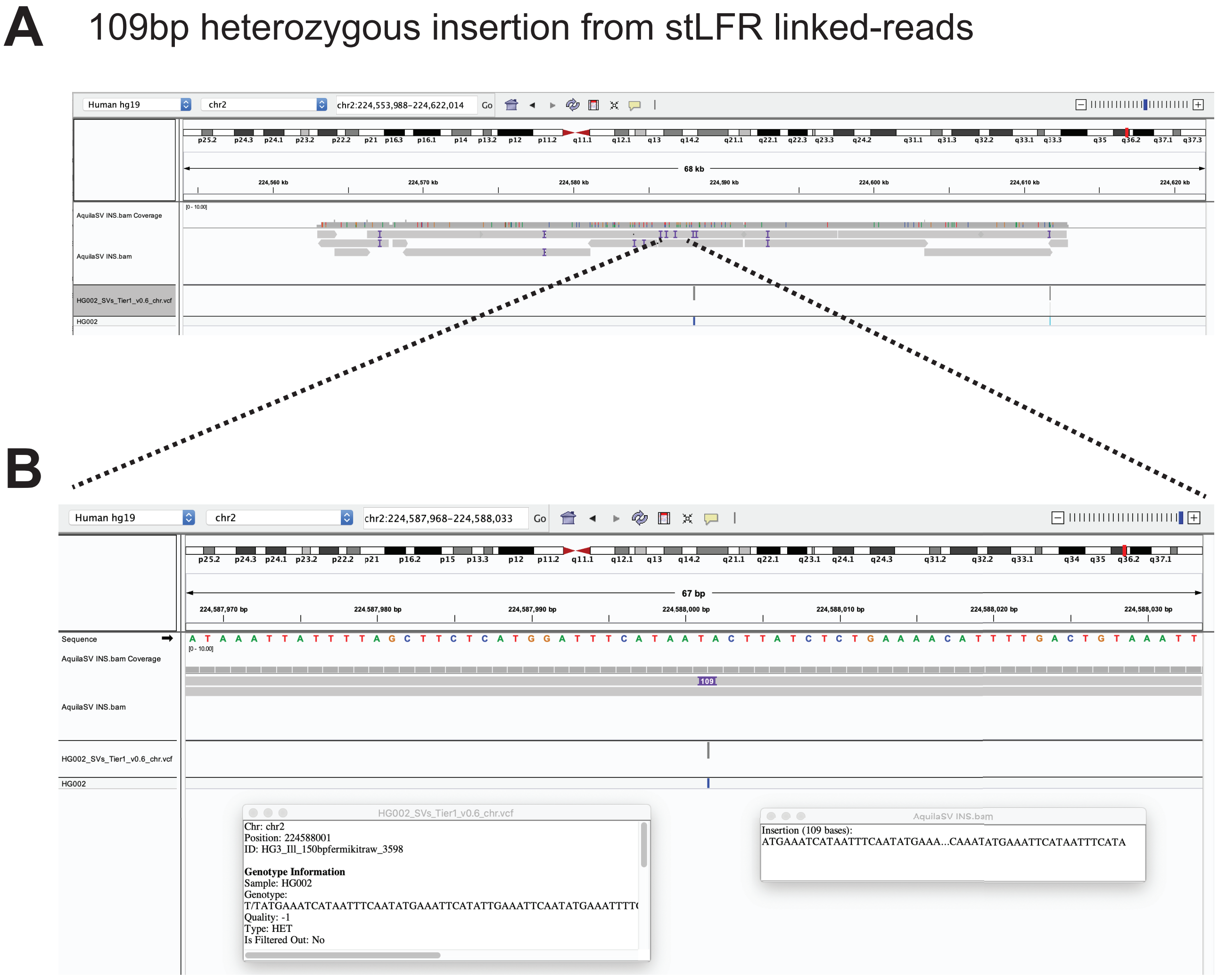
A
5.75kb homozygous deletion from 10X linked-reads
Human hg19
chr1
chr1:84,486,978-84,548,829
61 kb
AquilaSV DEL.bam Coverage
AquilaSV DEL.bam
HG002_SVs_Tier1_v0.6_chr.vcf
HG002
6,704
B
Human hg19
chr1
chr1:84,517,905-84,517,944
Go
41 bp
84,517,910 bp
84,517,920 bp
84,517,930 bp
84,517,940 bp
Sequence
AquilaSV DEL.bam Coverage
AquilaSV DEL.bam
HG002_SVs_Tier1_v0.6_chr.vcf
HG002
6,704
RefSeq Genes
HG002_SVs_Tier1_v0.6_chr.vcf
AquilaSV DEL.bam

```
Chr: chr1
Position: 84517925-84524629
ID: HG2_PB_SVrefine2Falcon1plusDovetail_434

Genotype Information
Sample: HG002
Genotype: T/T
Quality: -1
Type: HOM_VAR
Is Filtered Out: No
```

```
Read name = 2_PS84493199_hp1
Read length = 43,664bp
----------------------
Mapping = Primary @ MAPQ 60
Reference span = chr1:84,492,806-84,542,988 (-) = 50,183bp
Cigar =
395M39I7352M2D9284M1D1263M1D415M1D285M1D2176M2I3944M6704D3233M1D8
Clipping = Right 148 soft
----------------------
SupplementaryAlignments
1:84,535,904-84,535,965 (+) = 61bp @MAPQ 13 NM0
----------------------
s1 = 41673
s2 = 389
NM = 6824
AS = 79677
de = 0.0018
rl = 2949
cm = 7589
nn = 0
tp = P
ms = 72974
Hidden tags: SA
```

