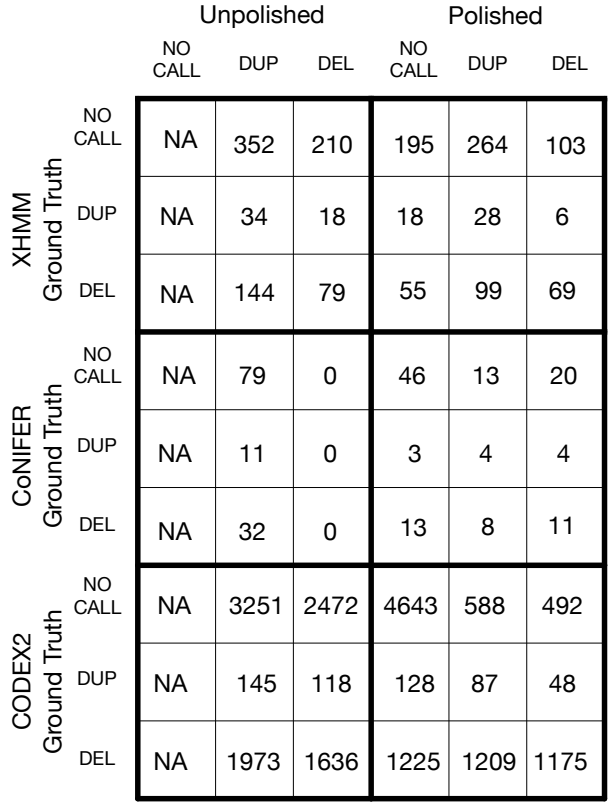
Supplementary Material
for
Polishing Copy Number Variant Calls on Exome Sequencing Data via Deep Learning

1 Supplementary Figures

		Unpolished			DECoNT-XHMM Polished			DECoNT-CoNIFER Polished			DECoNT-CODEX2 Polished		
		NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DEL	DEL
XHMM Ground Truth	NO CALL	NA	1708	1447	1974	790	391	2104	310	741	2647	505	3
	DUP	NA	1587	508	360	1589	146	1384	263	448	1666	422	7
	DEL	NA	198	1384	217	84	1281	715	208	659	1359	212	11
CoNIFER Ground Truth	NO CALL	NA	77	8	15	53	17	76	3	6	69	15	1
	DUP	NA	39	4	2	31	4	5	27	11	30	12	1
	DEL	NA	42	10	4	24	10	9	6	37	32	20	0
CODEX2 Ground Truth	NO CALL	NA	7955	6413	5337	3028	6003	1248	166	958	10974	1773	1621
	DUP	NA	2081	1786	368	1145	2354	339	130	561	896	1794	1177
	DEL	NA	7240	6774	1015	3723	9276	945	371	2282	3662	4216	6136

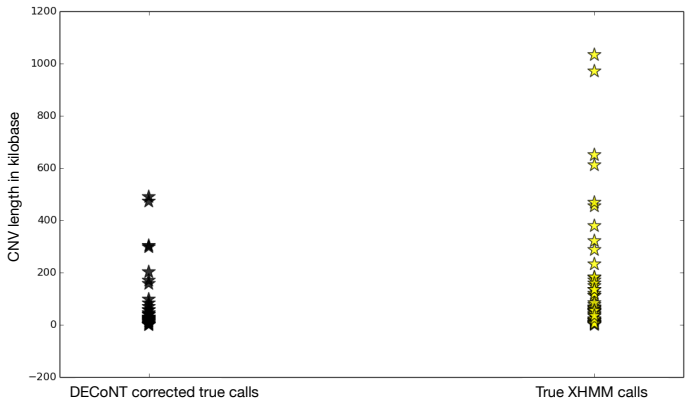
Supplementary Figure 1. The confusion matrices of the WES-based CNV callers before and after polishing with DECoNT on 1000 Genomes Data test samples. Confusion matrices given with blue borders represent unpolished predictions of corresponding WES-based CNV tools. Since DECoNT only operates on the calls made by a CNV caller, the first column for each unpolished confusion matrix is set as NA (i.e. Not Applicable). The red-bordered confusion matrices are the polished versions of a CNV caller with a DECoNT model trained on the calls made by the same caller (to produce Fig 2). Other confusion matrices are polised version of the CNV caller corrected by a DECoNT model trained on the calls made by a different caller (to produce Fig. 3). Notice the decrease in the number of false positives for both deletion and duplication calls in all platforms. For each polished tool, we used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for testing, This roughly corresponds to a test set size of 80 samples.



