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

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 S1: 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						
300-499	101	73 (**41**, **56%**)	324 (16, 5%)	156 (0, 0%)	-	-
500-999	138	99 (**60**, **61%**)	76 (12, 16%)	151 (1, 1%)	-	-
1000-1999	85	33 (**23**, **70%**)	5 (0, 0%)	89 (2, 2%)	-	-
≥2000	77	21 (**18**, **86%**)	2 (0, 0%)	21 (1, 5%)	-	-
Total(≥300)	401	226 (**142**, **63%**)	407 (28, 7%)	417 (4, 1%)	-	-
Max length (bp)	27836	**10436**	3168	4354	-	-
CHM13 dataset						
300-499	85	71 (**33**, **46%**)	-	156 (0, 0%)	-	-
500-999	126	100 (**53**, **53%**)	-	151 (4, 3%)	-	-
1000-1999	74	41 (**29**, **71%**)	-	89 (2, 2%)	-	-
≥2000	63	36 (**27**, **75%**)	-	21 (1, 5%)	-	-
Total(≥300)	348	248 (**142**, **57%**)	-	428 (7, 2%)	-	-
Max length (bp)	20444	**12146**	-	4346	-	-
NA19240 dataset						
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						
300-499	138	103 (**65**, **63%**)	-	-	9 (1, 11%)	92 (41, 44%)
500-999	198	120 (**73**, **61%**)	-	-	17 (6, 35%)	123 (58, 47%)
1000-1999	96	35 (**20**, **57%**)	-	-	2 (0, 0%)	41 (17, 41%)
≥2000	94	43 (27, **63%**)	-	-	3 (0, 0%)	74 (**42**, 57%)
Total(≥300)	526	301 (**185**, **61%**)	-	-	31 (7, 22%)	330 (158, 48%)
Max length (bp)	20442	**27836**	-	-	3721	4935
HG002 dataset						
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 S2: 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 all 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.

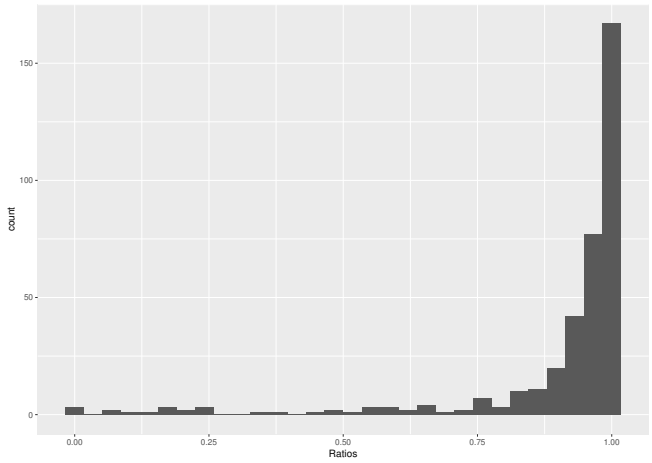

Figure S1: Histogram of intersection of barcodes associated with insertions and found on the reference to associated with insertion barcodes ratio.

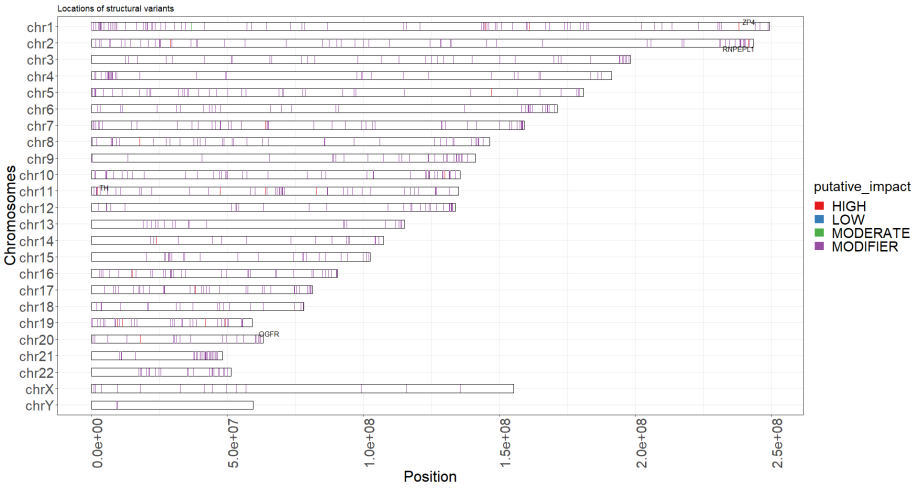

Figure S2: Chromatin location of insertion hotspots and highly disrupted genes. Vertical marks show insertion locations across chromosomes, colored by putative impact. Highly disrupted genes located at insertion hotspots (more than 30 samples have insertions) are annotated.

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 S3: List of ancestry specific novel sequence insertions. Last column shows the percentage of samples of a given population with this insertion.