Fig. S1: One example of 5.75kb homozygous deletion from the 10X linked-reads data by RegionIndel. A. The IGV displays assembled diploid contigs from AquiaSV. It includes tracks for the contig BAM file from RegionIndel and the GiaB benchmark VCF file. The contigs also show there are several variants in the nearby region. B. It is a zoom-in view of the target SV in the center. The left inset box has the description about this benchmark SV, and the right inset box has the description about the contig information generated by RegionIndel. The gold standard supports this homozygous deletion. The contigs indicates this 6704bp deletion is identified by contigs from both haplotypes.


A
109bp heterozygous insertion from stLFR linked-reads
Human hg19
chr2
chr2:224,553,988-224,622,014
Go
68 kb
AquilaSV INS.bam Coverage
AquilaSV INS.bam
HG002_SVs_Tier1_v0.6_chr.vcf
HG002
B
Human hg19
chr2
chr2:224,587,968-224,588,033
Go
67 bp
224,587,970 bp
224,587,980 bp
224,587,990 bp
224,588,000 bp
224,588,010 bp
224,588,020 bp
224,588,030 bp
Sequence
A T A A A T T A T T T T A G C T T C T C A T G G A T T T C A T A A T A C T T A T C T C T G A A A A C A T T T T G A C T G T A A A T T
AquilaSV INS.bam Coverage
[0 - 10.00]
AquilaSV INS.bam
109
HG002_SVs_Tier1_v0.6_chr.vcf
HG002
HG002_SVs_Tier1_v0.6_chr.vcf
AquilaSV INS.bam

Chr: chr2
Position: 224588001
ID: HG3_Ill_150bpfermikitraw_3598

Genotype Information
Sample: HG002
Genotype:
T/TATGAAATCATAATTTCAATATGAAATTCATATTGAAATTCAATATGAAATTTT
Quality: -1
Type: HET
Is Filtered Out: No

Insertion (109 bases):
ATGAAATCATAATTTCAATATGAAA...CAAATATGAAATTCATAATTTCATA

Fig. S2: One example of 109bp heterozygous insertion from the stLFR linked-reads data by RegionIndel. A. The IGV displays assembled diploid contigs from AquiaSV. It includes tracks for the contig BAM file from RegionIndel and the GiaB benchmark VCF file. The contigs also show that there are several variants in the nearby region. B. It is a zoom-in view of the target SV in the center. The left inset box has the description about this benchmark SV, and the right inset box has the description about the contig information generated by RegionIndel. It indicates that this insertion is identified from the contig of haplotype1. The gold standard supports this heterozygous insertion.