ÉCOLE: Learning to call copy number variants on whole exome sequencing data

Berk Mandiracioglu[1†], Furkan Özden[2†], Can Alkan[3], and A. Ercüment Çiçek[3,4‡]

1. Department of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

2. Department of Computer Science, Oxford University, Oxford, UK

3. Department of Computer Engineering, Bilkent University, Ankara, Turkey

4. Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA

† Equal contribution, ‡ correspondence: `cicek@cs.bilkent.edu.tr`

Abstract. Copy number variants (CNV) are shown to contribute to the etiology of several genetic disorders. Accurate detection of CNVs on whole exome sequencing (WES) data has been a long sought after goal for use in clinic. This was not possible despite recent improvements in performance because algorithms mostly suffer from low precision and even lower recall on expert-curated gold standard call sets. Here, we present a deep learning-based somatic and germline CNV caller for WES data, named *ÉCOLE*. Based on a variant of the transformer architecture, the model learns to call CNVs per exon, using high confidence calls made on matched WGS samples. We further train and fine-tune the model with a small set of expert calls via transfer learning. We show that ÉCOLE is able mimic the expert labeling for the first time with 68.7% precision and 49.6% recall. This corresponds to precision and recall improvements of 18.7% and 30.8% over the next best performing methods, respectively. ÉCOLE is the first method to achieve high precision and recall in mimicking human expert CNV calling. We also show that same fine tuning strategy using tumor samples enables ÉCOLE to detect RT-qPCR validated variations in bladder cancer samples without the need for a control sample. We think these features of ÉCOLE make CNV calling on WES data feasible for clinical use. ÉCOLE is available at https://github.com/ciceklab/ECOLE.

Keywords: CNV calling · deep learning · exome sequencing.

1 Introduction

Copy number variants (CNVs) are well-known and important risk factors for many conditions such as cancer [32,25], schizophrenia [33,1] and autism [14]. High throughput sequencing (HTS) has been the standard technique for the detection of CNVs over the the last decade. Various CNV detection algorithms that use whole genome sequencing (WGS) data have been very successful [19,26,4,40,22,39,18] with sensitivity and precision values reaching up to 96% and 97%, respectively [2]. This is in contrast to such algorithms working on the whole exome sequencing (WES) data, which suffer from very low precision [35,41,31]. WGS is a more accommodating platform for this task because it does not employ targeting probes which introduce length, GC and reference biases [28,23,21].On the other hand, WES has been more appealing in the clinic due to being more compact, interpretable and affordable than WGS. Unfortunately, WES technology has no clinical use for CNV detection due to these limitations.

A recent polishing approach has proven useful to correct the calls of many state-of-the-art WES-based germline CNV callers using more trustworthy calls made on the matched WGS samples [31]. While this was an important step forward, there are still bottlenecks to make it a feasible option for the use in the clinic. First problem is with the sensitivity of the results. The polisher can only work on the calls (e.g., deletion) returned by the base algorithm. It either changes these calls (e.g., to duplication) or neutralizes them (e.g., to no-call). While this helps to reduce the false discovery rate, it has limited effect on sensitivity as a polisher cannot make new calls (e.g., convert a no-call to deletion/duplication). Unfortunately, sensitivity has mostly been out of the scope of WES-based CNV calling domain due to very low performance. Second problem is that even precision performance after polishing is limited on expert-curated CNV call sets which are regarded as the golden ground truth (up to 35%). This is because the polisher uses automated WGS-based CNV calls as labels for model training but these labels (calls) have a very different distribution compared to human expert decisions. Unfortunately, such manually curated call sets are extremely small in size, which prohibits training machine learning models. Thus, a caller that is able to mimic human expert reasoning with high precision and recall would enable broad use of WES-based germline CNV detection in clinic.

Here, we present the first deep learning-based method (ÉCOLE: Exome-based COpy number variation calling LEarner) which can independently learn to perform somatic and germline CNV calling on WES data.

Our model is based on a novel variant of the *transformer* model [37] which is the state-of-the-art approach to process sequence data in the natural language processing domain [30,10]. ÉCOLE processes the read-depth signal over each exon. It learns which parts of the signal needs to be focused on and in which context (i.e., chromosome) to call a CNV. It uses the high-confidence calls (i.e., labels) obtained on the matched WGS samples as the semi-ground truth. ÉCOLE improves the exon-wise precision and also the recall of the next best method's performance substantially on a benchmark of automated WGS calls (13.5% and 17.9% improvements, respectively). It is the only method with balanced precision and recall. Moreover, for the first time, we also propose using transfer learning and fine tune the model parameters using a small number of human expert-labeled samples. We show that this approach improves the precision and recall by ∼18% and 30% respectively in predicting human labels. Thus, ÉCOLE is able to act as a human expert who is calling germline CNVs with substantial performance. Similarly, we use the fine tuning method to adapt ÉCOLE to call somatic variations using cancer samples. We show that we are able to detect PCR-validated copy number aberrations in 13 out of 16 bladder cancer samples while the state-of-the-art method can only detect validated calls in 2 samples even after polishing. With the ability to act as both a germline and a somatic CNV caller and being flexible to adapt to diseases and human experts easily with fine tuning, we propose ÉCOLE as a feasible option to use in clinic for CNV detection.

2 Results

2.1 ÉCOLE overview

Our model ÉCOLE is a deep neural network model which uses a novel variant of the transformer architecture [37] at its core. The transformer is a parallelizable encoder-decoder model that receives an input and applies alternating layers of multi-headed self attention, multi-layer perceptron (MLP) and layer normalization layers to it. Transformer architecture has achieved the state-of-the-art results in signal processing over recurrent neural networks in the natural language processing domain [37] as well as recently over the convolutional based models in the computer vision domain [13].

Figure 1 shows an overview of the system architecture. ÉCOLE takes the read depth over an exon at the base pair resolution. This information is transformed into a read depth embedding using a multi-layered perceptron.

