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

Furkan Özden[1], Can Alkan[1,*], and A. Ercüment Çiçek[1,2,*]

Department of Computer Engineering, Bilkent University, Ankara, Turkey
Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA
*{calkan, cicek}@cs.bilkent.edu.tr

Abstract. Accurate and efficient detection of copy number variants (CNVs) is of critical importance due to their significant association with complex genetic diseases. Although algorithms working on whole genome sequencing (WGS) data provide stable results with mostly-valid statistical assumptions, copy number detection on whole exome sequencing (WES) data has mostly been a losing game with extremely high false discovery rates. This is unfortunate as WES data is cost efficient, compact and is relatively ubiquitous. The bottleneck is primarily due to non-contiguous nature of the targeted capture: biases in targeted genomic hybridization, GC content, targeting probes, and sample batching during sequencing. Here, we present a novel deep learning model, *DECoNT*, which uses the matched WES and WGS data and learns to correct the copy number variations reported by any over-the-shelf WES-based germline CNV caller. We train DECoNT on the 1000 Genomes Project data, and we show that (i) we can efficiently triple the duplication call precision and double the deletion call precisions of the state-of-the-art algorithms. We also show that model consistently improves the performance in a (i) sequencing technology, (ii) exome capture kit and (iii) CNV caller independent manner. Using DECoNT as a universal exome CNV call polisher has the potential to improve the reliability of germline CNV detection on WES data sets and surge its application. The code and the models are available at https://github.com/ciceklab/DECoNT.

Keywords: Copy Number Variation · Whole Exome Sequencing · Deep Learning.

1 Introduction

Gene copy number polymorphism in a population due to deletions and duplications of genomic segments substantially derive genetic diversity [29, 14], affecting roughly 7% of the genome [32]. This class of structural variations (SVs), called copy number variations (CNVs), have also been associated with several genetic diseases and disorders such as neurodevelopmental/neurodegenerative disorders [26, 10, 23, 6, 39] and various cancers such as breast, ovary, and pancreas cancers [11, 22, 27, 25]. Karyotyping and microarray analyses have been the standard clinical testing for disease-causing CNVs for many years [36], but High Throughput Sequencing (HTS) has all but replaced these techniques with the ability to theoretically capture all forms of genomic variation. Numerous CNV detection algorithms have enjoyed success by analyzing whole genome sequencing (WGS) data using different sequence signatures such as read depth, discordant paired-end read mappings, and split reads [12]. WGS is a convenient resource for CNV callers as it provides near-Poisson depth of coverage [3]. On the other hand, accurate CNV detection on whole exome sequencing (WES) data has mostly been lacking. The algorithms which call CNVs on the WES data have notoriously high false discovery rates (FDR) reaching up to ∼60% which renders them impractical for clinical use [38, 33]. This is mainly due to several problems associated with the WES technology such as non-uniform read-depth distribution among exons caused by biases in (i) sample batches, (ii) GC content, and (iii) targeting probes [16, 21, 18]. This is unfortunate as WES data size is ten times smaller (i.e., ∼10GB vs ∼100GB) and it costs three times less compared to WGS which makes it highly abundant and a common choice for analyzing complex genetic disorders [31, 7, 28, 30]. For instance, the Genome Aggregation Database (gnomAD) contains around 125K WES samples as opposed to 70K WGS samples [17]. Thus, currently, such a rich resource of large scale WES data cannot be fully utilized to investigate the contribution of copy number variation to disease etiology.

Here, we present the first of its kind, exome CNV call *polisher* named *DECoNT* (Deep Exome Copy Number Tuner) to improve the performance of any off-the-shelf WES-based germline CNV detection algorithm. DECoNT is a deep learner that utilizes matched WES and WGS samples present in the 1000 Genomes Project [35] data set to learn the association between (i) calls made by any CNV caller working on the WES data and (ii) ground truth calls made on the WGS data for the same sample. Based on a bidirectional long short-term memory (Bi-LSTM) based architecture, it uses only the WES read depth along with the calls from the third party caller and learns to correct noisy predictions (Figure 1). We show that DECoNT can improve the duplication and deletion call precision of the state-of-the-art algorithms by up to 3-fold and 2-fold, respectively. The performance gain is consistent among CNV callers that output integer copy number predictions and categorical predictions (i.e., deletion, duplication, or no call). As the training phase is offline, polishing procedure is memory and time efficient and takes only a few seconds on average per sample. Furthermore, we show that the models learned are universal in the sense that they are (i) sequencing platform, (ii) exome sequencing kit, and (iii) CNV caller independent. For instance, using models learned on 1000 Genomes Project data set that uses Illumina as the sequencing platform and various capture kits such as Agilent and NimbleGen, DECoNT can correct calls made by any of the state-of-the-art CNV caller, for samples obtained from (i) other capture kits like Agilent SureSelect and Illumina Nextera Exome Enrichment Kit or (ii) other sequencing platforms like Illumina HiSeq 4000, Illumina NovaSeq 6000, and MGI; that are "unseen" during training. Thus, DECoNT is highly flexible and scalable and makes exome based CNV detection practical by boosting the performance of virtually any WES-based CNV caller algorithm. The tool and the models are available at https://github.com/ciceklab/DECoNT.

