Supplementary Data: Efficient detection of novel sequence insertions using low-input linked-read sequencing

Dmitry Meleshko[1,2], Rui Yang[1], Patrick Marks[3], Stephen Williams[3], and Iman Hajirasouliha[2,4,*]

[1]Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medicine of Cornell University, NY, 10021, USA
[2]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, NY, 10021, USA
[3]10x Genomics Inc., Pleasanton, California, 94566, USA
[4]Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, 10021, USA
*Corresponding author

Method	TP	FP	Missed	Sensitivity	Precision	F1
Coverage = 13X						
Novel-X	763	2	1237	0.381	0.997	0.552
PopIns	134	2610	1866	0.07	0.05	0.06
NUI	1362	3	638	0.68	0.998	0.81
Coverage = 26X						
Novel-X	1803	3	197	0.902	0.998	0.948
PopIns	504	2775	1496	0.25	0.182	0.211
NUI	1725	2	275	0.863	0.999	0.923
Coverage = 39X						
Novel-X	1842	1	158	0.921	0.9995	0.959
PopIns	779	2342	1221	0.390	0.250	0.304
NUI	1754	2	246	0.88	0.999	0.934
Coverage = 52X						
Novel-X	1855	1	145	0.928	0.999	0.962
PopIns	1011	2261	989	0.505	0.309	0.384
NUI	1746	3	254	0.873	0.998	0.931

Table S1: Comparison of Novel-X, PopIns and NUI-pipeline performance on downsampled data.

Assembly	Insertion reassembly	TP	Missed	FP
Velvet	SPAdes	1789	211	1
SPAdes	SPAdes	1680	320	5
Velvet	Supernova	1150	850	2
Supernova	SPAdes	1705	295	7
Velvet	Velvet	1386	614	0

Table S2: Performance of different assembly strategies on simulated data.

Novel insertions detection results for insertions longer than 300 bp						
Length (bp)	Validation set	Novel-X	Pamir	PopIns	NUI	Supernova/ Paftools
CHM1 dataset (40×)						
300-499	101	62 (**37**, **60%**)	61 (13, 21%)	163 (0, 0%)	-	-
500-999	138	90 (**54**, **60%**)	34 (13, 38%)	123 (1, 1%)	-	-
1000-1999	85	29 (**25**, **86%**)	6 (5, 83%)	82 (2, 2%)	-	-
≥2000	77	19 (**18**, **95%**)	1 (0, 0%)	28 (1, 5%)	-	-
Total(≥300)	401	200 (**134**, **67%**)	102 (31, 30%)	396 (4, 1%)	-	-
Max length (bp)	27836	**6822**	2072	2150	-	-
CHM13 dataset (40×)						
300-499	85	63 (**32**, **51%**)	68 (12, 18%)	175 (0, 0%)	-	-
500-999	126	80 (**45**, **56%**)	29 (7, 24%)	168 (3, 4%)	-	-
1000-1999	74	35 (**27**, **77%**)	8 (6, 75%)	83 (3, 2%)	-	-
≥2000	63	23 (**20**, **87%**)	4 (4, 100%)	34 (2, 6%)	-	-
Total(≥300)	348	201 (**124**, **62%**)	109 (29, 26%)	460 (8, 2%)	-	-
Max length (bp)	20444	**7175**	4136	5006	-	-
NA19240 dataset (73×)						
300-499	167	118 (**57**, **48%**)	69 (10, 14%)	232 (1, 0%)	162 (8, 5%)	122 (53, 43%)
500-999	238	156 (**94**, 60%)	42 (15, 36%)	185 (1, 1%)	66 (13, 20%)	145 (90, **62%**)
1000-1999	109	53 (32, **60%**)	14 (4, 29%)	76 (1, 1%)	76 (13, 17%)	64 (**35**, 55%)
≥2000	113	68 (49, **72%**)	4 (1, 25%)	14 (0, 0%)	77 (8, 10%)	86 (**57**, 66%)
Total(≥300)	627	395 (232, **59%**)	129 (30, 23%)	507 (3, 1%)	381 (40, 10%)	417 (**235**, 56%)
Max length (bp)	27821	**27836**	3630	5114	14919	19815
NA12878 dataset (60×)						
300-499	138	103 (**65**, 63%)	-	28 (1, 4%)	9 (1, 11%)	92 (41, 44%)
500-999	198	120 (**73**, 61%)	-	20 (0, 0%)	17 (6, 35%)	123 (58, 47%)
1000-1999	96	35 (**20**, 57%)	-	7 (0, 0%)	2 (0, 0%)	41 (17, 41%)
≥2000	94	43 (27, 63%)	-	1 (0, 0%)	3 (0, 0%)	74 (**42**, 57%)
Total(≥300)	526	301 (**185**, 61%)	-	56 (1, 2%)	31 (7, 22%)	330 (158, 48%)
Max length (bp)	20442	**27836**	-	308	3721	4935
HG002 dataset (59×)						
300-499	133	99 (**42**, **42%**)	-	177 (0, 0%)	-	117 (36, 30%)
500-999	215	150 (**58**, 39%)	-	89 (1, 1%)	-	159 (48, 30%)
1000-1999	103	48 (14, 29%)	-	24 (0, 0%)	-	64 (**20**, 31%)
≥2000	62	18 (7, 39%)	-	2 (0, 0%)	-	42 (**14**, 33%)
Total(≥300)	513	315 (**121**, 38%)	-	292 (1, 0%)	-	382 (118, 31%)
Max length (bp)	24323	5690	-	573	-	**10203**