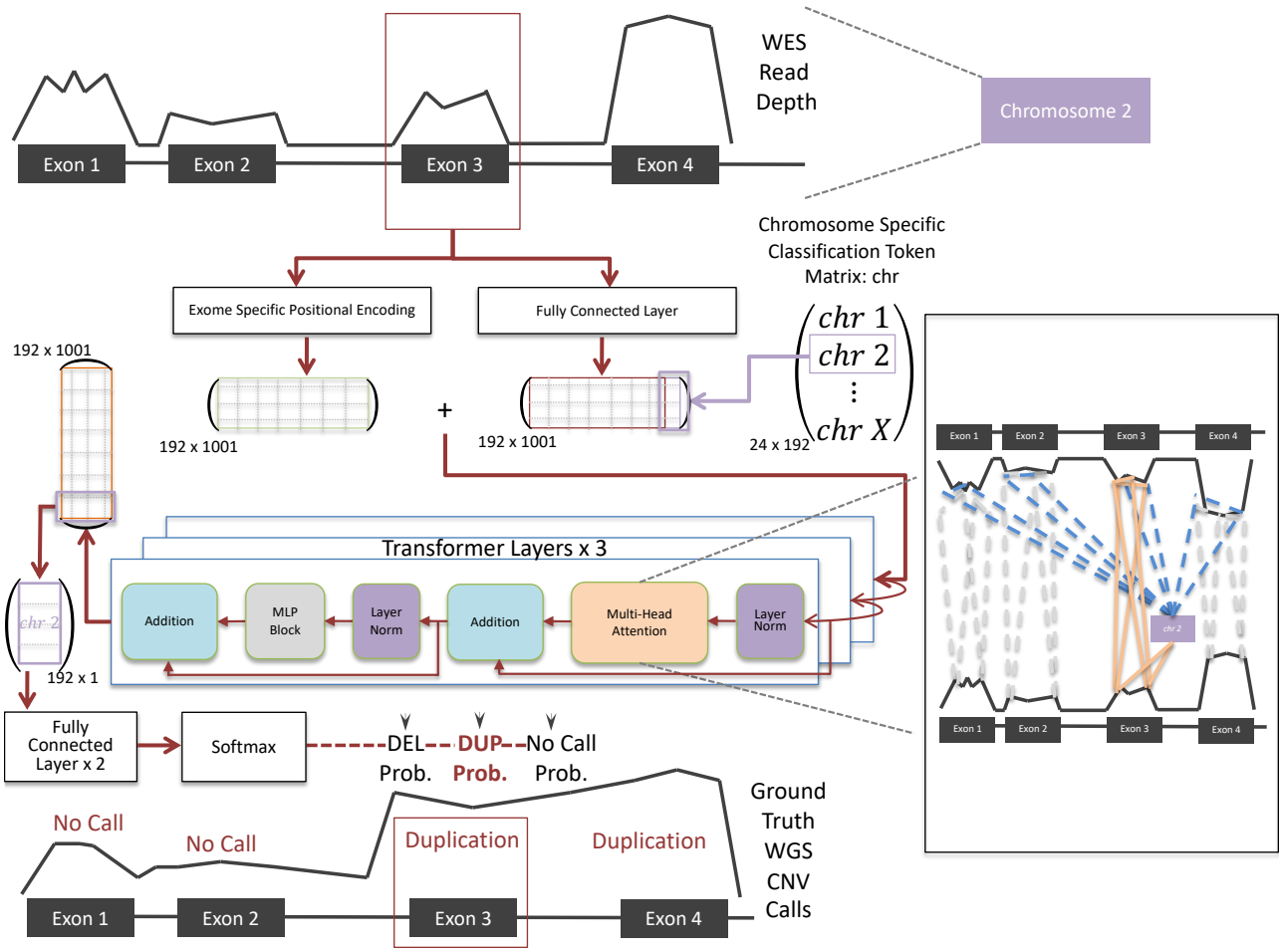

Figure. 1: ÉCOLE's system overview. The model inputs per exon (i) the read depth signal (length 1000, padded and masked), (ii) chromosome number, and (iii) start and end coordinates of the region. It maps the each 1000 read depth value to a higher dimensional vector $\in \mathbb{R}^{192}$ (input embedding) using a fully connected (FC) layer, which is concatenated with a chromosome specific classification token vector of $\in \mathbb{R}^{192}$. These chromosome specific tokens enable the model to learn the chromosome context of the exon samples to perform calls. Transformer layers use multi-head attention mechanism which learns the connections of each read depth value of base pairs with respect to all others base pairs in the given exon sample. Therefore, attention mechanism also learns to which read depth values the classification token needs to pay attention for the respective CNV call. To further learn positional context of the base pairs within the chromosome, the start and end coordinates of the sample are used to calculate the exon specific positional encoding $\in \mathbb{R}^{192\times1001}$. Two matrices are concatenated and input to a cascade of 3 transformer layers which generates an output vector $\in \mathbb{R}^{192\times1001}$. Then, the mapped transformation of chromosome specific classification token is fetched, which has the size $\mathbb{R}^{192}$. Finally, for final decision (DEL, DUP or no-call), we use 2 FC layers followed by softmax activation.

We use a classification token to be learnt, which is concatenated with the read depth embedding as also done in [13]. However, in our setting, this token is chromosome specific to add further context in to the classification task. Finally, the model uses a positional encoding vector which is summed up with the transformed read depth encoding and the classification token. This encoding informs the model on the absolute position of the considered exon. ÉCOLE applies 3 transformer blocks to this vector. Doing so, it learns the importance of the read depth over a specific base pair, with respect to the read depth on other base pairs, within the same exon region. That is, ÉCOLE uses an attention mechanism to learn to focus on which base pairs in which context (i.e., deletion, duplication or no-call). This is in analogy to natural languages where the same word (read depth) having a different stress in different sentences (exons) and in different paragraphs of a text (chromosomes). Finally, we perform classification using a two-layered perceptron which uses the output of the final transformer block. ÉCOLE uses higher confidence CNV calls obtained on 1000 Genomes WGS data as the "semi"-ground truth (i.e., compared to WES) to train the model. We use the CNVnator algorithm as the WGS-based germline CNV caller which provides has high sensitivity (86%-96%), and high precision (80%-97%) [3].

ÉCOLE is able to transfer the highly accurate decision making of a WGS-based CNV caller into the WES domain to achieve state-of-the-art performance. Yet, no algorithm in the literature is able to mimic the decision making of a human expert as such labeled data is available for a very small number of samples which is insufficient for training a complex model like ÉCOLE. Here, we apply transfer learning for the first time in the CNV calling domain. That is, we further tune the parameters of the ÉCOLE model (trained with the semi-ground truth) using a very small number of human expert labeled samples (dubbed $\text{ÉCOLE}^{\text{FT-EXPERT}}$). Similarly, to enable the model to call somatic CNVs, fine tune the parameters of the ÉCOLE model with bladder cancer samples with semi-ground-truth labels (CNVnator). We call this model $\text{ÉCOLE}^{\text{FT-SOMATIC}}$.

2.2 ÉCOLE achieves high performance in WES-based germline CNV calling

Evaluation Criteria. We consider calls per exon as our fixed evaluation unit. That is, for each exon, ÉCOLE makes a CNV prediction. For compared methods, we intersect their CNV call segments with the exons, if they report CNVs for larger regions than exons (e.g., merged bins, exons etc). Each exon has a unique semi-ground truth label (i.e., deletion, duplication or no-call) assigned with respect to the call made on WGS data of the same sample. See Supplementary Figure 5 for the visual demonstration of this procedure.

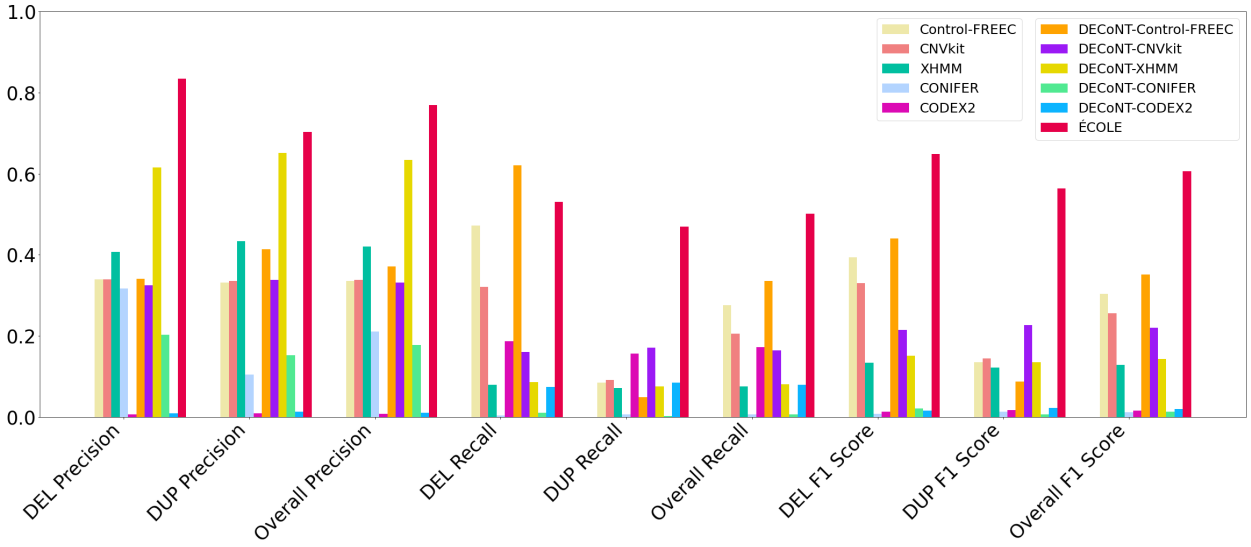

Figure. 2: The performance comparison of the WES-based CNV callers on the 1000 Genomes test set which contains 157 samples. CNVnator calls on the matched WGS samples are used as the ground truth. CNVkit and Control-FREEC return exact (integer) copy number predictions, which are discretized into deletion, duplication and no-call. We also used the DECoNT tool to polish call sets of all considered tools which are denoted by DECONT-*tool_name*

CNV calling performance of ÉCOLE on a WGS-based semi-ground truth call set. We compare the performance of ÉCOLE with the state-of-the-art germline CNV callers from the literature on the 1000 Genomes WES samples (test split, see Section 4.1 for data set details). The semi-ground truth CNV calls are obtained using CNVnator on the WGS samples of the same individuals. Compared methods are XHMM, CODEX2, CONIFER, CNV-kit, Control-FREEC [15,20,27,34,5]. Among those, CNV-kit and Control-FREEC predict integer copy numbers while the others report the CNV (i.e., deletion or duplication). To be able to fairly compare the performance with them, we discretize their predictions. We also polish the callsets of each algorithm using the CNV call polisher DECoNT and compare ÉCOLE with the polished versions of the call sets of these algorithms (See Section 4.2 for details of compared methods).

Figure 2 shows the precision, recall and F1 score results for each algorithm, and Supplementary Table 1 shows the corresponding values, see Supplementary Table 10 for the respective confusion matrices. ÉCOLE achieves the best average precision values over even polished versions of the other algorithms and provides 13.5% improvement over the next best performance by DECoNT polished XHMM callset (DECoNT-XHMM). Also in terms of deletion and duplication precision, we provide the best results with 21.9% and 5.2% improvements,

respectively. Studies in the literature focus on precision and not recall for mainly two reasons: (i) Not having false positives is the primary focus for the use in clinic, and (ii) due to (i), recall is very low to the point that it is not even reported. ÉCOLE achieves 50.1% overall recall which is a 18% improvement over the second best model, DECoNT-XHMM. While ÉCOLE is able to achieve high recall and it is also the first method that is able to balance precision and recall. ÉCOLE yields an F1-score of 60.7% which corresponds to an improvement of 46.3% over the second best result obtained by the DECoNT-XHMM call set. For all other methods, if the precision is high, the recall is low due to the small number of calls made and if the recall is high, precision is low due to the large number of predictions made.

