FINDEL: A DEEP LEARNING APPROACH TO EFFICIENT ARTIFACT REMOVAL FROM CANCER GENOMES

Denis Tan[1,4], Pengfei Zhou[2,4], Shaoting Zhang[2,3], VicPearly Wong[1], Jie Zhang[2,3,*], and Edwin Long[1,*]

[1]New Silkroutes Group Ltd, 119962, Singapore.
[2]SenseTime Research, Shanghai 200233, China.
[3]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China.
[4]These authors contributed equally.
[*]Corresponding authors: zhangjie1@sensetime.com, ai.genomics@healthsciences.sg

Abstract

Next-generation sequencing technologies have increased sequencing throughput by 100-1000 folds and subsequently reduced the cost of sequencing a human genome to approximately US$1,000. However, the existence of sequencing artifacts can cause erroneous identification of variants and adversely impact the downstream analyses. Currently, the manual inspection of variants for additional refinement is still necessary for high-quality variant calls. The inspection is usually done on large binary alignment map (BAM) files which consume a huge amount of labor and time. It also suffers from a lack of standardization and reproducibility. Here we show that the use of mutational signatures coupled with deep learning can replace the current standards in the bioinformatics workflow. This software, called FINDEL, can efficiently remove sequencing artifacts from cancer samples. It queries the variant call format file which is much more compact than BAM files. The software automates the variant refinement process and produces high-quality variant calls.

***K*eywords** bioinformatics software, cancer genomics, deep learning, machine learning, mutational signatures, sequencing artifacts, somatic mutations, variant refinement

1 Introduction

Since the discovery of the DNA structure in 1953 [1], significant progress in genetic understanding has been made over the past few decades. The advent of DNA sequencing instruments made possible the Human Genome Project (HGP) which was a global research initiative in 1990 to sequence the entire human genome consisting of 3 billion nucleotide base pairs [2, 3]. The international and cross-disciplinary nature of HGP has popularized the idea of open-source software and data accessibility where scientists and researchers all over the world have equal access to the ever-increasing repository of human genomes. The availability of individual genome data through cloud infrastructure and technologies has revolutionized the field of precision medicine where scientists can mine valuable information and recommend personalized treatment [4]. This can greatly improve the healthcare standards of the current and future generations.

The completion of the HGP has led to higher demand for sequencing technologies and datasets to solve more sophisticated biological problems. However, those efforts were hampered by the relatively low throughput and high costs of sequencing at that time [5]. This was changed in the mid-2000s with the invention of the high-throughput sequencing platform which resulted in a 50,000 times decrease in the cost of sequencing the entire human genome [6]. The promising results led to the term “next-generation sequencing” (NGS) which refers to platforms capable of massively parallel sequencing technologies that enable ultra-high throughput, scalability, and speed [7]. NGS technologies have increased sequencing throughput by 100-1000 folds [8] and subsequently reduced the cost of sequencing a human genome to approximately US$1,000 . These advances have made possible the use of sequencing technologies in clinical settings [6]. NGS technologies have found large-scale usage in de novo sequencing [9], mapping of diseases [10], quantifying expression levels using RNA sequencing [11, 12, 13], and conducting population genetic studies [14, 15, 16].

NGS technologies generate millions to billions of short-read sequences with read lengths of around 75-300 base pairs (bp) . More advanced NGS technologies (PacBio, Nanopore, 10x Genomics) are capable of much longer read sequences of more than 10 kilobases [17]. NGS technologies, while cheap and fast, are not flawless. The sequencing procedure is just the initial step of a typical bioinformatics pipeline that spans multiple phases. The raw sequence reads produced by the sequencing machine is stored in a FASTQ or unaligned Binary Alignment Map (uBAM) format. These formats are text-based which store information on the sequence reads such as read identifiers and base quality scores. Next, the reads are aligned to a reference genome and the related metadata are stored in either a Sequence Alignment Mapping (SAM), Binary Alignment Mapping (BAM), or CRAM file formats [18]. These files contain alignment characteristics such as matches, mismatches, and gaps represented in the Concise Idiosyncratic Gapped Alignment Report (CIGAR) format [19]. The BAM or equivalent files are consumed downstream through variant calling algorithms to identify genetic mutations such as single nucleotide polymorphisms (SNPs) consisting of transversions and transitions, insertions and deletions (InDels), and tumor mutation burden [20, 21]. The identified variants are stored in a tab-delimited text file format commonly known as the Variant Call Format (VCF).

Due to the downstream implications of NGS data, the inaccurate reads can cause erroneous identification of variants, specifically false positives (FP) and false negatives (FN). NGS data are susceptible to errors due to a myriad of factors such as base-calling and alignment errors [22]. This is further exacerbated by the lack of standardization in bioinformatics pipelines which are crucial to obtaining the correct data interpretation and clinical insights [23]. Errors in variant calling can result in missed detection of mutations which can be disastrous. Mutations are the cause of several diseases especially cancer [23].

A major shortfall of NGS technology is the frequent incorrect scoring of bases attributed to the existence of artifacts during the sample preparation and sequencing stage [24]. Heterogeneous mixtures can suffer from amplification bias during the polymerase chain reaction (PCR) process which produces skewed populations [25]. Other types of polymerase mistakes during the PCR stage such as base misincorporations and rearrangements because of template switching can also cause inaccurate variant calls. Cluster amplification, sequencing cycles, and image analysis are also prone to mistakes contributing to roughly 0.1-1% of bases being erroneously called [26]. These artifacts present a challenge for calling rare genetic variants as deep sequencing is ineffective when the base call error rate is high. This effectively limits the application of NGS in fields such as metagenomics [27, 28] , forensics [27], and human genetics [29, 30]. The accuracy demands of some clinical applications are even higher where the base calling error rate has to fall below 1 in 10,000. Examples of these clinical applications include detection of circulating tumor DNA [31], monitoring of response to chemotherapy using personalized tumor biomarkers [32], and prenatal screening for fetal aneuploidy [33, 34]. Due to these artifacts, the common NGS technologies suffer from a base-calling error rate of around 1 in 100 which severely falls short of the standard required by these clinical applications.

Automated pipelines to facilitate reads-to-variants workflow are widely available in this day and age, a prominent one being the Genome Analysis Toolkit (GATK) Best Practices Workflows which contain step-by-step recommendations for conducting variant discovery analysis using NGS data. These pipelines have built-in filters to remove false variant calls from sequencing errors, read misalignments, and other types of errors. However, there is little support for the efficient removal of artifacts due to variant caller inaccuracies [35]. According to the Association for Molecular Pathology (AMP) guidelines for interpretation and annotation of somatic variation, it is crucial to perform additional refinement of somatic variants to remove variant caller inaccuracies as these can result in suboptimal patient management and therapeutic opportunities [36, 37]. As of now, the standard practice for somatic variant refinement is to manually inspect the variants and this is usually performed by trained personnel such as bioinformaticians. Manual inspection is useful for incorporating domain knowledge commonly excluded by automated variant callers such as Mutect [38], SomaticSniper [39], Strelka [40], and VarScan2 [41]. They can detect inaccurate variant calls due to amplification bias of small fragments, errors in sequencing reads, and poor alignment in certain areas [35]. Efforts have been made to develop automated methods for variant refinement but progress is slow due to computational limitations. Manual inspection of variants for additional refinement is still necessary for high-quality variant calls and subsequent downstream analyses [42].