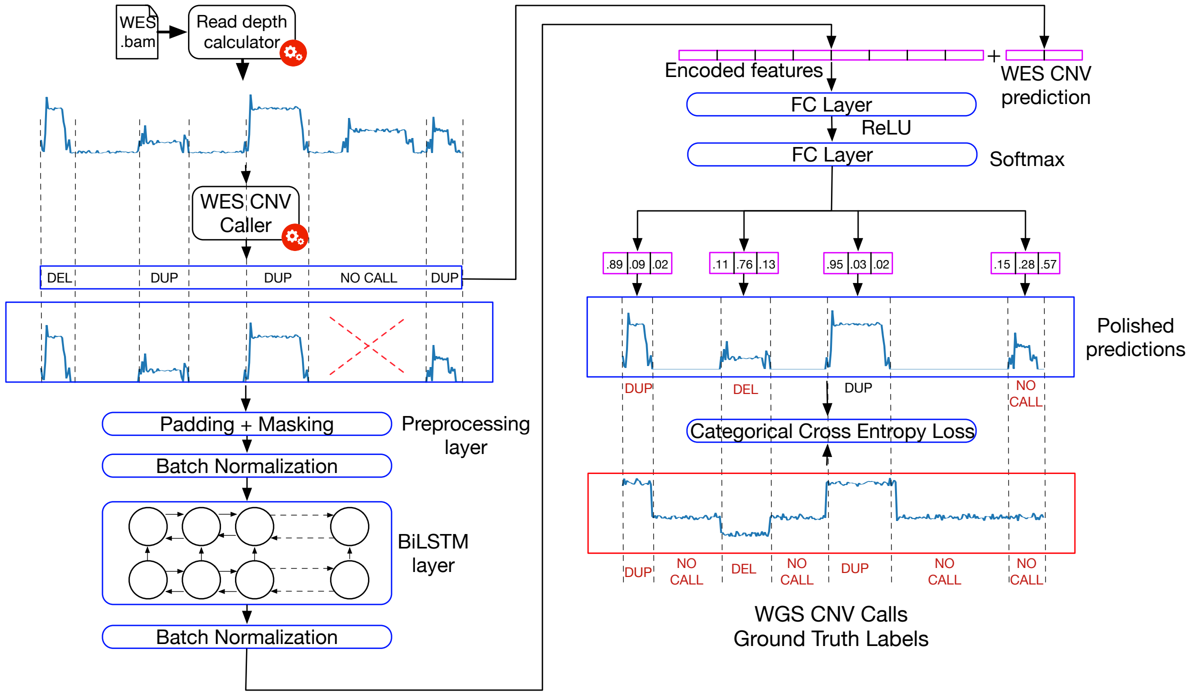

Fig. 1. Learning workflow of DECoNT. First, WES .bam data file from 1000 Genomes data set is used to calculate exome-wide read depth which is input to a third party WES-based CNV caller. The caller generates the calls for various regions which could be (i) a binary prediction like duplication, deletion (e.g., XHMM [8]) as shown in the figure, or (ii) an integer value that indicates the exact copy number (i.e., Control-FREEC [4]). The read depth of the regions for which a call has been made is input to a Bi-LSTM model. Encoded features are passed from a series of fully connected layers along with the original prediction of the caller algorithm. Using the ground truth calls from the WGS data of the same sample the method learns to predict (correct) the calls using cross entropy loss for the binary outputs (as shown in the figure) and using mean squared loss for integral calls.

2 Results

2.1 Bi-LSTM based Neural Network Learns to Correct False Positive Germline WES CNV Calls

A Bidirectional long short-term memory network [13] is a type of recurrent neural network which learns a representation (i.e., embedding) of a sequence by processing it character by character in the forward and the backward directions. While doing so, it remembers a summary of the sequence observed so far to capture the context for each character. RNNs and LSTM-based architectures have been widely and successfully used in natural language processing domain to process sequence data [37].

DECoNT uses a single hidden layered Bi-LSTM architecture with 128 hidden neurons in each direction to process the read depth signal (Methods). First, WES-based germline CNV caller result is obtained along with the read depth signal in those event regions. Bi-LSTM subnetwork learns a transformed representation for the read depth sequence (Figure 1). This embedding and the corresponding CNV call are input to a fully connected (FC) layer feed forward neural network. The FC layers predict the polished result for call. DECoNT makes use of the calls made on the WGS data of the same sample as the ground truth for the learning procedure. We use matched WGS data to obtain the ground truth calls for the CNV events called on the WES samples of the same individuals in the 1000 Genomes data set [35].

We polish state-of-the-art WES-based germline CNV callers. There are two types of such algorithms. The first type makes discrete predictions for CNVs (i.e., deletion and duplication). We consider three methods in this category: (i) XHMM [8], (ii) CoNIFER [20], and (iii) CODEX2 [15]. The second type predicts the exact copy number as an integer value. The sole example we consider of this type is Control-FREEC [4]. DECoNT architecture is flexible and can be easily modified to polish both types of algorithms (Methods). We train a DECoNT model for every above-mentioned tool using 3 NVIDIA GeForce RTX 2080 Ti and 1 NVIDIA TITAN RTX GPUs in parallel with training times ranging from ~ 1 to ~ 4 days (Methods).

We find that DECoNT is able to substantially improve the performance of all algorithms in almost all comparisons. For algorithms that make discrete predictions, we observe improvements in both duplication and deletion call precisions (Figure 2a). The largest gain is in duplication call precision for CoNIFER which is improved by 3-fold (i.e., 24.68% to 75%). The largest gain in deletion call precision is again obtained for CoNIFER which is improved by 1.5-fold (i.e., 45.45% to 68.51%). Also, overall precision is improved by 2.6-fold (i.e., 27.22% to 71.11%) for CoNIFER. This improvement is especially striking as CoNIFER is relatively conservative compared to other algorithms and seldom make calls despite relaxation of its parameters. For XHMM, we observe $1.4, 1.7$, and 1.5 fold increases which correspond to $20\%, 29\%$, and 24% improvements in duplication, deletion and overall precision, respectively. We see a similar trend for CODEX2. Before polishing with DECoNT, CODEX2 achieves 12% duplication precision, 45% deletion precision, and 27% overall precision. DECoNT provides 1.9 fold increase in duplication call precision, 1.5 fold increase in deletion call precision, and 1.75 fold increase in overall precision, respectively. These correspond to 11%, 23%, and 20% improvements in each respective metric. Confusion matrices before and after polishing are shown in Supplementary Figure 1 for all tools. We would like to note that these improvements are obtained in seconds per sample at the test time. Increased precision is an important result for the life scientists who work with these calls as the reliability of the calls are substantially increased as the number of false positives are substantially decreased.

As for the Control-FREEC, which outputs exact copy number values, we evaluate its performance (i.e., absolute error) on $20,482$ CNV events called (Figure 2b, Methods). DECoNT improves the absolute error in 74.58% of the test samples for an average AE improvement of 47.39 and deteriorates the performance in 25.35% of the test samples for an average AE deterioration of only 1.2. While unpolished Control-FREEC predictions have a Spearman correlation coefficient of 0.227 with matched ground truth copy numbers, DECoNT-polished predictions have a Spearman correlation coefficient of 0.568 (Figure 2c). DECoNT-polished predictisons highly resemble the distribution of the ground truth calls. Overall, DECoNT substantially improves the calls made by all four algorithms.