Table S3: Length breakdown and comparison between the validation set (constructed using SMRT-SV callset or multiple orthogonal methods callset), short-read methods Pamir and PopIns, and linked-read methods Novel-X, NUI, and Supernova/Paftools for 10X Genomics datasets. The numbers in brackets indicate the count of overlaps with SMRT-SV calls and the percentage of the overlapping calls. As it can be seen, the number of validated novel insertion calls using our method is significantly higher than those obtained by short-read methods.

Novel insertions detection results for insertions longer than 300 bp (UST Tell-Seq and stLFR datasets)					
Length (bp)	Validation set	Novel-X	PopIns	NUI	Supernova/ Paftools
NA12878 - UST Tell-Seq (41×)					
300-499	138	62 (**44**, **71%**)	503 (4, 1%)	322 (1, 0%)	131 (36, 27%)
500-999	198	58 (43, **74%**)	327 (3, 1%)	228 (2, 1%)	138 (**60**, 43%)
1000-1999	96	21 (17, **81%**)	157 (1, 1%)	124 (1, 1%)	55 (**31**, 56%)
≥2000	94	16 (14, **88%**)	140 (3, 2%)	124 (1, 1%)	41 (**27**, 66%)
Total(≥300)	526	157 (118, **75%**)	1127 (11, 1%)	798 (5, 1%)	367 (**154**, 42%)
Max length (bp)	20442	12143	4127	12667	**19696**
NA12878 - stLFR (40×)					
300-499	138	74 (50, **68%**)	385 (2, 1%)	0 (0, 0%)	128 (**68**, 53%)
500-999	198	65 (44, **68%**)	270 (5, 2%)	0 (0, 0%)	171 (**91**, 53%)
1000-1999	96	27 (18, **67%**)	112 (4, 4%)	0 (0, 0%)	70 (**35**, 50%)
≥2000	94	22 (14, **64%**)	62 (3, 5%)	0 (0, 0%)	45 (**28**, 62%)
Total(≥300)	526	188 (126, **67%**)	829 (14, 2%)	0 (0, 0%)	414 (**222**, 54%)
Max length (bp)	20442	12074	4957	0	**12667**
HG002 - UST Tell-Seq (42×)					
300-499	133	68 (36, **75%**)	104 (0, 0%)	410 (**52**, 13%)	99 (29, 29%)
500-999	215	77 (34, **44%**)	82 (1, 1%)	308 (**84**, 27%)	141 (48, 34%)
1000-1999	103	11 (5, **45%**)	82 (0, 0%)	182 (**37**, 20%)	45 (14, 10%)
≥2000	62	3 (2, **67%**)	63 (0, 0%)	182 (**14**, 8%)	7 (3, 43%)
Total(≥300)	513	159 (77, **48%**)	249 (1, 0%)	1082 (**187**, 17%)	292 (94, 32%)
Max length (bp)	24323	3630	537	**13470**	3631
HG002 - stLFR (35×)					
300-499	133	45 (21, **47%**)	365 (2, 1%)	259 (**39**, 15%)	91 (32, 35%)
500-999	215	66 (39, **59%**)	185 (4, 2%)	170 (**44**, 26%)	96 (31, 32%)
1000-1999	103	9 (5, **56%**)	17 (1, 6%)	94 (**16**, 17%)	35 (8, 23%)
≥2000	62	3 (2, **67%**)	2 (0, 0%)	91 (**6**, 7%)	8 (2, 25%)
Total(≥300)	513	123 (67, **54%**)	569 (7, 1%)	614 (**105**, 17%)	230 (73, 32%)
Max length (bp)	24323	3714	1033	**10549**	7360

Table S4: Length breakdown and comparison between the validation set (constructed using SMRT-SV callset or multiple orthogonal methods callset), short-read method PopIns, and linked-read methods Novel-X, NUI, and Supernova/Paftools for UST Tell-Seq and stLFR datasets. The numbers in brackets indicate the count of overlaps with SMRT-SV calls and the percentage of the overlapping calls.

