Thesis for Master of Science

High-confidence identification of *de novo* structural variation using a *k*-mer-based approach

Hyun Woo Kim

Graduate School of Hanyang University

August 2024

Thesis for Master of Science

High-confidence identification of *de novo* structural variation using a *k*-mer-based approach

Thesis Supervisor: Jin-Wu Nam

A thesis submitted to the graduate school of Hanyang University in partial fulfillment of the requirements for the degree of Master of Science

Hyun Woo Kim

August 2024

Department of Life Science

Graduate School of Hanyang University

This thesis, written by Hyun Woo Kim,
has been approved as a thesis for the Master of Science.

August 2024

Committee chairman: Heon Seok Kim, Ph.D. (Signature)

Committee member: Jin-Wu Nam, Ph.D. (Signature)

Committee member: Kiwon Jang, Ph.D. (Signature)

Graduate School of Hanyang University

Table of Contents

List of Figures and Tables

Abstract

High-confidence identification of *de novo* structural variation using a *k*-mer-based approach

Hyun Woo Kim

Department of Life Science

Graduate School of

Hanyang University

High-throughput whole genome sequencing (WGS) has revolutionized the field of genomics by enabling comprehensive interrogation of an individual's genetic makeup. This recent advance in sequencing technologies showed an outstanding importance to identify genetic variants in the purpose of predicting disease prognosis and devising clinical strategies. Among the vast landscape of human genetic variants, structural variations (SVs) represent a significant component of genomic diversity and are involved in various human diseases. Among SVs present in an individual's genome, identifying *de novo* SVs (dnSVs), those of which refer to SVs present in an individual but absent in their parents, is crucial for understanding the genetic basis of disease susceptibility and developmental disorders. Despite the advance in sequencing technologies, accurately identifying dnSVs remains challenging due to the complexity of the human genome and technical limitations. Current SV detection methods adopt a strategy to move along the process of aligning the whole sequencing data to the reference genome, coupled with

complicated preprocessing steps to generate a quality-controlled data for variant identification. However, this strategy is substantially inefficient as it analyzes non-informative data given the fact that less than 1% of the WGS data contain variant supporting information. Analyzing all the non-informative data would be a burdensome process leading to a large waste of computational resource and research time. Especially, given the fact that dnSVs occur in an extremely low frequency, dnSV identification requires a large size of datasets. These circumstances lead to an inevitable demand of developing a method that can identify dnSVs fast and efficiently.

This thesis presents a novel strategy to efficiently identify dnSVs utilizing a *k*-mer-based filtering approach with variant allele frequency estimation to selectively discover germline dnSVs. To validate the effectiveness of this strategy, large-scale WGS datasets from healthy individuals, Korean atomic bomb survivors, and patients with rare diseases were used to demonstrate the capability of this approach to accurately pinpoint dnSVs with high sensitivity and specificity. This thesis would offer a robust and practical strategy that can minimize the time and effort to validate spurious dnSV calls. By elucidating problematic dnSVs fast and efficiently, this method would lay the groundwork for future studies aimed at developing targeted strategies for population-scale large cohort research, precision medicine and genomic diagnostics.

I. Introduction

High-throughput whole genome sequencing (WGS) has emerged as a transformative technology in genomics, facilitating comprehensive examination of an individual's entire genomic landscape with unprecedented accuracy and efficiency. This powerful tool has revolutionized our understanding of genetic variation, enabling the detection of a wide range of genomic alterations, including single nucleotide variations (SNVs), small insertions and deletions (INDELs), and structural variations (SVs) [1]. Among these, SVs, which refer to genomic alterations ranging from 50 base pair (bp) to mega-bases, represent a substantial component of genomic diversity and have been involved in various human diseases and phenotypic traits [2]. These events encompass a diverse array of genomic rearrangements, including deletions (DELs), duplications (DUPs), insertions (INSs), inversions (INVs), and translocations (TRAs). By its nature of impacting a larger portion of the genome compared to SNVs and INDELs, SVs can disrupt a gene's function, alter gene dosage, and perturb regulatory elements, thereby contributing to disease susceptibility and phenotypic variation [3].

Of particular interest are *de novo* SVs (dnSVs), included as a category of *de novo* mutations (DNMs) which refer to genomic alterations that arise specifically in an individual and are absent in the individual's parents. DNMs play a crucial role in the etiology of Mendelian disorders, congenital anomalies, and several rare diseases [4]. Previous studies have focused on DNMs occurring as a single nucleotide (dnSNVs) or small INDEL mutations (dnINDELs) to establish a known knowledge of an individual having approximately 40 to 70 events and its rate increasing proportionately with parental age [5, 6]. However, in contrast to dnSNVs or dnINDELs, little is known about the characteristics of *de novo* SVs in terms of its prevalence, mechanism of generation and

impact on diseases. Since dnSVs are known to occur far less frequently than dnSNVs or dnINDELs, accurate estimation of its occurrence rate requires a much larger sample size [7]. Furthermore, although dnSVs occur less frequently than dnSNVs or dnINDEL, its nature of altering a large genomic segment connotes its deleterious effect on one's genome. There have been a handful of studies using microarrays and WGS presenting a controversial estimate of dnSV rate ranging from once in every 98 births to once in 5-6 births [8, 9]. Additionally, a recent study using a larger population-based sequencing analysis presented a higher rate of one dnSV per 3.5 births [10]. These contentious reports of the characteristics of dnSV highlight the importance of analyzing dnSVs in a large cohort with high confidence.

In terms of its association with diseases, several studies revealed dnSVs playing an important role in the genetic etiology of autism spectrum disorder (ASD) [8]. One study that analyzed individuals with rare disease suggested that dnSVs were enriched in individuals without any diagnostic SNVs or INDEL compared to those with diagnostic SNVs or INDELs explaining the disease [11]. This implies the possibility of dnSVs being a potential factor of unveiling the mechanism of disease occurrence from patients with previously unexplained genetic diseases. As there are growing evidence of *de novo* variants being an important factor of several diseases, it is likely to suspect the potential association of dnSVs with other mendelian diseases. Yet this remains inconclusive by the small numbers of dnSVs found due to the current limitations of accurately identifying them.

Current strategies for dnSV detection rely on comparative analysis of WGS data by calling SVs separately on each member of the trio and simply comparing the genotype of the family members [12-14]. Since there are no bioinformatic tools specifically designed

to identify dnSVs yet, analytical workflows of dnSV identification accompany unnecessary and burdensome steps to call child-specific SVs including a high proportion of false positives in its final output. Furthermore, finding dnSVs is somewhat finding a needle in a haystack and non-informative sequencing reads hinders the fast identification of dnSVs leading to a process taking up to several days [15]. This could be a major bottleneck especially in the clinical diagnostics of pediatric patients, as infants harboring genetic disorders show rapid progression of the disease [16]. Therefore, the development of a fast and efficient dnSV identification framework would be pertinent not only for personal genomics but also for population-scale studies.

This thesis presents a strategy to identify dnSVs with reduced time and computational cost of analyzing WGS datasets (Figure 1). This approach achieves a fast and accurate dnSV calling by extracting informative sequencing reads that are truly relevant to variants of interest. In addition, a novel strategy to estimate variant allele frequency of dnSV using a machine learning method is applied to selectively collect germline dnSVs. I anticipate that this method would offer a practical usage in population-scale research, personalized genomics and clinics.

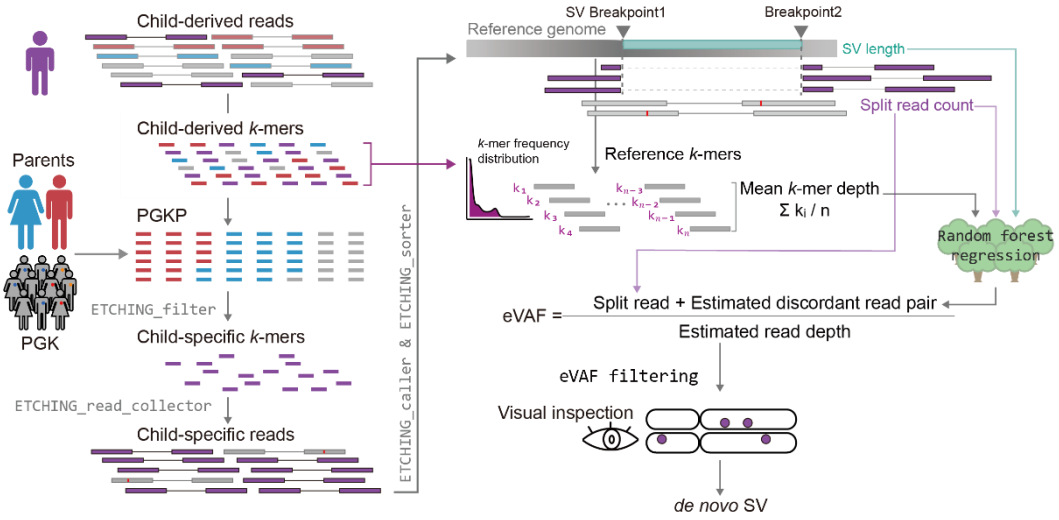


Figure 1. Graphical abstract of ETCHING-trio, a *k*-mer-based dnSV identification method.

II. Materials and Methods

II-1. Dataset

Illumina short-read WGS and PacBio CCS long-read data of the Ashkenazim Jewish trio collected as part of the Personal Genome Project were provided by the Genome In A Bottle (GIAB) consortium [17]. A high-confidence benchmark SV set from the Ashkenazim Jewish HG002 child was acquired from the GIAB FTP server (https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/). For performance evaluation, synthetic simulation data was generated as described in section II-5. Short-read WGS data from the 33 CEPH-Utah families were obtained from dbGaP by the study accession phs001872.v1. For the Korean rare disease cohort, fastq files from 150 quartets were provided by the Korea Bioinformation Center (KOBIC).

II-2. Preprocessing and alignment of WGS data

The quality of the WGS data was examined using FastQC (version 0.11.8) and low base quality (Phred score < 20) reads were trimmed using sickle (version 1.33) [19, 20]. Reads shorter than 70 base pairs (bp) after trimming were discarded using the -l 70 parameter. Filtered reads were then aligned to the human reference genome (hg19 or hg38) using BWA-MEM (version 0.7.17) with the -M -t 30 parameters [21]. Following read alignment, duplicate reads were marked and removed from the resulting bam file using Picard MarkDuplicates (version 2.21.4) [22]. After removing duplicated reads, the bam file was coordinate sorted using samtools (version 1.16.1) and was subsequently subjected to GATK (version 4.1.7) BaseRecalibrator and ApplyBQSR with known SNP

sites provided by the GATK Resource Bundle [23, 24] (https://console.cloud.google.com/storage/browser/gcp-public-data--broad-references/hg38/v0).

II-3. Development of ETCHING-trio

In order to identify child-specific dnSVs within a family trio WGS dataset, a *k*-mer filtering strategy of ETCHING (version 1.4.2) was employed [25]. A database comprising *k*-mers with a length of 31 (31mer) was constructed on each of the parents' WGS data using KMC (version 3.2.1) with -k31, -ci2 parameters [26]. Subsequently, the paternal and maternal *k*-mer databases were merged using kmc_tools with -ocsum parameter. Finally, the resulting parental database was integrated with Pan-genome *k*-mer (PGK) database (http://big.hanyang.ac.kr/ETCHING/PGK2.tar), which is a *k*-mer database comprising 894 human genome assemblies and eight human references provided by ETCHING, using kmc_tools with the same parameters described above, with the exception of an additional -ci5 parameter. The resulting Pan-genome *k*-mer with parent (PGKP) database was used as an input for ETCHING-trio to filter potential parental germline variants.