2.2 Polishing Performance on a Validated CNV call set

In order to further test the polishing performance of DECoNT, we also use a highly validated CNV call set published by Chaisson et. al., [5]. This data set contains the WGS CNV calls of 9 individuals from the 1000 Genomes data set for which a consensus call set is obtained using 15 different WGS CNV callers with

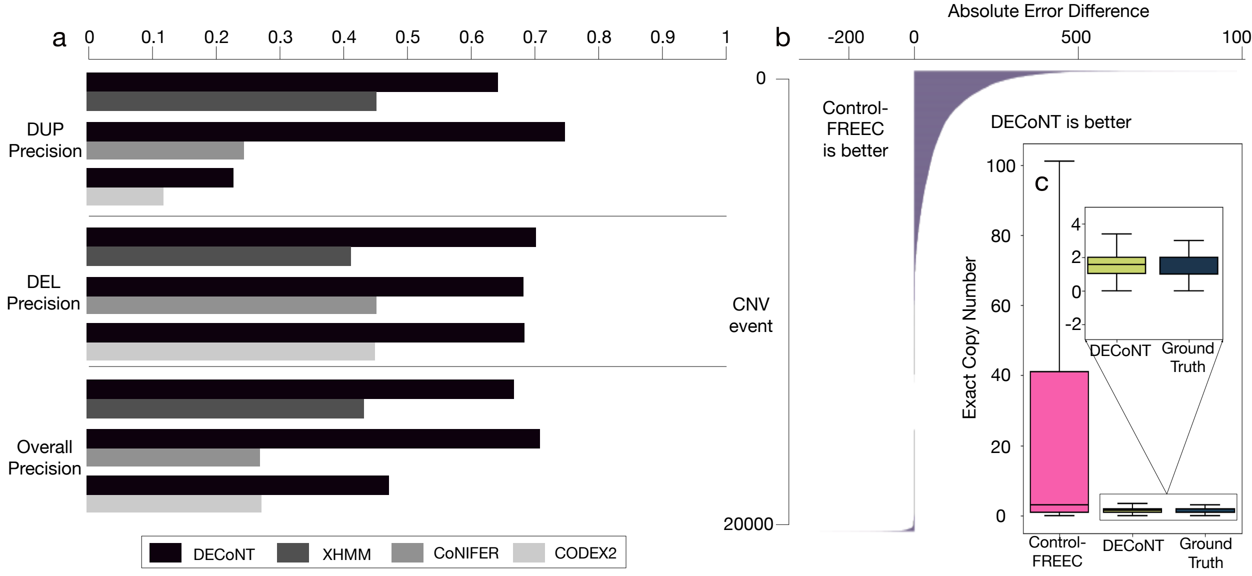

Fig. 2. The performance comparison of the WES-based CNV caller's before and after polishing with DECoNT. a) For the tools which predict existence of a CNV event (XHMM, CoNIFER and CODEX2) are evaluated with respect to duplication call precision, deletion call precision and overall precision. DECoNT improves the performance for all tools in all settings and results in drastic improvements. Different tones of gray represent different tools and the attached black bars represent the DECoNT-polished version of those tools. b) The sole algorithm which predicts an integer value for CNV events is Control-FREEC. We compare Control-FREEC and the DECoNT-polished with respect to Absolute Error (AE) difference on each of the 20, 482 samples (i.e., events). Bars to the right indicate the magnitude of the improvement due to polishing of DECoNT. For more than half of the samples, DECoNT results in an improvement. For only 29.21% of the samples the performance deteriorates. c) The distribution of the unpolished Control-FREEC predictions in the test samples (pink) is quite different than the ground truth copy number variation distribution (green; Spearman correlation: 0.227). On the other hand, DECoNT polished versions of the same events (dark blue) highly resemble the distribution of the ground truth calls (Spearman correlation: 0.586).

comparisons against high quality PB-SVs that have single base breakpoint resolution (Methods). We use 8 individuals who have matched WES data.

Using the same models explained in Section 2.1, we correct the CNV calls made on WES data of the 8 samples made by XHMM, CoNIFER and CODEX2. Note that none of the DECoNT models have seen the data of these individuals during training. We validate the performance using this call set. Table 1 summarizes the performances before and after polishing with DECoNT, with respect to WGS validated calls.

Again, DECoNT improves the performance of all three algorithms in all comparisons. The most substantial improvements are observed for CoNIFER. 7%, 31.4% and 16% improvements are observed for duplication, deletion, and overall precision, respectively. It is noteworthy that while CoNIFER does not report any deletion events, DECoNT was able to correct incorrect duplication calls into correct deletion calls and increase the precision to 31.4% in this category. See Supplementary Figure 2 for the confusion matrices obtained before and after polishing by DECoNT. For XHMM and CODEX2, we see consistent improvements reaching up to nearly 2-fold for CODEX2.

2.3 Polishing Performance Generalizes to Unseen Sequencing Platforms

We obtained the training data from the 1000 Genomes data set which is produced using Illumina Genome Analyzer II and Illumina HiSeq 2000. While data from these platforms is abundant and sufficient training data set size can be met, for users using other sequencing platforms, it might not be possible to train DECoNT due to lack of matched WES and WGS samples. We therefore evaluated whether models trained on the available 1000 Genomes data can be used to polish CNV calls made on WES samples obtained using other sequencing platforms or capture kits that have not been seen by DECoNT (Methods).

Table 1. The performances of the WES-based CNV caller algorithms before and after polishing are shown (DEL, DUP and overall precision). Validated WGS CNV call set of Chaisson *et al.* [5] is used as the ground truth CNV call set. We first use matched WES reads to call WES CNVs using CoNIFER, CODEX2, and XHMM. Then, we use DECoNT to polish obtained CNV calls. Table shows the DEL, DUP and overall precision of the methods

Tool	DUP Precision		DEL Precision		Overall Precision	
	default	polished	default	polished	default	polished
XHMM	0.064	**0.071**	0.257	**0.387**	0.135	**0.170**
CoNIFER	0.090	**0.160**	0.0*	**0.314**	0.090	**0.250**
CODEX2	0.027	**0.046**	0.387	**0.685**	0.185	**0.350**

*CoNIFER does not report any deletion events on this set of WES samples.

We obtain the WES data for the sample NA12878, sequenced using four different platforms: (i) Illumina NovaSeq 6000; (ii) Illumina HiSeq 4000; (iii) BGISEQ-500; and (iv) MGISEQ-2000. We use these four samples only for testing. All considered WES-based CNV callers are used to call CNV events on these four WES samples.

Table 2. The performance of discrete germline WES-based CNV callers on NA12878 data before and after being polished by DECoNT.