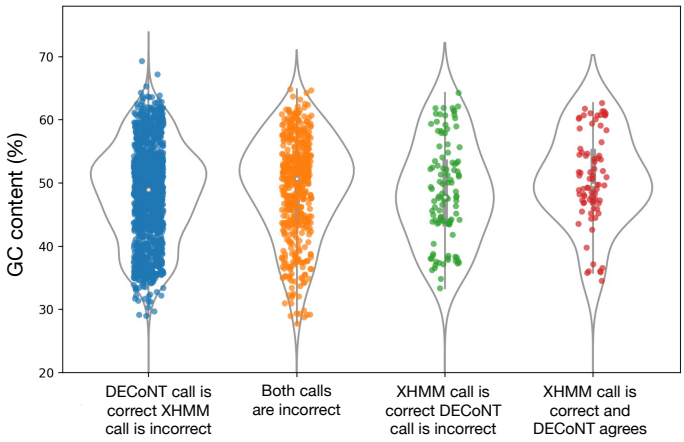
Supplementary Figure 2. Confusion matrices of the WES-based CNV callers before and after polishing with DECoNT on highly validated CNV callset published in Chaisson et. al. [1]. Similar to Fig. 1 tool provides great false discovery correction with slight true positive deterioration for both deletion and duplication calls, yielding much better performance metric results. 90% of the calls made on the 1000 Genomes dataset are used for training the models and 9 samples from Chaisson et al. are used for validation for these results.

		NovaSeq6000						HiSeq4000						BGI500						MGI2000					
		Unpolished			Polished			Unpolished			Polished			Unpolished			Polished			Unpolished			Polished		
		NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL	NO CALL	DUP	DEL
XHMM Ground Truth	NO CALL	NA	0	65	24	4	37	NA	0	79	27	6	46	NA	21	10	10	12	9	NA	21	10	10	12	9
	DUP	NA	2	17	6	2	11	NA	1	18	10	1	8	NA	1	6	1	1	5	NA	1	6	1	1	5
	DEL	NA	1	7	2	0	6	NA	1	10	0	1	10	NA	0	3	0	0	3	NA	0	3	0	0	3
CoNIFER Ground Truth	NO CALL	NA	0	0	0	0	0	NA	0	24	9	1	14	NA	0	5299	3088	846	1365	NA	0	5299	3088	846	1365
	DUP	NA	0	0	0	0	0	NA	0	14	6	1	7	NA	0	67	23	10	34	NA	0	67	23	10	34
	DEL	NA	0	0	0	0	0	NA	0	9	4	0	5	NA	0	299	115	58	126	NA	0	299	115	58	126
CODEX2 Ground Truth	NO CALL	NA	613	442	840	105	110	NA	503	377	607	164	109	NA	320	250	478	33	59	NA	320	250	478	33	59
	DUP	NA	32	33	14	31	20	NA	19	19	9	19	10	NA	21	13	12	13	9	NA	21	13	12	13	9
	DEL	NA	98	118	43	87	86	NA	107	92	54	69	76	NA	63	72	32	37	66	NA	63	72	32	37	66