II-4. Estimating variant allele frequency using *k*-mer

For a given SV, a 30 bp flanking region on each breakpoint (BP) was defined as the region of interest (ROI) window. The reference genomic sequence corresponding to the 61bp ROI window was converted into 31mers. The average reference read depth of

coverage was estimated by retrieving 31 mer counts from the sample's *k*-mer database and applying the following equation.

$$\mu = \frac{\sum_{i=1}^{n} (k * m)}{w}$$

where k is a *k*-mer frequency from the input sample, m is a mappability score of the corresponding genomic position, and w is the ROI window size. Mappability score of the reference genome was calculated using GenMap (version 1.3.0) with -K 31 –E 0 parameters [27]. For each *k*-mer, mappability score m was assigned according to the following criteria.

$$m = \begin{cases} m\ (j \in penalization\ region\ OR\ K \equiv repeat\ sequence) \\ 1\ (else) \end{cases}$$

where *j* is the genomic position of the corresponding *k*-mer, K is the *k*-mer's sequence, and $penalization\ region$ is defined as a union of repetitive genomic regions annotated by repeatmasker [28] and low-mappability score regions annotated from the GIAB genome stratification [29]. The output bam file of ETCHING was utilized to retrieve proband-specific split reads via an in-house script.

A Random Forest regression model was constructed to predict discordant read pairs supporting the SV, using split read count (*s*), average read coverage (*cov*), SV length (*l*) and VAF prior as explanatory variables. For the VAF prior, a context-dependent prior for each SV's breakpoint was calculated as s/cov.

Let $X = (X_1, X_2, \ldots, X_n)$ represent the vector of input features on each data point, and *d* denote the output variable, discordant read pair count. A forest of decision trees, designated $T = (T_1, T_2, \ldots, T_n)$ was generated, where each tree was trained on a bootstrapped sample of the dataset. The prediction of each tree is represented as $\widehat{d_i}$

where $i = 1, 2, \ldots, n$. At each node of the decision trees, a random subset of four features was drawn with replacement for splitting. Mean squared error (MSE) was applied to optimize the model's prediction splitting accuracy. Let $Q(T)$ represents the impurity measure *I* of each leaf node *t* of tree *T*.

$$Q(T) = \ Q_{left}(T) + \ Q_{right}(T)$$

$$where\ Q_{left}(T) = \frac{N_{left}}{N} \cdot I_{left}\ and\ Q_{right}(T) = \frac{N_{right}}{N} \cdot I_{right}$$

At each node of the tree, the feature X_i was selected and the split threshold was optimized to minimize the MSE of the node. To make a regression model, predictions from each tree were aggregated by computing the mean value of their outputs and was used as the final prediction for discordant read pair count *d*.

$$\hat{d}_{ensemble} = \frac{1}{N}\sum_{i=1}^{N} \hat{d}_i$$

For hyperparameter tuning, a grid search was applied to find the optimal number of trees ranging from 100 to 1,000 and the number of features ranging from two to four, minimum number of samples in each leaf node after node splitting ranging from 20 to 500. A simulated dataset comprising of 2,000 SV breakpoints was used to train the regression model, with the prediction accuracy tested on a separate validation set comprising 696 breakpoints. After predicting the discordant read pair count on each breakpoint, an estimated variant allele frequency (eVAF) for a given SV's BP is calculated as the following equation.

$$eVAF = \frac{s + d}{\mu + s + d}$$

dnSV candidates called by ETCHING-trio were filtered according to their eVAF value with a minimum cutoff of 0.3 and maximum cutoff of 0.7. dnSV candidates having both

breakpoints in the target range were selected as the final call set. For the three large cohort datasets, a more lenient eVAF cutoff of minimum of 0.2 and maximum of 0.8 was applied to increase the sensitivity.

II-5. Generating simulated data

For performance evaluation, a simulation dataset containing SVs with known VAF was incorporated into a synthetic reference genome. First, NGSNGS (version 0.9.0) [30] was employed with -c 30 -seq PE -l 500 -cl 150 parameters. This generated a 30x synthetic paired-end WGS dataset from the hg38 human reference genome having a read length of 150 bp and a fragment length of 500 bp. For the simulation data with varying read length, differing -l and -cl parameter was chosen as following, -l 350 -cl 75, -l 400 -cl 100, -l 450 -cl 125, -l 500 -cl 150. Next, SV sites ranging from 50 bp to 5 kb were randomly selected from the genome. For random site selection, repeat regions and regions having read depth less than five in the reference synthetic genome were excluded. Subsequently, random variant allele frequencies (VAFs) were assigned to each of the random sites, ranging from 0.1 to 0.29 for somatic SVs and 0.3 to 0.7 for germline and de novo SV. For the trio-simulation, the random variant sites were then concatenated with randomly selected set of 500 common SVs from dbVAR which have population allele frequency greater than 0.01, to create the final variant site list. Finally, Bamsurgeon was used to insert the input variants into the reference synthetic genome with --aligner mem --alignopts M: --maxdfrac 0.4 --minctglen 5000 --keepsecondary --require_exact parameters [31].

II-6. Generating artificial benchmark genome

To create an artificial benchmark genome containing multiple dnSVs from multiple individuals within a single genome, a synthetic reference genome was used as the initial data. Given a list of candidate dnSVs with their chromosome, start and end position, a 1 kb flanking region was selected as the target site. First, reads aligning to these target sites from the synthetic reference genome were removed. Next, the aforementioned empty dnSV regions were filled with reads derived from the individuals possessing target dnSV. Along with constructing a benchmark genome containing authentic dnSVs, the same process was carried out with each of the individuals' parents, resulting in three artificial benchmark genomes corresponding to the child, maternal and paternal data. All the processes, including the removing, sampling, and merging of sequencing reads were performed using samtools (version 1.16.1). The F1 generation individuals of CEPH-Utah families who had previously been reported to possess dnSVs were used as target samples [37]. The dnSVs were validated by third-generation transmission and a separate visual inspection was conducted to make a total of five dnSV candidates for insertion.

II-7. Visual inspection of *de novo* structural variant

The aligned bam files of the three family members were used to validate dnSVs called by ETCING-trio through a thorough inspection of the variant-supporting reads in all of the family members. This inspection was performed by juxtaposing the aligned bam file on IGV [18] to check whether the breakpoints of a dnSV possess true dnSV supporting reads, in accordance with the following criteria (Figure 2).

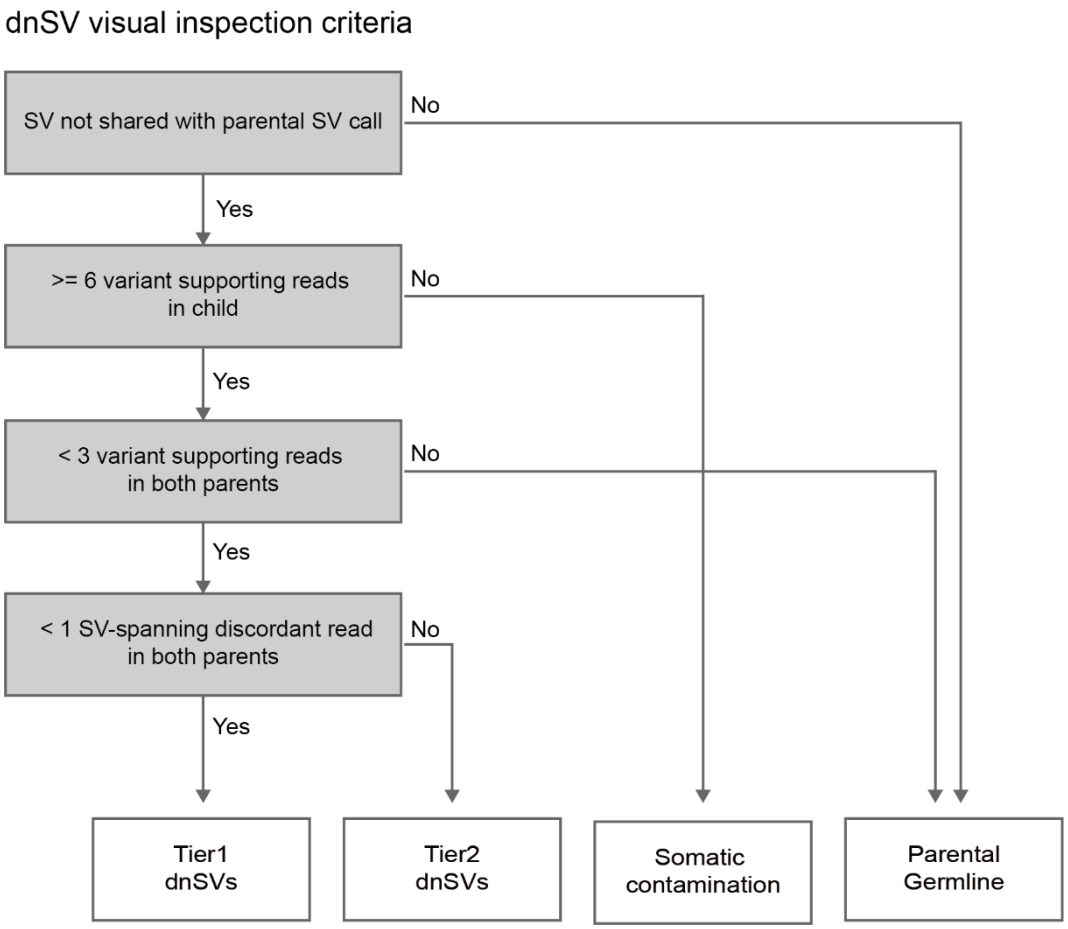


Figure 2. Workflow of dnSV visual inspection.

During the inspection process, only reads having mapping quality higher than Phred-scaled score 20 were used, with secondary alignment reads being discarded. First, parents of the trio were subjected to a separate SV calling process performed by ETCHING using only the PGK *k*-mer database. This process was conducted to ascertain whether the same SV was called in any parents. If any of the dnSVs matched with parental SV call, they were regarded as “Parental germline”. Then, variant-supporting soft-clipped reads and discordant read pairs were examined at each breakpoint in the child’s bam file. dnSVs having fewer than six variant-supporting reads were regarded as “Somatic contamination”

and were discarded. Once the child's variant-supporting reads have been checked, the same process was carried out on each parent to check whether they have variant-supporting reads less than three. dnSVs having less than three variant-supporting reads in both of the parents were used for the next step, while those not satisfying this criterion were regarded as "Parental germline". Subsequently, if any of the parents' discordant read pairs were exactly spanning the SV breakpoints, this was regarded to be unreliable to call as a true dnSV and was classified as "Tier 2 dnSVs". Candidate dnSVs satisfying all of the aforementioned criteria were classified as "Tier 1 dnSVs" and were used for further analysis. For cohorts having singling data available, an additional classification of Tier 1 dnSVs was conducted by examining whether the dnSV was shared between siblings. If a given variant was shared, this could possibly indicate false negative germline variants being missed in their parents or dnSVs occurred in the early developmental stage of the parent's germ cell. Since correctly discerning these two cases was not possible with the current data, these dnSVs were classified as "Possibly germline / early developmental dnSV".

II-8. Benchmark of structural variant callers

Four SV callers were applied to the HG002 and simulated trio data to facilitate the comparative analysis of the performance of dnSV identification. For each individual in a trio, quality-controlled bam file was generated following the procedure descripted in Method II-2. For each caller, the following commands were applied to identify candidate dnSVs.

