Supplementary Material: Discovery of tandem and interspersed segmental duplications using high throughput sequencing

Arda Soylev[1,2 ‡], Thong Le[3,4 ‡], Hajar Amini[5], Can Alkan[1,6,7*] and Fereydoun Hormozdiari[3,8,9*]

[1] Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey

[2] Department of Computer Engineering, Konya Food and Agriculture University, Konya, 42080, Turkey.

[3] UC-Davis Genome Center, University of California, Davis, CA, USA.

[4] Department of Computer Science, University of California, Davis, CA, USA.

[5] Department of Neurology, School of Medicine, University of California, Davis, CA, USA.

[6] Bilkent-Hacettepe Health Sciences and Technologies Program, Ankara, 06800, Turkey

[7] Department of Computer Science, ETH Zürich, 8006, Switzerland

[8] Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA.

[9] MIND Institute, University of California, Davis, CA, USA.

1 Command lines

1.1 Simulation using VarSim

In order to simulate SVs including deletions, inversions and tandem duplications we used VarSim, which inserts known genomic variants into a given reference genome. However, it is unable to simulate interspersed duplications, thus we developed a new simulator called CNVSim to include interspersed duplications in direct and inverted orientations to the simulated genome. We additionally added some fixed real inversions to make the simulation more realistic. Finally, we created a VCF file including the interspersed duplications and some of the real inversions and used it as input to VarSim to generate the FASTQ files encompassing all the genomic variants.

```
vc_in_vcf=/share/varsim_files/All.vcf.gz
sv_insert_seq=/share/varsim_files/insert_seq.txt
sv_dgv=/share/varsim_files/GRCh37_hg19_supportingvariants_2013-07-23.txt
reference=human_g1k_v37_gatk.fasta
simulator_executable=/share/varsim_files/ART/art_bin_VanillaIceCream/art_illumina
vcf=invdup_simu.vcf

varsim.sh --reference $reference --id simu --read_length 100 --sv_num_ins 0 --sv_num_del 500 \
--sv_num_dup 500 --sv_num_inv 500 --sv_percent_novel 0.01  --mean_fragment_size 350 \
--sd_fragment_size 50 --sv_min_length_lim 50 --sv_max_length_lim 10000 \
--sv_insert_seq $sv_insert_seq \
--vc_in_vcf $vc_in_vcf --sv_dgv $sv_dgv --nlanes 1 --total_coverage $coverage \
--simulator_executable $simulator_executable --out_dir $out --log_dir $log --work_dir $work \
--simulator art --vcfs $vcf
```

1.2 SV Discovery Tools

TARDIS:

```
tardis -i CHM1.bam --ref human_g1k_v37_gatk.fasta --sonic human_g1k_v37.sonic --out chm1
```

LUMPY:

```
lumpyexpress -B CHM1.bam -o chm1.vcf
```

DELLY:

```
delly call -o chm1 -g human_g1k_v37_gatk.fasta -x excludeTemplates/human.hg19.excl.tsv CHM1.bam
```

TIDDIT:

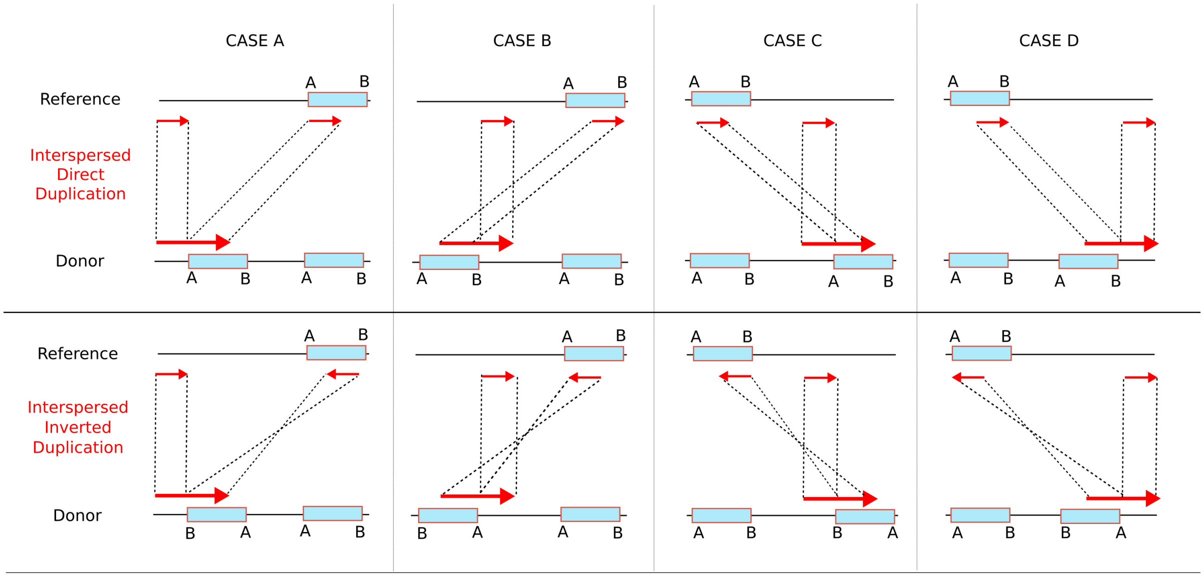
```
python TIDDIT.py --sv --bam CHM1.bam --ref human_g1k_v37_gatk.fasta -o chm1
```

SoftSV:

```
SoftSV --input CHM1.bam --output chm1
```

SVelter:

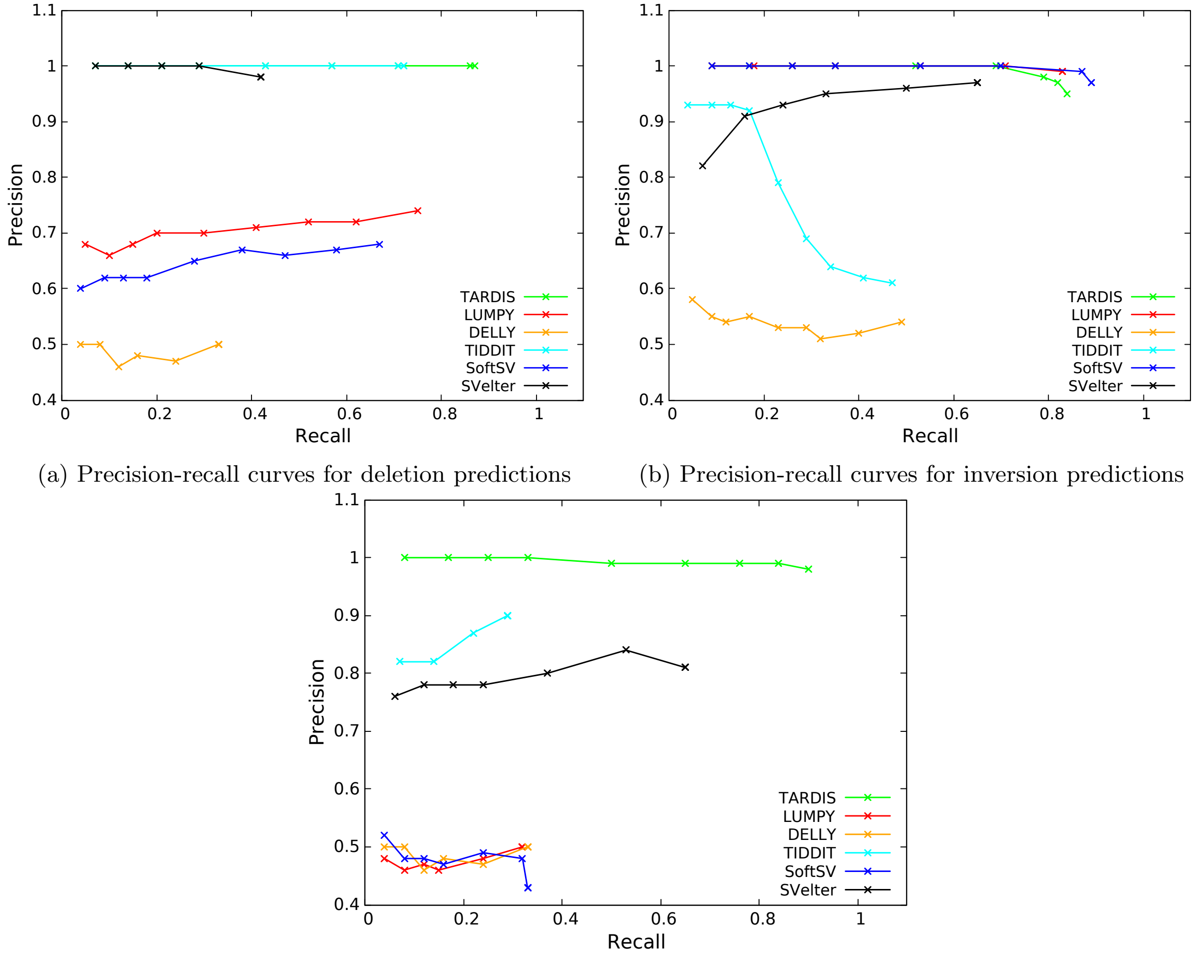
```
svelter.py Setup --reference human_g1k_v37_gatk.fasta --workdir SV/ \
--support /svelter/Support/GRCh37/

svelter.py -sample CHM1.bam --workdir SV/
```


CASE A
CASE B
CASE C
CASE D
Reference
Interspersed Direct Duplication
Donor
Reference
Interspersed Inverted Duplication
Donor
A
B

Supplementary Fig. 1: Split Read signatures for inverted and direct interspersed segmental duplications. In case A and B, duplicated part is inserted on the left and in C and D, on the right of the original segment. A) Soft clip is at the end of the read and is mapped after the primary mapping. B) Soft clip is at the beginning of the read and is mapped after the primary mapping. C) Soft clip is at the end of the read and is mapped before the primary mapping. D) Soft clip is at the beginning of the read and is mapped before the primary mapping.

Supplementary Table S1: Large segmental duplications found in chromosome Y simulation.

chromosome	**start**	**end**	**length**	**type**	**genotype**	**distance to insertion locus**
Y	5,580,120	5,648,940	68,820	direct interspersed	homozygous	204,780
Y	15,349,440	15,442,800	93,360	tandem	homozygous	-
Y	17,107,980	17,171,160	63,180	tandem	homozygous	-
Y	18,553,380	18,670,200	116,820	tandem	heterozygous	-