We also compare ÉCOLE with CNLearn which is a random forest based method that creates an ensemble of four WES-based callers (See Section 4.2 for details). We compare our results on the 28 samples for which we obtained results via personal communication with Santhosh Girirajan. As shown in Table 1 ÉCOLE performs substantially better in all metrics considered, and see Supplementary Table 11 for the corresponding confusion matrix.

Table 1: Performance comparison of ÉCOLE with CNLearn on 28 samples from the 1000 Genomes Project.

Tool	DEL Precision	DUP Precision	Overall Precision	DEL Recall	DUP Recall	Overall Recall	DEL F1 Score	DUP F1 Score	Overall F1 Score
CNLearn	0.084	0.221	0.152	0.002	0.010	0.006	0.004	0.019	0.012
ÉCOLE	0.834	0.679	0.757	0.541	0.500	0.520	0.656	0.675	0.617

Note that these 28 samples are not included in the training set of ÉCOLE and the predictions are obtained via personal communication with the authors.

CNV calling performance generalizes to other sequencing platforms The WES data we used to train the ÉCOLE model were obtained using Illumina HiSeq 2000 and Illumina Genome Analyzer II platforms. Here, we show that ÉCOLE's performance generalizes to other sequencing platforms that are not used during training. Here, we test the ÉCOLE model using the sequencing data of the NA12828 sample obtained using (i) BGISEQ 500, (ii) HiSeq 4000, (iii) NovaSeq 6000, and (iv) MGISEQ 2000. We did not use any related data for this sample during the training process.

The results are shown in Supplementary Table 3 and Supplementary Figure 3. See Supplementary Table 13-16 for the corresponding confusion matrices. We observe that ÉCOLE is the best performing method in all categories with overall F1 scores ranging between 49.9% and 58.6%. Note that the performance for BGISEQ and MGISEQ platforms are relatively more important for these set of experiments as these platforms are built by an entirely different manufacturer. In BGISEQ and MGISEQ, we observe that the ÉCOLE remains to be the best performing tool with respect to all considered benchmarks, providing an at least ∼14% overall F1 score improvements over the second best method, DECoNT-Control-FREEC. Once again, ÉCOLE is the only method with balanced precision and recall. Similarly, in NovaSeq 6000 and HiSeq 4000 platforms we observe ∼40% and ∼30% overall F1 score improvements.

These results demonstrate the robustness of our model in dealing with systematic biases and noise introduced by different systems. We show that our model can be used across platforms when there is not enough WGS-matched data samples to train a ÉCOLE model obtained on the platform of interest.

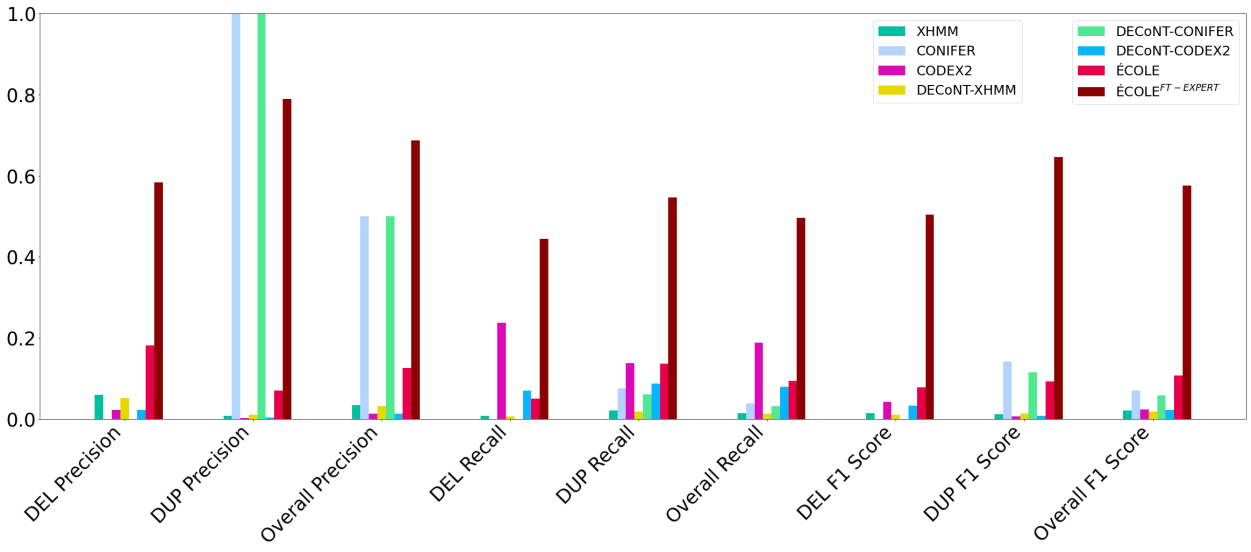

Figure. 3: The performance comparison of the WES-based CNV callers on the 1000 Genomes test set which contains 157 samples. Chaisson et al.'s human expert curated calls on the matched WGS samples are used as the ground truth [7]. We also used the DECoNT tool to polish call sets of all considered tools which are denoted by DECONT-*tool_name*. ÉCOLE$^{\text{FT-EXPERT}}$ corresponds to the fine-tuned version of ÉCOLE model with human expert calls.

CNV calling performance on mimicking human expert calls. Here, we use the highly validated CNV call set produced by Chaisson et al. [7] as the ground truth to test the performance of the WES-based CNV callers. Note that this call set contains CNV calls for 9 individuals from the 1000 Genomes Project WGS samples. This is a consensus call set which relies on the results of 15 WGS-based CNV callers compared against structural variatons generated using PacBio with single basepair breakpoint resolution. Thus, this dataset is the closest we can get to a true ground truth set curated by human experts. We use 8 samples from this call set who have the matching WES data. Calls on 4 of these samples are used for training and the rest are used for benchmarking (see Methods for details).

Results are shown in Figure 3. Please see Supplementary Table 2 for the values in this figure and Supplementary Table 12 for the corresponding confusion matrices. All compared CNV callers and their polished versions have much lower F1 score performance on predicting human expert calls compared to predicting the WGS-based semi ground truth labels (i.e., CNVnator calls). The top F1 score performance reaches up to ∼10% as opposed to ∼20%, and no algorithm shows balanced precision and recall. These are inline with the observations in [31].

We also observe that ÉCOLE also have lower performance and provides only 3.7% overall F1 score improvement over the next best method, CONIFER. This is expected as the call set is a consensus of many callers which is cross referenced with complementary long read data and further filtered by human experts. Thus, the label distribution on this data set is different than what ÉCOLE is trained with. This call set is more than two orders of magnitude smaller than what we use to train ÉCOLE which prohibits training an ÉCOLE model from scratch.

To address this issue, we use transfer learning and use the left-out 4 samples from Chaisson et al. to fine tune the parameters of the trained ÉCOLE model. That is, we further train the final ÉCOLE model using the human expert labeled samples and adjust the model weights to be able to mimic human reasoning. We call this fine-tuned model $\text{ÉCOLE}^{\text{FT-EXPERT}}$. Note that none of the other methods have a way of incorporating this information.

We observe that $\text{ÉCOLE}^{\text{FT-EXPERT}}$ outperforms all other methods including the baseline ÉCOLE with overall precision of 68.7% and overall recall of 49.6%. It effectively balances precision and recall and obtains the top F1 score in all categories. It provides substantial improvements in F1 scores with 42.6%, 50.5%, and 46.8% increases over the next best method in deletion, duplication and overall F1 scores, respectively.

Somatic CNV calling performance. ÉCOLE is a germline CNV caller by design as it is trained with normal tissue samples. Similar to the difference between the automated WGS-based calls and the human expert calls, germline CNV calls and the somatic CNV calls have different distributions. This is due to the difference between the WES read depth signal of tumor and control samples. For this reason, specific callers or specific modes of callers are designed for somatic CNV calling which often require paired control and tumor samples to account for the difference which increases the computing and sequencing cost.

Here, using the same fine-tuning strategy, we update the parameters of the ÉCOLE model with cancer samples from [17] (SRA: SRP017787). This study reports matched WES and WGS samples of 16 bladder cancer samples and RT-qPCR validated CNVs in 4 regions. These events coincide with the following genes and affect the corresponding samples: A deletion in *CDKN2A/B* (samples B63, B112 and B80-0), a duplication in *CCDN1* (samples B37 and B103), a duplication in *DHFR* (samples B15, B18, B19, B24, B34 and B50) and a duplication in *ERBB2* (samples B9, B23, B80, B80-5, and B86) genes.