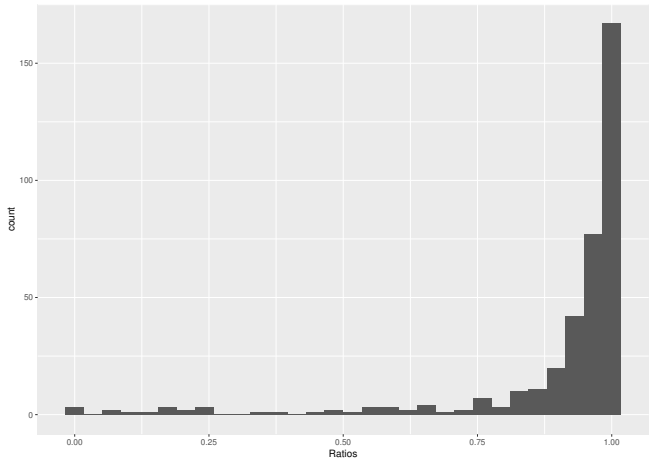

Figure S1: Histogram of intersection of barcodes associated with insertions and found on the reference to associated with insertion barcodes ratio.

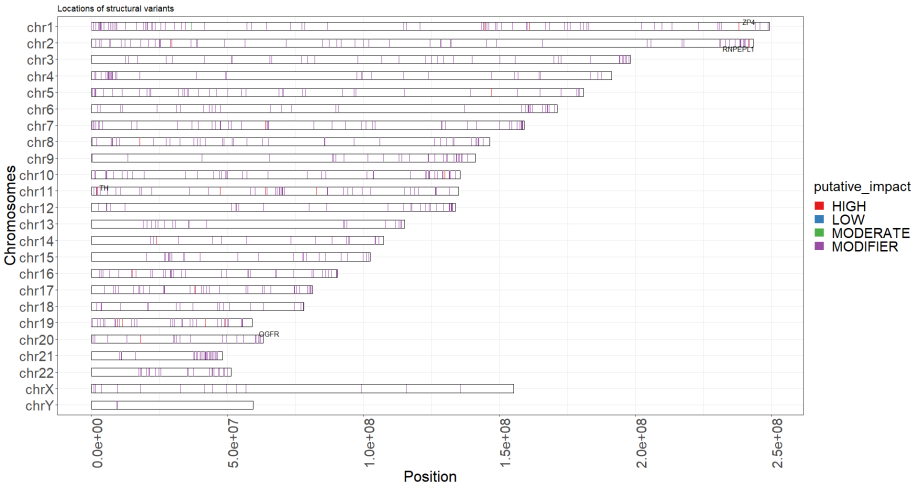

Figure S2: Chromatin location of insertion hotspots and highly disrupted genes. Vertical marks show insertion locations across chromosomes, colored by putative impact. Highly disrupted genes located at insertion hotspots (more than 30 samples have insertions) are annotated.

Ancestry	Location	Samples with insertion	Total samples	Percentage
African	chr20 11312665	22	24	0.9167
African	chr16 71384688	20	24	0.8333
African	chr10 131512438	23	24	0.9583
European	chr20 61444279	21	22	0.9545
European	chr21 38914031	19	22	0.8636
European	chr2 218082962	22	22	1.0
European	chr2 234514767	22	22	1.0
European	chr2 236887596	20	22	0.9091
European	chr2 241517117	20	22	0.9091
European	chr2 2750405	22	22	1.0
European	chr3 126539876	18	22	0.8182