Although manual inspection of variants has been utilized for several years in clinical diagnostic and molecular pathology applications [43, 44, 45], somatic variant refinement strategies are not documented extensively and reported minimally by studies involving postprocessing of automated variant calls [45, 46, 47, 48, 49]. These results in a lack of standardized protocol for somatic variant refinement, increased variability of work performed by different labs, and difficulty in reproducing the work of others [49]. Additionally, manual inspections are usually conducted on BAM files using genomic viewers such as Integrative Genomics Viewer (IGV) [50, 51], Savant [52], Trackster [53], and BamView [54]. These BAM files are extremely large and require sophisticated storage solutions. This is because the format stores data on a per-read basis and the space requirement grows almost linearly with the number of reads [55]. A BAM file for a 30x whole genome requires about 80-90 gigabytes of storage. A lab or research institution which handles around 1000 samples needs to secure approximately 80 terabytes of disk space [56]. The sheer size of these BAM files impedes the progress of the manual inspection process which will require more labor and time. A more efficient and scalable variant refinement solution with light processing and storage requirements while eliminating variability in the downstream analyses needs to be developed.

A promising area of research beneficial to variant refinement lies in the study of mutational signatures. Mutational signatures have been linked to mutational processes which drive somatic mutations in cancer genomes [57]. The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium [58] of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) analyzed 84,729,690 somatic mutations identified from 4,645 whole-genome and 19,184 exome sequences that cover a wide range of cancer types and generated many mutational signatures [59]. The mutational signatures consist of 49 single-base-substitution (SBS), 11 doublet-base-substitution (DBS), 4 clustered-base-substitution, and 17 small insertion-and-deletion (Indel) signatures. Further classification of each mutation type was performed whereby SBSs consisted of 96 classes , DBSs consisted of 78 classes , and Indels consisted of 83 classes . These mutational signatures can be found in the Catalogue of Somatic Mutations in Cancer (COSMIC) database [60]. The PCAWG study also discovered that DNA sequencing artifacts, analysis artifacts, technical artifacts, variability in NGS technologies, and usage of different variant calling methods can generate characteristic mutational signatures as well [59]. These discoveries open up the possibility of using mutational signatures to filter artifact-mediated variant calls from true variants.

Our team proposes FINDEL (Find & Delete), a variant refinement software written in Python programming language that removes sequencing artifacts from human cancer samples using mutational signatures and deep learning. Along with each sample, FINDEL generates a Hypertext Markup Language (HTML) report and 2 separate VCF files containing refined and artifactual mutations respectively. FINDEL is light on computational requirements and is capable of running on a simple local machine. The performance and speed of FINDEL are validated using open-source datasets.

2 Methods

2.1 Overview of FINDEL

The main function of FINDEL is to automate the process of removing sequencing artifacts using mutational signatures from the COSMIC database coupled with a deep learning approach. The software is intended

for use in the variant refinement stage and is not meant to replace the variant calling algorithms. FINDEL can remove variability in the bioinformatics pipeline within the post variant calling phase. Currently, FINDEL focuses on artifacts involving single-base substitutions (SBSs) or also known as single-nucleotide polymorphisms (SNPs). SNPs form the majority of the genetic variations found in the human genome. In a typical human genome, more than 99.9% of variants are characterized as SNPs and short indels with SNPs being 6-7x higher in frequency than short indels [61]. SNPs are also responsible for the onset of multiple cancer types including breast cancer [62], chronic lymphocytic leukemia [63, 64, 65, 66, 67, 68], neuroblastoma [69], gastric carcinogenesis [70, 71, 72, 73, 74, 75, 76], prostate cancer [77], and others [78, 79].

2.2 Identification of Single-Base Substitutions Mutational Signatures Linked to Sequencing Artifacts

The initial step in building the artifact removal algorithm is to identify the SBS mutational signatures that are linked to the sequencing artifacts. These can be found in the COSMIC database. A total of 18 SBS mutational signatures have been classified as possible sequencing artifacts: SBS27, SBS43, and SBS45-SBS60 . The remaining signatures either have valid proposed aetiologies or unknown causes. As of now, SBS signatures can be segregated into 96 different contexts. There are 6 possible base substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. Each base substitution can be further analyzed in its 5' and 3' nucleotide context forming a total of 96 trinucleotide contexts [80].

2.3 Input Requirements for Algorithm

FINDEL takes in a VCF file containing a single sample. PyVCF module in Python is used to parse the VCF file for further processing steps. The other input required by the algorithm is the reference sequence. The reference sequence is used to extract the mutation type and reference context.

2.4 Output Files

FINDEL outputs 2 separate VCF files containing the refined and artifactual mutations respectively. The refined VCF file can be used for downstream analyses.

2.5 Supervised Deep Learning

After the information from the mutational signatures are obtained, we propose a supervised deep neural network-based method to directly predict the mutation types from the VCF file. We devised the specific modules to deal with a variety of features in addition to the mutational signatures and utilized a complex network to fuse the diverse information for the mutation site identification. The deep learning architecture has a comprehensive embedding representation to enhance the algorithm performance. Upon training completion, FINDEL can conduct inference efficiently without the need for a customized variant refinement optimization process for every new cancer sample.

2.5.1 Data Preprocessing

Firstly, we constructed the features of mutation types. All point mutations are classified into one of the 96 SBS nucleotide contexts, represented by an index ranging from 0 to 95. Secondly, we parsed the site features. Specifically, we used the "INFO" and "SAMPLE" columns from the provided input VCF file. One-hot encoding and standard scaling were used to preprocess the discrete and continuous features respectively. Thirdly, we built context features. We selected 10 bases at the left and right of each mutation site (a total of 21 bases). Each base is represented by a number (T:0, C:1, A:2, G:3, others:4), and the context feature was represented as a 21-dimensional vector. Finally, we adopted the Hap.py tool to process the original annotation file as the ground-truth labels in this study.

2.5.2 Neural Network Architecture

Fig. 1 shows the proposed neural network architecture. We used a learnable embedding layer to map the 96-dimensional mutation type into a high-dimensional space. Each mutation type is converted into a point in the high-dimensional space, the coordinates of which constitute an embedding vector for the type. The parsed site features were combined and fed into a linear layer. The reference context is passed into another

embedding layer, converting the 21 bases in the mutation neighborhood into a two-dimensional context matrix. We used a two-layer 1D-convolution to model the multiple sequences and extract the backward and forward relationships in the context matrix. The context matrix was transformed into a vector.

After extracting the mutation types, site features, and context features, we concatenated the embedding vectors and fed them into a five-layer fully connected network. Each layer consisted of dropout linear transformation, ELU (Exponential Linear Unit) activation function, and highway module. The dropout was used at the beginning of each layer to randomly set the unit weights to 0, helping the model avoid overfitting specific features. Equations (1), Equations (2), and Equations (3) show the formulas for the linear transformation, activation function, and highway module respectively:

$$Linear_i(X) = W_i X + b_i \tag{1}$$

$$ELU(x) = \begin{cases} x, & x > 0 \\ \alpha(e^x - 1), & x \leq 0 \end{cases} \tag{2}$$

$$H_{i+1} = H_i + ELU(Linear(H_i)) \tag{3}$$

where W and b are learnable parameters. The ELU hyperparameter α controls the value at which the activation function saturates for negative inputs. The highway module connects the input of each layer to the end, solving the gradient vanishing problem caused by overly deep networks.

Equation (4) defines the cost function.

$$L = L_{BCE} + \gamma L_{L2} \tag{4}$$

L_{BCE} refers to the binary cross-entropy loss while γL_{L2} is the L2 regularization which adds a penalty term to reduce overfitting. We adopted the mini-batch gradient descent algorithm for training and each batch consisted of 5000 variants. The Adam optimizer was used to update the network weights. All hyperparameters were tuned using grid search and evaluated on the validation set. The GIAB consortium reference samples (HG001-HG007) were used for training and testing. Two different training methods were used. The first method built 1 model using 1 VCF file. Within each file, the SNPs were split such that 60% of them were part of the train set, 20% of them were part of the validation set, and 20% of them were part of the test set. The second method used all the SNPs from 1 VCF file and tested them on all the remaining files. For example, a model was trained on all the SNPs from the HG001 VCF file and evaluated on the SNPs from the HG002-HG007 files. Table I shows the final values chosen for the hyperparameters.