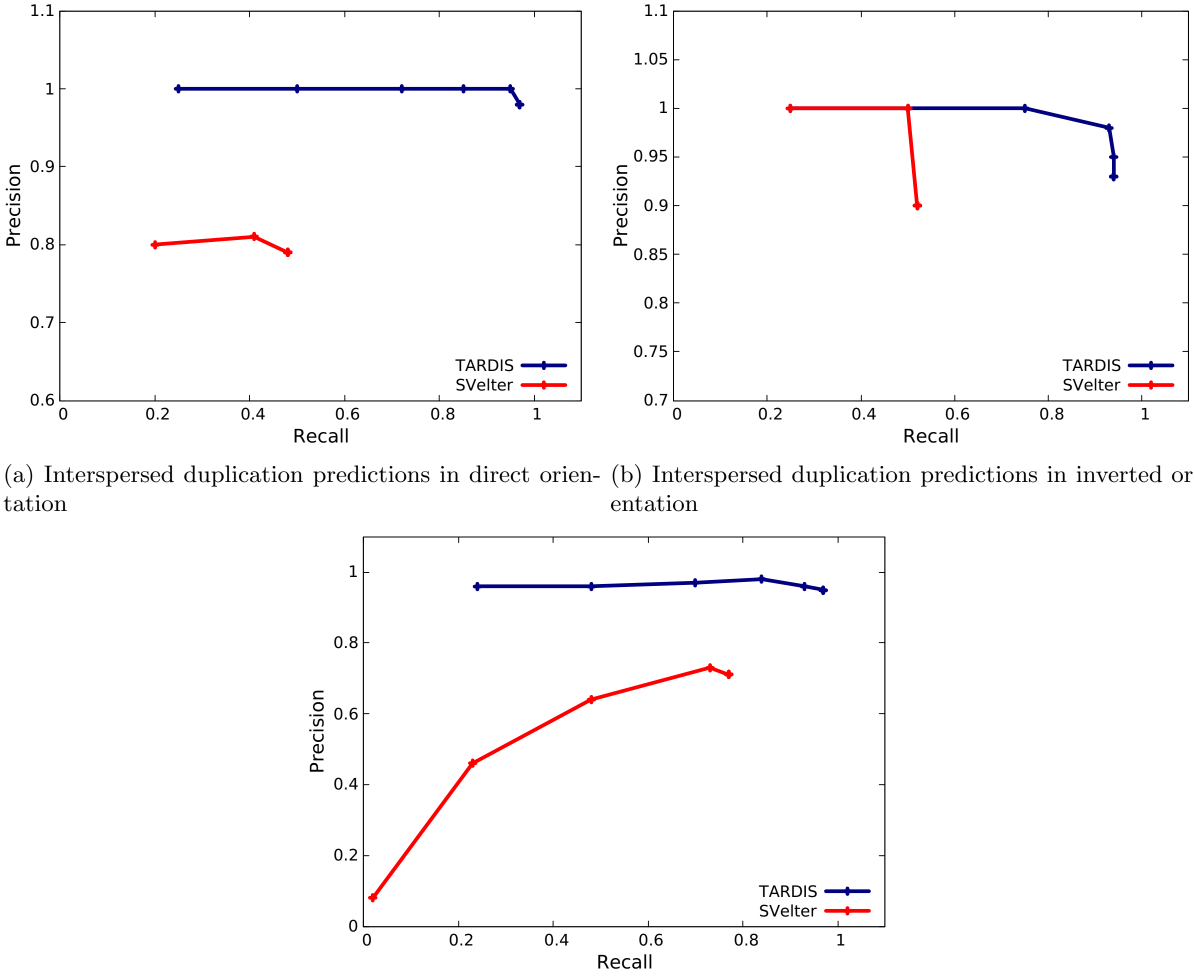


(a) Precision-recall curves for deletion predictions

(b) Precision-recall curves for inversion predictions

(c) Precision-recall curves for duplication predictions

Supplementary Fig. 2: Precision-Recall curves for the comparison of deletion (a) inversion (b) and deletion (c) duplication predictions on the simulated dataset for 30X coverage using TARDIS, TIDDIT, LUMPY, DELLY, SVelter and SoftSV.

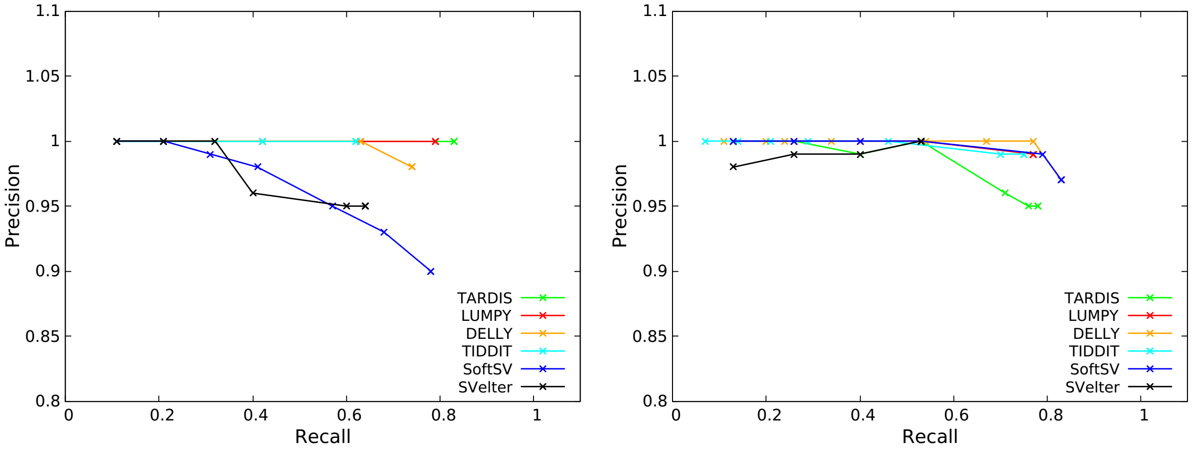


(a) Interspersed duplication predictions in direct orientation

(b) Interspersed duplication predictions in inverted orientation

(c) Tandem duplication predictions

Supplementary Fig. 3: Precision-Recall curves for the comparison of (a) interspersed duplications in direct orientation (b) interspersed duplications in inverted orientation (c) and tandem duplications on the simulated dataset for 30X coverage using TARDIS and SVelter



(a) Precision-recall curves for deletion predictions

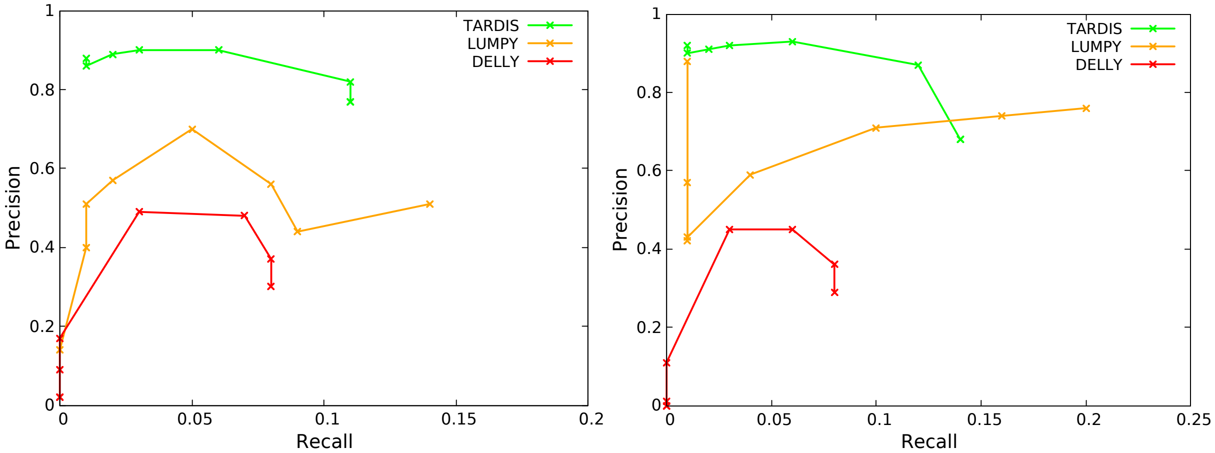
(b) Precision-recall curves for inversion predictions

Supplementary Fig. 4: Precision-Recall curves for the comparison of deletion (a) and inversion (b) predictions excluding the duplication regions on the simulated dataset for 30X coverage using TARDIS, TIDDIT, LUMPY, DELLY, SVelter and SoftSV.