Platform	Tool	DUP Precision		DEL Precision		Overall Precision	
		default	polished	default	polished	default	polished
NovaSeq 6000	XHMM	**0.660**	0.330	0.078	**0.111**	0.097	**0.133**
	CoNIFER	NA*	NA*	NA*	NA*	NA*	NA*
	CODEX2	0.043	**0.139**	0.198	**0.398**	0.112	**0.266**
HiSeq 4000	XHMM	**0.500**	0.125	0.093	**0.156**	0.100	**0.152**
	CoNIFER	0.0**	**0.500**	0.191	**0.192**	0.191	**0.214**
	CODEX2	0.032	**0.075**	0.188	**0.389**	0.099	**0.212**
BGISEQ-500	XHMM	0.045	**0.076**	0.157	**0.176**	0.088	**0.200**
	CoNIFER	0.0**	**0.010**	0.052	**0.082**	0.052	**0.055**
	CODEX2	0.051	**0.156**	0.214	**0.492**	0.125	**0.364**
MGISEQ-2000	XHMM	0.045	**0.076**	0.157	**0.176**	0.088	**0.200**
	CoNIFER	0.0**	**0.010**	0.052	**0.082**	0.052	**0.055**
	CODEX2	0.051	**0.156**	0.214	**0.492**	0.125	**0.364**

We evaluated caller performance on NA12878 data obtained using four different sequencing platforms: (i) NovaSeq 6000; (ii) HiSeq 4000; (iii) BGISEQ-500; (iv) MGISEQ-2000. Note that DECoNT models did not train on the data sequenced with any of these sequencing platforms or on NA12878 sequencing data of any form. DUP precision, DEL precision and Overall precision results are shown. In all comparisons, DECoNT provides substantial improvements showing the generalizability of our models trained on 1000 Genomes data set. *CoNIFER does not report any CNV calls on NA12878 WES data sequenced with NovaSeq 6000. For that reason DECoNT has no input to correct and thus that comparison is not applicable. **CoNIFER does not report ant duplication events in the unpolished case.

Even though DECoNT has not seen the read depth information or the CNV events on these sequencing platforms, it still can generalize from the training on the 1000 Genomes data and still can substantially improve the performances of XHMM, CoNIFER, and CODEX2 (Table 2 and Supplementary Figure 3). We observe improvements in 34 out of 36 tests.

The most substantial improvement is observed for CODEX2 $\sim$ 18% improvement on average which corresponds to an average 2.6-fold increase in performance. This even exceeds testing performance on the same platform as training (i.e., $\sim$ 2-fold improvement). For XHMM, the performance is improved 10 out of 12 tests, reaching up to doubling the performance in overall precision performance for BGISEQ and MGISEQ platforms. For NovaSeq 6000 and HiSeq 4000, the performance deteriorates in duplication precision. However, XHMM makes a few duplication calls: 3 and 2, respectively. While DECoNT keeps the true positives, it adds a few false positives and this results in the performance decrease in these settings. CoNIFER does not report any events on the NovaSeq 6000 platform despite tuning its parameters to more relaxed settings. On BGISEQ-500 and MGISEQ-2000 platforms, even though CoNIFER does not report any duplication calls, DECoNT finds some deletion calls and increases duplication precision from 0 to 1%. While it does not report ant duplication calls for the HiSeq 4000 platform, DECoNT is able to increase the precision to 50% but by reporting a true positive and a false positive. The trend in deletion precision performance is similar. Finally, overall precision performance is consistently increased in all tests and the improvement ranges from 0.3% to 2.3%.

For Control-FREEC, ∼65 to ∼74 percent of the CNV calls have been improved as opposed to only ∼7 to ∼8 percent of the calls have been deteriorated by DECoNT. We observe a decrease in average absolute error after polishing in all four platforms which ranges from 0.94 to 1.0 (Table 3).

We note that the improvements provided by DECoNT on the BGI and MGI platforms are important as these systems belong to a completely different manufacturer. Since these platforms are expected to have different systematic biases and read depth distributions compared to the training data of DECoNT, we would also expect a lower testing performance. Yet, DECoNT is able to generalize well and consistently proves to be useful across a diverse set of technologies. Overall, the performance is on par with the tests obtained on Illumina Genome Analyzer II and Illumina HiSeq 2000. Polishing procedure consistently improves the performance in a platform-independent manner.

Table 3. The performance of Control-FREEC on NA12878 data before and after being polished by DECoNT. We evaluate caller performance on NA12878 data obtained using four different sequencing platforms: (i) NovaSeq 6000; (ii) HiSeq 4000; (iii) BGISEQ-500; (iv) MGISEQ-2000. Note that DECoNT models did not train on the data sequenced with any of these sequencing platforms or on NA12878 sequencing data of any form. Table shows the number of CNVs reported on each sample, the percentage of improved and deteriorated events, the average decrease in absolute error after being polished by DECoNT. In all comparisons, DECoNT provides substantial improvements showing the generalizability of our models trained on 1000 Genomes data set.

Platform	# of Events	% of Improved Events	% of Deteriorated Events	Mean Absulte Error (MAE) Difference Decreased by
NovaSeq 6000	329	**73.85%**	7.59%	**0.9392**
HiSeq 4000	437	**70.94%**	6.86%	**1.0022**
BGISEQ-500	367	**64.57%**	8.17%	**0.9809**
MGISEQ-2000	367	**64.57%**	8.17%	**0.9809**

2.4 Polishing Performance on Calls From Unseen CNV Callers

A distinct DECoNT model is trained for every WES-based germline CNV caller. This makes sense as the call regions and numbers substantially differ among algorithms in their recommended settings (e.g., CODEX2 calls 10 times more events than XHMM). We check if a DECoNT model trained using calls made by one algorithm can be used to polish the calls made by others in the absence of a trained model (e.g., due to time constraints in training).

We use the same DECoNT models trained for XHMM, CoNIFER, and CODEX2 on 1000 Genomes Data. For each tool-specific DECoNT model, we polish the calls made by others on samples not seen during training. For instance, we polish the calls made by CODEX2, using the DECoNT model trained on XHMM calls. This experiment results in 6 tests (i.e., for two-way comparison among every tool pair). We measure the performance of the polishing procedure using duplication call precision, deletion call precision and overall precision to obtain 18 performance results in total (Methods).

We observe that DECoNT improves the performance metric in 10 out of the 18 comparisons, XHMM-trained DECoNT consistently improves the other tools' performance in all metrics, except DEL precision when polishing calls reported by CoNIFER - ranging from 2% to 13% (Figure 3 and Supplementary Figure 1). Duplication precision is improved in most of the cases with the exception of CoNIFER-trained and CODEX2-trained DECoNT models deteriorating the performance of XHMM by 11% and 8% respectively. For deletion precision, this is not the case. Deletion precision is improved for CODEX2 for both DECoNT models. However, for CoNIFER, deletion precision is deteriorated by 13% and 45% when polished with XHMM-trained and CODEX2-trained DECoNT models respectively. This is due to very limited number of deletion CNV predictions of CoNIFER as even a small perturbation to the true positives of deletion calls yield large differences in precision. Also, CoNIFER-trained DECoNT model very slightly deteriorates deletion precision of XHMM calls by 5%.