0.1 Supplementary text A: Benchmark results for the CHM1 dataset

In summary, for the CHM1 dataset, Novel-X identified 226 insertions longer than 300 bp with a mean length of 1,011 bp and a total length of 228 kbp. The average sum of the left and right anchors of insertion length equals 3,183 bp. The maximum sum of two anchors length for a single insertion that we were able to achieve equals 15,149 bp. We also identified insertion sites that technically cannot be located on the reference genome with standard short-read data. Novel insertion detection tools for short-read data produce anchors that are indeed limited by the short insertion size (e.g. only 300 bp for an anchor). Mapping such short sequences to repetitive regions is often a challenging task that cannot be resolved unambiguously. In order to show that such insertion sites are ubiquitous, we extracted 300 bp upstream and downstream regions of insertion sites and searched these regions for repetitive sequences using RepeatMasker (Smit *et al.* (2004)). In total from the call set produced by Novel-X, 40 insertions were located in repetitive regions of the reference genomes. We confirmed these results using UCSC mappability tracks. For k100 and k50 Umap Multitrack we found 24 and 93 insertions in repetitive regions. PopIns and Pamir were able to call only 6 and 3 insertions identified with k100 Umap Multitrack, respectively.

To compare results for different novel insertion callers we compared short-read novel sequence insertion callers with the validation set obtained with the SMRT-SV (PacBio) algorithm (see Table S2 and Figure 4). While Pamir finds more insertions than Novel-X and PopIns, it tends to call shorter insertions. Novel-X finds more insertion of size greater than 500 bp compared to other callers. These findings are consistent with our theory because longer insertion sequences recruit more barcodes than short ones and their assembly will more likely produce long anchors during the assembly step. Another encouraging validation of our contribution is that about 91% of Novel-X calls overlap with SMRT-SV calls while the amount of agreement with Pamir and PopIns is only 44% and 18%, respectively. Pamir and PopIns are also more likely to produce

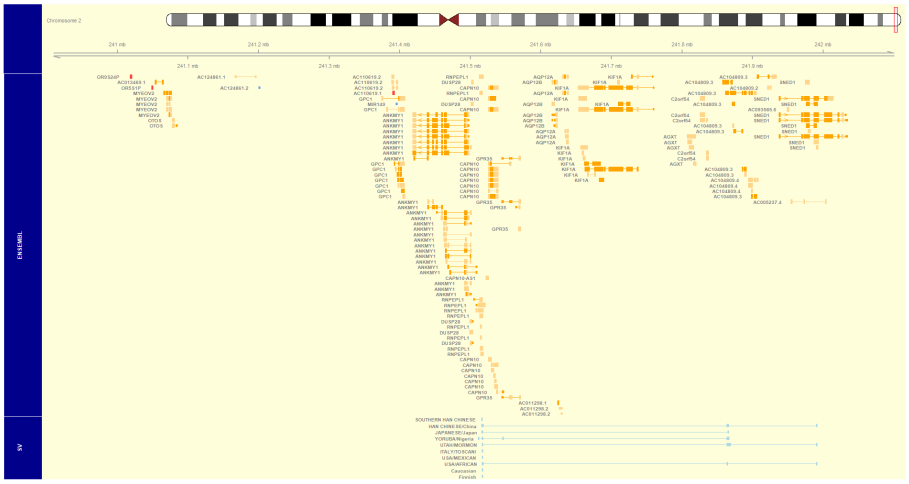

Figure S3: Overlapped region of insertion hotspot and highly disrupted gene RNPEPL1. Top track represents Chromosome 2 and corresponding location of overlap regions. ENSEMBL track represents genes located the region. Insertion track shows length and position of inserted contigs, linked within population.

false-positive calls. We also tried Manta (Chen *et al.* (2016)) as a popular tool for the general short-read data SV calling, but it was not able to find any novel insertions longer than 300 bp. Manta can theoretically only call small insertions (up to maximum two insert size), and stops being practical for insertions even before this size.

For 124 insertions reported in (Huddleston and Eichler (2016)) as non-template, we also checked if the insertion sequence content is similar between short or linked read callers and SMRT-SV. We globally aligned overlapping insertion sequences called by Novel-X and SMRT-SV, PopIns and SMRT-SV, Pamir, and SMRT-SV, maximizing the number of matches. We inspected identity scores and the visual representation of alignments. For all methods and for the majority of insertion pairs, the sequence percent identity tends to vary in the 98-100% range, except for a few cases when the insertion was truncated or extended. In the case of Novel-X, we have two insertions extended for 31 and 20 nucleotides respectively. Pamir has an insertion truncated by 226 nucleotides, both PopIns insertions are correct.

Gene overlap analysis: We checked if our novel insertion calls overlap with known genes and their coding sequences. We compared our reported breakpoint positions on the reference with gencode annotation v.24 (Harrow *et al.* (2012)). 111 out of 226 insertions fall inside known gene sequences, but only two of those falls into the exonic region of known protein-coding genes. In general, it appears that novel sequence insertions may contain novel exons or important non-coding regions but they do not necessarily disrupt known exonic sequences.

References

Chen, X. *et al.* (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**(8), 1220–1222.

Harrow, J. *et al.* (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, **22**(9), 1760–1774.

Huddleston, J. and Eichler, E. E. (2016). An incomplete understanding of human genetic variation. *Genetics*, **202**(4), 1251–1254.

Smit, A. F. A. *et al.* (1996-2004). RepeatMasker Open-3.0.