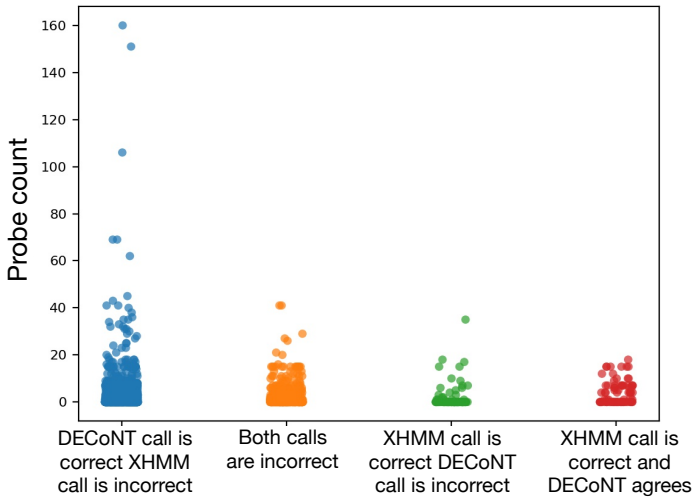
Supplementary Figure 3. Confusion matrices of the WES-based CNV callers before and after polishing with DECoNT on NA12878 data obtained from different sequencing platforms: (i) NovaSeq6000; (ii) HiSeq4000; (iii) BGI500; (iv) MGI2000. Since DECoNT only operates on the calls made by a CNV caller, the first column for each unpolished confusion matrix is set as NA (i.e. Not Applicable). Since CoNIFER does not report any calls on NovaSeq6000 platform, DECoNT has no input to polish and thus the comparison is not applicable. Similar to Figures 1 and 2, we observe that DECoNT substantially decreases the number of false discoveries with slight true positive deterioration for both deletion and duplication calls. 90% of the calls made on the 1000 Genomes dataset are used for training the models and only NA12878 sample is used for validation for these results.



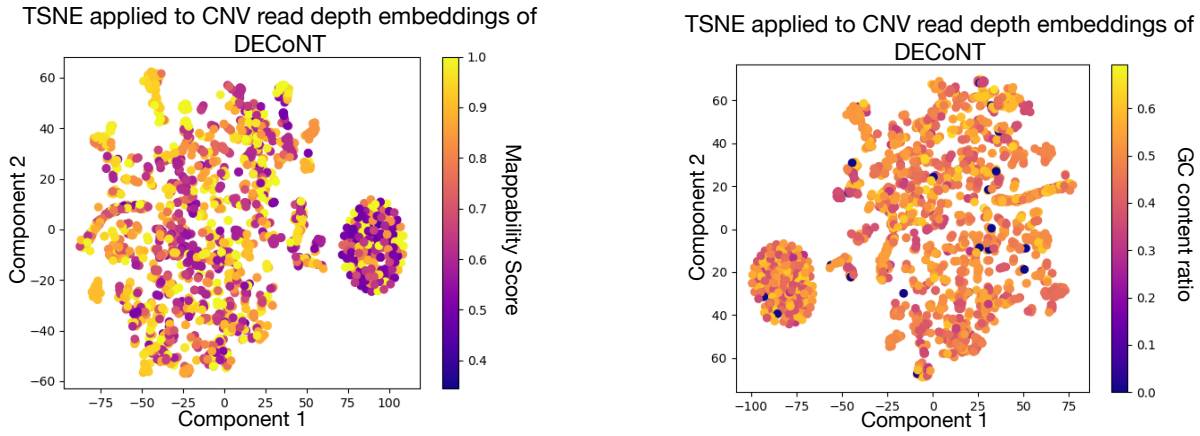
Supplementary Figure 4. This figure shows the length distribution of true raw XHMM calls and true DECoNT-corrected XHMM calls obtained on the 1000 Genomes WES data set test samples. The ground truth is the CNV calls made by CNVnator on the corresponding WGS samples. We see that for smaller size CNVs XHMM requires more correction by DECoNT. However, again, vast majority of the CNVs cannot be distinguished by the CNV length to decide whether it needs a DECoNT correction. For each polished tool, we used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for testing, This roughly corresponds to a test set size of 80 samples.