Supplementary Table S2: Effect of Read-Pair Support for SV discovery for TARDIS

RP support cut-off	SV Type	# SVs	True	False	Miss	FDR	TPR
5	Deletion	700	624	5	76	0.01	0.89
	Inversion	579	496	34	83	0.06	0.86
	Interspersed Dups.	200	195	9	5	0.04	0.98
	Inverted Dups.	200	191	33	9	0.15	0.96
	Tandem Dups.	200	194	26	6	0.12	0.97
8	Deletion	700	615	2	85	0.00	0.88
	Inversion	579	493	32	86	0.06	0.85
	Interspersed Dups.	200	195	17	5	0.08	0.98
	Inverted Dups.	200	191	28	9	0.13	0.96
	Tandem Dups.	200	191	22	9	0.10	0.96
9	Deletion	700	609	2	91	0.00	0.87
	Inversion	579	491	31	88	0.06	0.85
	Interspersed Dups.	200	194	8	6	0.04	0.97
	Inverted Dups.	200	190	28	9	0.13	0.96
	Tandem Dups.	200	191	20	9	0.09	0.96
10	Deletion	700	608	1	92	0.00	0.87
	Inversion	579	491	31	88	0.06	0.85
	Interspersed Dups.	200	194	8	6	0.04	0.97
	Inverted Dups.	200	190	27	10	0.12	0.95
	Tandem Dups.	200	190	20	10	0.10	0.95
20	Deletion	700	581	1	119	0.00	0.83
	Inversion	579	476	25	103	0.05	0.82
	Interspersed Dups.	200	193	5	7	0.03	0.97
	Inverted Dups.	200	188	20	12	0.10	0.94
	Tandem Dups.	200	176	20	24	0.10	0.88

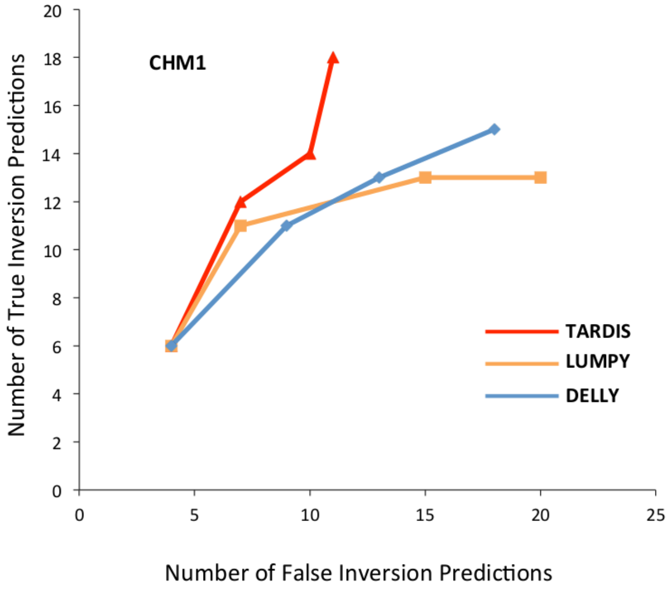
Table shows the effect of read-pair support cut-off value in SV discovery accuracy, that is, the number of minimum supporting read-pairs for an SV to be selected. The analysis were done using the simulated data of 60x coverage. FDR and TPR denotes false discovery and true positive/recall rates respectively.



(a) Precision-recall curves for CHM1

(b) Precision-recall curves for CHM13

Supplementary Fig. 5: Precision - Recall curve for the comparison of deletion predictions on (a) CHM1 and (b) CHM13 genomes. Here we compare against predicted inversions using PacBio reads based on BLASR mappings. Overall TARDIS achieves better accuracy than the two other approaches.



Supplementary Fig. 6: Validation of top predicted inversions of different tools using local assembly of the PacBio reads for CHM1.

Supplementary Table S3: 50 highest scoring segmental duplications predicted by TARDIS in the CHM1 genome.