2.6 Evaluation Data Sources and Processing

The performance and speed of FINDEL are evaluated using high-quality open-source datasets. For the proof of concept training and validation, we will be using 4 different datasets containing a total of 33 VCF files.

2.6.1 HCC1954 Cell Line

This breast cancer cell line whole genome sequencing (WGS) data from a 61 years old Asian female was initiated on October 30 1995 and took around 4 months to establish [81]. We downloaded the VCF file and the benchmarking results from the International Cancer Genome Consortium Data Portal [82]. The reads were aligned to the hs37d5 reference genome. The benchmarking results containing the true positive and false positive variants were stored in 2 mutation annotation format (MAF) files which are tab-delimited text files with aggregated mutation information on a project level . No processing was required since the VCF file was provided directly. A total of 1 VCF file was obtained.

2.6.2 Formalin-Fixed Paraffin-Embedded (FFPE) and Fresh Frozen Whole Exome Sequencing (WES) Samples

This study was part of an attempt by Samsung Medical Center to compare the WES data between FFPE and fresh frozen samples. The datasets were obtained from the Sequence Read Archive (SRA) [83] with the accession number PRJNA301548 . The samples were sequenced using the Illumina HiSeq 2000 sequencing

system . We used the DNA sequencing data from the following 3 runs: SRR2911437, SRR2911438, and SRR2911453. The paired-end reads were aligned to the HG19 reference genome using the Burrows-Wheeler Alignment tool [84] with the Maximal Exact Match option, more commonly know as the BWA-MEM algorithm. After the reads were aligned, Mutect2 was chosen for the variant calling process. The tumor-only mode with default settings was used. A total of 3 VCF files was obtained.

2.6.3 Whole Exome Sequencing of Biological and Technical Replicates in Breast Cancer Samples

This study was conducted by the Memorial Sloan Kettering Cancer Center on 3 Feb 2016. The datasets were obtained from the SRA with the accession number SRP070662. The samples were sequenced using the Ion PGM sequencing system We used the DNA sequencing data from the following 22 runs: SRR3182418, SRR3182419, SRR3182421, SRR3182423, SRR3182424, SRR3182427, SRR3182429, SRR3182430, SRR3182431, SRR3182432, SRR3182436, SRR3182437, SRR3182440, SRR3182442, SRR3182444, SRR3182446, SRR3182447, SRR3182474, SRR3182475, SRR3182476, SRR3182477, and SRR3182478. The processing steps are the same as the FFPE and fresh frozen WES datasets. A total of 22 VCF files was obtained.

2.6.4 Genome in a Bottle (GIAB) Consortium Reference Samples

The GIAB consortium is jointly hosted by the National Institute of Standards and Technology (NIST) and the Joint Initiative for Metrology in Biology (JIMB) to develop the technical infrastructure and improve clinical insights from whole human genome sequencing. GIAB has a total of 7 human genomes: 1 pilot genome of Utah/European ancestry (NA12878/HG001) from the HapMap project , and 2 son/father/mother trios of Ashkenazi Jewish (HG002/HG003/HG004) and Han Chinese (HG005/HG006/HG007) ancestry from the Personal Genome Project. Each human genome is also accompanied by benchmark variant calls and regions to validate variant calling pipelines. These are considered the gold standard benchmarks within the bioinformatics community. The processing steps are mostly the same as the FFPE and fresh frozen WES datasets with some differences. The b37 reference genome was used in place of the HG19 reference genome. Both originate from the GRCh37 reference genome with some minor differences. However, from a bioinformatics pipeline technical perspective, these variations of reference genomes are not directly interchangeable due to contig name changes. The discrepancies are further elaborated in .The change was due to the addition of the germline resource and the panel of normals (PoN) during the Mutect2 variant calling process. Both VCF files used the b37 reference genome. Therefore, the alignment phase required the use of the same reference genome as well. A total of 7 VCF files was obtained.

The HCC1954 cell line, FFPE and fresh frozen WES, and WES of biological and technical replicates in breast cancer datasets lack the labeling of true mutations. DL_v11 used an unsupervised training method with a negative objective function as the loss function. In the GIAB annotated dataset, we used chromosomes 1 to 19 as the training set, chromosomes 21 and 22 as the validation set, and chromosome 20 as the test set. We used cross-entropy as the loss function,

3 Results

FINDEL will be evaluated based on its performance and speed on the evaluation data. For performance evaluation, benchmark VCF files are required as they contain the ground truth variant calls. The datasets without benchmark VCF files will be used for speed evaluation. The following 5 metrics will be used for performance evaluation:

$$precision = TP/(TP + FP) \tag{5}$$

$$recall = TP/(TP + FN) \tag{6}$$

$$F1 = (2 * precision * recall)/(precision + recall) \tag{7}$$

$$accuracy = (TP + TN)/(TP + FN + TN + FP) \tag{8}$$

$$specificity = TN/(TN + FP) \tag{9}$$

where TP refers to true positives, TN refers to true negatives, FP refers to false positives, and FN refers to false negatives. Elapsed time, in either minutes or seconds, will be used for comparing the speed of the models. From now on, "r_artifact_removal" and "py_artifact_removal" will be used to refer to the R and Python model respectively before applying deep learning while "DL_v11" will be used to refer to the Python model after applying deep learning. These terms are mainly used in the comparison of the model results pre and post-deep learning. The software, which outputs additional analyses and VCF files beyond the final trained model, is still named FINDEL.

The initial version of the DL_v11 model was trained on the HCC1954 cell line dataset. Although both performance and speed results are shown in Table II, the main focus is on the difference in speed between the models written in Python vs R programming language. The elapsed time of py_artifact_removal is less than 3 minutes while the elapsed time of r_artifact_removal is 305 minutes. This is a significant speedup of more than 100 times while maintaining similar performance. The objective function value of py_artifact_removal is 80.34 while the objective function value of r_artifact_removal is 80.67. The higher the value, the better the performance. The difference is negligible in this case. Meanwhile, DL_v11 achieves a significantly higher objective function value of 89.21 with a training and inference time of 20 and less than 3 minutes respectively. However, the performance improvement based on this dataset alone is not robust enough. As seen from the table, there is very little labeled data available. Only 214 variants are labeled with 213 of them being classified as positive SNPs and only 1 negative SNP. From a supervised deep-learning perspective, this is an imbalanced classification problem and the performance metrics might be biased. Even though DL_v11 correctly predicted 211 out of 213 positive SNPs, it failed to predict the only negative SNP. It can maximize the objective function value simply by predicting all SNPs to be positive as they are the majority of the labels. This poses a problem when we want to minimize false positives. On the other hand, py_artifact_removal can detect the only negative SNP but misses much more positive SNPs, achieving an overall lower objective function value. Therefore, the other labeled datasets are used to provide a more robust estimate of the model performances. The key takeaway here is the superior speed achieved by the models written in Python as compared to R. The subsequent evaluation comparison will be made between the 2 python models: py_artifact_removal and DL_v11.

The subsequent versions of the DL_v11 models were trained on the GIAB consortium reference samples (HG001-HG007). The following results were based on the first method of training using the same VCF file for both training and testing. The performance comparison between py_artifact_removal and DL_v11 based on the 5 metrics is given in Appendix A. DL_v11 performed better on all the metrics (F1, accuracy, precision, specificity) except recall.