DELLY [32]

delly call -t ALL -g genome.fasta -x delly_exclude.tsv child.bam -o child.bcf

The command above was performed on the child, mother, and father each to generate *child.bcf, mother.bcf, father.bcf* .

delly merge -o output_merge.bcf child.bcf mother.bcf father.bcf -m 50

delly call -t ALL -g genome.fasta -x delly_exclude.txv output_merge.bcf -o child_genotyped.bcf child.bam

The command above was performed on the child, mother, and father each to generate *child_genotyped.bcf mother_genotyped.bcf father_genotyped.bcf ,*

bcftools merge -0 -f PASS -m id -O v -o final_output.vcf child_genotyped.bcf mother_genotyped.bcf father_genotyped.bcf

Lumpy [33]

sambamba sort --sort-by-name -o child_sort.bam child.bam

samtools view -h child_sort.bam | samblaster --ignoreUnmated --excludeDups --addMateTags --maxSplitCount 2 --minNonOverlap 20 | samtools view -hSb > child_lumpyReady.bam

sambamba sort -m 8G -o child_lumpyReady_sort.bam child_lumpyReady.bam

sambamba index child_lumpyReady_sort.bam

samtools view -h -b -F 1294 -o child_discordants.bam child_lumpyReady_sort.bam

sambamba sort -m 8G -o child_discordants_sort.bam child_discordants.bam

sambamba index child_discordants_sort.bam

samtools view -h child_lumpyReady_sort.bam | extractSplitReads_BwaMem -i stdin | samtools view -h -Sb > child_splits.bam

sambamba sort -m 8G -o child_splits_sort.bam child_splits.bam

sambamba index child_splits_sort.bam

All the commands above was performed on the child, mother, and father each.

lumpyexpress -B child_lumpyReady_sort.bam, mother_lumpyReady_sort.bam, father_lumpyReady_sort.bam -S child_splits_sort.bam, mother_splits_sort.bam, father_splits_sort.bam -D child_discordants_sort.bam, mother_discordants_sort.bam, father_discordants_sort.bam -R genome.fasta -o output_lumpyexpress.vcf

svtyper -i output_lumpyexpress.vcf -B child_lumpyReady_sort.bam, mother_lumpyReady_sort.bam, father_lumpyReady_sort.bam -T genome.fasta -o output_lumpyexpress_genotyped.vcf

Manta [34]

configManta.py --bam child.bam --bam mother.bam --bam father.bam --referenceFasta genome.fasta --runDir manta_analysis_dir

manta_analysis_dir/runWorkflow.py

convertInversion.py path_to_samtools genome.fa manta_diploidSV.vcf > diploidSV_convertINV.vcf

SvABA [35]

svaba run -a analysis_id -G genome.fasta -t child.bam -t mother.bam -t father.bam -D Homo_sapiens_assembly38.dbsnp.vcf –germline

In order to draw a PR curve and calculate auPR for the five SV callers, differing thresholds were applied on each of the SV callers. For ETCHING-trio, differing eVAF filtering threshold was applied in the range starting from 0.0–1.0 to 0.45–0.55, narrowing down the range by 0.05. For DELLY and Lumpy, the sum of DR and RV values and the SU value in the INFO column, respectively, was used for setting the threshold. For

SvABA, the score on the QUAL column was used. The thresholds for each of the SV callers were split into ten ranges using the minimum and maximum score of a given sample.

After acquiring SV candidates from each caller, dnSVs with the genotype GT = (0/1, 0/0, 0/0) for the child, mother, and father, respectively, were extracted using an in-house script. For the golden standard benchmark of HG002 trio, SVs having genotype GT = (0/1, 0/0, 0/0) in the GIAB benchmark set [17] were compared with dnSV calls from five SV callers. dnSVs detected by the majority of the five SV callers were defined as a golden standard benchmark and were visually validated by Illumina short-read and PacBio long-read data.

III. Results

III-1. A fast *k*-mer filtering strategy can selectively identify child-specific variants

To build a *de novo* SV (dnSV) calling pipeline, which I call ETCHING-trio, ETCHING's *k*-mer filtering strategy was implemented to selectively call child-specific variants using whole genome sequencing data from a family trio [25]. First, parental sequencing reads were converted into 31 mers to construct a parental *k*-mer database. The resulting database was merged with a Pan-genome reference *k*-mer (PGK) database to create a Pan-genome reference *k*-mer with parent (PGKP) database. PGKP database was used as input for *k*-mer filtering to discard reference *k*-mers and *k*-mers derived from parental germline and population common variants. These child-specific *k*-mers were used to collect child-specific reads that are utilized to call candidate dnSVs.

With ETCHING-trio, WGS data from the HG002 Ashkenazim Jewish trio were used to call candidate dnSVs. 300X and 100X WGS data from HG002 son and HG003, HG004 parents were obtained from the Genome In A Bottle (GIAB) consortium with its complementary benchmark SV set [17]. These data were subsampled to generate a 30x data for further analysis. To evaluate the effectiveness of including the parental data in *k*-mer filtering, SV calls using only PGK and using PGKP were compared (Figure 3). A significant number of SVs were filtered out by including the parental data in the input database, indicating the effectiveness of using parental *k*-mer data to filter out possible parental germline and population common SVs.

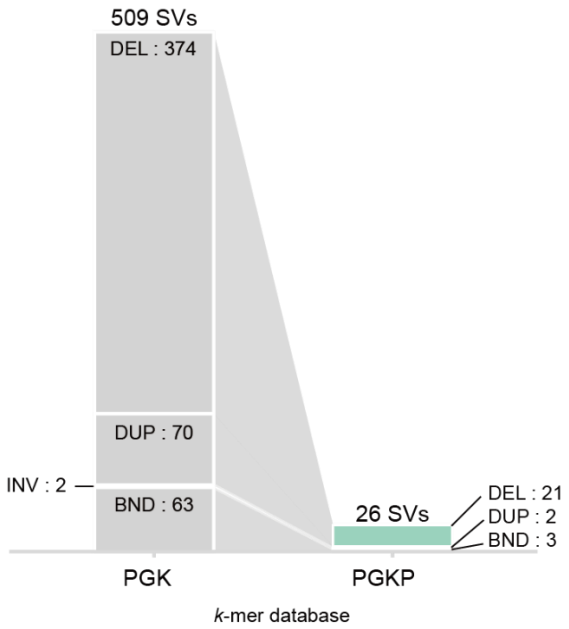
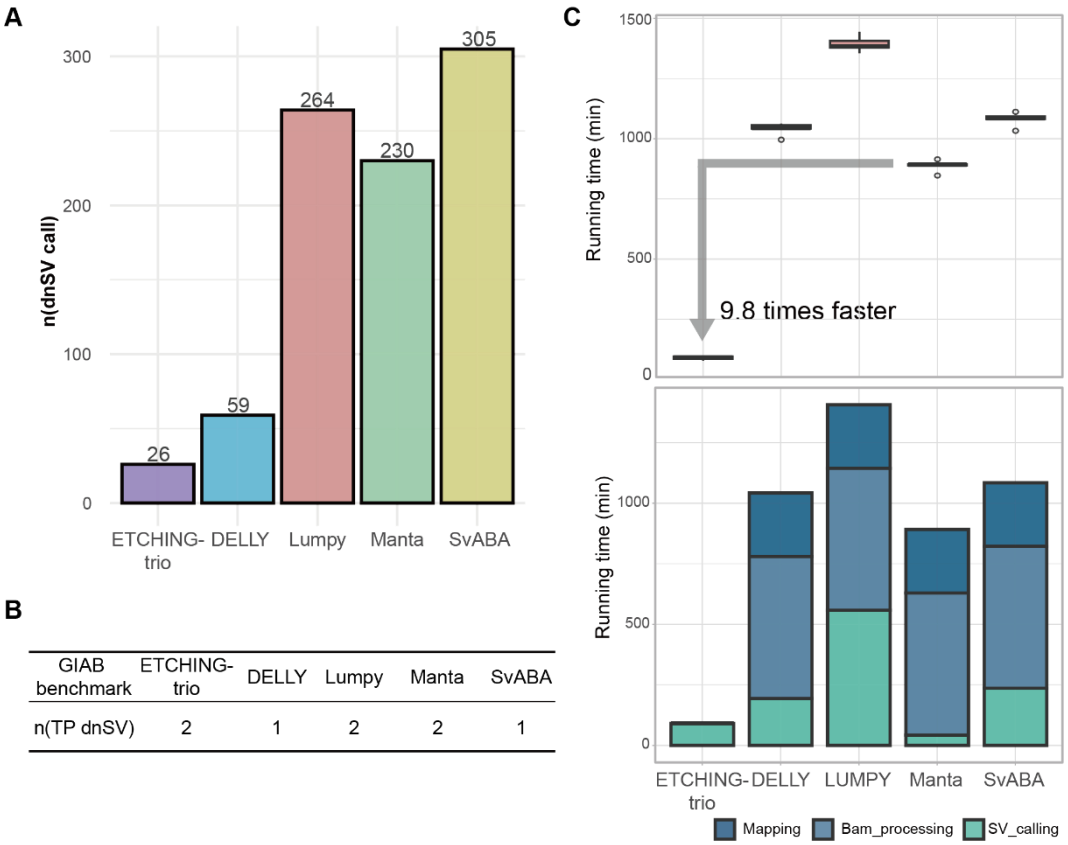


Figure 3. Number of SVs called by ETCHING-trio before and after including parental data in the reference *k*-mer database. PGK : Pan-genome *k*-mer database without parental data, PGKP : Pan-genome *k*-mer with parent database.

Using HG002 trio, dnSV calls were compared to four conventionally used SV callers (Figure 4). The large number of calls from the four SV callers suggests that there is a high number of false positives (FPs) given the fact that dnSVs occur at a very low frequency. ETCHNG-trio called the fewest dnSVs among the five SV callers and was the fastest caller, taking about only 1h 30min to analyze a whole trio. This was approximately 9.8 times faster than Manta, which was the second fastest SV caller. Since all other SV callers require WGS data to be aligned to the reference genome, read alignment and preprocessing steps were the major bottlenecks. In addition, by comparing the dnSV calls with HG002's benchmark SV set with its genotype GT = (0/1, 0/0, 0/0), ETCHING-trio identified two of the TP dnSVs same as Lumpy and Manta, demonstrating that ETCHING-trio's strategy can significantly reduce FPs without losing TPs.



B

GIAB benchmark	ETCHING-trio	DELLY	Lumpy	Manta	SvABA
n(TP dnSV)	2	1	2	2	1

Figure 4. dnSV calls from five SV callers on HG002 Ashkenazim Jewish trio. (A) Number of dnsV calls of each SV callers. (B) The number of true positive dnSVs from GIAB germline benchmark SVs found by each SV caller. (C) Running time of each SV callers. The running time of each SV caller were repeatedly checked five times (upper). The median time of performing read alignment (Mapping), bam file preprocessing (Bam_processing) and SV calling is plotted in the bar plot (lower).

III-2. A *k*-mer-based estimation of variant allele frequency can be used to filter out false positive dnSV calls

Since dnSVs derive from parental germ cells, an ideal variant allele frequency (VAF) is expected to be 0.5. As dnSV calling approach selectively collects variants that are specifically present in the child, these calls could potentially include somatic variants that would have low VAF. To eliminate somatic contamination, a VAF filtering strategy was applied to dnSVs called by ETCHING-trio (Figure 5).

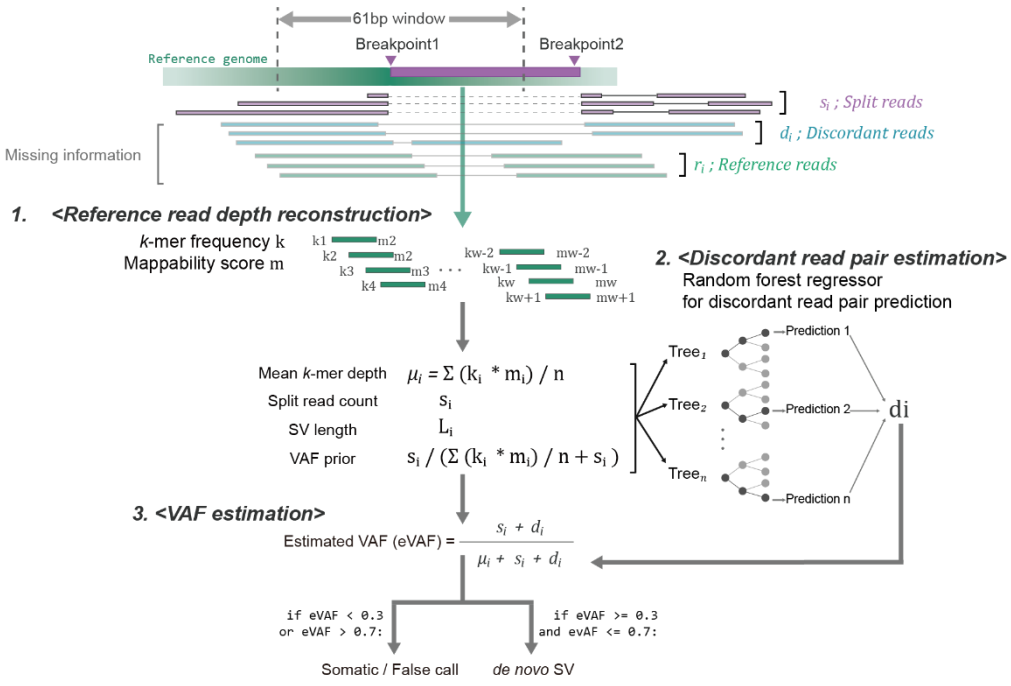


Figure 5. Schematic workflow of estimating VAF. Reference read and discordant read pair depth are estimated to reconstruct the total depth of coverage for a breakpoint window. Estimated VAF was calculated as variant supporting read count over estimated depth of coverage.