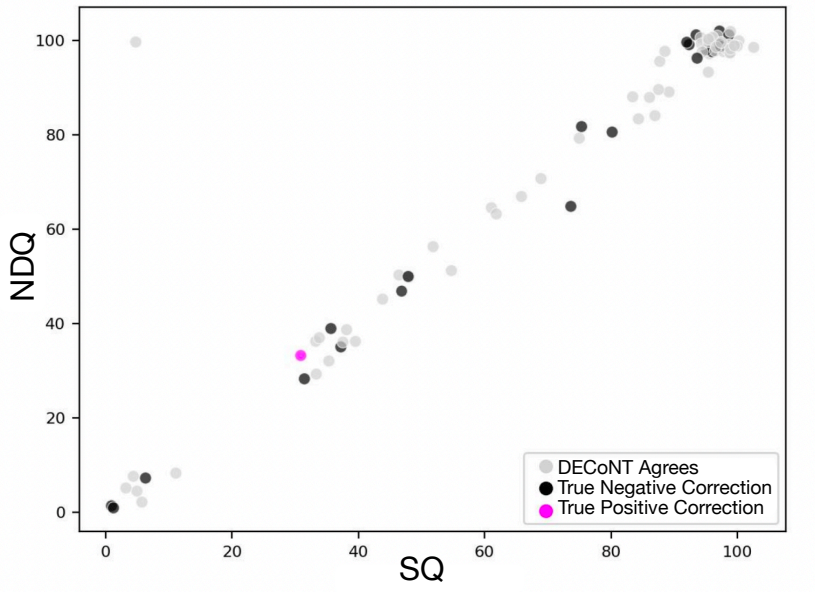
Supplementary Figure 5. This figure shows the probe content distribution of the XHMM calls obtained on the 1000 Genomes WES data set test samples. The ground truth is the CNV calls made by CNVnator on the corresponding WGS samples. The ground truth is the CNV calls made by CNVnator on the corresponding WGS samples. Blue dots indicate that the original XHMM call is changed by DECoNT and the changed prediction matches the ground truth (correct). Green dots indicate that the original XHMM call is changed by DECoNT and the changed prediction does not match the ground truth (incorrect). Red dots indicate original XHMM call is correct and DECoNT agreed. Finally, yellow dots indicate that both DECoNT's and XHMM's calls are incorrect. For each category, a random jitter is added to the x-axis for better visualization. For each polished tool, we used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for testing, This roughly corresponds to a test set size of 80 samples.



Supplementary Figure 6. This figure shows the probe content distribution of the XHMM calls obtained on the 1000 Genomes WES data set test samples. The ground truth is the CNV calls made by CNVnator on the corresponding WGS samples. Blue dots indicate that the original XHMM call is changed by DECoNT and the changed prediction matches the ground truth (correct). Green dots indicate that the original XHMM call is changed by DECoNT and the changed prediction does not match the ground truth (incorrect). Red dots indicate original XHMM call is correct and DECoNT agreed. Finally, yellow dots indicate that both DECoNT's and XHMM's calls are incorrect. For each category, a random jitter is added to the x-axis for better visualization. For each polished tool, we used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for testing, This roughly corresponds to a test set size of 80 samples.



Supplementary Figure 7. These figures show the distribution of the 1000 Genomes WES test set CNV calls, obtained using XHMM, on a 2D t-SNE space. To check whether hidden Bi-LSTM encodings of DECoNT correlates with sequence features, we annotated each point in the t-SNE space both with mappability score (left figure) and GC ratio (right figure) using a colour-map. We do not observe any obvious clustering pattern in neither of the figures. This suggests that predictions of DECoNT does not directly depend on either of these features. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting, This roughly corresponds to a test set size of 80 samples.



Supplementary Figure 8. This figure shows the distribution of the *Some deletion* (SQ) and *Not diploid* (NDQ) scores reported by the XHMM software for the WES CNVs of the 1000 Genomes WES test sample HG00733. We used 90% of the calls made on 802 of the 1000 Genomes data set samples for training. We mark the corrected calls by DECoNT. True Negative correction means a DEL or DUP call is converted to NO-CALL and the ground truth is NO-CALL. True positive correction is a DEL or DUP call is converted to a DUP or DEL call, respectovely and it matches the ground truth after correction. We polish the calls using DECoNT and label the polished calls by comparison with the WGS CNV calls made by CNVNator for the same sample. Figure indicates that quality filtering of XHMM is not sufficient as corrections are made regardless of the quality values.