The performance comparison within py_artifact_removal is given in Appendix B. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish (HG002/HG003/HG004) ancestry than the Utah/European (HG001) and Han Chinese (HG005/HG006/HG007) ancestries for F1, accuracy, and precision. For recall, similar scores were achieved on the human genomes from all the ancestries. Results were mixed for specificity. The performance comparison within DL_v11 is given in Appendix C. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish ancestry than the Utah/European and Han ancestries for F1, accuracy, and precision, and recall. Results were mixed for specificity.

The next set of results used the second training method where the train and test set consisted of different human genomes. The F1, accuracy, and precision performance comparisons within DL_v11 for different train sets are given in Appendix D, E, and F respectively. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish ancestry than the Utah/European and Han Chinese ancestries regardless of the train set used. The recall performance comparison is given in Appendix G. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish and Utah/European ancestries than the Han Chinese ancestry regardless of the train set used. The human genome from the Utah/European ancestry (HG001) had the best recall for all the train sets. The specificity performance comparison is given in Appendix H. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish ancestry than the Han Chinese ancestry regardless of the train set used. The human genome from the Utah/European ancestry (HG001) had the worst specificity for all the train sets. The average performance comparison using different train sets is given in Appendix I. Generally, higher average scores on all 5 metrics were achieved when the human genomes from the Han Chinese ancestry were used as the train samples. The time comparison between the 3 models is given in Appendix J. All datasets used here are from the FFPE and fresh frozen WES samples. On both datasets, SRR2911438 and SRR2911453, a significantly

shorter time was used by the 2 models written in Python (py_artifact_removal and DL_v11) as compared to the model written in R (r_artifact_removal). The correlation between elapsed time and the number of SNPs for py_artifact_removal is given in Appendix K. All datasets used here are from the FFPE and fresh frozen WES samples and the breast cancer WES samples. A strong positive linear correlation can be observed. As the number of SNPs increases, the elapsed time taken by py_artifact_removal increases as well.

4 Discussion

FINDEL, a variant refinement software infused with deep learning, has demonstrated its ability to efficiently remove sequencing artifacts from cancer samples using mutational signatures and other features. We have evaluated its performance and speed on high-quality open-source datasets. The performance is largely boosted by the supervised deep learning approach in addition to the use of mutational signatures. False positives are greatly reduced through the variant refinement process. The algorithm speed is mostly due to the optimal choice of programming language and code refactoring. The overall software runtime is low as our approach queries VCF files instead of BAM files where the former has a much lower memory footprint. Moreover, the software can be easily run on a local machine. FINDEL eliminates the need for users to manually inspect BAM files using some form of genomics viewer program. This saves both labor and time, increasing the turnover of cancer sample interpretation. Doctors require less time for the clinical diagnosis of patients and can increase the range of therapeutic opportunities for them. More importantly, the automation and standardization of the variant refinement process increase reproducibility by different entities. The usage of mutational signatures to partition the unrefined mutation set into refined and artifactual mutation subsets was possible as various artifacts had characteristic mutational signatures that were different from the true variants. These mutational signatures are constantly updated and can be found in the COSMIC database. This essentially removes the need for bioinformaticians to manually inspect large BAM files and search for patterns that are representative of artifacts using domain knowledge. The manual inspection process is a major bottleneck impeding the progress of the downstream analyses after the variant calling process. Eliminating this laborious task greatly reduces the entire bioinformatics workflow time. As the research on mutational signatures become more comprehensive in the future, we expect the algorithm to further improve in performance as well. FINDEL only requires a user-provided VCF file and an SBS mutational signature data file. No manual preprocessing on the user part is required as long as the standard format is used. Subsequently, FINDEL outputs a VCF file representing the refined mutations. The infusion of supervised deep learning to FINDEL mainly aims to improve algorithmic performance and further reduce the number of false positives. It also utilizes the information from the VCF files beyond just the mutational signatures. Specifically, additional features are engineered from the “INFO” and “SAMPLE” columns of the VCF files. As seen from the results, the addition of supervised deep learning greatly improves the score of most of the metrics. The disadvantage is that the initial training of the neural network might take quite some time. However, subsequent usage of the model, aka inference, takes little time. The deep learning model only needs to be retrained if there is evidence of model degradation. Also, the performance of the model is affected by the quality of the data. Generally, higher scores were achieved on the human genomes from the Ashkenazi Jewish ancestry than the Utah/European and Han Chinese ancestries due to the former having higher sequencing depth.

5 Conclusions and Future Work

We have developed a deep learning-based bioinformatics software, FINDEL, that can efficiently remove sequencing artifacts from cancer samples using mutational signatures and other features. It eliminates the laborious process of manually inspecting large BAM files by querying much smaller VCF files. The algorithm automates and standardizes the variant refinement process, reducing required labor and time while increasing reproducibility. The user only needs to provide the input files and the software does the rest of the work. Within minutes, the unrefined VCF file containing the artifactual variants are removed and the refined VCF file can be used for subsequent downstream analyses. The potential of FINDEL to redefine the bioinformatics workflow through its superior performance and speed allows better patient management and therapeutic opportunities. As of now, the software mainly focuses on single nucleotide polymorphisms (SNPs). In the future, the scope can be broadened to insertions and deletions, copy number variations, and even structural variations. Overall, we show that the use of mutational signatures coupled with deep learning can replace the manual inspection process, potentially setting a new standard for the variant refinement process.

6 Data availability

The datasets can be downloaded at their original sites.

7 Code availability

The source codes are only available under research or commercial collaboration.

8 Web service

We set up a web service hosting our trained model.

9 Acknowledgements

Not applicable.

10 Author contributions

11 Competing interests

The authors declare no competing interests.

Fig. 1

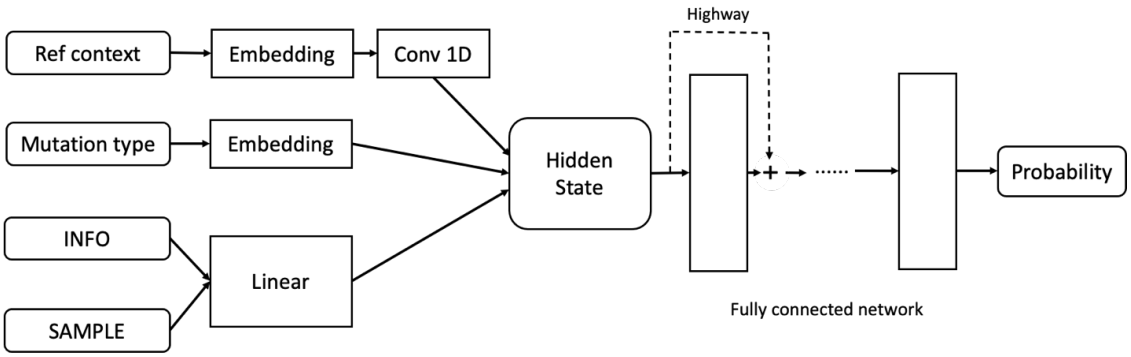

Fig. 1. Neural network architecture.