We fine-tune ÉCOLE to ÉCOLE$^{\text{FT-SOMATIC}}$ using (i) the CNVnator semi-ground truth labels obtained on the WGS data of samples B112, B24, B80 and (ii) the corresponding WES read depth signal obtained on the matched WES data of samples B112, B24, B80. We use the remaining 13 cancer samples to test if we can detect the RT-qPCR validated CNVs for each sample. We compare ÉCOLE$^{\text{FT-SOMATIC}}$ with XHMM which consistently obtains the highest precision, its polished call set DECoNT-XHMM and ÉCOLE.

As shown in Table 2, XHMM is able to detect the validated deletion event in the *CDKN2A/B* gene for one sample (B112) and does not return any calls for the remaining 10 samples. The polished version of XHMM's call set verifies these calls. ÉCOLE does not make any calls for any of the samples in the validated regions. On the other hand, ÉCOLE$^{\text{FT-SOMATIC}}$ is able to detect all of the 13 validated CNVs in the corresponding 13 test samples (all samples except the samples used in the fine-tuning). This shows that the model is flexible and can be easily configured to make somatic calls even without the need for a control sample.

We also computed the genome-wide precision, recall and F1 score performances with respect to the semi-ground truth labels obtained on the matched WGS data of the 13 test samples obtained using CNVnator. Please refer to Supplementary Table 24 for the corresponding confusion matrices. We find that ÉCOLE has both lower precision and lower recall than others. Table 3 and Supplementary Figure 8 show that ÉCOLE$^{\text{FT-SOMATIC}}$ outperforms others and provides an F1 score improvement of 25.2% over the next best method which shows that

fine tuning improves the performance (See Supplementary Table 24 for the corresponding confusion matrix). ÉCOLE$^{\text{FT-SOMATIC}}$ trades some precision of ÉCOLE for a large gain in recall. We wanted to make sure that fine tuning does not act as a simple threshold which is relaxed so that ÉCOLE$^{\text{FT-SOMATIC}}$ makes more calls than ÉCOLE to achieve higher recall. For this, we relaxed the call threshold of ÉCOLE to make it more liberal (i.e., it makes a call even if the probability is less than 0.33). Despite the increase in recall in this case, ÉCOLE was not able to a make call for any of the validated regions. This shows us fine tuning effectively teaches the algorithm about making calls in somatic samples and does not serve as a simple filtering mechanism.

Table 2: CNV calls for the RT-qPCR validated regions of 16 bladder cancer samples from Guo et al. [17]

Gene	Chromosome	Region Start	Region End	Call	Sample Name	XHMM	DECoNT-XHMM	ÉCOLE	ÉCOLE$^{\text{FT-SOMATIC}}$
CDKN2A/B	9	20,3m	24,1m	DEL	B63_Cancer	No	No	No	**Yes**
					B112_Cancer	**Yes**	**Yes**	No	N/A
					B80-0_Cancer	No	No	No	**Yes**
CCDN1	11	69.8m	69.8m	DUP	B37_Cancer	No	No	No	**Yes**
					B103_Cancer	No	No	No	**Yes**
DHFR	5	79.9m	80m	DUP	B15_Cancer	No	No	No	**Yes**
					B18_Cancer	No	No	No	**Yes**
					B19_Cancer	No	No	No	**Yes**
					B24_Cancer	No	No	No	N/A
					B34_Cancer	No	No	No	**Yes**
					B50_Cancer	No	No	No	**Yes**
ERBB2	17	35m	35.2m	DUP	B9_Cancer	No	No	No	**Yes**
					B23_Cancer	No	No	No	**Yes**
					B80_Cancer	No	No	No	N/A
					B80-5_Cancer	No	No	No	**Yes**
					B86_Cancer	No	No	No	**Yes**

Table lists the genes, regions, validated calls and the predictions of each method. Note that ÉCOLE$^{\text{FT-SOMATIC}}$ is fine-tuned on samples B112, B24, and B80. The calls of ÉCOLE$^{\text{FT-SOMATIC}}$ for these samples are denoted as N/A as they are used during training.

Table 3: Somatic CNV calling performance comparison on 13 bladder cancer test samples from Guo et al. [17]

Tool	DEL Precision	DUP Precision	Overall Precision	DEL Recall	DUP Recall	Overall Recall	DEL F1 Score	DUP F1 Score	Overall F1 Score
XHMM	0.235	0.962	0.698	0.012	0.028	0.020	0.023	0.054	0.038
DECoNT-XHMM	0.193	0.950	0.572	0.010	0.023	0.017	0.019	0.045	0.033
ÉCOLE	0.373	0.673	0.523	0.019	0.010	0.015	0.036	0.020	0.029
ÉCOLE$^{\text{FT-SOMATIC}}$	0.243	0.423	0.333	0.147	0.372	0.260	0.183	0.395	0.292

CNVnator calls are used as the semi-ground truth to calculate the metrics.

2.3 CNV calling performance on merged CNV segments

Evaluation Criteria. WES-based CNV callers often make calls for exons or bins which sometimes exceed exon bounds and then use a segmentation method to merge the subsequent calls into a larger call region. On the other hand, the ground truth calls on the WGS data are often shorter. A merged call on the exome can span multiple WGS-based calls. To assign a WGS-based semi-ground truth label to the WES-based call, the covered calls made on the WGS data are merged and a consensus label is assigned [31]. Supplementary Figure 6 exhibits this procedure visually for further reference. This procedure comes with the following problems: First, it reduces the resolution in the ground truth due to smoothing. Second, this results in the ground truth to change with respect to the break points of calls made by each WES-based caller. This makes it impossible to form a global ground truth call set to calculate recall. It was not a problem earlier in the literature as methods were mostly focused on the precision. Here, we compare the precision of ÉCOLE with others when we merge the exon-level calls to obtain larger call segments that also cover noncoding regions. Note that, ÉCOLE works at a base-pair resolution and makes a call per exon. Here, we merge subsequent exons with the same call to obtain a merged CNV segment to compare with other algorithms which often rely on a segmentation step and compare the precision performance.

Supplementary Table 4 and Supplementary Figure 1 shows the precision of each algorithm for the samples in the 1000 Genomes Dataset test split. We use the merged CNV segments as predictions for all algorithms and use merged the semi-ground truth labels obtained on the WGS data for the same samples. We can observe

that ÉCOLE is able to perform comparably to the top performing tool (DECoNT-XHMM) with 1% overall precision improvement, while providing 3.3% duplication precision improvement. This is still important as ÉCOLE achieves this precision quality while maintaining over 18% improvement in average recall metric. Evidently, ÉCOLE is able to make calls on a greater scale (merged CNV segments) just as it is able to perform on high resolution (i.e, exon-level).

The comparison of precision performances with CNLearn is provided in Supplementary Table 5. We observe that ÉCOLE has better precision than CNLearn. It provides a 49.3% improvement on average precision while providing a substantial average recall improvement as discussed before.

Supplementary Table 6 and Supplementary Figure 2 shows the precision performances of every method when using the human expert-curated labels as the ground truth [7]. See Supplementary Table 19 for the corresponding confusion matrices. We obtain a 14.3% average precision improvement over the next best method, CONIFER. While CONIFER achieves perfect precision in DUP category, it has zero precision in DEL category and it only makes a handful of calls. The actual second best performing method with an acceptable number of calls is polished CODEX2 which is 30% behind ÉCOLE$^{\text{FT-EXPERT}}$.

Supplementary Table 7 and Supplementary Figure 4 shows the performances of the tools on the NA12878 sample which was sequenced on various platforms. ÉCOLE is able to maintain its preeminence over all performance metrics when merged CNV segments considered. We observe that our model is providing at least $\sim 28\%$ average precision improvement over the second best performing method in all the sequencing platforms considered.

We conclude that ÉCOLE improves the state-of-the-art CNV calling precision even when outputting merged CNV calls instead of exon-level calls. Note that this is a disadvantageous benchmark setting for our approach as our approach works on a base-pair resolution and merge process decreases the resolution of our calls.

2.4 Ablation study and the need for a complex model like ÉCOLE

We use a linear SVM classifier as a baseline method to show the need for a deep learning method to call CNVs on WES data. The SVM model is trained using the same training set used by ÉCOLE and we use the default parameters in the scikit-learn implementation. We input the same read depth signal per exon and use the exon

level semi-ground truth CNVnator labels for performance comparison on the 1000 Genomes Dataset test set as done in Section 2.2.

The precision and recall performances are shown in Supplementary Table 8. We observe that while SVM achieves 8.5% recall improvement over tool ÉCOLE, it only is able to yield close to 3% precision. ÉCOLE has 50% recall and 77% precision which corresponds to a 55.1% F1-score improvement over the baseline model. This shows the need for a complex model like ÉCOLE to learn an attention based embedding on the read depth signal to be able to accurately classify exons as deleted or duplicated.