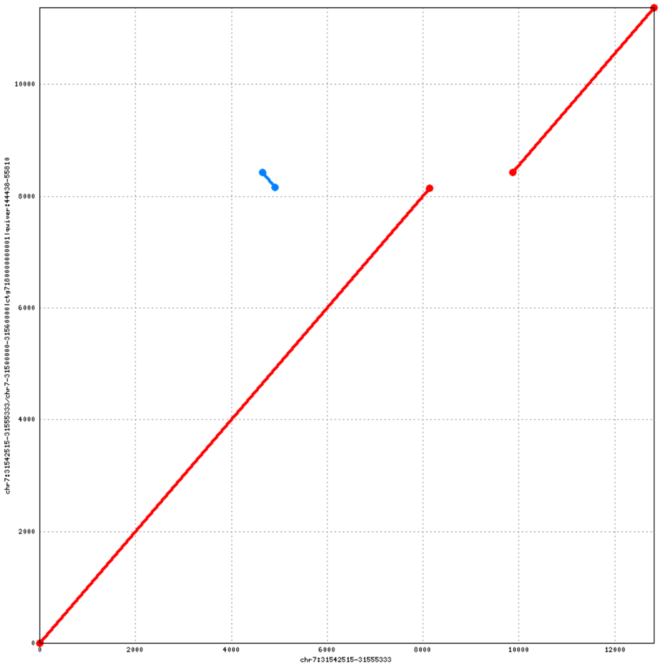
Duplication	TARDIS		Validation	Duplication	TARDIS		Validation
Insertion Locus	Dup. Type	Score	(PacBio)	Insertion Locus	Dup. Type	Score	(PacBio)
chr11 63,701,552 - 63,702,044	Direct	0.000096	True	chr2 37,928,294 - 38,101,823	Tandem	0.000073	N/A
chr3 194,546,158 - 194,546,552	Direct	0.000100	True	chr20 60,032,848 - 60,033,403	Tandem	0.000080	True
chr5 143,512,369 - 143,512,967	Direct	0.000139	True	chr5 3,323,855 - 3,324,309	Tandem	0.000106	N/A
chr2 240,640,651 - 240,641,122	Direct	0.000199	True	chr7 2,554,464 - 2,554,791	Tandem	0.000111	True
chr20 2,359,605 - 2,360,003	Direct	0.000271	True	chr12 110,099,340 - 110,099,746	Tandem	0.000117	True
chr9 112,285,747 - 112,286,145	Direct	0.000300	True	chr6 168,052,194 - 168,052,468	Tandem	0.000117	True
chr8 2,215,143 - 2,215,392	Direct	0.000310	N/A	chr1 207,097,489 - 207,097,910	Tandem	0.000123	True
chr18 69,711,702 - 69,712,115	Direct	0.000323	True	chr16 86,008,734 - 86,009,147	Tandem	0.000127	True
chr17 46,615,512 - 46,615,903	Direct	0.000326	True	chr17 80,317,607 - 80,318,019	Tandem	0.000127	N/A
chr6 160,877,582 - 160,878,047	Direct	0.000342	N/A	chr10 127,513,435 - 127,513,672	Tandem	0.000129	True
chr2 125,052,915 - 125,053,261	Inverted	0.000088	True	chr14 106,049,125 - 106,049,349	Tandem	0.000129	True
chr3 43,834,996 - 43,835,748	Inverted	0.000089	True	chr6 44,012,338 - 44,012,957	Tandem	0.000129	True
chr14 67,171,710 - 67,172,020	Inverted	0.000092	True	chr9 132,158,817 - 132,159,088	Tandem	0.000129	N/A
chr2 72,440,071 - 72,440,597	Inverted	0.000105	True	chr12 13,164,470 - 13,164,800	Tandem	0.000136	True
chr9 107,816,537 - 107,817,079	Inverted	0.000140	True	chr20 62,720,020 - 62,720,215	Tandem	0.000136	True
chr17 36,405,748 - 36,407,397	Inverted	0.000149	False	chr10 132,974,754 - 132,975,320	Tandem	0.000144	True
chr1 114,645,858 - 114,646,155	Inverted	0.000235	True	chr8 2,215,817 - 2,216,236	Tandem	0.000144	N/A
chr5 115,350,905 - 115,351,086	Inverted	0.000236	True	chr9 34,681,581 - 34,681,899	Tandem	0.000194	True
chr12 71,532,699 - 71,533,378	Inverted	0.000245	True	chr6 35,754,661 - 35,766,731	Tandem	0.000255	True
chr7 31,586,861 - 31,587,129	Inverted	0.000278	True	chr20 62,123,612 - 62,124,210	Tandem	0.000257	True
chr18 11,511,287 - 11,511,480	Inverted	0.000280	True	chr20 59,567,884 - 59,590,251	Tandem	0.000268	True
				chr18 77,831,329 - 77,831,784	Tandem	0.000273	N/A
				chrX 417,958 - 418,361	Tandem	0.000273	True
				chr20 42,325,214 - 42,325,573	Tandem	0.000290	True
				chr19 34,882,471 - 34,883,258	Tandem	0.000310	True
				chr2 3,184,299 - 3,185,046	Tandem	0.000310	N/A
				chr3 197,117,159 - 197,117,807	Tandem	0.000318	N/A

Here we list the insertion locations of the top 50 scoring segmental duplications in CHM1 genome. All predictions are sorted by the SV score (lower is better). If the validation is N/A, that means the incorrect prediction from PacBio data, which will be skipped in the comparison. TARDIS only gives one false call and three interspersed duplications that are wrongly assigned to tandem duplications.

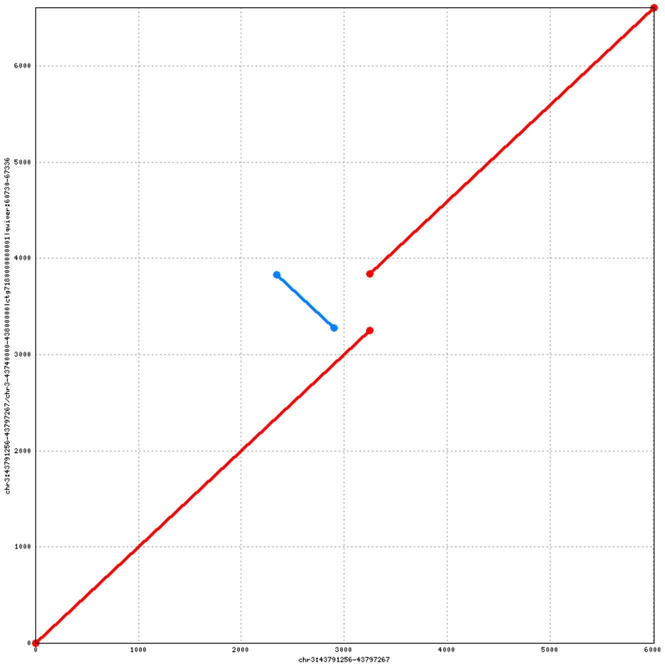
Supplementary Table S4: 20 highest scoring segmental duplications predicted by TARDIS in the CHM13 genome.