To calculate VAF for a variant, variant supporting read count was divided by the total read depth at the variant's site. However, due to the nature of SVs spanning a large genomic interval, short-read WGS data cannot cover the entire event with a single read. Therefore, each breakpoint (BP) of an SV was treated independently in the VAF calculation and those with VAF of both BPs in the range of 0.3 to 0.7 (0.3–0.7) were selected as a final dnSV set. Following the *k*-mer filtering strategy of ETCHING-trio, non-variant supporting reference reads and discordant read pairs become eliminated. This makes it impossible to calculate the total read depth of coverage. Therefore, a *k*-mer-based read depth of coverage estimation was applied to reconstruct the original depth of a given genomic interval. For each BP, a variant window of 61bp size, including the 30bp flanking region of a given BP, was used as the region of interest (ROI) for depth estimation. The *k*-mer sequences of this ROI were extracted from the human reference genome to create an ROI reference *k*-mer set. The frequencies of these *k*-mers were retrieved from the sample's *k*-mer database and were multiplied by the corresponding mappability score. The mean value was used as an estimate of the reference read depth (Materials and Methods II-4). This estimate was concatenated with the variant supporting split reads to create an initial read coverage of the ROI. As with the reference reads, discordant read pairs get eliminated in ETCHING-trio's *k*-mer filtering step. To fully reconstruct the variant supporting reads including discordant read pairs, a simple machine learning approach was applied to predict the number of discordant read pairs present in the ROI window. A random forest regression was used to predict the discordant read pair count given the SV length, split read count, estimated read coverage prior and a VAF prior. The VAF prior, calculated as the ratio of split read count to the estimated read coverage prior, was used as a feature to discriminate between somatic and germline

variants. The regression model was trained on 2,000 simulated data generated by adding artificial SVs using Bamsurgeon on a 60x coverage synthetic genome. [31]. The predictive performance showed a Pearson correlation coefficient of approximately 0.7 on a separate validation set consisting of 696 simulated breakpoints (Figure 6).

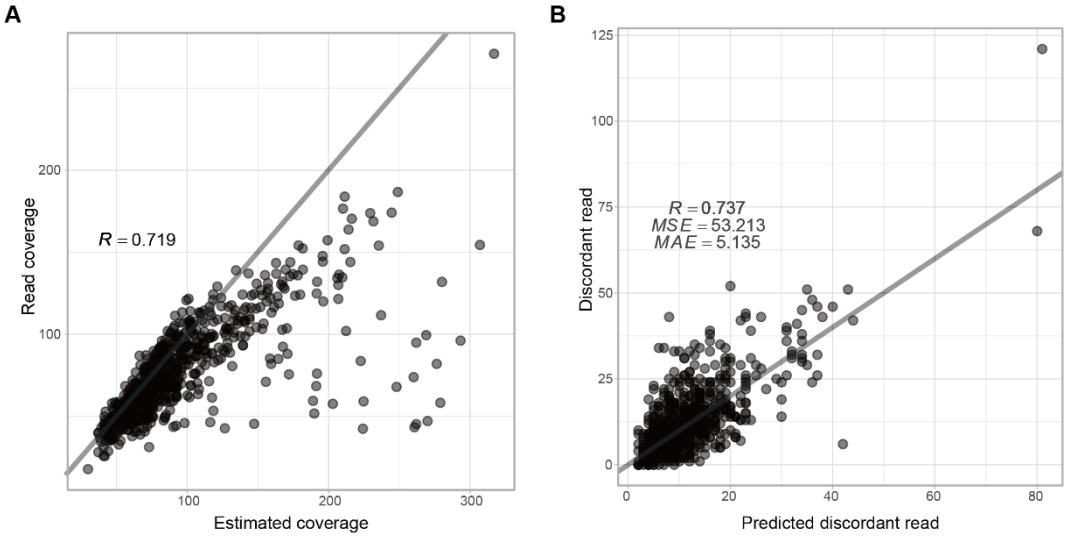


Figure 6. Accuracy of average read depth of coverage and discordant read pair estimation. (A) Estimation of average read depth of coverage. (B) Prediction of the number of discordant read pairs. Pearson correlation coefficient of the estimated value and the real value is depicted in the scatter plot. MSE : Mean squared error, MAE : Mean absolute error.

After predicting discordant read pairs, variant supporting reads including split reads and predicted discordant reads were divided by the estimated read coverage to generate an estimated VAF (eVAF） for a given BP window. The eVAF was used to filter dnSV candidates and a categorical test performance showed approximately 0.7 of the F1-score in the validation set (Figure 7).

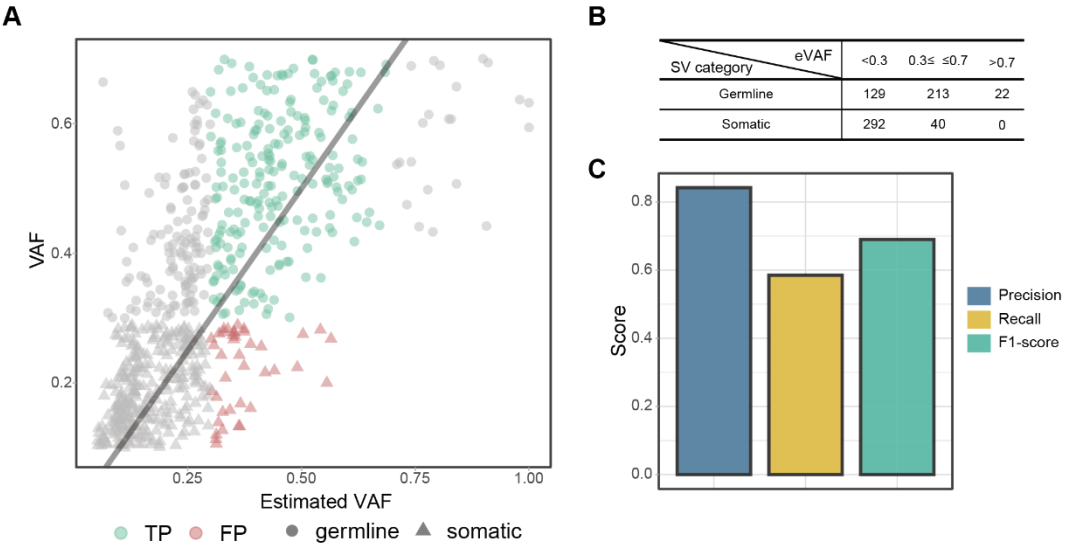


Figure 7. Categorical testing performance of ETCHING-trio's eVAF filtering strategy. (A) The accuracy of estimated VAF and its true VAF of 696 simulated breakpoints. (B) The number of simulated germline and somatic SVs called in each eVAF range with the cutoff of 0.3–0.7. (C) Performance with precision, recall and F1-score plotted as a bar plot.

To test the performance of dnSV calling and eVAF filtering on the scenario of having a trio data, a simulated child dataset containing 679 somatic and germline dnSVs was generated. This data was generated to reflect the nature of variant inheritance by additionally including 454 population common SVs, 618 maternally-shared and 606 paternally-shared SVs along with 346 somatic and 333 dnSVs. The performance of calling true dnSV was compared with four other SV callers (Figure 8). ETCHING-trio with eVAF filtering showed the best performance in terms of precision and F1-score reflecting its low false positives and efficiency in finding true dnSVs. In terms of recall, DELLY and Lumpy outperformed ETCHING-trio with a compromised precision. Detailed dnSV call statistics for each of the callers are listed in Table 1.

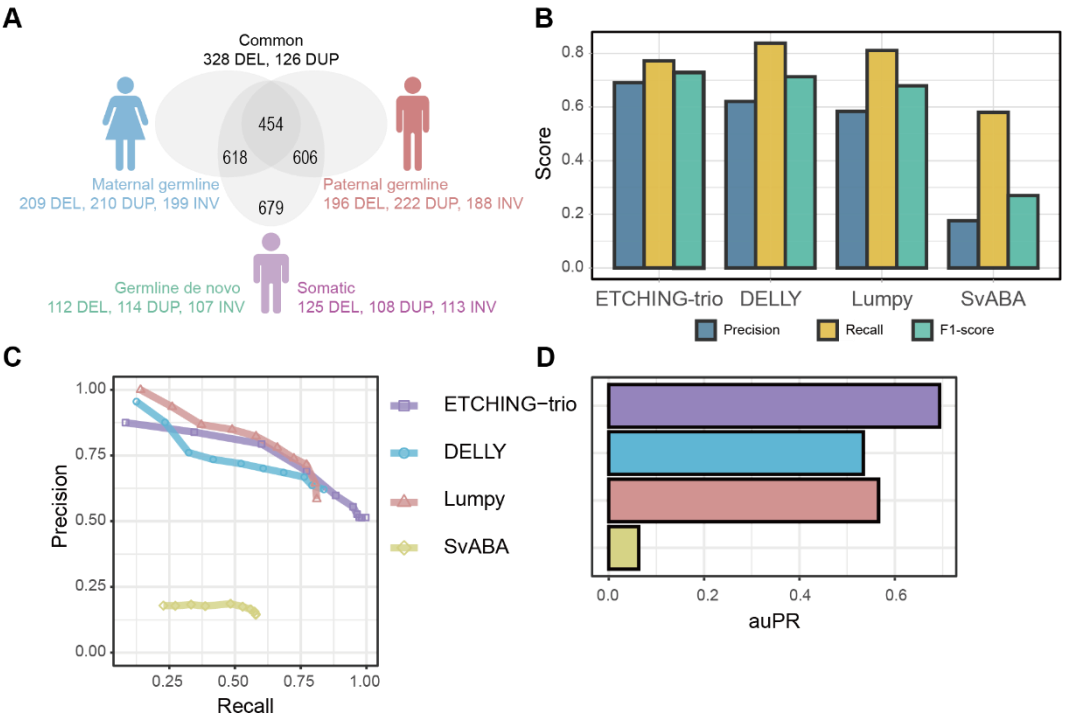


Figure 8. dnSV calling performance on a simulated trio data. (A) A simulated trio was generated including population-common SVs, maternally- and paternally-inherited SVs, and child-specific somatic and *de novo* SVs. (B) dnSV calling performance on each SV callers plotted in a bar plot with its precision, recall and F1-score. (C, D) Performance plotted as a PR curve with differing SV filtering threshold on each SV caller and the area under PR curve (auPR).

Simulated SVs	DEL	INV	DUP
Simulated dnSVs	112	107	114
Simulated somatic SVs	125	113	108
Sum	237	220	222
Called dnSVs	**DEL**	**INV**	**DUP**
ETCHING-trio call	120	123	129
ETCHING-trio TP	100	70	87
DELLY call	91	163	195
DELLY TP	78	93	108
Lumpy call	114	173	145
Lumpy TP	82	103	85
SvABA call	395	407	323
SvABA TP	51	94	48

Table 1. The number of simulated SVs generated in the child and the number of SVs and TP SVs called by each SV caller. The number of dnSVs called and the number of true positives (TPs) are listed for each SV caller.

To assess the performance of ETCHING-trio on data with varying read lengths, five trio datasets were simulated with paired-end read lengths ranging from 75bp to 150bp (Figure 9). The results demonstrated a slight decline in performance for datasets with shorter read lengths and showed a saturation of performance as the read length exceeds 100bp. Given that ETCHING-trio's model was optimized on paired-end data with a 150bp read length, further optimization on data with varying read lengths might be necessary.