ÉCOLE uses a rather customized model of the Transformer architecture which was first introduced in [12] and proved its success in various domains from NLP to computer vision. Our transformer encoder uses (i) a chromosome specific classification token, instead of a fixed classification token, and (ii) an exon-specific positional encoding instead of a fixed positional encoding to learn which parts of the signal is important for a CNV call and in which context (i.e., chromosome). To show the need for these context specific techniques, we train a standard transformer architecture which does not have the aforementioned customized methods and which is otherwise identical to our model. That is, we use a standard 3-layered transformer model with fixed positional encoding. The model generates an output vector, which is the standard classification token (i.e, not chromosome specific) of size $\in \mathbb{R}^{192}$. We start the training with the learning rate of $5 \cdot 10^{-5}$ and use cosine annealing scheduler. We used Adam optimizer during the training of the base model, and the model converged after 6 epochs.

The Supplementary Table 9 shows the precision and recall performances of the baseline transformer model and ÉCOLE. We observe that ÉCOLE is able to outperform the baseline by a large margin, providing 30% average precision and 49% average recall improvements. The chromosome specific classification token provides a good prior for the model to learn the relevance of each read depth value in the context of their respective chromosome. Moreover, exon-specific positional encoding renders the model to differentiate the absolute position of the exon samples along the exome. Hence it gives the model the capacity to learn the context of the read-depth values along with the chromosome-specific classification tokens. As the read-depth samples can have variant distributions depending on the context (i.e, absolute position and the chromosome), the model is able to learn context-dependent sample distributions.

2.5 Interpretability of the CNV calls

Transformer-based neural networks are inherently interpretable as they incorporate an attention mechanism. The attention component of the network learns which parts of the read depth signal has to be focused on by the model to make the decision, similar to humans selectively focusing on certain parts of an image to recognize. However, it is not straightforward to visualize the parts of the read depth signal focused by ÉCOLE since the model uses a multi-head attention mechanism which means multiple attentions are calculated over the signal which are then concatenated and transformed (linear) into same dimensions as input (192 x 1001). Therefore, there is an implicitly learned complex relationship between these attention maps that the model uses to get the final decision. As Voita et al. demonstrate that every attention head carry different importance for the final classification and a simple average over the multiple heads cause noisy relevance maps for visualization [38].

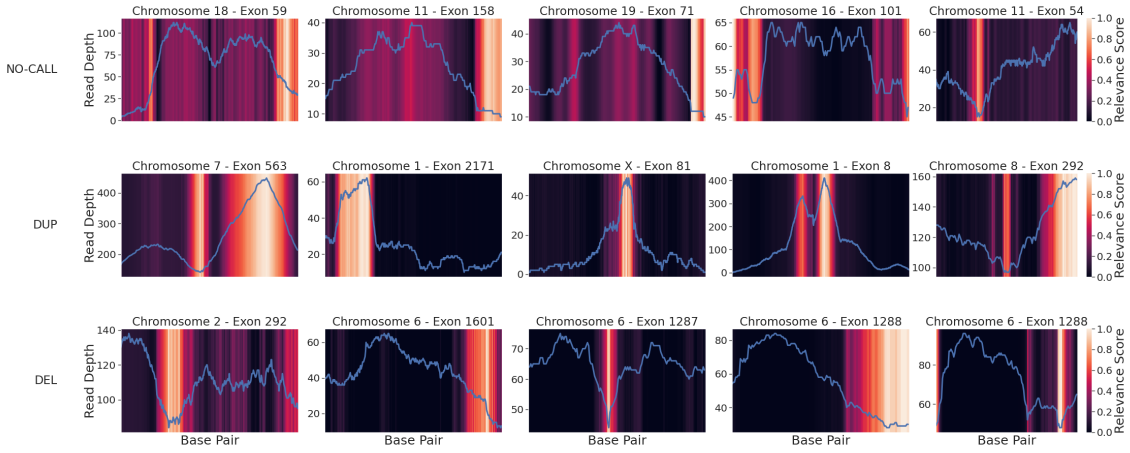

Figure. 4: The figure shows the read depth signal over 15 exons with 5 NO-CALLs (row 1), 5 DUP calls (row 2) and 5 DEL calls (row 3). Heatmaps in the background denote the relevance score of the corresponding part of the signal assigned by the model. The brighter the color the higher the attention devoted to that region. For each panel, x-axis denotes the index of the base pair, left y-axis denotes the read depth value, and right y-axis denotes the relevance score (attention).

We use Generic Attention-model Explainability method proposed by Chefer et al. to visualize the parts of the signal that are deemed important for making the CNV calls [9]. Figure 4 shows the read depth signal observed over 15 exons. The background heatmap indicates which parts of the signal are attended by the model

where brighter color indicates more attention. ÉCOLE classifies the examples in the first row as NO-CALL, the second row as duplication, and the last row as deletion. For the duplication calls, the sharp shifts in read depth signals, mostly elevations, were focused by the model. Likewise, for deletion calls, we can observe that the model focused on locations that have sharp downfalls of read depth values. For both cases, rest of the signal receives almost no attention and is ignored by the model. For the exons with no calls, we observe that the model still focuses on the inclines and declines in the read depth signal, but other parts of the signal receive relatively more attention compared to the exons with calls. Since the model cannot detect a concrete pattern and is not confident enough, it opts for a no call.

2.6 Insights from ÉCOLE's CNV calls

First, we focus on ÉCOLE's calls made on pseudoautosomal regions of Chromosome X - PAR1 and PAR2 which are diploid regions and are usually problematic for CNV callers. We compare the performance with XHMM. Polished XHMM callset has a precision of 0.37 and 0.5 on these regions, respectively. On the other hand, ÉCOLE achieves precision of 73.6% and 73.8%, respectively. On the X chromosome as a whole, ÉCOLE has an exon-wise precision of 65% where polished XHMM has a precision of 16%. These results show that ÉCOLE performs well in this challenging setting.

We then checked whether the performance of ÉCOLE varies across chromosomes and with exon length. Figure 5 shows the chromosome-wise stratification of the calls where each dot represents a call made by ÉCOLE, colored by one of the possibilities: (i) ÉCOLE calls an event and it is correct - True Positive (i.e., matching WGS-based call, the semi ground-truth); (ii) ÉCOLE calls an event and it is incorrect False Positive - (i.e., not matching WGS-based call), and (iii) ÉCOLE does not make a call and it is incorrect - False Negative (i.e., not matching WGS-based call), and (iii) . We observe that method's performance is better in longer exons whereas most of the mistakes are made on shorter exons, which are less than 500 bps. This is expected as the read length in these regions are shorter and is more prone to noise as the method is input with less information. We also observe that the success of the method varies across chromosomes. The method performs well in chromosomes 14, 21 and Y with accuracy reaching up to 80%. On the other hand, the performance is lower on chromosomes 9, 10 and 13, where the accuracy is below 10%. Except chromosome 9, these are chromosomes with short exons

and relatively low number of calls which might explain why model had difficulty in learning the true distribution of the calls.

3 Discussion

Copy number variants have a large spectrum of phenotypic affects from just playing a role in genetic diversity to underlying complex genetic disorders by affecting roughly 10% of the genome [42]. Accurate CNV calling on WES data for use in clinic has been a long sought after goal due to cost, size and time advantages compared to WGS. Indeed with its high diagnostic yield, WES has been a mainstream tool in routine practice in genomic medicine [8]. Yet, WES-based CNV callers have suffered from low precision and concordance [41,35]. As we have recently shown, it has been possible to transfer the satisfactory CNV calling performance of WGS-based CNV callers to the WES-based callers, using a deep learning-based polishing approach [31]. Polishing selectively prunes out false positive CNV calls and substantially improves precision. However, by design, a polisher cannot make new calls as it is dependent on the calls of the base caller. While it is possible to change a false positive to a true positive call, it is rare and it is not possible to change a false negative call (i.e., no call) to a true positive call. This hinders improving the recall. Here, we show that it is possible to use the deep learning techniques to process the read depth signal and train a stand-alone WES-based CNV caller which is able to achieve WGS-level precision and recall performance at the same time.