Duplication	TARDIS		Duplication	TARDIS	
Insertion Locus	Dup. Type	Score	Insertion Locus	Dup. Type	Score
chr8 58,116,437 - 58,118,469	Direct	0.000188	chr17 80,864,373 - 81,006,658	Tandem	0.000054
chr6 119,011,599 - 119,012,388	Direct	0.000217	chr16 81,798,809 - 81,799,175	Tandem	0.000121
chr6 57,209,065 - 57,297,292	Direct	0.000233	chr2 87,623,860 - 87,642,147	Tandem	0.000125
chr5 143,512,394 - 143,512,967	Direct	0.000234	chr19 34,882,364 - 34,882,984	Tandem	0.000216
chr11 63,701,560 - 63,702,044	Direct	0.000234	chr22 49,780,535 - 49,780,919	Tandem	0.000273
chr8 76,769,879 - 76,770,323	Direct	0.000247	chr5 1,044,880 - 1,045,357	Tandem	0.000273
chr3 194,546,160 - 194,546,552	Direct	0.000261	chr6 44,012,353 - 44,012,977	Tandem	0.000273
chr5 140,859,762 - 140,860,171	Direct	0.000269			
chr14 48,325,266 - 48,325,533	Inverted	0.000208			
chr11 98,844,907 - 98,845,328	Inverted	0.000210			
chr2 61,703,139 - 61,703,479	Inverted	0.000210			
chr5 169,597,408 - 169,597,797	Inverted	0.000220			
chr12 78,387,851 - 78,388,292	Inverted	0.000246			

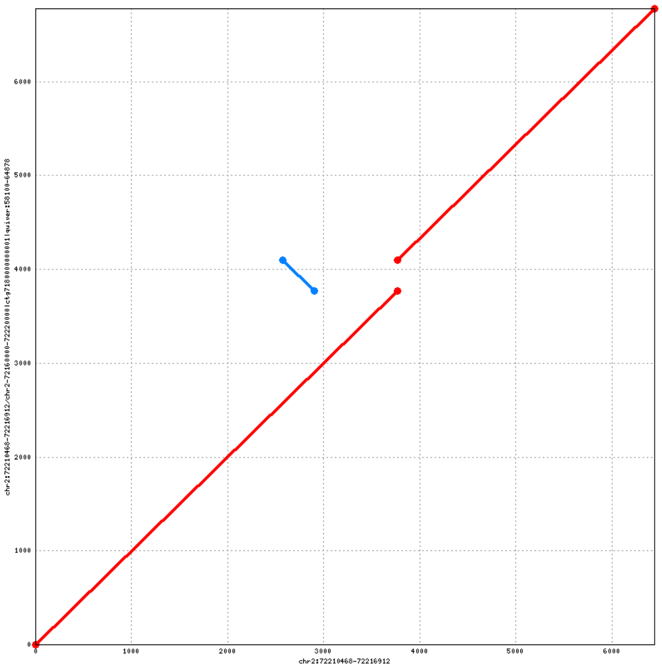
Here we list the insertion locations of the top 20 scoring segmental duplications in CHM13 genome. All predictions are sorted by the SV score (lower is better).



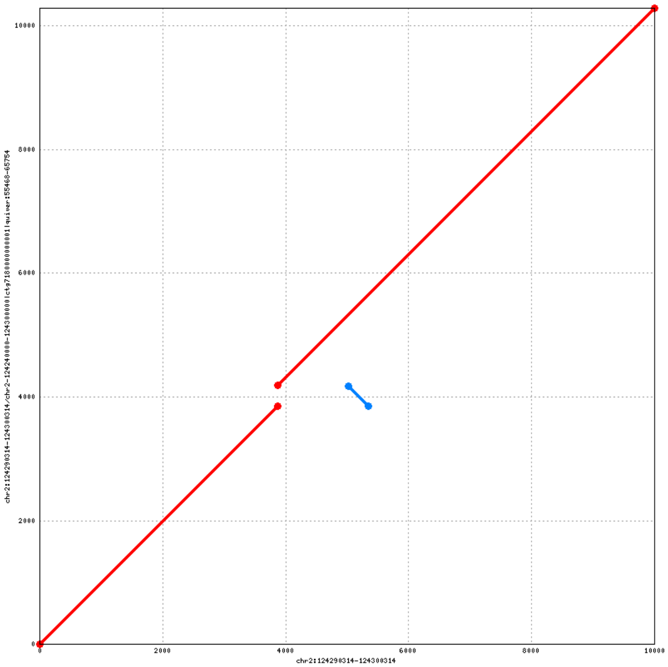
Supplementary Fig. 7: Inversion predicted within 7:31,586,823-31,590394.



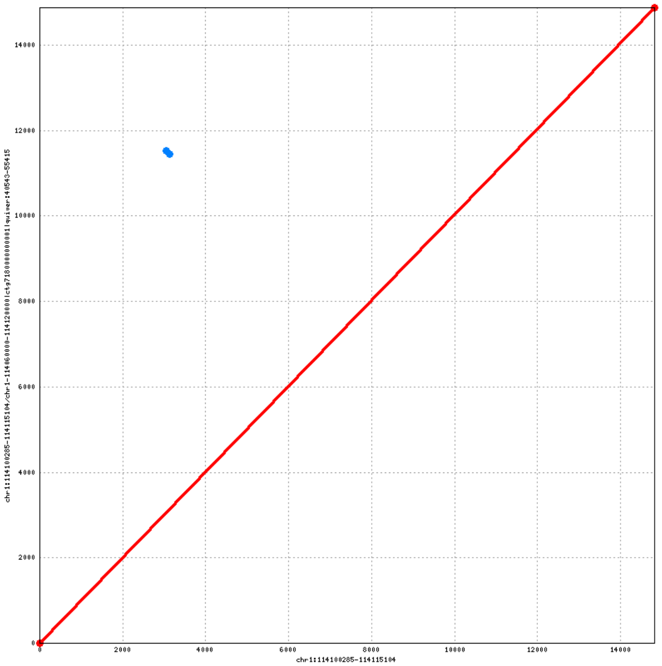
Supplementary Fig. 8: Inverted Duplication predicted within 3:43,834,994-43,836,299.