While XHMM improves overall precision for both other methods, in half of the overall precision comparisons, the performance is decreased.

Overall, DECoNT is still somewhat effective despite being trained using a different call set. The training process uses the read depth information for the event regions which enables DECoNT to generalize to polish other tools. While, arguably, it can be used to polish other tools' calls, a DECoNT model trained on the calls of the to-be-polished WES-based caller is suggested, as the improvements in the discussed performance metrics are larger in this case, which is expected.

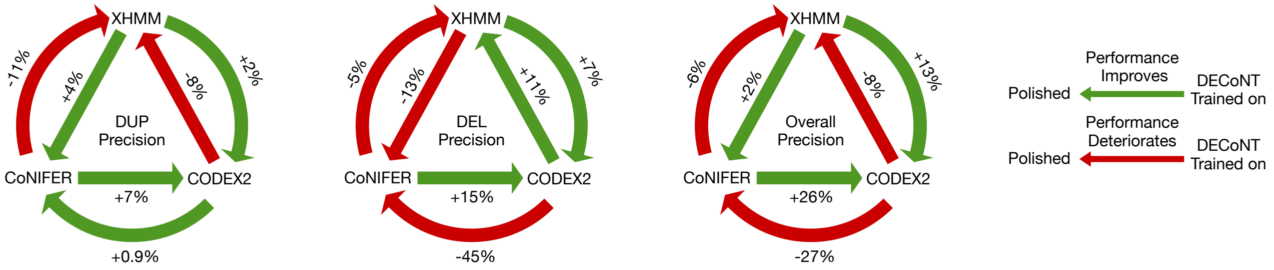

Fig. 3. Performance of DECoNT when polishing calls from unseen CNV callers. DECoNT learns a different set of weights and a different model for each WES based CNV caller. In order to demonstrate the cross-model performance, we used DECoNT to correct CNV calls made by tools other than the ones used for training. We try every pair combination. Tools being pointed by an arrow are call generating tools (i.e., being corrected). Tools at source of the arrow are the tools that are used to train the DECoNT model. Green arrows indicate improvement and red arrows indicate deterioration in the corresponding performance metric.

3 Discussion

High throughput sequencing platforms, since their inception in 2007, have now become the dominant source of data generation for biological and medical research and on their way to be routinely used for diagnosis and treatment guidance. Although whole human genome sequencing cost is now reduced below the $1,000 mark, whole exome sequencing will likely remain the main workhorse in clinical settings due to i) lower cost, ii) capturing almost all actionable genetic defects within exons, and iii) smaller data size that reduces computational burden for analysis. However, the main drawback of WES has been discovery and genotyping of CNVs. First, depth of coverage among exons are not uniform, making it very difficult to apply read depth based methods. Second, the reads often do not span CNV breakpoints which is a must for read pair and split read based approaches. Therefore it is often necessary to complement WES studies with alternative approaches such as array comparative genomic hybridization or quantitative RT-PCR.

We specifically designed our new algorithm, DECoNT, to address this limitation as a CNV call *polisher*. Using a deep learning approach, we were able to boost both the precision of several widely-used state-of-the-art algorithms that use WES data for CNV discovery. Although we trained DECoNT using matched WGS and WES samples from the 1000 Genomes Project, we also demonstrated that the performance gain is independent from the training data, the capture kit, and the sequencing platform. Therefore, the trained model is portable, and it can be applied to data sets regardless of the data generation protocol without requiring new samples to train DECoNT.

Copy number variation is an important cause of genetic diseases that may be difficult to characterize in clinical settings without specific assays. WES is a powerful method to genotype small mutations but so far it has been unsuccessful to discover large CNVs that have a more direct effect in gene losses. DECoNT aims to help ameliorate high false discovery rate problems related to CNV characterization using WES, also including integer copy number prediction. Therefore DECoNT adds an important type of genomic variation discovery to the capabilities of WES and enhances the genome analysis arsenal in the clinic. The next challange will be relieving DECoNT from dependence of existing variation callers, and make it a standalone, highly accurate CNV discovery tool using whole exome sequencing.

4 Methods

4.1 Data set

For training and testing of DECoNT, we used $1,000$ samples (i.e., HG00096 to HG02356, when sample IDs are alphabetically ordered) from the 1000 Genomes Project. For these samples we obtain both WES and WGS data. WES samples were captured using the NimbleGen SeqCap EZ Exome v3 as capture kit, and sequenced to an average of 50× depth with Illumina Genome Analyzer II and Illumina HiSeq 2000 platforms. The average read length is 76 bps. Reads were aligned to the GRCh38 using the BWA-MEM aligner [24]. WGS samples were also sequenced using the same platforms with an average read length of 100bps. Average depth coverage for this set is 30×. For XHMM, CoNIFER and CODEX2, the ground truth CNV calls are obtained using CNVNator [1] tool. For Control-FREEC (tool iv), the ground truth exact copy number variation events are obtained using mrCaNaVaR [2].

For tools that output a categorical prediction of a CNV, we also use a highly validated CNV call set published in Chaisson *et al.*, [5] as another validation source. The WGS CNV calls in this call set are thoroughly validated. That is, they were obtained via a consensus of 15 different WGS CNV callers with comparisons against high quality PB-SVs that have single base breakpoint resolution. We obtain WGS CNV calls for these 9 samples from 1000 Genomes data set. (i.e., HG00512, HG00513, HG00514, HG00731, HG00732, HG00733, NA19238, NA19239, NA19240). We also obtained aligned WES reads of these samples, with the exception of HG00514 for which no WES data was available. This data set is only used for testing.