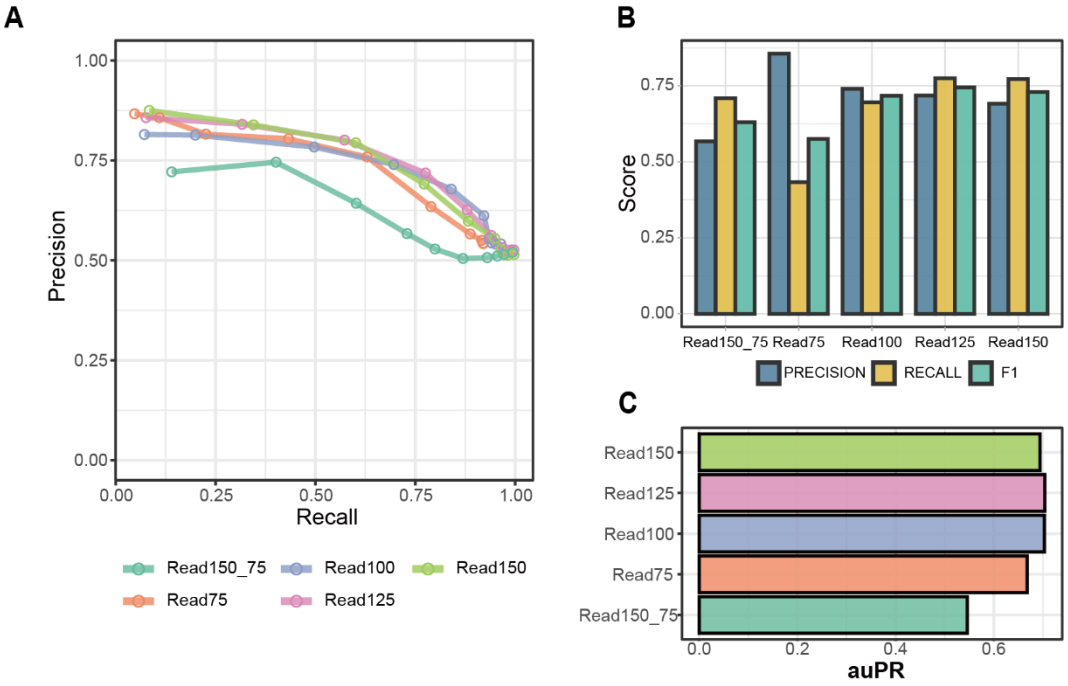


Figure 9. dnSV calling performance on simulated trio data with varying read length. (A) Performance plotted as a PR curve with differing eVAF filtering threshold on ETCHING-trio. (B) Precision, recall and F1-score on each read length data using 0.3–0.7 VAF filtering threshold. (C) Performance calculated as the area under PR curve (auPR).

In order to validate the performance of eVAF filtering on real data, a golden standard DEL set from HG002 trio was used (Figure 10). This gold standard set was defined based on the germline benchmark SV set from GIAB accompanied by orthogonal validation with five SV callers. (Materials and Methods II-6). Two DELs were identified in the majority of five SV callers and were additionally validated to be true dnSVs via visual inspection using Illumina short-read and PacBio CCS long-read data. After filtering dnSV candidates based on their eVAF, 26 dnSV candidates from ETCHING-trio were reduced to six. Among the identified calls, the two true DELs were called as positive while the remaining 24 false positives were reduced to four, which is approximately 17% of the

initial call. eVAFs of the two true positives were estimated to be near 0.5, which is an ideal value for germline dnSVs, thereby demonstrating the accuracy of VAF estimation.

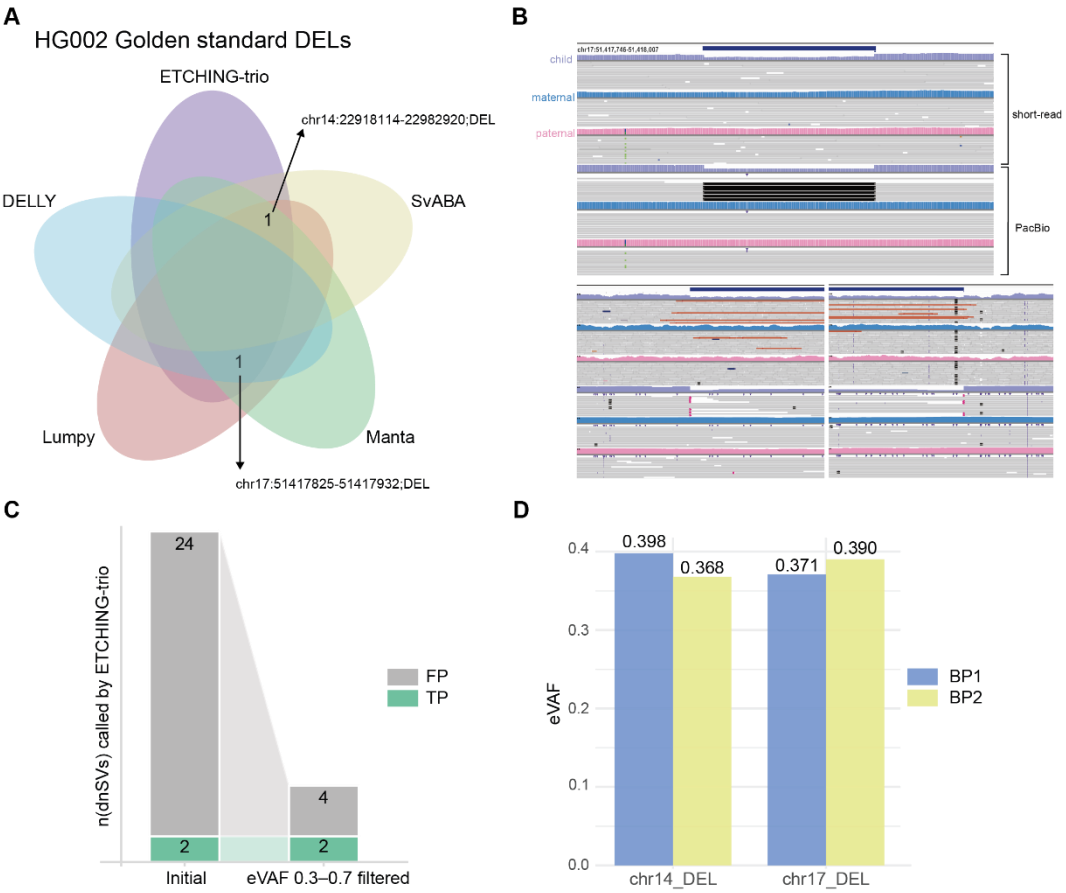


Figure 10. VAF estimation and filtering performance on HG002 trio. (A) Two true positive DELs defined as a golden standard from the HG002 dnSV benchmark set defined by orthogonal validation. (B) IGV images of the two true positives validated by Illumina short-read and PacBio long-read data. (C) The number of dnSVs called by ETCHING-trio before and after eVAF filtering. (D) The estimated VAF for the two true positives.

To further validate the performance of ETCHING-trio with various samples, an artificial benchmark genome including previously known dnSVs was generated (Figure 11). This benchmark genome comprises sequencing reads aligned to the 1 kb flanking region of the target position from the samples known to have previously reported dnSVs. As this genome incorporates real sequencing reads from multiple individuals, using this single genome would have the equal effect of analyzing multiple trio samples. Individuals from the F1 generation of CEPH-Utah families with previously reported dnSVs were utilized as target samples. A total of five dnSVs were selected as candidate variants [37]. Following the insertion of reads from each individual into a synthetic reference genome, this genome was subjected to dnSV calling with ETCHING-trio. By employing an eVAF filter with a range of 0.2 and 0.8, all five dnSVs were successfully identified with no false positives. The VAF estimate of the dnSVs showed a minimum of 0.230 and a maximum of 0.643.

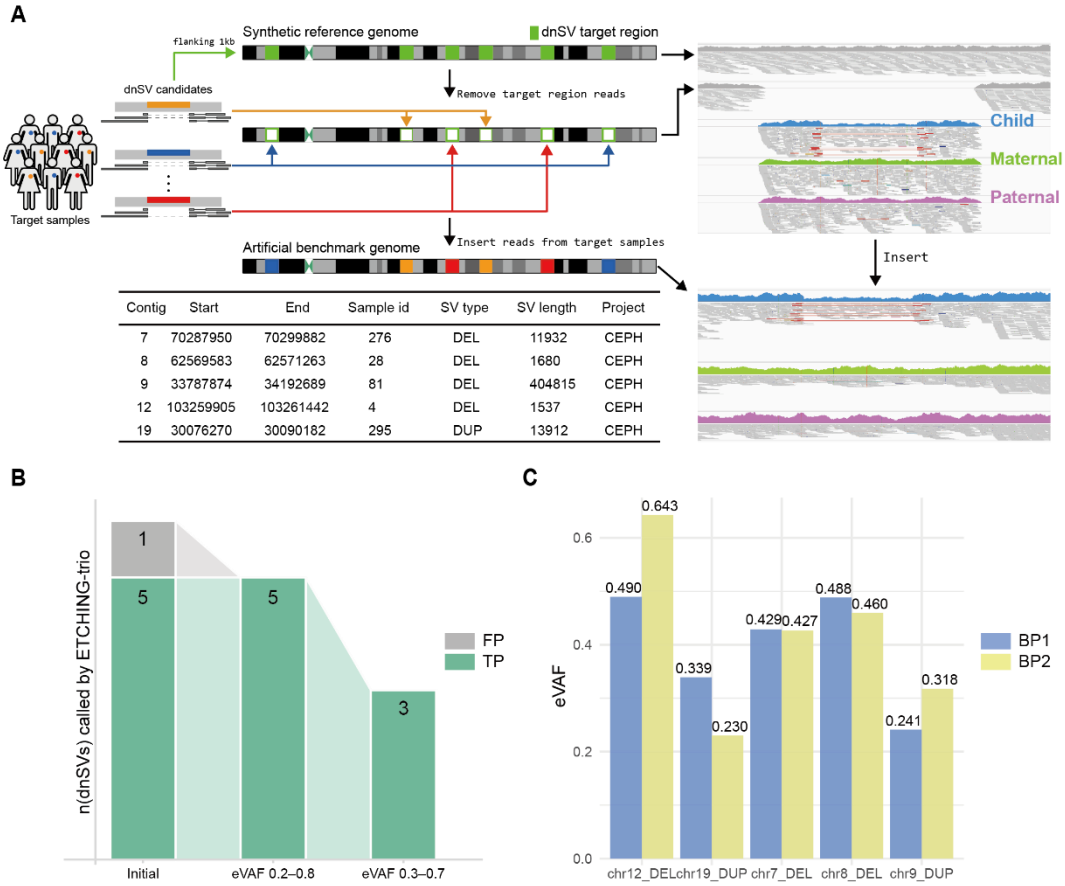


Figure 11. ETCHING-trio's performance on artificial benchmark genome using CEPH samples with known dnSVs. (A) A schematic figure of generating an artificial benchmark genome with known dnSVs (left upper left). A list of five dnSVs previously reported in CEPH cohort (lower left). An example IGV image showing the process of generating artificial benchmark genome (right). (B) The number of dnSV calls before, and after eVAF filtering with the range of 0.2–0.8 and 0.3–0.7. (C) The estimated VAF value of the five dnSVs called by ETCHING-trio.

III-3. ETCHING-trio can identify de novo structural variants in three large cohort datasets

After developing the aforementioned strategy, three large cohort datasets, including Korean Atomic Bomb survivor Cohort (ABC), CEPH-Utah families with third-generation grandchildren (CEPH), and Korean Rare Disease cohort (RD), were employed to identify true germline dnSVs. A total of 48, 29 and 200 second-generation individuals from the ABC, CEPH, and RD cohorts, respectively, were subjected to dnSV calling using ETCHING-trio. (Figure 12). In the ABC cohort, a median of two dnSVs per individual were identified, showing a maximum of seven dnSVs observed in an individual. In the CEPH cohort, a median of six dnSVs per individuals were identified with a maximum of 53 dnSV observed in a single individual. Finally, the RD cohort exhibited a median of one dnSV per individual, with a maximum of 40 dnSVs in a single individual.