2 Supplementary Tables

Supplementary Table 1. This table summarizes the polishing performance of DECoNT on the X chromosome, PAR1 and PAR2 regions of the males in the test split obtained from 1000 Genomes WES samples. The base caller in this analysis is XHMM. The results are obtained on the test samples from the 1000 Genomes dataset and the ground truth is obtained from the CNVnator calls on WGS of the same samples. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting, This roughly corresponds to a test set size of 80 samples.

	Deletion Precision (Unpolished-Polished)	Duplication Precision (Unpolished-Polished)	Overall Precision (Unpolished-Polished)
X Chromosome	0.0350 - 0.1153	0.3018 - 0.4753	0.1702 - 0.4072
PAR1	0.1667 - 0.2500	0.7083 - 0.6112	0.5278 - 0.5455
PAR2	0.3334 - 0.6667	0.25 - 0.50	0.2871 - 0.4

Supplementary Table 2. Table shows the classification performance improvement of CNVkit after polishing. We discretized the predictions of CNVkit as also done for Control-FREEC and explained in Supplementary Note 1. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting, This roughly corresponds to a test set size of 80 samples.

CNVKit	Deletion Precision	Duplication Precision	Overall Precision
Before DECoNT	0.0940	0.1525	0.1234
After DECoNT	0.1234	0.5497	0.2527

Supplementary Table 3. The 8 WES CNV calls that DECoNT and CNLearn does not agree are presented. The ground truth CNV calls are obtained through CNVNator WGS CNV calls. Note that, CNLearn samples are polished with a DECoNT model trained with XHMM data. Training a DECoNT model with consensus calls made by CNLearn would increase performance. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting, This roughly corresponds to a test set size of 80 samples.

Sample	Chromosome	CNV Start	CNV End	CNLearn Prediction	DECoNT Prediction	Ground Truth (CNVNator WGS Calls)
NA19144	11	6128771	6170380	DEL	NO-CALL	NO-CALL
NA19144	chr14	73541473	73573608	DUP	DEL	NO-CALL
NA11832	chr6	32519300	32666612	DUP	DEL	DEL
NA11832	chr15	34386562	34528116	DUP	DEL	DUP
NA18968	chr6	29889285	29945317	DUP	DEL	DEL
NA18968	chr6	32519300	32579157	DUP	DEL	DEL
NA18968	chr6	32584060	32665112	DUP	NO-CALL	DEL
NA12249	chr16	55810440	55826342	DUP	NO-CALL	DUP

Supplementary Table 4. CNV calls of XHMM on 16 bladder cancer patient samples - tumor and normal tissue - obtained from Guo et al. Only calls in the qPCR validated regions are shown. XHMM reports calls only in the following validated region: chr9:20,305,364 - 24,115,910. The last column shows the DECoNT-polished versions of each call.

SampleID	IndividualID_SampleType	Chromosome	CNV Start	CNV End	XHMM Call	DECoNT-polished Call
SRR645432	B112_Cancer	chr9	19,116,209	21,862,054	DEL	DEL
SRR645432	B112_Cancer	chr9	23,692,541	23,765,104	DEL	DEL
SRR645579	B63_Cancer	chr9	21,077,288	21,862,054	DEL	DEL
SRR645629	B80-0_Cancer	chr9	21,409,162	21,862,054	DEL	DEL
SRR645629	B80-0_Cancer	chr9	21,409,162	21,854,930	DEL	DEL
SRR645631	B80-0_Normal	chr9	21,304,672	21,854,930	DUP	NO-CALL
SRR645631	B80-0_Normal	chr9	21,029,287	21,854,930	DUP	NO-CALL