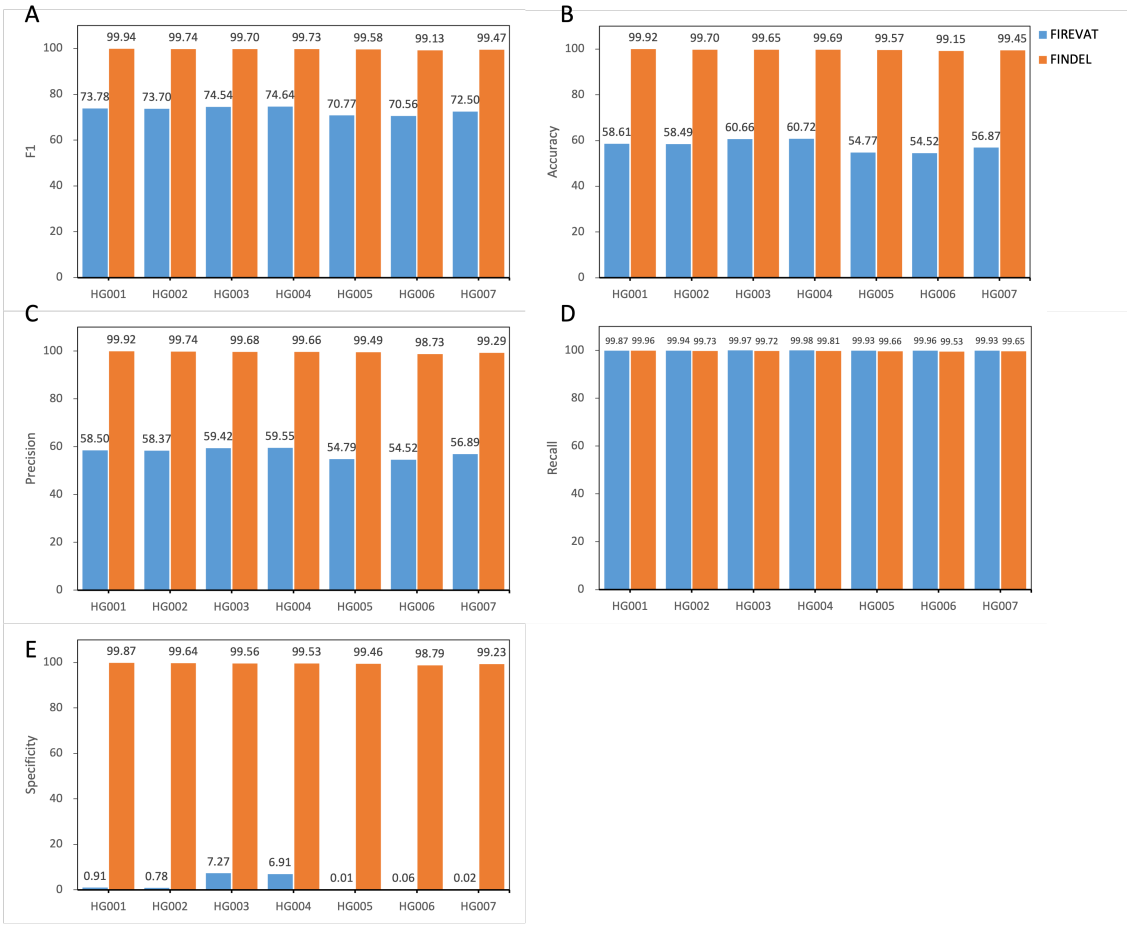

Fig. 2. Performance comparison between py_artifact_removal (before deep learning) and DL_v11 (after deep learning) based on 5 metrics. All values are in percentages unless otherwise indicated. All datasets used here are from the GIAB consortium reference samples (HG001-HG007). The PY_ARTIFACT_REMOVAL method is based on mutation signatures, which tends to produce positive results. The supervised learning-based DL- V11 method is able to accurately identify artificial mutations. In all four metrics except recall, dl-v11 has a huge improvement.

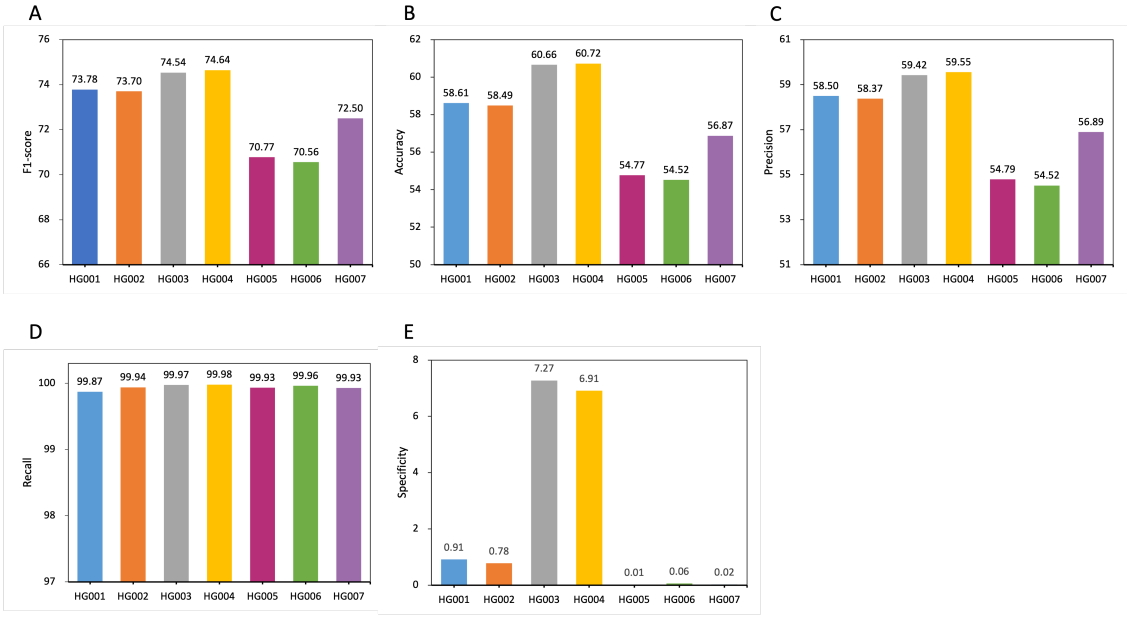

Fig. 3. Performance comparison within py_artifact_removal (before deep learning) based on 5 metrics. All values are in percentages unless otherwise indicated. All datasets used here are from the GIAB consortium reference samples (HG001-HG007). Different races will affect the performance of the model. PY_ARTIFACT_REMOVAL has better performance in Utah/European (HG001) and Ashkenazi Jewish (HG002/HG003/HG004) than Han Chinese (HG005/HG006/HG007) ancestries.

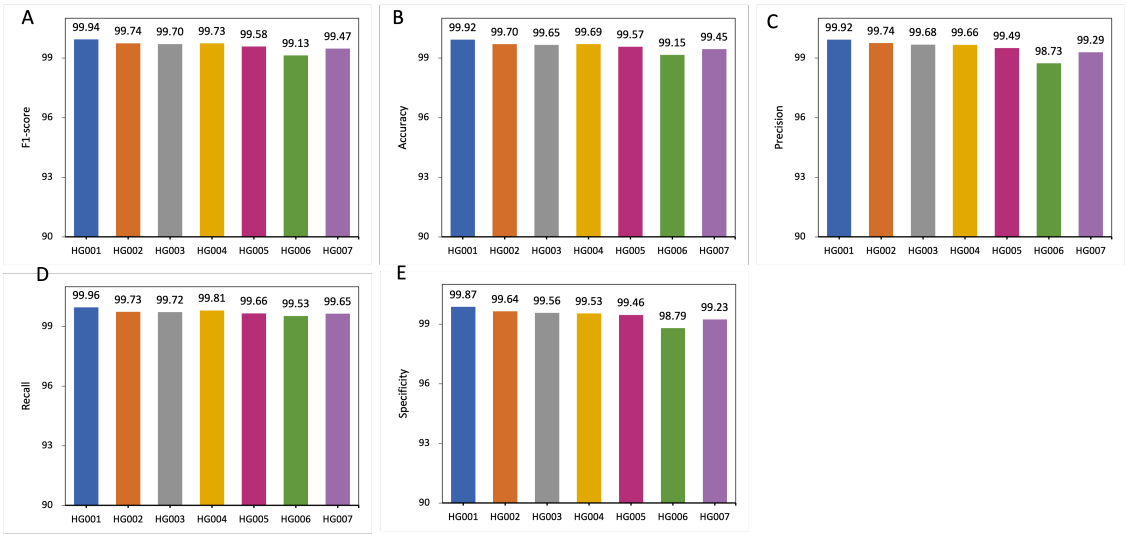

Fig. 4. Performance comparison within DL_v11 (after deep learning) based on 5 metrics. All values are in percentages unless otherwise indicated. All datasets used here are from the GIAB consortium reference samples (HG001-HG007). An independent model is trained for each sample. DL_v11 achieves the best performance on HG001, with an f1-score of 99.94. Lower scores were achieved on the other human genomes with the Han Chinese(HG006) ancestry genome performing the poorest.