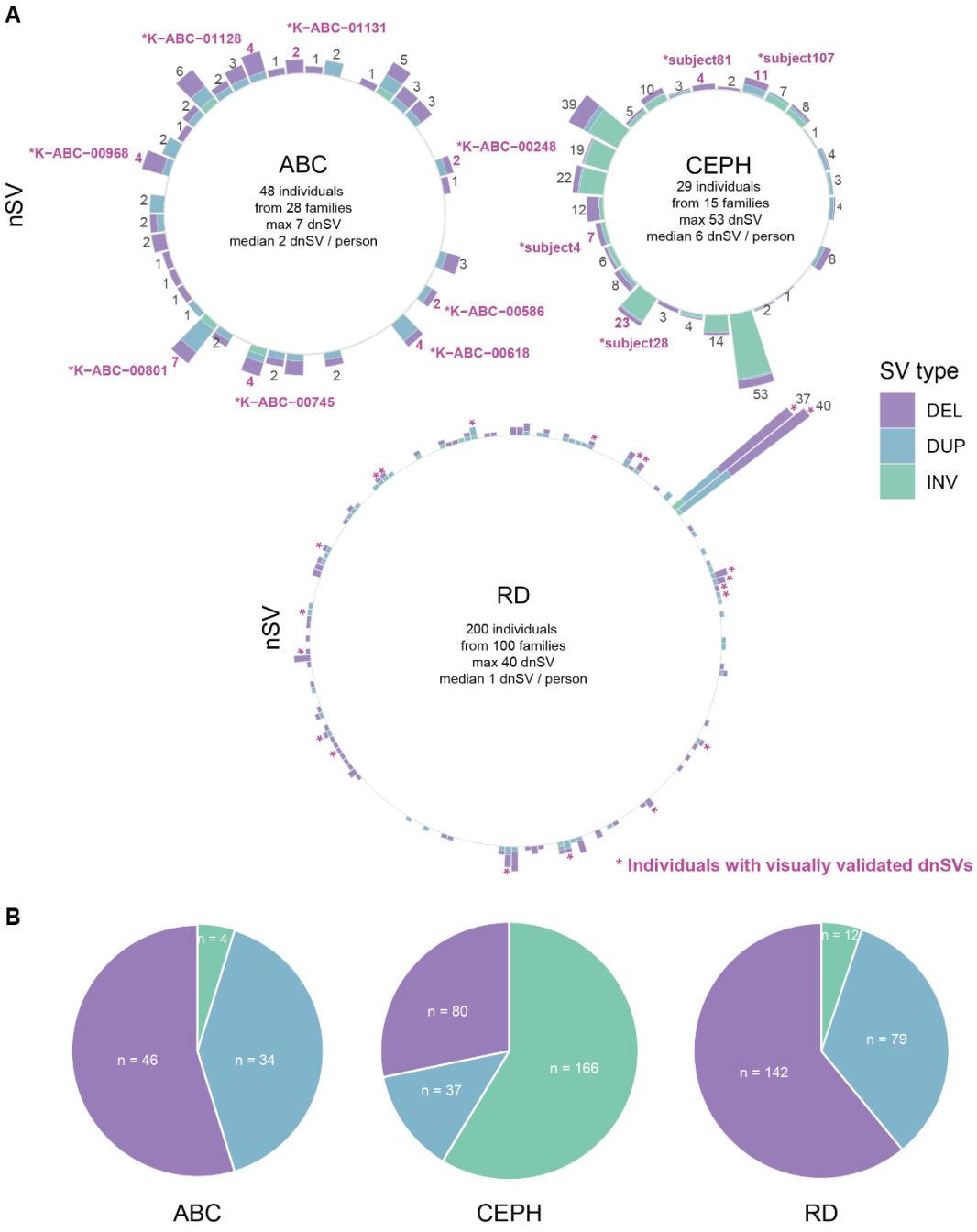
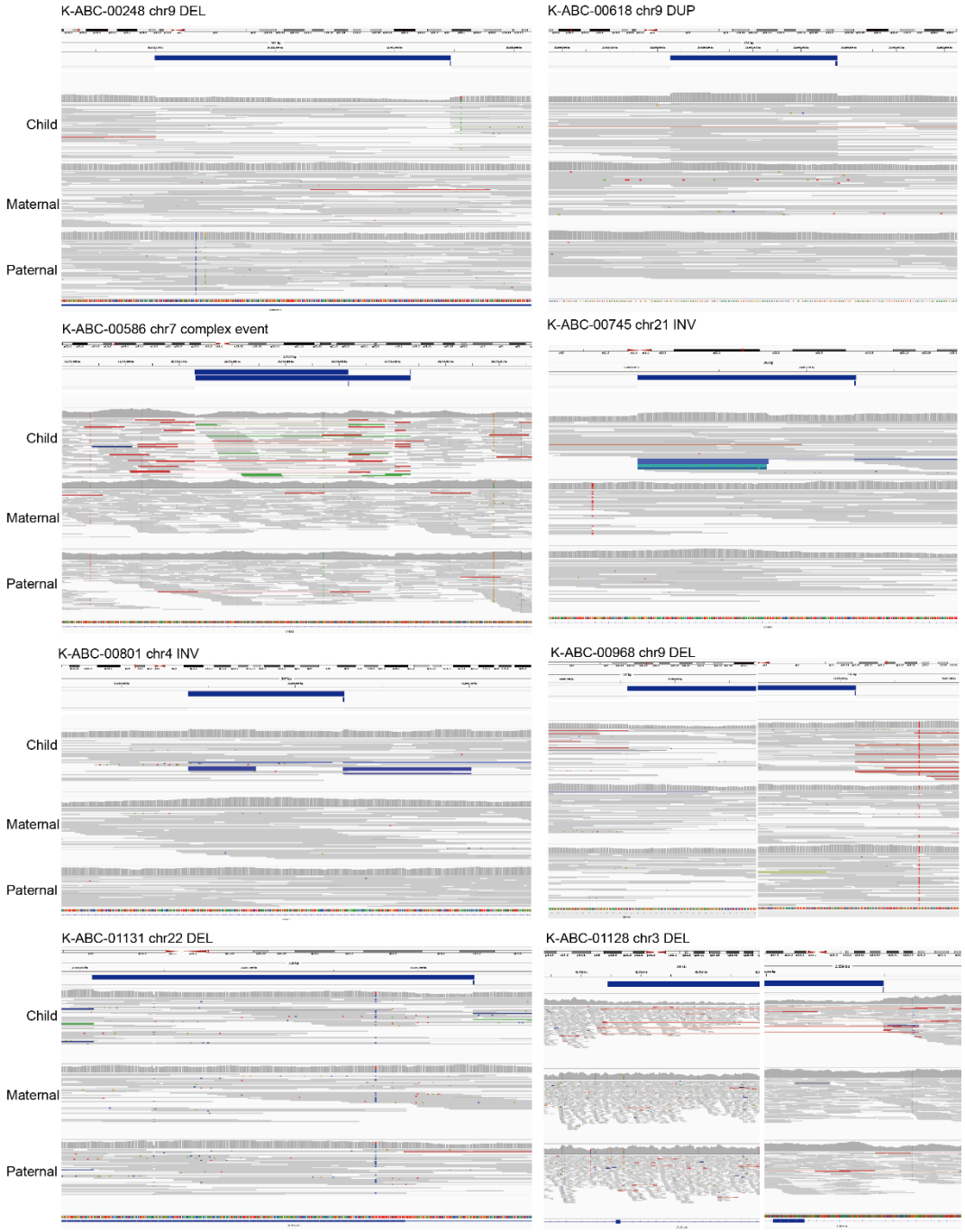


Figure 12. dnSV calls on three large cohorts. (A) Each bar, color-labeled by SV type, represents the total number of dnSV called on each individual. Individuals having the magenta asterisk are the ones having validated dnSVs by visual inspection. (B) The total number of dnSVs called from three cohorts is plotted as a pie chart.

After calling dnSV on three cohorts, a meticulous visual inspection was conducted on each dnSV to ascertain the presence in the child and the absence in both parents. This inspection was conducted by juxtaposing the aligned bam file of the child and parents on IGV. Furthermore, to ascertain the confidence of visually inspected dnSVs, parental SVs were separately called and were used to filter potential false negative germline variants in the parents. (Materials and Methods II-5). In the ABC cohort, eight dnSVs from eight individuals were validated as true dnSVs. These include four DELs, two INVs, one DUP and one complex event, which was identified as a large DUP harboring a smaller DEL within it. This case is suspected to be a dispersed DUP as it was corroborated by both discordantly mapped read pairs exhibiting abnormally large insert sizes and R2R1 pairs, which are read pairs having the second read mapped to an earlier genomic coordinate. The dnSV rate of ABC was determined to be as one mutation for every 0.17 births (8 dnSV / 48 individuals). In the CEPH cohort, six dnSVs composed of three DELs, two DUPs and one INV were validated to have true dnSV signatures. As this cohort includes third-generation grandchildren, the dnSVs were further validated by checking the transmission to their offsprings. Any dnSV being absent in all third-generation offsprings was considered as a false positive somatic contamination and was discarded. Consequently, five out of six dnSVs were identified to be transmitted to at least one of their offspring, which makes up the rate of one mutation for every 1.73 births. The visually inspected dnSVs identified in the ABC and CEPH cohorts are listed in the table and figure below (Table 2, Figure 13).

Cohort	Contig	Start	End	SV type	Genes	Region
ABC	chr3	42712890	42786496	DEL	Multiple	Exonic
ABC	chr4	41089939	41090028	INV	*APBB2*	Intronic
ABC	chr7	28771064	28771869	Complex	*CREB5*	Intronic
ABC	chr9	74865157	79071614	DEL	Multiple	Exonic
ABC	chr9	18032700	18032947	DEL	*ADAMTSL1*	Intronic
ABC	chr9	23060906	23060975	DUP	-	Intergenic
ABC	chr21	21072604	21072728	INV	*NCAM2*	Intronic
ABC	chr22	17589008	17589233	DEL	*SLC25A18*	Exonic
CEPH	chr8	78725370	78725443	DEL	*IL7*	Intronic
CEPH	chr8	62569584	62571264	DEL	*NKAIN3*	Intronic
CEPH	chr8	116397177	116397480	DUP	-	Intergenic
CEPH	chr9	33787876	34192689	DUP	Multiple	Exonic
CEPH	chr8	96250303	96250541	DEL	*MTERF*	Intronic

Table 2. List of dnSVs identified in ABC and CEPH cohorts. Each dnSV is listed with its genomic position, SV type, overlapping genes and its genomic region of the gene if present.


K-ABC-00248 chr9 DEL
Child
Maternal
Paternal
K-ABC-00618 chr9 DUP
K-ABC-00586 chr7 complex event
Child
Maternal
Paternal
K-ABC-00745 chr21 INV
K-ABC-00801 chr4 INV
Child
Maternal
Paternal
K-ABC-00968 chr9 DEL
K-ABC-01131 chr22 DEL
Child
Maternal
Paternal
K-ABC-01128 chr3 DEL