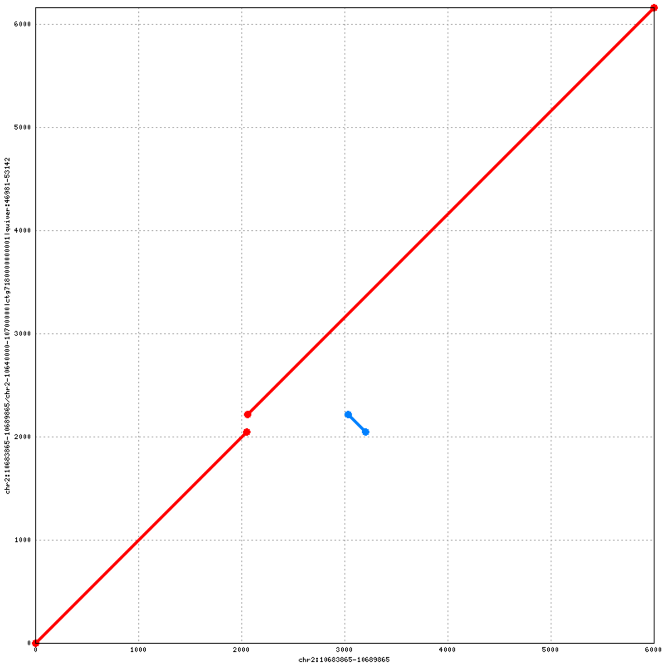
Supplementary Fig. 9: Inverted Duplication predicted within 2:72,440,066-72,441,647.



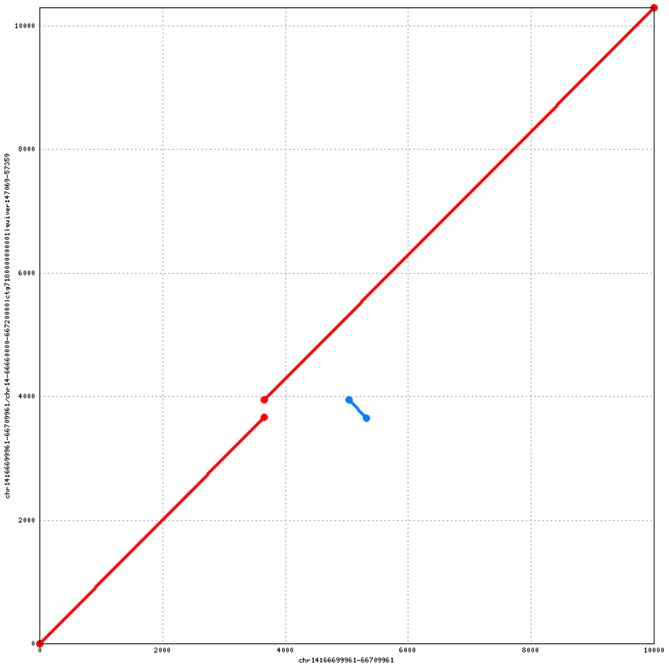
Supplementary Fig. 10: Inverted Duplication predicted within 2:125051481-125053239



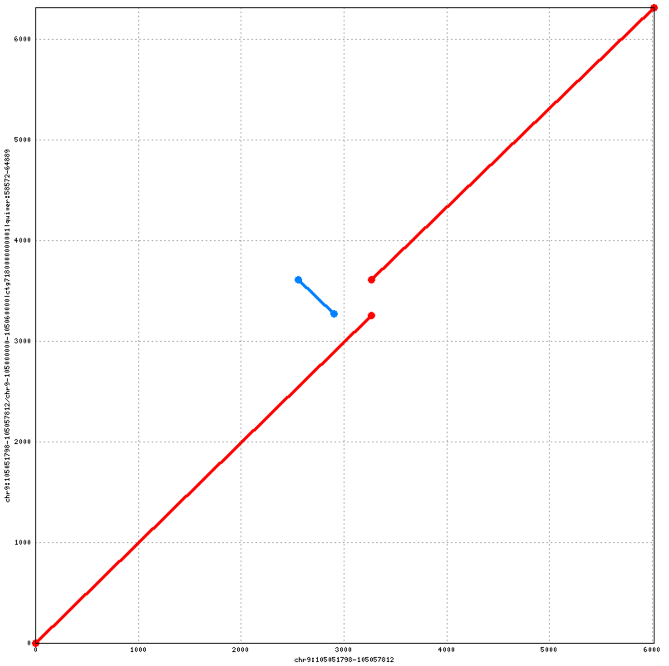
Supplementary Fig. 11: Inverted Duplication predicted within 1:114,645,854-114,654,623.



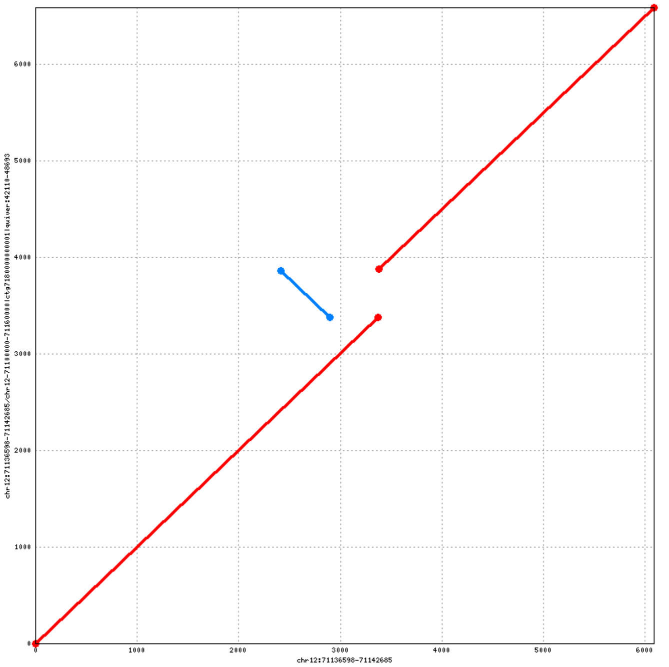
Supplementary Fig. 12: Inverted Duplication predicted within 2:10,825,652-10,827,218.