4.2 DECoNT Model

Problem Formulation Let X denote the set of CNV events detected on the WES data set by a WES-based CNV caller, and $X^{(i)}$ denote the i^{th} event. F_i denotes the set of features we use for $X^{(i)}$ which contains the following information: (i) the chromosome that the CNV event occured ($X^{(i)}_{chr}$); (ii) the start coordinate of the CNV event ($X^{(i)}_{start}$); (iii) the end coordinate of the CNV event; (iv) the type (e.g., deletion) of the called event ($X^{(i)}_{call}$); and (v) the read depth vector between $X^{(i)}_{start}$ to $X^{(i)}_{end}$ ($X^{(i)}_{RDSeq}$). Let $Y^{(i)}_{gt}$ denote the ground truth label obtained from the WGS CNV call for $X^{(i)}$. There are two cases: (i) For the tools that predict the existence of an event $X^{(i)}_{gt} \in \{0, 1, 2\}$, denoting no call, deletion or duplication, respectively; and (ii) For the tools that predict the copy number $Y^{(i)}_{gt} \in \mathbb{Z}^{\geq}$. Then, the problem at hand is formulated as a classification task for (i), and as a regression task for (ii). That is, our goal is to learn a function f such that $f(F_1, \cdots, F_n) \rightarrow (Y^{(1)}_{pr}, \cdots, Y^{(n)}_{pr})$ such that the difference is between $(Y^{(1)}_{pr}, \cdots, Y^{(n)}_{pr})$ and $(Y^{(1)}_{gt}, \cdots, Y^{(n)}_{gt})$ is minimized with respect to a loss function. Here, $n = |X|$ and $Y^{(i)}_{pr}$ is the predicted label for $X^{(i)}$ and it is in the same domain as $Y^{(i)}_{gt}$ in respective tasks.

DECoNT Architecture DECoNT is an end-to-end multi-input neural network designed for polishing and improving the performance of the WES-based germline CNV callers. It is capable of improving accuracy of WES CNV calling for both exact CNV prediction (i.e., integer) and categorical CNV prediction cases (i.e., deletion, duplication or no call). For each CNV caller, a distinct network is trained.

DECoNT's pipeline for the categorical CNV prediction case can be divided into three main building blocks: (i) a data preprocessing step that extracts the read depth for genomic regions of interest (i.e., CNV call regions made by the CNV caller). It also normalizes the read depth sequence and acts as a regularizer for the model. Resulting read depth information is -1 padded to the length of the longest call sequence and masked; (ii) a bidirectional LSTM network (BiLSTM) that inputs the read depth sequence and extracts the required encoded features (i.e., embeddings). This subnetwork has 128 neurons in each direction and is followed by a batch normalization layer; and (iii) a 2-layered fully connected (FC) neural network that inputs the embedding calculated by Bi-LSTM, concatenated with the prior CNV prediction of the CNV caller (a one-hot-encoded vector). The first FC layer has 100 neurons and uses ReLU activation. The output layer has 3 neurons and it calculates the posterior probability of each event via softmax activation: no call, deletion, or duplication. We use weighted cross-entropy as the loss function. This architecture has a total of $160,351$ parameters with $159,837$ of which are trainable. The rest are the batch normalization parameters.

For a training data set of N samples, the formulation of DECoNT can be summarized as follows:

$$X^{(1:N)}_{encoding1} = \text{BatchNorm}(\text{BiLSTM}^{(128)}(\text{BatchNorm}(\text{Mask}(X^{(1:N)}_{RDSeq})))) \tag{1}$$

$$X^{(1:N)}_{encoding2} = \text{CAT}(X^{(1:N)}_{encoding1}, X^{(1:N)}_{call}) \tag{2}$$

$$X^{(1:N)}_{encoding3} = \text{ReLU}(\text{FC}^{(100)}(X^{(1:N)}_{encoding2})) \tag{3}$$

$$Y^{(1:N)}_{pr} = \text{Softmax}(\text{FC}^{(3)}(X^{(1:N)}_{encoding3})) \tag{4}$$

where BiLSTM$^{(\cdot)}$ represents bi-direcitonal LSTM layer with $\cdot$ hidden units in each direction. Similarly, FC$^{(\cdot)}$ represents a dense layer with $\cdot$ neurons. ReLU and BatchNorm stand for rectified linear unit activation function and batch normalization respectively.

Using $Y^{(1:N)}_{pr}$ and $Y^{(1:N)}_{gt}$ training phase minimizes the categorical cross-entropy loss. We use Adam optimizer [19] with a mini batch size of 128 samples. All weights in the network are initialized using Xavier initialization [9].

DECoNT's pipeline for the exact (i.e., integer) CNV prediction is almost the same as the one described above. The first difference is instead of taking the one-hot encoded version of the CNV call, it inputs an integer value representing the called copy number. The second difference is at the output layer. Instead of 3 neurons with softmax activation, this version has a single neuron with ReLU activation to perform regression instead of classification. It has a total of $160,149$ parameters, $159,635$ of which are trainable. Again, the rest are the batch normalization parameters. So, the last layer in the formulation above (Eq. 4) is replaced by the following layer and in this case $Y^{(1:N)}_{pr} \in \mathbb{Z}^{\geq}$.

$$X^{(1:N)}_{pr} = \text{ReLU}(\text{FC}^{(1)}(X^{(1:N)}_{encoding3})) \tag{5}$$

Using $Y^{(1:N)}_{pr}$ and $Y^{(1:N)}_{gt}$ training phase now minimizes the mean absolute error loss. Again, we use Adam optimizer with a mini batch size of 128 samples and we use Xavier initialization for weights.

4.3 Polishing the State-of-the-art WES-based Germline CNV Callers

We polish the CNV calls made by four state-of-the-art WES-based germline CNV callers (i) XHMM [8], (ii) CoNIFER [20], (iii) CODEX2 [15], and (iv) Control-FREEC [4]. Tools (i - iii) perform categorical CNV prediction and (iv) performs exact CNV prediction. We use calls made on the WGS samples by CNVNator [1] as the ground truth call set for discrete predictions and the exact copy number predictions made by mrCaNaVaR as the ground truth call set for integral prediction (i.e., Control-FREEC). First, DECoNT obtains the results of these tools (i - iv). Then, it learns to correct these calls on a portion of the 1000 Genomes data set using ground truth calls. Finally, on the left out test portion of the data, we compare the performance of the CNV callers before and after polishing by DECoNT.

Settings for the WES-based CNV Callers We follow the recommended settings for the WES-based Callers. For XHMM, the parameters are set as follows: (i) $Pr(\text{start DEL}) = Pr(\text{start DUP}) = 1e{-}08$, (ii) mean number of targets in CNV (geometric distribution) = 6, (iii) mean distance between targets within CNV (exponential decay) = $70kb$, and (iv) DEL, DIP, DUP read depth distributions modeled as $\sim \mathcal{N}(-3,1)$, $\sim \mathcal{N}(0,1)$ and $\sim \mathcal{N}(3,1)$, respectively. For CODEX2, minimum read coverage of 20 was enforced at the filtering step. Then, the algorithm automatically chooses its parameter, K, using BIC (i.e. Bayesian Information Criterion) and AIC (i.e. Akaike Information Criterion). CoNIFER performs SVD on the data matrix and then removes n singular vectors with n largest singular values. We set n to 6. Control-FREEC has 45 parameters which were all set to default values as stated in [4].