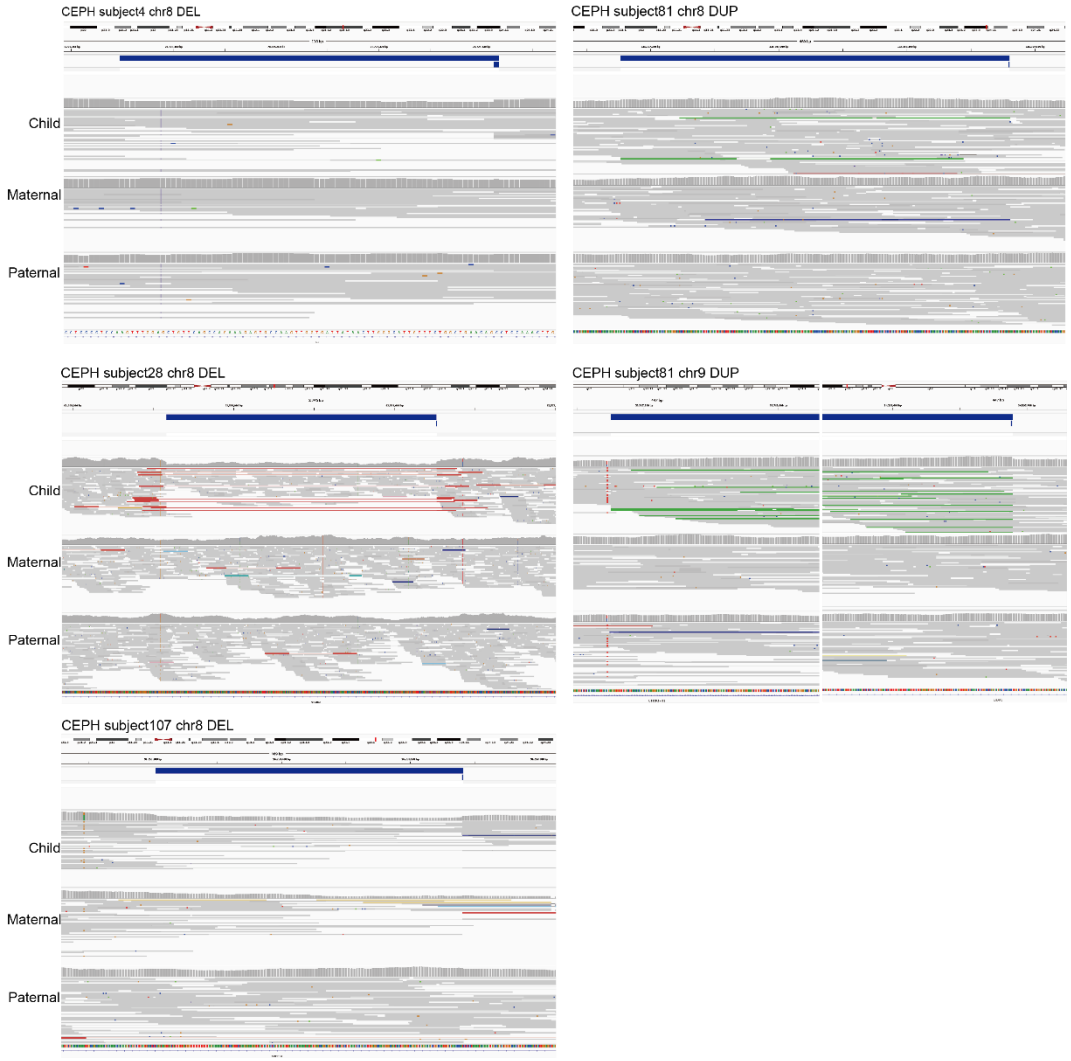


Figure 13. IGV images of visually validated dnSVs in ABC and CEPH cohorts. Top eight images correspond to ABC cohort. Bottom five images correspond to the CEPH cohort. In the image, each IGV track corresponds to the bam file of a trio in the order of child – maternal – paternal starting from the top. The dark blue box located in the top of each image represents the dnSV region called by ETCHING-trio.

In the RD cohort, 34 dnSVs were validated by visual inspection including 22 DELs and 12 DUPs. The rate of dnSVs in the RD cohort was therefore, one mutation for every 0.17 births. This sums up to a total of 47 dnSV across all three cohorts. Among the validated dnSVs, 31 events occurred in the genic region including 10 events overlapping the exonic region and 21 overlapping the intronic region of the corresponding gene. Additionally, since the RD cohort is composed of 100 quartets, each family includes two offspring. Therefore, the validated dnSVs were further examined to check whether they are shared between siblings. 11 events demonstrated evidence of dnSV being shared between siblings. These shared dnSVs may be either false negative calls of parental germline variants or dnSVs that occurred during early developmental stage of parental germ cells. Further investigation would be necessary to discern these dnSVs. dnSV calls categorized into six categories are depicted in the alluvial plot below (Figure 14).

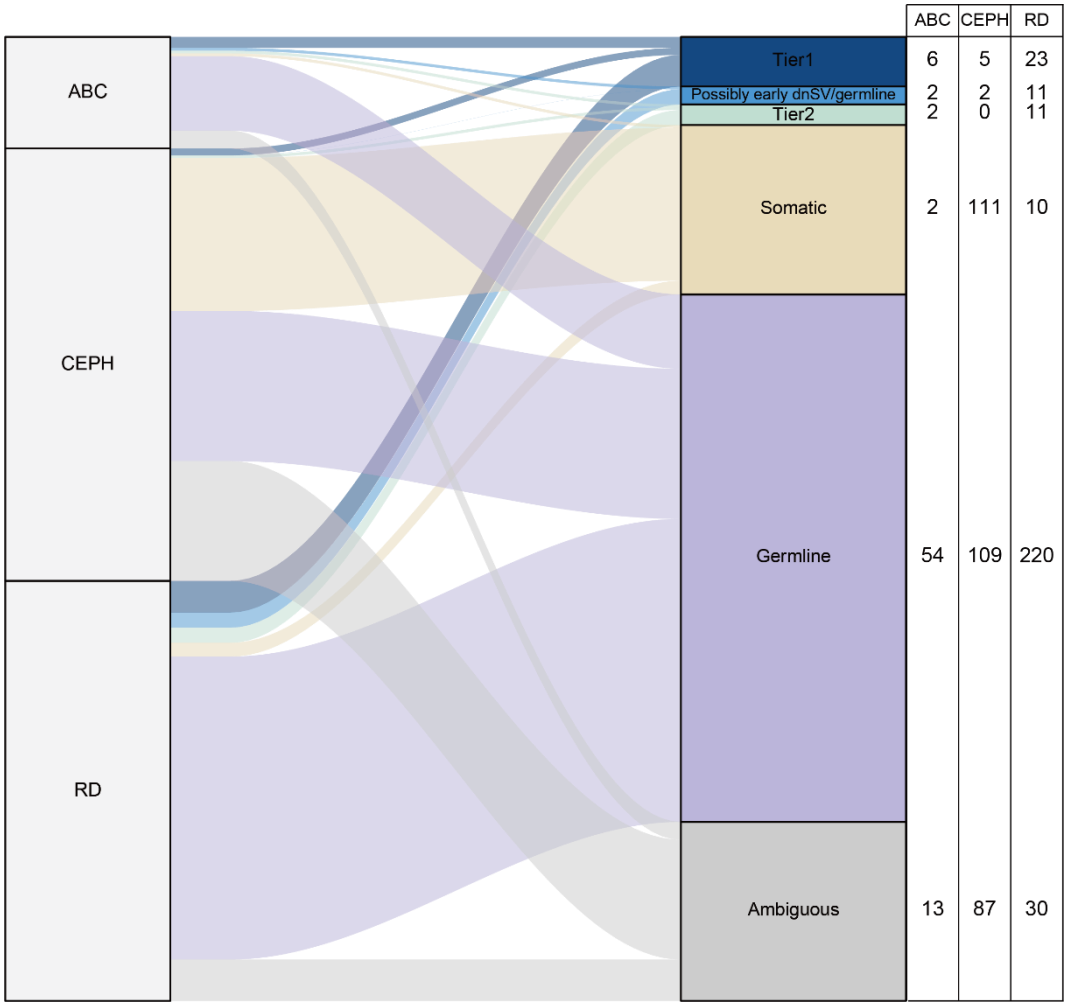


Figure 14. dnSV calls classified into six categories in the three cohorts. dnSV calls of each cohort are on the left side and their corresponding categories are on the right side of the plot. The number of dnSVs for each category is listed in the right side of the plot.

Previous studies have reported a positive correlation between paternal age and the presence of DNMs [36, 37]. In order to elucidate the parental age effect on dnSV occurrence, individuals from the ABC and CEPH cohorts were divided into two groups; those with at least one dnSV and those without. Parental age at pregnancy for the ABC cohort and the age at birth for the CEPH cohort was compared between the two groups

using a one-sided Wilcoxon rank-sum test. There was no significant difference between the two groups possibly due to the limited sample size. It is anticipated that an expanded cohort size would elucidate a more deliberate features associated with dnSV (Figure 15).

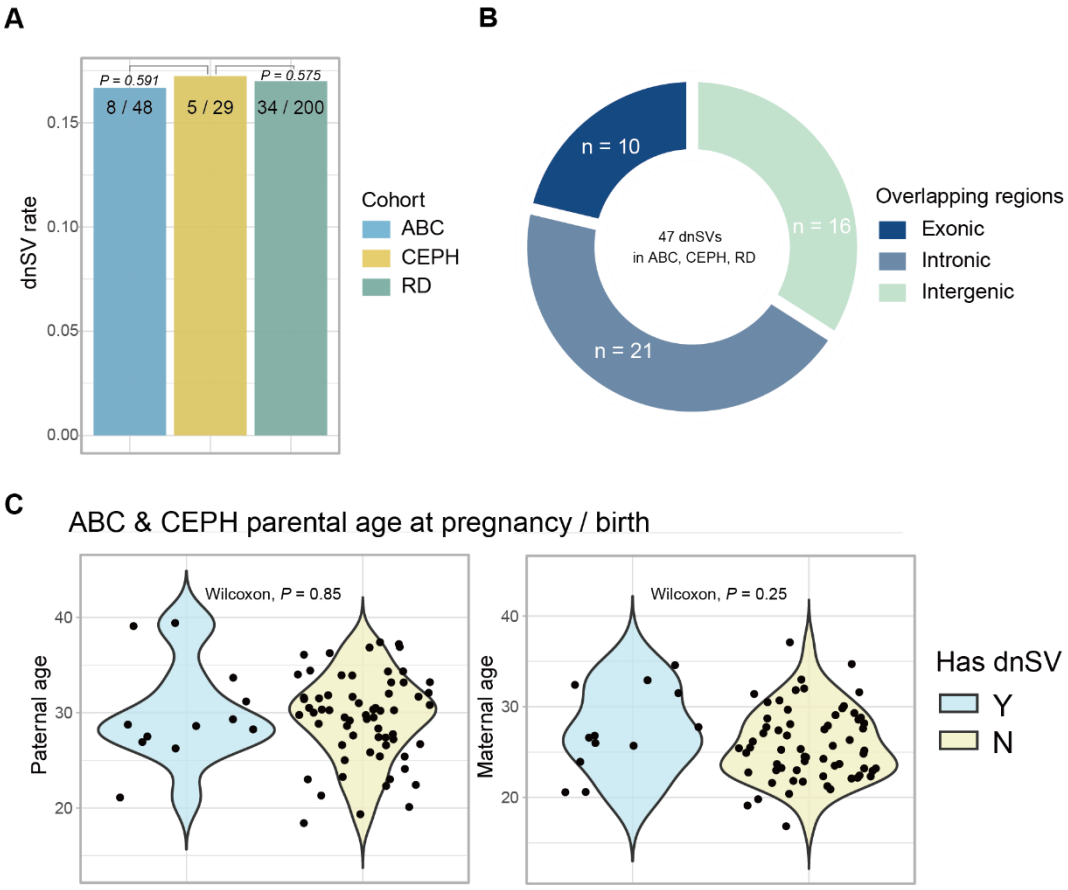


Figure 15. Characteristics of visually validated dnSVs in three cohorts. (A) Comparison of dnSV rates in three cohorts was performed by Fisher's exact test. (B) Number of dnSVs categorized by the genomic locus harboring the event. (C) Parental age effect on dnSV compared by one-sided Wilcoxon rank-sum test for increased paternal (left) or maternal (right) age.

IV. Discussions

This study presents the development of a novel pipeline to efficiently identify dnSVs using a *k*-mer-based approach, which was subsequently applied to three large cohort datasets. ETCHING-trio leverages a *k*-mer filtration strategy to facilitate the selective collection of child-specific variants. Moreover, a strategy to estimate variant allele frequency (VAF) was developed to filter out potential somatic SV contamination. *k*-mer frequency was used to approximate the read depth of coverage and a Random forest regression model was developed to predict variant supporting discordant read pairs given the split read, SV length, and an estimated value of read coverage. This estimation strategy verified the availability of utilizing *k*-mer frequency to reconstruct the sequencing read depth. The results indicated a high degree of accuracy in predicting depth of coverage and discordant read pairs. One of the primary strengths of this pipeline is its capacity to selectively collect relevant sequencing reads that support child-specific variants. This results in a substantial reduction in the number of erroneously called dnSVs, a reduction in running time and computational costs to find candidate dnSVs. Moreover, the calculation of VAF on SVs in short-read sequencing data broadens the opportunity to discern somatic and germline SVs, thereby reducing the number of false positive calls. This method showed enhanced performance in both simulated data and HG002 trio data when compared to conventionally used SV callers, thereby substantiating the efficacy of identifying dnSVs. Applying this method to three large cohort datasets yielded a total of 47 dnSVs. Among them, 31 dnSVs were located on the genic region, suggesting the possibility that dnSVs may be involved in phenotypic variation and diseases. The observed dnSV rate in each cohort was comparable to that reported in previous studies. However, no significant association was observed between

parental age and the presence of dnSVs. This may be attributed to the limited sample size, and further analysis of these cohorts is expected to confirm a more valid result. Furthermore, additional validation of the dnSVs identified in these cohorts should be performed using additional platforms such as long-read sequencing or PCR validation. The results of the identified dnSVs would be further consolidated by cross-validating them with multi-platform data.