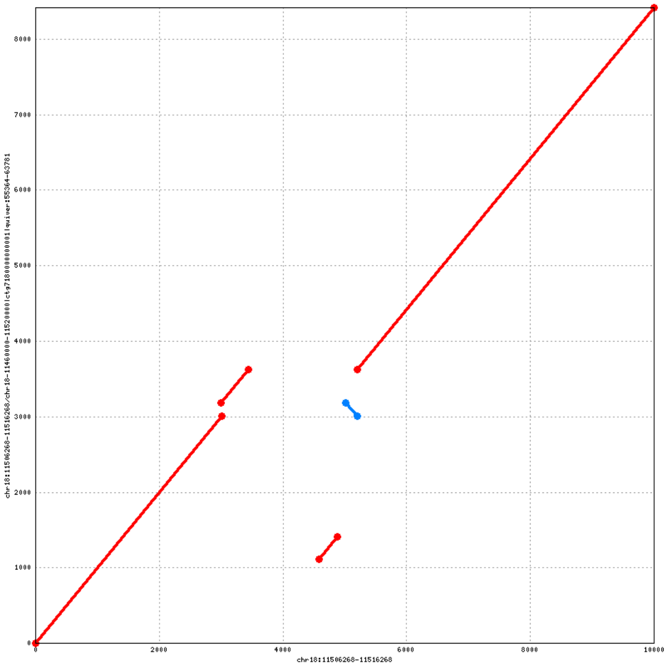
Supplementary Fig. 13: Inverted Duplication predicted within 14:67,169,917-67,171,999.



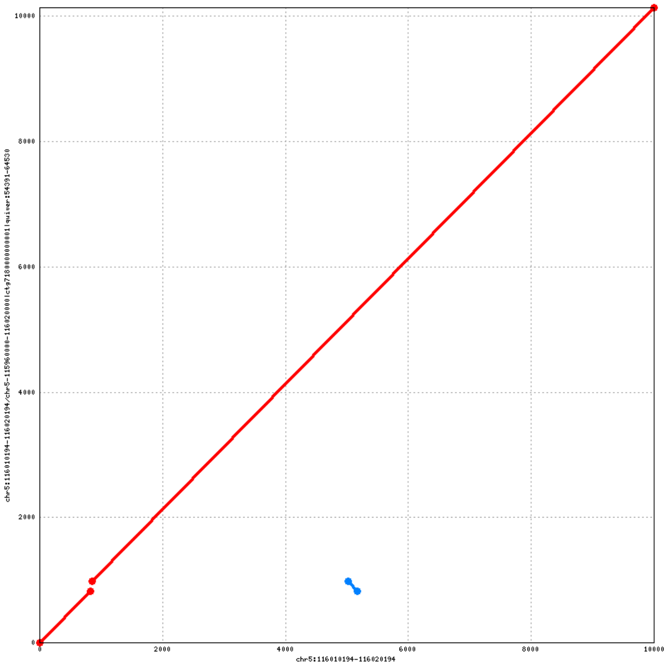
Supplementary Fig. 14: Inverted Duplication predicted within 9:107,816,536-107,817,623.



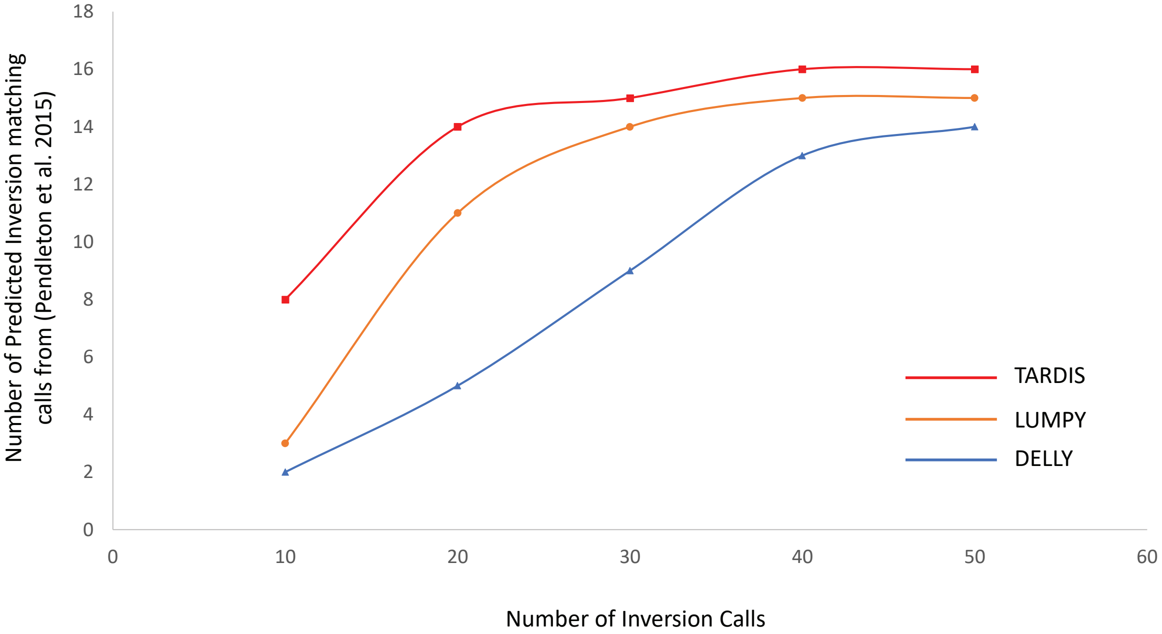
Supplementary Fig. 15: Inverted Duplication predicted within 12:71,532,693-71,534,000.



Supplementary Fig. 16: Inverted Duplication predicted within 18:11,508,829-11,511,479.



Supplementary Fig. 17: Inverted Duplication predicted within 5:115,346,294-115,351,084.



Supplementary Fig. 18: Comparison of top inversion predictions on NA12878 sample against predicted and validated set of inversion of the same samples using PacBio data from [1].

Supplementary Table S5: Performance comparison in terms of time and memory usage for CHM1 and NA12878 genomes.

	CHM1		NA12878	
	Time	Memory	Time	Memory
TARDIS	2h 05m	9 GB	3h 56m	9 GB
TIDDIT	1h 31m	5 GB	2h 25m	5 GB
LUMPY	7h 38m	7 GB	11h 05m	8 GB
DELLY	64h 22m	7 GB	33h 05m	9 GB
SoftSV	175h 26m	4 GB	137h 35m	2 GB

We benchmarked TARDIS, TIDDIT, LUMPY, DELLY and SoftSV using a haploid (CHM1) and a diploid (NA12878) genome with the same computing resources (Intel(R) Xeon(R) CPU E7- 4830 @ 2.13GHz : 4 CPUs × 8 cores each = 32 cores total 512 GB RAM)

References

[1] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12:780–786, August 2015.