We use WGS-based calls as labels to train our model obtained using CNVnator. These must be regarded as a semi-ground truth rather than an absolute ground truth data as CNVnator reports 86%-96% recall and 3%-20% false discovery rate. The ideal case is using a human expert curated set of calls to train ÉCOLE. However, such a call set is only available for 9 samples from the 1000 Genomes Dataset [7]. Unfortunately, it is orders of magnitude smaller compared to the CNVnator callset and it is not possible to be able train a complex model like ÉCOLE. As human expert decision making does not resemble the decision making of automated tools, the overall precision in predicting human expert calls even after polishing was limited at 35% [31]. Here, we show that it is possible to use a pretrained ÉCOLE model and further train it using this limited set of human calls. This is called fine-tuning in machine learning literature. That is, we take the model trained with large-but-not-fully-confident WGS-based calls and then continue training with small-but-confident human expert calls. We show that fine-tuned ÉCOLE (ÉCOLE$^{\text{FT-EXPERT}}$) is the first method to achieve human-like

performance. Similarly, germline CNV calling and somatic CNV calling differ due to the difference in typical read depth signatures between a control and a tumor sample. We use the fine-tuning strategy to convert ÉCOLE, which is a germline CNV caller, into a somatic CNV caller using matched WES and WGS tumor samples. ÉCOLE$^{\text{FT-SOMATIC}}$ is specific to bladder cancer as the samples we used were as such. However, the storage, computational and time cost of configuring ÉCOLE into any cancer type of interest is very low as the model requires only a few samples and only a few epochs for the model update. We think with human expert level performance and the ability to perform accurate somatic CNV calling, ÉCOLE is the first candidate CNV caller to use in clinic.

We observe that short exons are more difficult for ÉCOLE to generalize as well as some chromosomes with small number of examples. While ÉCOLE is released as an organism or condition agnostic tool for broad use, it is possible to incorporate prior condition-specific knowledge into the model to make it work in a more optimized fashion for such regions or conditions. For instance one can group certain nearby exons to have a longer and more informative read depth signal or one could let the chromosome specific tokens to be shared across some chromosomes to increase the performance in relatively low-performing chromosomes.

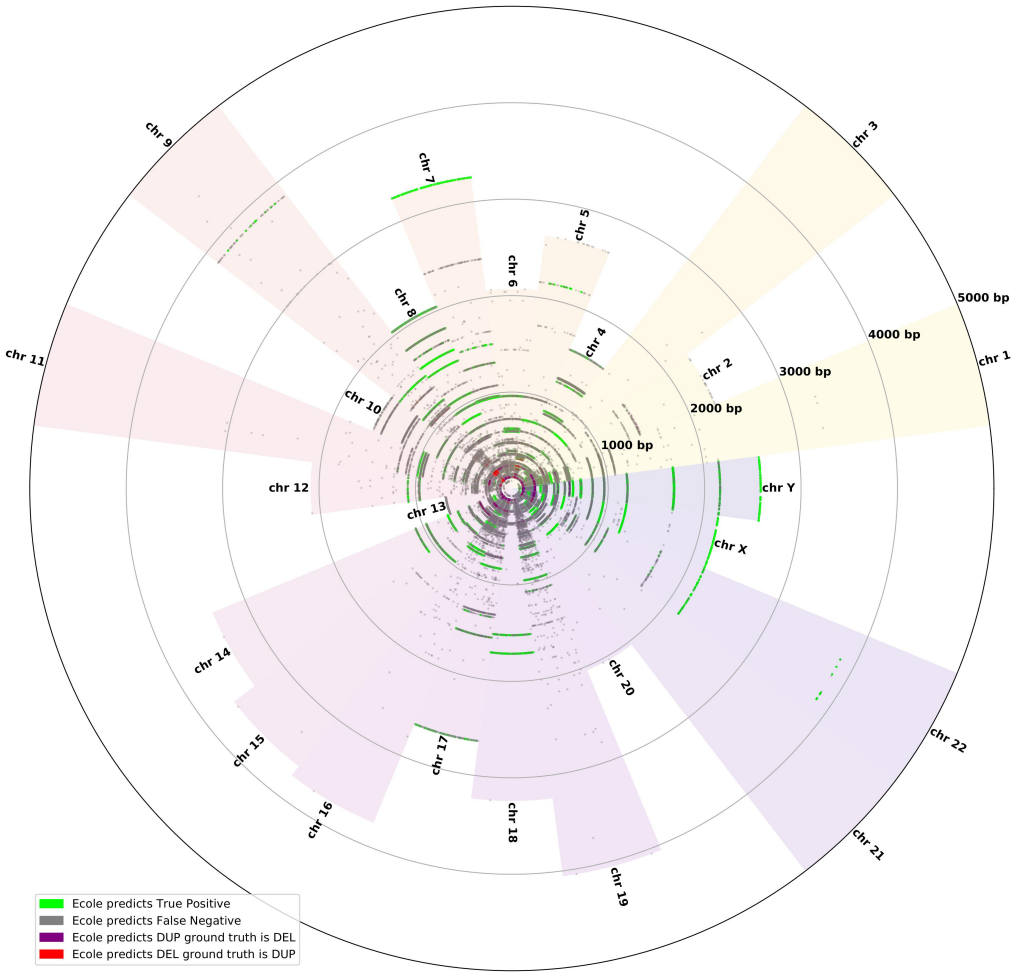

Figure. 5: The figure shows the CNV predictions of ÉCOLE on the 1000 Genomes Project test set. The plot is divided into equal partitions for each chromosome. For each chromosome, there is a scatter plot denoting the CNVs (exons) predicted by of ÉCOLE for that chromosome, the CNV calls are scattered along the radius of the circle. The radius denotes the length (in base pairs) of the exon for the respective CNV call. The bars denote the maximum length of exon within the respective chromosome.

4 Methods

4.1 Data Sets

Training and test sets from the samples in the 1000 Genomes Project. We use 707 samples from the the 1000 Genomes Project [6] to train and test ÉCOLE. This corresponds to samples HG00096 to HG02356 in alphabetical order. We use both the WES and the WGS data for each sample. The WES data were sequenced with Illumina Genome Analyzer II and Illumina HiSeq 2000 while the WGS data were generated using NovaSeq [6] and for WES data NimbleGen SeqCap v3 capture kit was used. Average read depth is 50× for WES and 30× for WGS data with average read lengths of 76 bps and 100 bps, respectively. BWA-MEM is used for alignment on GRCh38 [29]. We use the CNVnator [3] tool to call CNVs on the WGS data of each sample to obtain the semi-ground truth labels. The training and test sets consist of 550 and 157 samples, respectively. We use the training set to train and obtain the final ÉCOLE model where the WES read depth is used as the input and the WGS-based CNVnator labels are used as the semi-ground truth. The test set is used to evaluate the performance as shown in Figure 2 and Table 1. Supplementary Table 25 lists the corresponding names of the samples.

NA12878 samples for generalizability tests. We use the calls made for the NA12878 sample to test the generalizability of ÉCOLE to various sequencing platforms. This sample has WES data provided by the following platforms: BGISEQ 500, Illumina HiSeq 4000, MGISEQ 2000, and NovaSeq 6000. We use this sample only for testing and its data is not included in the training set by any means. Again, we use the CNVnator calls on the WGS sample of NA12878 to obtain the semi-ground truth labels per exon for the evaluation.

Fine tuning and test sets from the samples in Chaisson et al. call set. Chaisson et al. [7] provide human expert-validated consensus calls of 15 CNV WGS callers on 9 samples from the 1000 Genomes project. We obtained the calls made for the 8 samples, for which there is also matching WES data in the 1000 Genomes dataset, namely: HG00512, HG00513, HG00731, HG00732, HG00733, NA19238, NA19239, NA19240. The calls made by Chaisson et al. on the WGS data were used as the golden standard ground truth for all compared algorithms and ÉCOLE. We used the ground truth calls made for 4 samples (NA19238, NA19239, HG00731, HG00512) to fine tune the parameters of ÉCOLE when applying transfer learning. We used the remaining 4

samples (HG00513, HG00732, HG00733, NA19240) for the test (inference) and comparison with other tools. See Supplementary Table 25 for the corresponding names of the samples.

Fine tuning and test sets from the samples in Guo et al. bladder cancer call set. Guo et al. report matched WES and WGS samples of 16 bladder cancer patients (accession number: SRP017787). We fine tune the ÉCOLE model with 3 cancer samples (samples B112, B24, B80) from this data set [17]. We use the semi-ground truth labels obtained on the matched WGS samples for these 3 patients for fine tuning.

We use the remaining 13 cancer samples for testing in two ways. First, we check if tools make calls in the RT-qPCR-validated regions in these samples. Then, we use CNVnator to obtain the semi-ground truth labels on the matched WGS samples for these 13 patients to compute precision, recall and F1 scores. See Supplementary Table 25 for the respective names of the samples.

4.2 Experimental Setup

Compared Methods We compared ÉCOLE with the following state-of-the-art WES-based germline CNV callers: XHMM, CODEX2, CONIFER [15,20,27]. These report categorical CNV predictions like ÉCOLE (i.e., deletion, duplication or no-call). CNV-kit and Control-FREEC [34,5] are also WES-based germline CNV callers but they report exact (i.e., integer) CNVs. To be able to also compare with these two tools, we discretize their predictions. That is, if the predicted copy number is larger than 2, it is classified as duplication; if it is less than 2 it is classified as deletion and no-call if it is equal to 2. We polished the calls made by all of the aforementioned tools using DECoNT as described in [31] and used the DECoNT models released on GitHub (accessed in Nov 2021). Polished call sets of these methods are called DECoNT-*toolname* (e.g., DECoNT-XHMM). We also compared ÉCOLE with CNLearn which learns to aggregate the calls of other WES-based germline CNV callers (CANOES, XHMM, CONIFER and CLAMMS). Through personal communication, we obtained the calls of this tool on 28 samples in our test set.