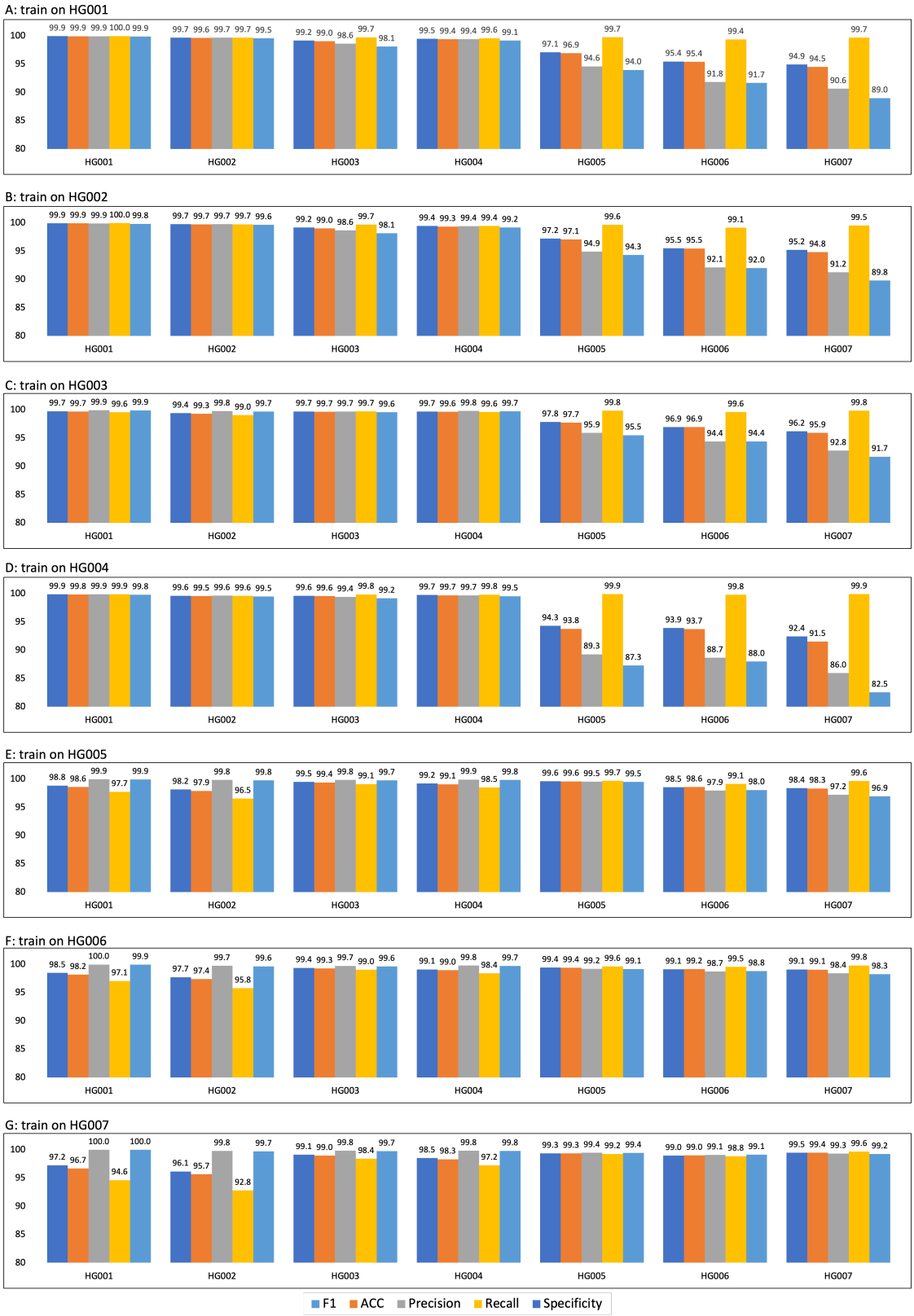

Fig. 5. Performance comparison within DL_v11 (after deep learning) for different trainsets. The model achieves better performance when the test and training data come from the same sample. When the training and test data come from different sources, the algorithm performance will be affected. For example, in A Barplot, the training data is chromosome 1 to 19 of HG001(Utah/European), and the test data is chromosome 20 of HG001 to HG007. On HG001 all five metrics achieve a performance of 99.9, but the model predicts more false-positive mutations on HG005/HG006/HG007 (Han Chinese) of different ethnicities, resulting in a lower Specificity

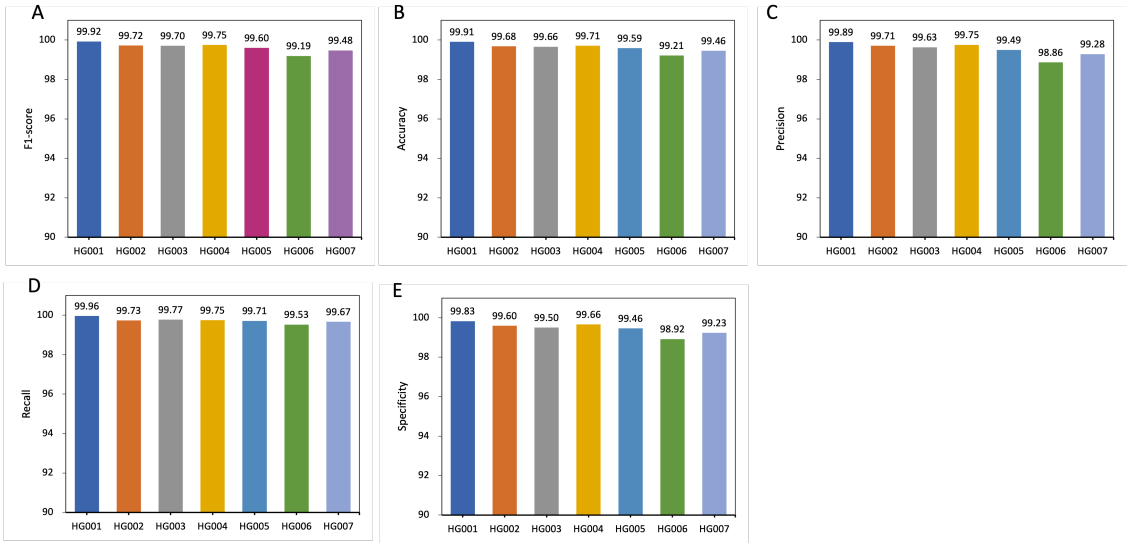

Fig. 6. Performance comparison within DL_v11 (after deep learning) for whole trainsets. Training using data from multiple samples was able to improve the robustness of the model and alleviate the effects of different sources of data on model performance. The DL_v11 model trained using chromosome 1 to 19 from HG001 to HG007, the model was able to achieve an f1-score of 99% on all samples.