Training Settings for DECoNT We train a DECoNT model for each of the above-mentioned tools. The set X of CNV calls per tool is shuffled and divided into training, validation and testing sets which contain $70\%, 20\%$, and 10% of the data, respectively. The number of events in the test sets are $6,832$ (3101 no-calls, 2098 duplications, 1633 deletions); $81,761$ (67885 no-calls, 3042 duplications, 10834 deletions); 180 (85 no-calls, 43 duplications, 52 deletions) and $20,482$ (minimum copy number is 0, maximum copy number is 585); for XHMM, CODEX2, CoNIFER, and Control-FREEC, respectively. The second input of the algorithm is the read depth for the CNV-associated regions on the WES data. We calculate it using the Sambamba tool [34]. For all tools other than CODEX2, DECoNT is trained up to 30 epochs with early stopping by checking the loss on the validation fold. Training for CODEX2 has a maximum epoch number 60.

Performance Metrics Tools (i - iii) predict CNVs either as deletion or duplication. The main problem of these callers are false discovery rates [38, 33]. Given a deletion or duplication call by tools (i - iii), DECoNT outputs a probability for the call to be deletion, duplication or no call (i.e., false discovery). The option with the highest probability is returned as the prediction.

In order to assess the performance of tools (i - iii) before and after being polished, we calculate the following performance metrics using $Y_{pr}^{(1:N)}$ and $Y_{gt}^{(1:N)}$: (i) duplication call precision; (ii) deletion call precision, and (iii) overall precision. We first define the following variables: TP_1 := number of duplications correctly identified; TP_2 := number of deletions correctly identified; FP_1 := number of duplications incorrectly identified; FP_2 := number of deletions incorrectly identified.

Then, the performance metrics are defined as follows:

$$\text{Duplication call precision} = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FP}_1} \tag{6}$$

$$\text{Deletion call precision} = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FP}_2} \tag{7}$$

$$\text{Overall precision} = \frac{\mathrm{TP}_1 + \mathrm{TP}_2}{\mathrm{TP}_1 + \mathrm{TP}_2 + \mathrm{FP}_1 + \mathrm{FP}_2} \tag{8}$$

In order to test DECoNT's performance on exact CNV prediction problem, which is a regression task, we use Absolute Error (AE) between the predicted and ground truth copy number values. For an event X_i, $AE^{(i)}$ is defined as follows:

$$AE^{(i)} = \left| Y_{pr}^{(i)} - Y_{gt}^{(i)} \right| \tag{9}$$

Time Performance All models are trained on a SuperMicro SuperServer 4029GP-TRT with 2 Intel Xeon Gold 6140 Processors (2.3GHz, 24.75M cache), 251GB RAM and 3 NVIDIA GeForce RTX 2080 Ti (11GB, 352Bit) and 1 NVIDIA TITAN RTX GPUs (24GB, 384Bit). We used 4 GPUs in parallel to train all 4 models and total training times were approximately as follows: $\sim 70, 12, 95$, and 50 hours for XHMM, CoNIFER, CODEX2, and Control-FREEC, respectively. Note that training is performed offline. The average polishing time per sample is in the order of seconds, for all models.

4.4 Polishing Samples from Other Sequencing Platforms

The training data we use is obtained using Illumina Genome Analyzer II and Illumina HiSeq 2000 machines. We check if models trained on these 1000 Genomes data can be used to polish CNV calls made on WES samples obtained using other sequencing platforms or capture kits that have not been seen by DECoNT.

We obtain the WES data for the sample NA12878, sequenced using four different platforms: (i) Illumina NovaSeq 6000; (ii) Illumina HiSeq 4000; (iii) BGISEQ-500; and (iv) MGISEQ-2000. Reads are aligned to the reference genome (GRCh38) using BWA [24] with *-mem* option and default parameters. Average depth coverage for these samples are 241×, 395×, 328×, and 129×, respectively. We use these four samples only for testing. All considered WES-based CNV callers are used to call CNV events on these four WES samples with default parameters. Using the CNVnator calls obtained on the WGS sample for NA12878 as the ground truth, we measure the performance the CNV callers before and after polishing with DECoNT. Note that NA12878 data is not included in the training data set in any form.

4.5 Polishing Other WES-based CNV Caller Algorithms

In our framework, a separate DECoNT model is trained for every WES-based germline CNV caller. We check if a DECoNT model trained using calls made by one algorithm can be used to polish the calls made by others in the absence of a trained model.

We use the same three models trained with the settings described in Section 4.3 for XHMM, CoNIFER, and CODEX2. For each tool-specific DECoNT model, we polish the calls made by others. Here the training and testing folds are again exclusive. For testing, we use the same test folds for each tool as described in in Section 4.3. This experiment results in 6 tests (i.e., for two-way comparison among every tool pair). We measure the performance of the polishing procedure using duplication precision, deletion precision and overall precision to obtain 18 performance results in total.

Competing Interests: Authors declare no competing interests.
Author Contributions: AEC and CA designed and supervised the study. AEC and FO designed the model. FO implemented the software and performed the experiments. AEC, CA and FO wrote the manuscript.

References

1. Abyzov, A., Urban, A.E., Snyder, M., Gerstein, M.: Cnvnator: An approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. Genome Research **21**(6), 974–984 (Jul 2011). https://doi.org/10.1101/gr.114876.110
2. Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O., et al.: Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics **41**(10), 1061 (2009)
3. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.L., Abel, L.: Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences **112**(17), 5473–5478 (2015)
4. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., Barillot, E.: Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics **28**(3), 423–425 (Jun 2011). https://doi.org/10.1093/bioinformatics/btr670
5. 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)
6. Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al.: A copy number variation morbidity map of developmental delay. Nature genetics **43**(9), 838 (2011)
7. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.: Synaptic, transcriptional and chromatin genes disrupted in autism. Nature **515**(7526), 209–215 (2014)
8. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., Mccarroll, S.A., O'Donovan, M.C., Owen, M.J., et al.: Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. The American Journal of Human Genetics **91**(4), 597–607 (2012). https://doi.org/10.1016/j.ajhg.2012.08.005
9. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256 (2010)
10. Heinzen, E.L., Need, A.C., Hayden, K.M., Chiba-Falek, O., Roses, A.D., Strittmatter, W.J., Burke, J.R., Hulette, C.M., Welsh-Bohmer, K.A., Goldstein, D.B.: Genome-wide scan of copy number variation in late-onset alzheimer's disease. Journal of Alzheimer's Disease **19**(1), 69–77 (2010)
11. Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., Berney, D., Scher, H., Carver, B., Scardino, P., et al.: Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Elife **7**, e37294 (2018)
12. Ho, S.S., Urban, A.E., Mills, R.E.: Structural variation in the sequencing era. Nature Reviews Genetics pp. 1–19 (2019)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C.: Detection of large-scale variation in the human genome. Nature genetics **36**(9), 949–951 (2004)
15. Jiang, Y., Wang, R., Urrutia, E., Anastopoulos, I.N., Nathanson, K.L., Zhang, N.R.: Codex2: full-spectrum copy number variation detection by high-throughput dna sequencing. Genome Biology **19**(1) (2018). https://doi.org/10.1186/s13059-018-1578-y