Parameter Settings For all samples we align, WES reads to the reference genome (GRCh38) using BWA with -mem option and default parameters. We calculate the read depth using the Sambamba tool [36] with the base -L option to align the reads in the exon regions. We ran the compared methods using their recommended settings. For XHMM, following parameter values were used: $Pr(\text{start DEL}) = Pr(\text{start DUP}) = 1e-08$; mean

number of targets is 6; mean distance between targets is 70kb, and DEL, DIP, DUP read depth distributions were modeled as $\sim \mathcal{N}(-3,1)$, $\sim \mathcal{N}(0,1)$ and $\sim \mathcal{N}(3,1)$, respectively. For CODEX2, minimum read coverage was set to 20. CoNIFER performs SVD on the data to remove top n singular vectors. We set n to 6. For Control-FREEC and CNV-kit, we set all parameters to default values.

Training ÉCOLE We trained our model using the WES data as the training set of 550 samples from the 1000 Genomes data set. We used the Adam optimizer [24] and the model converged in 4 epochs. We used Xavier weight initialization [16]. We started training with a learning rate of $5 \cdot 10^{-5}$ and used cosine annealing learning rate schedule.

To obtain the final ÉCOLE$^{\text{FT-EXPERT}}$ model, we further fine-tuned the ÉCOLE model with golden standard ground truth calls on 4 samples obtained from Chaisson et al. as explained in the Data Sets section. We again used Adam optimizer and cosine annealing schedule with an initial earning rate of $5 \cdot 10^{-5}$. The model converged in 11 epochs.

To obtain the final ÉCOLE$^{\text{FT-SOMATIC}}$ model, we further fine-tuned the ÉCOLE model with the semi-ground truth calls made on 3 cancer samples obtained from Guo et al. as explained in the Data Sets section. We have used Adam optimizer and cosine annealing learning rate scheduler with a base learning rate of $5 \cdot 10^{-5}$, fine-tuning lasted for 11 epochs.

All models are trained on a SuperMicro SuperServer 4029GP-TRT with 2 Intel Xeon Gold 6140 Processors (2.3GHz, 24.75M cache) and 256GB RAM. We used a single NVIDIA GeForce RTX 2080 Ti GPU (24GB, 384Bit) for training. The initial model took approximately 15 days to converge and each fine-tuning took approximately 4 hours. Note that users do not need to train a model from scratch and can use the released ÉCOLE model for inference which is rapid. The average time to call all CNVs per exome is ~ 5 mins.

Performance Metrics ÉCOLE assigns a pseudo probability scores for each call to be deletion, duplication or no-call where the event with the largest score is the final prediction. We measured the performance of all

compared methods and ÉCOLE using precision and recall which are defined as follows:

$$\text{Duplication precision } (PRE_{dup}) = \frac{TP_{dup}}{TP_{dup} + FP_{dup}} \tag{1}$$

$$\text{Deletion precision } (PRE_{del}) = \frac{TP_{del}}{TP_{del} + FP_{del}} \tag{2}$$

$$\text{Overall precision} = \frac{PRE_{dup} + PRE_{del}}{2} \tag{3}$$

$$\text{Duplication recall } (REC_{dup}) = \frac{TP_{dup}}{T_{dup}} \tag{4}$$

$$\text{Deletion recall } (REC_{del}) = \frac{TP_{del}}{T_{del}} \tag{5}$$

$$\text{Overall recall} = \frac{REC_{dup} + REC_{del}}{2} \tag{6}$$

$$\tag{7}$$

where TP_{dup} := the number of duplication calls that are correctly called; TP_{del} := the number of deletion calls that are correctly called; FP_{dup} := the number of duplication calls that are incorrectly called; FP_{del} := the number of deletion calls that are incorrectly called; T_{dup} := the number of ground truth duplication calls; T_{del} := the number of ground truth deletion calls.

4.3 ÉCOLE Architecture

Problem Formulation Let X be the set of all exons with available read depth signal and X^i indicate the i^{th} exon where $i \in \{1, 2, ..., N\}$ and $N = |X|$. Every X^i is associated with the following features: X^i_{chr}, X^i_{start}, X^i_{end} and X^i_{RDSeq}. X^i_{chr} is the chromosome of the exon where $chr \in \{1, 2, 3, ..., 24\}$. 23 and 24 represent chromosomes X and Y, respectively. X^i_{start}, X^i_{end} are the start and end coordinates of the exonic region. X^i_{RDSeq} is a standardized vector of read depth values at a basepair resolution. Standardization is performed for every read depth value using the global mean and standard deviation of read depth values in the training data. Every X^i_{RDSeq} is -1 padded from left to have the maximum length of 1000. For exons longer than 1000 bps, they are considered if the non-zero read-depth values in that exon are of length ¡ 1000. Y^i represents the corresponding ground truth label for exon i, either obtained from CNVnator or from Chaisson et al. depending on the application. Let $\hat{Y^i} = f(X^i, \theta)$ be the CNV prediction (i.e deletion, duplication, no-call) using the model f (a multi-class classifier) which is parameterized by θ. The goal is to find the model parameters θ that minimizes the difference between predicted exon-level CNV labels and their ground truth labels.

Model Description We illustrate the overview of the model in Figure 1. Each exon i is associated with the vector $X^i_{RDSeq} \in \mathbb{R}^{1000\times 1}$ which represents the read depth signal in that region.

First, ÉCOLE maps each read depth value j of the read depth vector for the i^{th} exon ($X^i_{RDSeq}[j]$) into a higher dimension $H = 192$ by using a fully connected neural network (see Eq. 8).

$$FFN(X^i_{RDSeq}[j]) = (X^i_{RDSeq}[j] \cdot W + b^T)^T, \qquad W \in \mathbb{R}^{192\times 1}, b \in \mathbb{R}^{192}, j \in [1, 1000] \tag{8}$$

We refer to the transformed form of the full vector X^i_{RDSeq} as the input embedding and denote it with $\mathbf{X}^i_{embed} \in \mathbb{R}^{192\times 1000}$ (see Eq. 9).

$$\mathbf{X}^i_{embed} = FFN(X^i_{RDSeq}[1])...FFN(X^i_{RDSeq}[1000]) \tag{9}$$

ÉCOLE employs two techniques to learn the context in which the read depth signal indicates a copy number variation. First, it learns a *Chromosome Specific Classification Token* matrix $C \in \mathbb{R}^{192\times 24}$ where each chromosome k is represented with a column vector c^k of size 192. That is, $c^k = C[:, k]$ where $c^k \in \mathbb{R}^{192}$. The vector for the chromosome in which exon i resides ($c^{X^i_{chr}}$) is concatenated with $\mathbf{X}^i_{embed}$ to obtain $\hat{\mathbf{X}}^i_{embed} \in \mathbb{R}^{192\times 1001}$ (see Eq. 10). This joint matrix lets the model to learn the meaning of the read depth vector in the context of different chromosomes to be able to distinguish chromosome specific read depth patterns and model the variance across chromosomes.

$$\hat{\mathbf{X}}^i_{embed} = \mathbf{X}^i_{embed} c^{X^i_{chr}} \tag{10}$$

The second technique is using a positional encoding which enables the model to learn the relative locations of the read depth values with respect to each other and absolute position in the entire exome sequence and extract the position meaning that contributes to calling CNVs. In this work, for an exon i, we create a location vector v of length 1001. We use sine and cosine functions of different radial frequencies similar to version in [37] to create the positional embedding matrix $\mathbf{E}^i_{pos} \in \mathbb{R}^{192\times 1001}$ as done in Eqs. 11 and 12).

$$\mathbf{E}^i_{pos}[loc, j] = \sin(loc/10^{9*2j/H}), \qquad loc \in \{1, ..., 1001\}, j \in \{1, ..., H/2\} \tag{11}$$

$$\mathbf{E}^i_{pos}[loc, 2j+1] = \cos(loc/10^{9*2j/H}), \qquad loc \in \{1, ..., 1001\}, j \in \{1, ..., H/2\} \tag{12}$$

To serve as an intuition, we could assume that positional encoding is a clock, then loc and j are hour and minute hands, respectively. Moving along the loc (i.e over read depth embedding) and j (i.e between 1 and latent dimension H) values is basically rotating the hour and minute hands with varying frequencies. The constant 10^9 allows the encoding to uniquely map the start and end coordinates,X^i_{start}, X^i_{end}, which have a range of $[14.6 \cdot 10^3, 290 \cdot 10^6)$. This encoding enables the model to get positional and deal with the inherent noise in varying read depth values. This matrix is summed with $\hat{\mathbf{X}}^i_{embed}$ to obtain the input to the transformer $\mathbf{O}^i_0 \in \mathbb{R}^{192\times1001}$ (see Eq. 14).