References

[1] James D Watson and Francis HC Crick. The structure of dna. In *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press, 1953.

[2] Elaine R Mardis. Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303, 2013.

[3] Leroy Hood and Lee Rowen. The human genome project: big science transforms biology and medicine. *Genome medicine*, 5(9):1–8, 2013.

[4] Bartha Maria Knoppers, Jennifer R Harris, Anne Marie Tassé, Isabelle Budin-Ljøsne, Jane Kaye, Mylène Deschênes, and H Zawati Ma'n. Towards a data sharing code of conduct for international genomic research. *Genome Medicine*, 3(7):1–4, 2011.

[5] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[6] Kris A Wetterstrand. The cost of sequencing a human genome. *National human genome research institute*, 2016.

[7] HPJ Buermans and JT Den Dunnen. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941, 2014.

[8] Martin Kircher and Janet Kelso. High-throughput dna sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[9] Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, Yinqi Bai, et al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, 2010.

[10] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.

[11] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, 2008.

[12] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510, 2010.

[13] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[14] Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, et al. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, 2009.

[15] Yingrui Li, Nicolas Vinckenbosch, Geng Tian, Emilia Huerta-Sanchez, Tao Jiang, Hui Jiang, Anders Albrechtsen, Gitte Andersen, Hongzhi Cao, Thorfinn Korneliussen, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics*, 42(11):969–972, 2010.

[16] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

[17] Tuomo Mantere, Simone Kersten, and Alexander Hoischen. Long-read sequencing emerging in medical genetics. *Frontiers in genetics*, 10:426, 2019.

[18] Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane, and Ewan Birney. Efficient storage of high throughput dna sequencing data using reference-based compression. *Genome research*, 21(5):734–740, 2011.

[19] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[20] Somak Roy, William A LaFramboise, Yuri E Nikiforov, Marina N Nikiforova, Mark J Routbort, John Pfeifer, Rakesh Nagarajan, Alexis B Carter, and Liron Pantanowitz. Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. *Archives of pathology & laboratory medicine*, 140(9):958–975, 2016.

[21] Sabah Kadri. Advances in next-generation sequencing bioinformatics for clinical diagnostics: Taking precision oncology to the next level. *Advances in Molecular Pathology*, 1(1):149–166, 2018.

[22] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[23] Lawrence A Loeb. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Reviews Cancer*, 11(6):450–457, 2011.

[24] Travis C Glenn. Field guide to next-generation dna sequencers. *Molecular ecology resources*, 11(5):759–769, 2011.

[25] Takahiro Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (pcr). *Journal of bioscience and bioengineering*, 96(4):317–323, 2003.

[26] Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1, 2014.

[27] Béatrice Lecroq, Franck Lejzerowicz, Dipankar Bachar, Richard Christen, Philippe Esling, Loïc Baerlocher, Magne Østerås, Laurent Farinelli, and Jan Pawlowski. Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proceedings of the National Academy of Sciences*, 108(32):13177–13182, 2011.

[28] Rachel Mackelprang, Mark P Waldrop, Kristen M DeAngelis, Maude M David, Krystle L Chavarria, Steven J Blazewicz, Edward M Rubin, and Janet K Jansson. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, 480(7377):368–371, 2011.

[29] Todd E Druley, Francesco LM Vallania, Daniel J Wegner, Katherine E Varley, Olivia L Knowles, Jacqueline A Bonds, Sarah W Robison, Scott W Doniger, Aaron Hamvas, F Sessions Cole, et al. Quantification of rare allelic variants from pooled genomic dna. *Nature methods*, 6(4):263–265, 2009.

[30] Astrid A Out, Ivonne JHM van Minderhout, Jelle J Goeman, Yavuz Ariyurek, Stephan Ossowski, Korbinian Schneeberger, Detlef Weigel, Michiel van Galen, Peter EM Taschner, Carli MJ Tops, et al. Deep sequencing to reveal new variants in pooled dna samples. *Human mutation*, 30(12):1703–1712, 2009.

[31] Julia Beck, Howard B Urnovitz, William M Mitchell, and Ekkehard Schütz. Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. *Molecular cancer research*, 8(3):335–342, 2010.

[32] Rebecca J Leary, Isaac Kinde, Frank Diehl, Kerstin Schmidt, Chris Clouser, Cisilya Duncan, Alena Antipova, Clarence Lee, Kevin McKernan, M Francisco, et al. Development of personalized tumor biomarkers using massively parallel sequencing. *Science translational medicine*, 2(20):20ra14–20ra14, 2010.

[33] H Christina Fan, Yair J Blumenfeld, Usha Chitkara, Louanne Hudgins, and Stephen R Quake. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing dna from maternal blood. *Proceedings of the National Academy of Sciences*, 105(42):16266–16271, 2008.

[34] Rossa WK Chiu, Ranjit Akolekar, Yama WL Zheng, Tak Y Leung, Hao Sun, KC Allen Chan, Fiona MF Lun, Attie TJI Go, Elizabeth T Lau, William WK To, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma dna sequencing: large scale validity study. *Bmj*, 342, 2011.

[35] Erica K Barnell, Peter Ronning, Katie M Campbell, Kilannin Krysiak, Benjamin J Ainscough, Lana M Sheta, Shahil P Pema, Alina D Schmidt, Megan Richters, Kelsy C Cotto, et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genetics in Medicine*, 21(4):972–981, 2019.

[36] Somak Roy, Christopher Coldren, Arivarasan Karunamurthy, Nefize S Kip, Eric W Klee, Stephen E Lincoln, Annette Leon, Mrudula Pullambhatla, Robyn L Temple-Smolkin, Karl V Voelkerding, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. *The Journal of Molecular Diagnostics*, 20(1):4–27, 2018.

[37] Marilyn M Li, Michael Datto, Eric J Duncavage, Shashikant Kulkarni, Neal I Lindeman, Somak Roy, Apostolia M Tsimberidou, Cindy L Vnencak-Jones, Daynna J Wolff, Anas Younes, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, american society of clinical oncology, and college of american pathologists. *The Journal of molecular diagnostics*, 19(1):4–23, 2017.

[38] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–219, 2013.

[39] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, 2012.

[40] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

[41] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012.

[42] Elaine R Mardis. The $1,000 genome, the$100,000 analysis? *Genome medicine*, 2(11):1–3, 2010.

[43] Samuel P Strom. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer biology & medicine*, 13(1):3, 2016.

[44] Jihun Kim, Woong-Yang Park, Nayoung KD Kim, Se Jin Jang, Sung-Min Chun, Chang-Ohk Sung, Jene Choi, Young-Hyeh Ko, Yoon-La Choi, Hyo Sup Shim, et al. Good laboratory standards for clinical next-generation sequencing cancer panel tests. *Journal of pathology and translational medicine*, 51(3):191, 2017.

[45] Ramaswamy Govindan, Li Ding, Malachi Griffith, Janakiraman Subramanian, Nathan D Dees, Krishna L Kanchi, Christopher A Maher, Robert Fulton, Lucinda Fulton, John Wallis, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6):1121–1134, 2012.

[46] Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, 2017.

[47] Esther Rheinbay, Prasanna Parasuraman, Jonna Grimsby, Grace Tiao, Jesse M Engreitz, Jaegil Kim, Michael S Lawrence, Amaro Taylor-Weiner, Sergio Rodriguez-Cuevas, Mara Rosenberg, et al. Recurrent and functional regulatory mutations in breast cancer. *Nature*, 547(7661):55–60, 2017.