Nevertheless, several limitations of this study remain. Firstly, although the estimation of VAF showed improved performance in distinguishing germline and somatic variants, it still exhibited a considerable number of false calls in three cohort datasets. This illustrates the possibility of somatic SVs being incorrectly called as dnSV or true dnSVs being overlooked by the VAF filter. Further refinement of the estimation model should be considered to improve the overall performance. Secondly, the absence of a high-confidence benchmark set for dnSVs precluded the performance being validated using various real data. The performance of ETCHING-trio was validated using several simulated data and a real data from the Ashkenazim Jewish trio, which is the only trio dataset that possesses high-confidence benchmark SVs available. However, a single dataset is insufficient for validating the performance of the method developed. To address this limitation, an artificial benchmark genome was constructed to include dnSVs from various individuals. As this data has the power of analyzing multiple samples with a single trio genome, additional data with high-confidence dnSVs should be further applied to cross-validate the performance of ETCHING-trio.

This study is currently in the process of identifying a simple rate of dnSVs across several cohorts. Additionally, the results suggest the potential association of dnSVs with disease based on its genomic location. Nevertheless, further analysis is necessary to

unveil the true nature of dnSVs in the human genome. One additional analysis that remains as a further study is to investigate the sequence similarity context of each dnSV's breakpoint in order to elucidate the causative mechanism of dnSV. This method and further study would broaden our understanding of the nature of dnSVs and facilitate further research to elucidate the biological mechanism and impact of dnSVs.

References

1. Seplyarskiy, Vladimir B., and Shamil Sunyaev. "The origin of human mutation in light of genomic data." Nature Reviews Genetics 22.10 (2021): 672-686.
2. Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. "Structural variation in the sequencing era." Nature Reviews Genetics 21.3 (2020): 171-189.
3. Weischenfeldt, Joachim, et al. "Phenotypic impact of genomic structural variation: insights from and for human disease." Nature Reviews Genetics 14.2 (2013): 125-138.
4. Veltman, Joris A., and Han G. Brunner. "De novo mutations in human genetic disease." Nature Reviews Genetics 13.8 (2012): 565-575.
5. Sasani, Thomas A., et al. "Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation." Elife 8 (2019): e46922.
6. Jónsson, Hákon, et al. "Parental influence on human germline de novo mutations in 1,548 trios from Iceland." Nature 549.7673 (2017): 519-522.
7. Werling, Donna M., et al. "An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder." Nature genetics 50.5 (2018): 727-736.
8. Sebat, Jonathan, et al. "Strong association of de novo copy number mutations with autism." Science 316.5823 (2007): 445-449.
9. Brandler, William M., et al. "Paternally inherited cis-regulatory structural variants are associated with autism." Science 360.6386 (2018): 327-331.
10. Collins, Ryan L., et al. "A structural variation reference for medical and population genetics." Nature 581.7809 (2020): 444-451.
11. Jung, Hyunchul, et al. "Deciphering the role of germline complex de novo structural variations in rare disorders." bioRxiv (2024): 2024-04.
12. Rausch, Tobias, et al. "DELLY: structural variant discovery by integrated paired-end and split-read analysis." Bioinformatics 28.18 (2012): i333-i339.
13. Layer, Ryan M., et al. "LUMPY: a probabilistic framework for structural variant discovery." Genome biology 15 (2014): 1-19.
14. Chen, Xiaoyu, et al. "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications." Bioinformatics 32.8 (2016): 1220-1222.
15. Maron, Jill L., et al. "Rapid whole-genomic sequencing and a targeted neonatal gene panel in infants with a suspected genetic disorder." Jama 330.2 (2023): 161-169.
16. Owen, Mallory J., et al. "Rapid sequencing-based diagnosis of thiamine metabolism dysfunction syndrome." New England Journal of Medicine 384.22 (2021): 2159-2161.
17. Zook, Justin M., et al. "A robust benchmark for detection of germline large deletions and insertions." Nature biotechnology 38.11 (2020): 1347-1355.
18. Robinson, James T., et al. "Integrative genomics viewer." *Nature biotechnology* 29.1 (2011): 24-26.
19. Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data." (2010): 1-1.

20. Joshi, N. A., and J. N. Fass. "Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files." (2011).
21. Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *bioinformatics*25.14 (2009): 1754-1760
22. http://broadinstitute.github.io/picard
23. Li, Heng, et al. "The sequence alignment/map format and SAMtools. " *bioinformatics* 25.16 (2009): 2078-2079.
24. McKenna, Aaron, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome research* 20.9 (2010): 1297-1303.
25. Sohn, Jang-il, et al. "Ultrafast prediction of somatic structural variations by filtering out reads matched to pan-genome k-mer sets." *Nature Biomedical Engineering* 7.7 (2023): 853-866.
26. Kokot, Marek, Maciej Długosz, and Sebastian Deorowicz. "KMC 3: counting and manipulating k-mer statistics." *Bioinformatics* 33.17 (2017): 2759-2761.
27. Pockrandt, Christopher, et al. "GenMap: ultra-fast computation of genome mappability." *Bioinformatics* 36.12 (2020): 3687-3692.
28. Smit, A. F. A., R. Hubley, and P. Green. "RepeatMasker Open-4.0. 2013–2015." (2015): 289-300.
29. Krusche, Peter, et al. "Best practices for benchmarking germline small-variant calls in human genomes." *Nature biotechnology* 37.5 (2019): 555-560.
30. Henriksen, Rasmus Amund, Lei Zhao, and Thorfinn Sand Korneliussen. "NGSNGS: next-generation simulator for next-generation sequencing data." *Bioinformatics* 39.1 (2023): btad041.
31. Ewing, Adam D., et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection." *Nature methods* 12.7 (2015): 623-630.
32. Rausch, Tobias, et al. "DELLY: structural variant discovery by integrated paired-end and split-read analysis." *Bioinformatics*28.18 (2012): i333-i339.
33. Layer, Ryan M., et al. "LUMPY: a probabilistic framework for structural variant discovery." *Genome biology* 15 (2014): 1-19.
34. Chen, Xiaoyu, et al. "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications." *Bioinformatics* 32.8 (2016): 1220-1222.
35. Wala, Jeremiah A., et al. "SvABA: genome-wide detection of structural variants and indels by local assembly." *Genome research* 28.4 (2018): 581-591.
36. Kloosterman, Wigard P., et al. "Characteristics of de novo structural changes in the human genome." Genome research 25.6 (2015): 792-801.
37. Belyeu, Jonathan R., et al. "De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families." The American Journal of Human Genetics 108.4 (2021): 597-607.

국문요지

k-mer 기반의 접근법을 활용한 드노보 구조변이 분석

김현우

자연과학대학 생명과학과

한양대학교 대학원

고효율 전장 유전체 시퀀싱은 개인의 유전체 구성을 포괄적으로 분석할 수 있게 함으로써 유전체학 분야에 혁명을 일으켰다. 최근 시퀀싱 기술의 발전은 질병 예후를 예측하고 임상 전략을 수립하는 데 있어 유전적 변이를 식별하는 것이 매우 중요하다는 것을 보여준 바 있다. 방대한 인간 유전체 변이 중 구조변이는 유전체 다양성의 중요한 구성 요소이며 신경 발달 장애와 암을 포함한 다양한 인간 질병에 관여하는 것이 선행 연구를 통해 알려져 있다. 다양한 유전체 변이 중 한 종류인 드노보 구조변이는 트리오 상황에서 한 개인에게는 존재하지만 부모에게는 없는 구조변이를 의미하며 이를 식별하는 것은 질병 감수성 및 발달 장애의 유전적 기반을 이해하는 데 매우 중요하다. 하지만 시퀀싱 기술의 발전에도 불구하고 인간 유전체의 복잡성과 기술적 한계로 인해 드노보 구조변이를 정확하게 식별하는 것은 여전히 어려운 과제로 남아있는 상황이다. 현재의 구조변이 검출 방법은 전장 시퀀싱 데이터 전체를 참조 유전체에 정렬하는 과정과 복잡한 전처리 단계를 거쳐 유전변이 식별을 위한 품질 관리된 전장 유전체 데이터를 생성하는 전략을 채택하고

있다. 그러나 이 전략은 전장 유전체 데이터의 1% 미만이 유전자 변이에 대한 정보를 포함하고 있다는 사실을 고려할 때 불필요한 비정보성 데이터를 분석하는 것이므로 상당히 비효율적이다. 불필요한 비정보성 데이터를 모두 분석하게 된다면 전산 자원의 낭비와 연구 시간의 손실로 이어지게 된다.

해당 논문에서는 구조변이 대립유전자 빈도 추정과 함께 *k*-mer 기반의 필터링 접근법을 활용하여 드노보 구조변이를 선택적으로 발견하는 새로운 전략을 제시한다. 이 전략의 효과를 검증하기 위해 건강한 개인, 한국인 원폭피해자 코호트, 한국인 희귀 질환 환자 코호트의 대규모 전장 유전체 데이터를 사용하였고 높은 민감도와 특이도로 드노보 구조변이를 정확하게 찾아낼 수 있는 해당 접근법의 능력을 입증했다. 이 논문을 통해 위양성 드노보 구조변이 호출을 추가 검증하는 데 드는 연구 시간과 연구자의 노력을 최소화하는 강력하고 실용적인 전략을 제시할 수 있을 것으로 기대한다. 이 방법은 드노보 구조변이를 빠르고 효율적으로 규명함으로써 정밀 의학 및 유전체 진단을 위한 표적 전략 개발을 목표로 하는 향후 연구의 토대를 마련할 수 있을 것이다.

Acknowledgement

3 년 반 동안의 학위 과정 동안 도움을 주신 많은 분들께 감사의 말을 전하며 논문을 마칩니다.

항상 지도해 주시며 바른길로 이끌어 주신 남진우 교수님께 깊은 감사의 말씀을 전합니다. 학위 과정 동안의 많은 방황과 흔들림에도 믿어 주시고 배려해 주신 덕분에 지금 이 자리까지 무사히 올 수 있었습니다.

학위 논문 심사를 맡아 주신 김헌석 교수님과 장기원 박사님께 깊이 감사드립니다. 심사위원님들의 아낌없는 조언과 과학적 통찰로 제 연구를 더욱 발전시켜 학위 논문을 마무리할 수 있었습니다.

함께 연구하며 아낌없는 조언과 도움, 응원을 해주신 연구실 동료분들께도 감사드립니다. 학위 과정 동안 만났던 많은 선배와 후배들, 손 박사님, 보현이형, 민학이형, 경우형, 경태형, 서원 누나, 석주형, 상호형, 효선 누나, 성진이형, 은경 누나, 우희 누나, 한지, 지훈, 가영, 준섭, 한솔, 민욱이형, 가문, Ngoc, 승은, 현석, 솔빈, Liam, Anna, 성우, 예원에게 감사의 말을 전하며 항상 도움 주시는 동은 선생님과 주연 선생님께도 감사드립니다. 덕분에 많이 배우고 성장했습니다.

학위 기간 동안 곁에서 지지하고 응원해 준 상현, 29 대 자연대 학생회 집부 친구들, 원준, 정원, 그리고 대일에게 감사의 말을 전합니다.

마지막으로 언제나 든든한 버팀목이 되어주는 사랑하는 가족들, 성일, 효숙, 대환에게 큰 사랑과 감사를 전합니다.

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

MAY 11, 2024

Degree :	Master
Department :	DEPARTMENT OF LIFE SCIENCE
Thesis Supervisor :	Nam, Jin-Wu
Name :	KIM Hyun Woo (Signature)

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2024년05월11일

학위명 : 석사

학과 : 생명과학과

지도교수 : 남진우

성명 : 김현우

한 양 대 학 교 대 학 원 장 귀 하