16. Kadalayil, L., Rafiq, S., Rose-Zerilli, M.J., Pengelly, R.J., Parker, H., Oscier, D., Strefford, J.C., Tapper, W.J., Gibson, J., Ennis, S., et al.: Exome sequence read depth methods for identifying copy number changes. Briefings in bioinformatics **16**(3), 380–392 (2015)
17. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.: Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv p. 531210 (2019)
18. Kebschull, J.M., Zador, A.M.: Sources of pcr-induced distortions in high-throughput sequencing data sets. Nucleic acids research **43**(21), e143–e143 (2015)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Krumm, N., Sudmant, P.H., Ko, A., Oroak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., Eichler, E.E.: Copy number variation detection and genotyping from exome sequence data. Genome Research **22**(8), 1525–1532 (2012). https://doi.org/10.1101/gr.138115.112
21. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., Eichler, E.E., Project, N.E.S., et al.: Copy number variation detection and genotyping from exome sequence data. Genome research **22**(8), 1525–1532 (2012)
22. Kumaran, M., Cass, C.E., Graham, K., Mackey, J.R., Hubaux, R., Lam, W., Yasui, Y., Damaraju, S.: Germline copy number variations are associated with breast cancer risk and prognosis. Scientific reports **7**(1), 1–15 (2017)
23. Levy, D., Ronemus, M., Yamrom, B., Lee, Y.h., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., et al.: Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron **70**(5), 886–897 (2011)
24. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem (2013)
25. Macintyre, G., Goranova, T.E., De Silva, D., Ennis, D., Piskorz, A.M., Eldridge, M., Sie, D., Lewsley, L.A., Hanif, A., Wilson, C., et al.: Copy number signatures and mutational processes in ovarian carcinoma. Nature genetics **50**(9), 1262–1270 (2018)
26. Pankratz, N., Dumitriu, A., Hetrick, K.N., Sun, M., Latourelle, J.C., Wilk, J.B., Halter, C., Doheny, K.F., Gusella, J.F., Nichols, W.C., et al.: Copy number variation in familial parkinson disease. PloS one **6**(8) (2011)
27. Reid, B.M., Permuth, J.B., Chen, Y.A., Fridley, B.L., Iversen, E.S., Chen, Z., Jim, H., Vierkant, R.A., Cunningham, J.M., Barnholtz-Sloan, J.S., et al.: Genome-wide analysis of common copy number variation and epithelial ovarian cancer risk. Cancer Epidemiology and Prevention Biomarkers **28**(7), 1117–1126 (2019)
28. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al.: Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell **180**(3), 568–584 (2020)
29. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al.: Large-scale copy number polymorphism in the human genome. Science **305**(5683), 525–528 (2004)
30. Singh, T., Neale, B., Daly, M., Consortium, S.E.M.A., et al.: Initial results from the meta-analysis of the whole-exomes of over 20,000 schizophrenia cases and 45,000 controls. European Neuropsychopharmacology **29**, S813–S814 (2019)
31. Study, T.D.D.D., Fitzgerald, T., Gerety, S., Jones, W., van Kogelenberg, M., King, D., McRae, J., Morley, K., Parthiban, V., Al-Turki, S., et al.: Large-scale discovery of novel genetic causes of developmental disorders. Nature **519**(7542), 223–228 (2015)
32. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L.B., Posukh, O.L., Sahakyan, H., Watkins, W.S., Yepiskoposyan, L., Abdullah, M.S., Bravi, C.M., Capelli, C., Hervig, T., Wee, J.T.S., Tyler-Smith, C., van Driem, G., Romero, I.G., Jha, A.R., Karachanak-Yankova, S., Toncheva, D., Comas, D., Henn, B., Kivisild, T., Ruiz-Linares, A., Sajantila, A., Metspalu, E., Parik, J., Villems, R., Starikovskaya, E.B., Ayodo, G., Beall, C.M., Di Rienzo, A., Hammer, M.F., Khusainova, R., Khusnutdinova, E., Klitz, W., Winkler, C., Labuda, D., Metspalu, M., Tishkoff, S.A., Dryomov, S., Sukernik, R., Patterson, N., Reich, D., Eichler, E.E.: Global diversity, population stratification, and selection of human copy-number variation. Science **349**, aab3761 (Sep 2015). https://doi.org/10.1126/science.aab3761
33. Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S., Zhu, M.: An evaluation of copy number variation detection tools from whole-exome sequencing data. Human mutation **35**(7), 899–907 (2014)
34. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., Prins, P.: Sambamba: fast processing of ngs alignment formats. Bioinformatics **31**(12), 2032–2034 (2015). https://doi.org/10.1093/bioinformatics/btv098
35. The 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature **526**(7571), 68–74 (Sep 2015). https://doi.org/10.1038/nature15393, http://dx.doi.org/10.1038/nature15393
36. Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W., Pereira, S.L., Whitney, J., Chan, A.J., Pellecchia, G., et al.: A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. The American Journal of Human Genetics **102**(1), 142–155 (2018)
37. Yu, Y., Si, X., Hu, C., Zhang, J.: A review of recurrent neural networks: Lstm cells and network architectures. Neural computation **31**(7), 1235–1270 (2019)
38. Zare, F., Dow, M., Monteleone, N., Hosny, A., Nabavi, S.: An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC bioinformatics **18**(1), 286 (2017)

39. Zarrei, M., Burton, C.L., Engchuan, W., Young, E.J., Higginbotham, E.J., MacDonald, J.R., Trost, B., Chan, A.J., Walker, S., Lamoureux, S., et al.: A large data resource of genomic copy number variation across neurodevelopmental disorders. NPJ genomic medicine **4**(1), 1–13 (2019)