European	chr4 138867027	19	22	0.8636
European	chr4 38734186	19	22	0.8636
European	chr4 38754378	19	22	0.8636
European	chr4 7131367	22	22	1.0
European	chr6 11491594	18	22	0.8182
European	chr6 116047605	18	22	0.8182
European	chr6 165678196	18	22	0.8182
European	chr6 2376247	20	22	0.9091
European	chr6 39845370	21	22	0.9545
European	chr7 131654753	20	22	0.9091
European	chr8 142432223	21	22	0.9545
European	chr8 142502428	20	22	0.9091
European	chr8 21766795	21	22	0.9545
European	chr8 40820766	20	22	0.9091
European	chr8 47764037	19	22	0.8636
European	chr8 48174023	21	22	0.9545
European	chr9 121895762	19	22	0.8636
European	chr20 25067132	18	22	0.8182
European	chr20 22471994	19	22	0.8636
European	chr1 204195225	18	22	0.8182
European	chr9 123347685	22	22	1.0
Asian	chr9 88974428	17	19	0.8947
Asian	chrX 140723716	16	19	0.8421
European	chr10 131512438	19	22	0.8636
European	chr10 46526439	19	22	0.8636
European	chr10 49742424	21	22	0.9545
European	chr10 84601962	18	22	0.8182
European	chr11 126799470	22	22	1.0
European	chr11 2187926	20	22	0.9091
European	chr12 110880724	18	22	0.8182
European	chr12 57191929	20	22	0.9091
European	chr13 111435138	20	22	0.9091
European	chr13 111621629	19	22	0.8636
European	chr1 27086496	19	22	0.8636
European	chr13 112972801	19	22	0.8636
European	chr14 82287525	19	22	0.8636
European	chr15 26893428	21	22	0.9545
European	chr16 10040003	21	22	0.9545
European	chr16 87502298	20	22	0.9091
European	chr18 37283950	18	22	0.8182
European	chr18 63103455	19	22	0.8636
European	chr18 79541908	20	22	0.9091
European	chr19 1550008	18	22	0.8182
European	chr19 33250333	22	22	1.0
European	chr19 33458810	19	22	0.8636
European	chr1 168395496	19	22	0.8636
European	chr1 182138477	18	22	0.8182
European	chr13 22679414	18	22	0.8182
European	chr9 36764431	20	22	0.9091
European	chr9 88974428	21	22	0.9545
European	chr9 89254524	18	22	0.8182

Hispanic	chr21 41672810	3	3	1.0
Hispanic	chr21 7915747	3	3	1.0
Hispanic	chr2 218082962	3	3	1.0
Hispanic	chr2 236887598	3	3	1.0
Hispanic	chr2 241517117	3	3	1.0
Hispanic	chr2 2750405	3	3	1.0
Hispanic	chr3 72211325	3	3	1.0
Hispanic	chr4 101164092	3	3	1.0
Hispanic	chr4 38754378	3	3	1.0
Hispanic	chr4 7131367	3	3	1.0
Hispanic	chr5 135778794	3	3	1.0
Hispanic	chr6 11491592	3	3	1.0
Hispanic	chr6 116047604	3	3	1.0
Hispanic	chr6 167686999	3	3	1.0
Hispanic	chr6 3776803	3	3	1.0
Hispanic	chr7 102932561	3	3	1.0
Hispanic	chr7 156101312	3	3	1.0
Hispanic	chr7 31657167	3	3	1.0
Hispanic	chr8 130897691	3	3	1.0
Hispanic	chr8 139157950	3	3	1.0
Hispanic	chr8 142432223	3	3	1.0
Hispanic	chr8 142502431	3	3	1.0
Hispanic	chr8 17796330	3	3	1.0
Hispanic	chr8 41900622	3	3	1.0
Hispanic	chr8 47764037	3	3	1.0
Hispanic	chr21 38914020	3	3	1.0
Hispanic	chr21 38059653	3	3	1.0
Hispanic	chr21 37707714	3	3	1.0
Hispanic	chr20 61444268	3	3	1.0
European	chr9 93618898	18	22	0.8182
European	chrX 153382567	21	22	0.9545
Hispanic	chr10 131512438	3	3	1.0
Hispanic	chr11 126799421	3	3	1.0
Hispanic	chr11 2187926	3	3	1.0
Hispanic	chr11 30928615	3	3	1.0
Hispanic	chr11 70803384	3	3	1.0
Hispanic	chr12 132166674	3	3	1.0
Hispanic	chr12 132742762	3	3	1.0
Hispanic	chr13 111435138	3	3	1.0
Hispanic	chr13 112972799	3	3	1.0
Hispanic	chr14 100526933	3	3	1.0
Asian	chr9 123347685	18	19	0.9474
Hispanic	chr14 34047463	3	3	1.0
Hispanic	chr16 86013934	3	3	1.0
Hispanic	chr16 87502298	3	3	1.0
Hispanic	chr18 13982054	3	3	1.0
Hispanic	chr18 854878	3	3	1.0
Hispanic	chr19 29010235	3	3	1.0
Hispanic	chr19 33250333	3	3	1.0
Hispanic	chr19 33458810	3	3	1.0
Hispanic	chr1 168395496	3	3	1.0