3 Supplementary Notes

Supplementary Note 1 For the integer CNV calls of Control-FREEC, we have categorized the calls such that Copy Number > 2 is Duplication, Copy Number < 2 is Deletion and Copy Number $= 2$ is No-Call. Then, we evaluated the polishing performance with the performance metrics defined in Section 4.3. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting, This roughly corresponds to a test set size of 80 samples. Performance Metrics with respect to the 1000 Genomes WGS CNV calls of CNVNator:

- Duplication Precision was increased from 0.1063 to 0.3932
- Deletion Precision was increased from 0.2578 to 0.5936
- Overall Precision was increased from 0.1277 to 0.4432

Supplementary Note 2 In order to show the need for a complex machine learning model like DECoNT for this polishing task, we also experimented with traditional machine learning methods such as Support Vector Machines (SVM), Logistic Regression and Polynomial Regression (degree = 2) as polishers. We used the scikit-learn implementations and the default parameters. These algorithms are run with the same settings we used for DECoNT. We worked on the 1000 Genomes dataset samples and same the train-test split. We input the same features into these models as we input to DECoNT: read depth and the call of the baseline caller. We used the corresponding WGS calls by CNVnator as the ground truth as we do for DECoNT. We used XHMM and FREEC as the baseline callers for this experiment. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting. This roughly corresponds to a test set size of 80 samples.

Below, we show that these models actually cannot polish the calls and deteriorate the results. See the notes below:

- XHMM predictions result in 0.4541 and 0.4144 precision for duplication and deletion calls, respectively.
- When correcting XHMM calls, SVM based model predictions result in 0.3562 and 0.3321 in precision for duplication and deletion calls, respectively.
- When correcting XHMM calls, Logistic Regression based model predictions result in 0.3334 and 0.2174 in precision for duplication and deletion calls, respectively.

- Control-FREEC predictions result in a MSE of 37.17 with standard deviation of 75.89
- When correcting Control-FREEC calls, Polynomial Regression polished model predictions result in a MSE of 58.10 with standard deviation of 18.91

Supplementary Note 3 In order to test our assumption that running the base callers in their suggested parameter settings is sound we performed an experiment with XHMM which is the best performing method in our benchmarks. We ran it in also conservative and liberal settings in addition to the suggested setting. The parameter values that correspond to these settings are given in the table below. We used XHMM and FREEC as the baseline callers for this experiment. We used 90% of the calls made on 802 1000 Genomes data set samples for training and the remaining 10% of the calls for plotting. This roughly corresponds to a test set size of 80 samples.

XHMM	minTarget Size	maxTarget Size	minMean TargetRD	maxMean TargetRD	minMean SampleRD	maxMean SampleRD	maxSd SampleRD
Conservative	10	1000	10	5000	25	2000	1500
Suggested	5	10000	5	5000	5	2000	1500
Liberal	0	100000	0	50000	0	20000	15000

The precision values before and after polishing with DECoNT are given in the table below.

XHMM	Dup Precision (Unpolished-Polished)	Del Precision (Unpolished-Polished)	Overall Precision (Unpolished-Polished)
Conservative	0.4758 - 0.6548	0.4572 - 0.7120	0.4665 - 0.6834
Suggested	0.4541 - 0.6451	0.4144 - 0.7046	0.4348 - 0.6704
Liberal	0.3785 - 0.5543	0.3028 - 0.5921	0.3406 - 0.5732

We observe that the liberal setting results in a worse polished precision $\sim 10\%$. Conservative and suggested setting results are similar. The improvement in precision values are stable across all runs. Thus, we suggest using the default parameter settings for the base callers unless they return insufficient number of calls which prohibit DECoNT training. Then, the parameter choices can be relaxed.

Supplementary Note 4 The IDs of the 4 samples taken from 1000 Genomes data set that were used in CNLearn analysis are as follows: NA11832, NA12249, NA18968, NA19144.

References

1. Chaisson, M.J., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al.: Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature communications **10**(1), 1–16 (2019)