$$\mathbf{O}^i_0 = \hat{\mathbf{X}}^i_{embed} + \mathbf{E}^i_{pos} \tag{13}$$

ÉCOLE uses an efficient variant of the Transformer architecture [11] (only the encoder part). The encoder consists of a sequence of a parallel attention block (Multi-head attention, $MHA^{(h)}$) followed by a multi-layered perceptron (MLP) block. The multi-head attention mechanism lets the model to learn the pertinence of read depth values in a chromosome in relation to deletion and duplication events (see Figure 1 and [37] for MHA details). That is, it learns which parts of the signal it needs to focus on. An MHA block uses 8 parallel attention layers (i.e heads). Firstly, the inputs for MHA are layer normalized (LN) and are propagated through the MHA block. Later, the outputs of these blocks are summed with the input of the respective LN block. The summed output is again layer normalized then passed through an MLP. The outputs of the MLP block are summed with the input of the respective LN block to produce $\mathbf{O}^i_1 \in \mathbb{R}^{192\times1001}$ (See Eqs. 14 and with the input of the respective LN block 15). This procedure is repeated L times where $L = 3$ in our application.

$$\mathbf{O}^{i^{'}}_\ell = MHA(LN(\mathbf{O}^i_{\ell-1})) + \mathbf{O}^i_{\ell-1}, \qquad \ell = 1, ..., L \tag{14}$$

$$\mathbf{O}_\ell = MLP(LN(\mathbf{O}^{i^{'}}_\ell)) + \mathbf{O}^{i^{'}}_\ell, \qquad \ell = 1, ..., L \tag{15}$$

ÉCOLE passes the column vector corresponding to the chromosome of exon i ($\mathbf{O}_3[:, X^i_{chr}] \in \mathbb{R}^{192\times1}$) through a MLP to obtain probabilities for deletion, duplication and no-call events and maximum among these is returned as the prediciton for that exon i (see Eq. 16).

$$\hat{Y}^i = \arg\max(Softmax(MLP(O^i_3[:, X^i_{chr}]))) \qquad \hat{Y}^i \in \{DEL, DUP, NOCALL\} \tag{16}$$

4.4 Processing exons with no read depth available

We developed ÉCOLE to perform CNV calling on exon target regions using read depth information, however it is important to note that about 20% of the exon targets do not contain read depth sequences on average per sample. In order to perform CNV calling to these regions, we have applied majority voting on the predictions of ÉCOLE based on the 3 nearest neighbor exon targets. Supplementary Figure 7 demonstrates this procedure visually for further reference.

4.5 Interpretability of the ÉCOLE

In order to explain the predictions of our model, we used Generic Attention-model Explainability method [9]. This method is class-dependent, uses label information, and it is a saliency method that highlights the relevant parts of the input for the classification that predicts the respective label. In addition, we use the Attention matrices $\mathbf{A}$ of the last Transformer blocks and obtain the specified relevance scores. The derivations for the relevancy scores for the efficient variant of Transformer block, i.e Performer [11] can be seen in the Supplementary Note 3.1.

Data Availability

Features: The input features of the models is the WES read depth data which is generated using the Sambamba tool (default options).

The 1000 Genome Project sample names we used to train and test the models are provided in the Methods section which are available at the 1000 Genomes Project. WES and WGS samples are available at the following link: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/`. Guo et al. samples are available at Sequence Reads Archive (SRP017787)

Labels: The labels we use for training, fine tuning and testing are available at the following link: `https://zenodo.org/record/7317266#.Y3F0jS8w1hE`.

Software Availability

Software: ÉCOLE is implemented and released at `https://github.com/ciceklab/ECOLE` under CC-BY-NC-ND 4.0 International license. The scripts used to generate the data for all figures and tables in the manuscript and the source code are provided at `https://zenodo.org/record/7317266#.Y3F0jS8w1hE`.

Competing Interests

Authors declare no competing interests.

Acknowledgments

We would like to thank Vijay Kumar and Dr. Santhosh Girirajan for their help with obtaining results on CNLearn. We also thank Dr. Girirajan for the feedback on our manuscript. AEC acknowledges the funding from TUBA GEBIP, Bilim Akademisi - BAGEP and TUSEB Aziz Sancar Research Incentive awards.

AEC and CA supervised the study. BM and FO designed the model. BM and FO implemented the software and performed the experiments. AEC, CA, BM and FO wrote the manuscript.

References

1. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature **455**(7210), 237–241 (2008). https://doi.org/10.1038/nature07239
2. Abyzov, A., Urban, A., 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**, 974–84 (02 2011). https://doi.org/10.1101/gr.114876.110
3. 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 (2011)
4. 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), 1061–1067 (2009). https://doi.org/10.1038/ng.437
5. 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
6. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, Germer, S., Brand, H., Hall, I.M., Talkowski, M.E., Narzisi, G., Zody, M.C.: High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. Cell **185**(18), 3426–3440.e19 (Sep 2022)
7. 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)
8. Chassagne, A., Pélissier, A., Houdayer, F., Cretin, E., Gautier, E., Salvi, D., Kidri, S., Godard, A., Thauvin-Robinet, C., Masurel, A., et al.: Exome sequencing in clinical settings: preferences and experiences of parents of children with rare diseases (sequapre study). European Journal of Human Genetics **27**(5), 701–710 (2019)
9. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers (2021)
10. Chen, M.X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z., Wu, Y., Hughes, M.: The best of both worlds: Combining recent advances in neural machine translation (2018)

11. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., Weller, A.: Rethinking attention with performers (2021)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
14. Durand, C.M., Betancur, C., Boeckers, T.M., Bockmann, J., Chaste, P., Fauchereau, F., Nygren, G., Rastam, M., Gillberg, I.C., Anckarsäter, H., et al.: Mutations in the gene encoding the synaptic scaffolding protein shank3 are associated with autism spectrum disorders. Nature Genetics **39**(1), 25–27 (2006). https://doi.org/10.1038/ng1933
15. 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
16. 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. JMLR Workshop and Conference Proceedings (2010)
17. Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., et al.: Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nature genetics **45**(12), 1459–1463 (2013)
18. Ho, S.S., Urban, A.E., Mills, R.E.: Structural variation in the sequencing era. Nature Reviews Genetics pp. 1–19 (2019)
19. Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C.: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Research **19**(7), 1270–1278 (2009). https://doi.org/10.1101/gr.088633.108
20. 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
21. 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)

22. Karakoc, E., Alkan, C., O'Roak, B.J., Dennis, M.Y., Vives, L., Mark, K., Rieder, M.J., Nickerson, D.A., Eichler, E.E.: Detection of structural variants and indels within exome data. Nature Methods **9**(2), 176–178 (2011). https://doi.org/10.1038/nmeth.1810

23. 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)

24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

25. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., Wilson, R.K., et al.: Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research **22**(3), 568–576 (2012). https://doi.org/10.1101/gr.129684.111

26. Korbel, J.O., Urban, A.E., Grubert, F., Du, J., Royce, T.E., Starr, P., Zhong, G., Emanuel, B.S., Weissman, S.M., Snyder, M., et al.: Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proceedings of the National Academy of Sciences **104**(24), 10110–10115 (2007). https://doi.org/10.1073/pnas.0703834104

27. 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

28. 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)

29. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem (2013)

30. Luo, H., Zhang, S., Lei, M., Xie, L.: Simplified self-attention for transformer-based end-to-end speech recognition (2020)

31. Özden, F., Alkan, C., Çiçek, A.E.: Polishing copy number variant calls on exome sequencing data via deep learning. bioRxiv pp. 2020–05 (2021)

32. Shlien, A., Malkin, D.: Copy number variations and cancer susceptibility. Current Opinion in Oncology **22**(1), 55–63 (2010). https://doi.org/10.1097/cco.0b013e328333dca4

33. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., et al.: Large recurrent microdeletions associated with schizophrenia. Nature **455**(7210), 232–236 (2008). https://doi.org/10.1038/nature07229

34. Talevich, E., Shain, A.H., Botton, T., Bastian, B.C.: Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. PLoS computational biology **12**(4), e1004873 (2016)

35. 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)
36. 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
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
38. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned (2019)
39. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z.: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics **25**(21), 2865–2871 (2009). https://doi.org/10.1093/bioinformatics/btp394
40. Yoon, S., Xuan, Z., Makarov, V., Ye, K., Sebat, J.: Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research **19**(9), 1586–1592 (2009). https://doi.org/10.1101/gr.092981.109
41. 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)
42. Zarrei, M., MacDonald, J.R., Merico, D., Scherer, S.W.: A copy number variation map of the human genome. Nature reviews genetics **16**(3), 172–183 (2015)