Hispanic	chr1 230332053	3	3	1.0
Hispanic	chr1 25529503	3	3	1.0
Hispanic	chr1 34541476	3	3	1.0
Hispanic	chr20 61337292	3	3	1.0
Hispanic	chr14 82287525	3	3	1.0
Asian	chr9 121895747	16	19	0.8421
Asian	chr8 48174023	17	19	0.8947
Asian	chr8 47764037	16	19	0.8421
African	chr2 2750405	24	24	1.0
African	chr2 98660051	23	24	0.9583
African	chr3 126539876	23	24	0.9583
African	chr4 101164092	20	24	0.8333
African	chr4 138867044	22	24	0.9167
African	chr4 156847734	20	24	0.8333
African	chr4 157982654	22	24	0.9167
African	chr4 38734186	21	24	0.875
African	chr4 7131367	23	24	0.9583
African	chr5 42089532	21	24	0.875
African	chr6 167686999	20	24	0.8333
African	chr6 2376247	21	24	0.875
African	chr6 39845371	20	24	0.8333
African	chr8 1337419	20	24	0.8333
African	chr8 142502498	22	24	0.9167
African	chr8 17796249	20	24	0.8333
African	chr8 21766795	21	24	0.875
African	chr8 29820462	21	24	0.875
African	chr8 40820766	23	24	0.9583
African	chr8 47764037	22	24	0.9167
African	chr9 121895747	20	24	0.8333
African	chr9 123347685	22	24	0.9167
African	chr9 36764431	20	24	0.8333
African	chr9 88974428	22	24	0.9167
African	chrX 140723716	22	24	0.9167
African	chr2 241517117	22	24	0.9167
African	chr2 236887598	23	24	0.9583
African	chr2 235169315	21	24	0.875
African	chr2 234514767	23	24	0.9583
African	chr10 46526439	23	24	0.9583
African	chr10 49742424	20	24	0.8333
African	chr10 63783076	20	24	0.8333
African	chr10 84601964	20	24	0.8333
African	chr11 132125738	21	24	0.875
African	chr12 57191936	22	24	0.9167
African	chr13 111621629	21	24	0.875
African	chr13 112972801	21	24	0.875
African	chr13 22679414	21	24	0.875
African	chr15 26893415	23	24	0.9583
African	chr15 83695234	22	24	0.9167
African	chr15 94051934	22	24	0.9167
African	chrX 153382567	21	24	0.875
African	chr17 5622802	20	24	0.8333

African	chr18 63103455	20	24	0.8333
African	chr18 79541908	20	24	0.8333
African	chr19 29010235	20	24	0.8333
African	chr19 33250333	21	24	0.875
African	chr1 14492689	22	24	0.9167
African	chr1 168395498	22	24	0.9167
African	chr1 182138477	20	24	0.8333
African	chr20 22471994	22	24	0.9167
African	chr20 25067132	21	24	0.875
African	chr20 61444279	24	24	1.0
African	chr21 7915747	20	24	0.8333
African	chr2 218082962	20	24	0.8333
African	chr18 61644856	22	24	0.9167
Hispanic	chr8 48174023	3	3	1.0
Asian	chr10 63783073	16	19	0.8421
Asian	chr11 132125738	17	19	0.8947
Asian	chr2 218082962	17	19	0.8947
Asian	chr2 234514767	19	19	1.0
Asian	chr2 236887598	18	19	0.9474
Asian	chr2 241517117	16	19	0.8421
Asian	chr2 2750405	19	19	1.0
Asian	chr3 126539876	16	19	0.8421
Asian	chr3 72211314	16	19	0.8421
Asian	chr3 85548236	16	19	0.8421
Asian	chr4 157982644	17	19	0.8947
Asian	chr4 38734186	19	19	1.0
Asian	chr4 38754378	18	19	0.9474
Asian	chr4 7131367	19	19	1.0
Asian	chr6 11491594	16	19	0.8421
Asian	chr6 116047605	17	19	0.8947
Asian	chr6 157238333	16	19	0.8421
Asian	chr6 2376247	18	19	0.9474
Asian	chr6 39845370	19	19	1.0
Asian	chr7 102932561	16	19	0.8421
Asian	chr7 31657167	17	19	0.8947
Asian	chr7 99219897	16	19	0.8421
Asian	chr8 142432223	19	19	1.0
Asian	chr8 142502470	18	19	0.9474
Asian	chr8 21766795	17	19	0.8947
Asian	chr8 29820462	18	19	0.9474
Asian	chr8 40820766	19	19	1.0
Asian	chr2 14274479	17	19	0.8947
Asian	chr22 49276647	17	19	0.8947
Asian	chr21 7915747	17	19	0.8947
Asian	chr21 38508035	17	19	0.8947
Asian	chr11 2187926	16	19	0.8421
Asian	chr11 69291997	17	19	0.8947
Asian	chr12 110880724	16	19	0.8421
Asian	chr12 57191929	17	19	0.8947
Asian	chr13 111435138	17	19	0.8947
Asian	chr13 111621629	16	19	0.8421

Asian	chr13 112972801	18	19	0.9474
Asian	chr13 22679414	16	19	0.8421
Asian	chr15 26893428	17	19	0.8947
Asian	chr15 94051934	16	19	0.8421
Asian	chr16 10040003	17	19	0.8947
Asian	chr17 5622822	17	19	0.8947
Asian	chr10 84601964	17	19	0.8947
Asian	chr18 13982054	17	19	0.8947
Asian	chr18 79541908	17	19	0.8947
Asian	chr18 854878	18	19	0.9474
Asian	chr19 1550008	17	19	0.8947
Asian	chr19 33250333	19	19	1.0
Asian	chr19 33458810	17	19	0.8947
Asian	chr1 14492712	18	19	0.9474
Asian	chr1 168395498	18	19	0.9474
Asian	chr20 25067142	18	19	0.9474
Asian	chr20 25601604	18	19	0.9474
Asian	chr20 44404780	18	19	0.9474
Asian	chr20 58520304	16	19	0.8421
Asian	chr20 61444279	18	19	0.9474
Asian	chr18 63103470	16	19	0.8421
Hispanic	chr9 36764431	3	3	1.0

Table S5: List of ancestry specific novel sequence insertions. Last column shows the percentage of samples of a given population with this insertion.

Supplementary text A: "Supplementary repository"

Supplementary repository is available at https://github.com/1dayac/novel_insertions_supplementary. VCFs are stored in results folder. Subfolders correspond to different datasets we used - simulated, NA19240, HG002, HG002_stlfr, HG002_tellseq, NA12878, NA12878_stlfr, NA12878_tellseq, CHM1, CHM13, simulated. Simulated folder also contains VCFs for downsample experiment and different assemblers experiment. To evaluated VCFs we use compare_vcf.py script with dataset parameter.

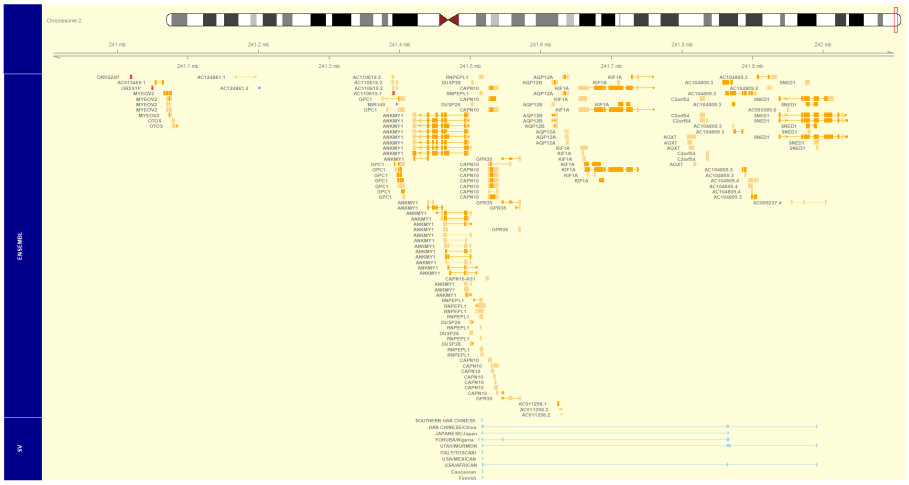

Figure S3: Overlapped region of insertion hotspot and highly disrupted gene RNPEPL1. Top track represents Chromosome 2 and corresponding location of overlap regions. ENSEMBL track represents genes located the region. Insertion track shows length and position of inserted contigs, linked within population.