[48] Marios Giannakis, Eran Hodis, Xinmeng Jasmine Mu, Mai Yamauchi, Joseph Rosenbluh, Kristian Cibulskis, Gordon Saksena, Michael S Lawrence, Zhi Rong Qian, Reiko Nishihara, et al. Rnf43 is frequently mutated in colorectal and endometrial cancers. *Nature genetics*, 46(12):1264–1266, 2014.

[49] Sarah Sandmann, Aniek O De Graaf, Mohsen Karimi, Bert A Van Der Reijden, Eva Hellström-Lindberg, Joop H Jansen, and Martin Dugas. Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific reports*, 7(1):1–12, 2017.

[50] James T Robinson, Helga Thorvaldsdóttir, Aaron M Wenger, Ahmet Zehir, and Jill P Mesirov. Variant review with the integrative genomics viewer. *Cancer research*, 77(21):e31–e34, 2017.

[51] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[52] Marc Fiume, Vanessa Williams, Andrew Brook, and Michael Brudno. Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, 26(16):1938–1944, 2010.

[53] Jeremy Goecks, Nate Coraor, Anton Nekrutenko, James Taylor, Galaxy Team, et al. Ngs analyses by visualization with trackster. *Nature biotechnology*, 30(11):1036–1039, 2012.

[54] Tim Carver, Simon R Harris, Thomas D Otto, Matthew Berriman, Julian Parkhill, and Jacqueline A McQuillan. Bamview: visualizing and interpretation of next-generation sequencing read alignments. *Briefings in bioinformatics*, 14(2):203–212, 2013.

[55] Jacob Pritt and Ben Langmead. Boiler: lossy compression of rna-seq alignments using coverage vectors. *Nucleic acids research*, 44(16):e133–e133, 2016.

[56] Adam J Adler, Graham B Wiley, and Patrick M Gaffney. Infinium assay for large-scale snp genotyping applications. *Journal of visualized experiments: JoVE*, (81), 2013.

[57] Ludmil B Alexandrov and Michael R Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development*, 24:52–60, 2014.

[58] ICGC The, TCGA Pan-Cancer Analysis of Whole, Genomes Consortium, et al. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82, 2020.

[59] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, et al. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020.

[60] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783, 2017.

[61] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[62] Peter Devilee and Matti A Rookus. A tiny step closer to personalized risk prediction for breast cancer, 2010.

[63] Dan A Landau, Eugen Tausch, Amaro N Taylor-Weiner, Chip Stewart, Johannes G Reiter, Jasmin Bahlo, Sandra Kluth, Ivana Bozic, Mike Lawrence, Sebastian Böttcher, et al. Mutations driving cll and their evolution in progression and relapse. *Nature*, 526(7574):525–530, 2015.

[64] Dan A Landau, Scott L Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S Lawrence, Carrie Sougnez, Chip Stewart, Andrey Sivachenko, Lili Wang, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726, 2013.

[65] Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood, The Journal of the American Society of Hematology*, 120(20):4191–4196, 2012.

[66] Michael Hallek, K Fischer, Gunter Fingerle-Rowson, Anne Michelle Fink, Raymonde Busch, Jiří Mayer, M Hensel, Georg Hopfinger, G Hess, U Von Grünhagen, et al. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *The Lancet*, 376(9747):1164–1174, 2010.

[67] Jennifer Edelmann, Karlheinz Holzmann, Florian Miller, Dirk Winkler, Andreas Bühler, Thorsten Zenz, Lars Bullinger, Michael WM Kuehn, Andreas Gerhardinger, Johannes Bloehdorn, et al. High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood, The Journal of the American Society of Hematology*, 120(24):4783–4794, 2012.

[68] Lorenzo De Paoli, Michaela Cerri, Sara Monti, Silvia Rasi, Valeria Spina, Alessio Bruscaggin, Mariangela Greco, Carmela Ciardullo, Rosella Famà, Stefania Cresta, et al. Mga, a suppressor of myc, is recurrently inactivated in high risk chronic lymphocytic leukemia. *Leukemia & lymphoma*, 54(5):1087–1090, 2013.

[69] Derek A Oldridge, Andrew C Wood, Nina Weichert-Leahey, Ian Crimmins, Robyn Sussman, Cynthia Winter, Lee D McDaniel, Maura Diamond, Lori S Hart, Shizhen Zhu, et al. Genetic predisposition to neuroblastoma mediated by a lmo1 super-enhancer polymorphism. *Nature*, 528(7582):418–421, 2015.

[70] Caiyun He, Huakang Tu, Liping Sun, Qian Xu, Yuehua Gong, Jingjing Jing, Nannan Dong, and Yuan Yuan. Snp interactions of helicobacter pylori-related host genes pgc, ptpn11, il1b, and tlr4 in susceptibility to gastric carcinogenesis. *Oncotarget*, 6(22):19017, 2015.

[71] Qian Xu, Jing-wei Liu, and Yuan Yuan. Comprehensive assessment of the association between mirna polymorphisms and gastric cancer risk. *Mutation Research/Reviews in Mutation Research*, 763:148–160, 2015.

[72] Jingwei Liu, Caiyun He, Chengzhong Xing, and Yuan Yuan. Nucleotide excision repair related gene polymorphisms and genetic susceptibility, chemotherapeutic sensitivity and prognosis of gastric cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 765:11–21, 2014.

[73] Ye Zhang, Li-Ping Sun, Cheng-Zhong Xing, Qian Xu, Cai-Yun He, Ping Li, Yue-Hua Gong, Yun-Peng Liu, and Yuan Yuan. Interaction between gstp1 val allele and h. pylori infection, smoking and alcohol consumption and risk of gastric cancer among the chinese population. 2012.

[74] Qian Xu, Qiguan Dong, Caiyun He, Wenjing Liu, Liping Sun, Jingwei Liu, Chengzhong Xing, Xiaohang Li, Bengang Wang, and Yuan Yuan. A new polymorphism biomarker rs629367 associated with increased risk and poor survival of gastric cancer in chinese by up-regulated mirna-let-7a expression. *PloS one*, 9(4):e95249, 2014.

[75] Jing-Wei Liu, Cai-Yun He, Li-Ping Sun, Qian Xu, Cheng-Zhong Xing, and Yuan Yuan. The dna repair gene ercc6 rs1917799 polymorphism is associated with gastric cancer risk in chinese. *Asian Pacific Journal of Cancer Prevention*, 14(10):6103–6108, 2013.

[76] Caiyun He, Huakang Tu, Liping Sun, Qian Xu, Ping Li, Yuehua Gong, Nannan Dong, and Yuan Yuan. Helicobacter pylori-related host gene polymorphisms associated with susceptibility of gastric carcinogenesis: a two-stage case-control study in chinese. *Carcinogenesis*, 34(7):1450–1457, 2013.

[77] Tao Peng, Yangyang Sun, Zhiwei Lv, Ze Zhang, Quanxin Su, Hao Wu, Wei Zhang, Wei Yuan, Li Zuo, Li Shi, et al. Effects of fgfr4 g388r, v10i polymorphisms on the likelihood of cancer. *Scientific reports*, 11(1):1–12, 2021.

[78] Jing-Jing Jing, Min Li, and Yuan Yuan. Toll-like receptor 4 asp299gly and thr399ile polymorphisms in cancer: a meta-analysis. *Gene*, 499(2):237–242, 2012.

[79] Vijay K Ulaganathan, Bianca Sperl, Ulf R Rapp, and Axel Ullrich. Germline variant fgfr4 p. g388r exposes a membrane-proximal stat3 binding site. *Nature*, 528(7583):570–574, 2015.

[80] Erik N Bergstrom, Mi Ni Huang, Uma Mahto, Mark Barnes, Michael R Stratton, Steven G Rozen, and Ludmil B Alexandrov. Sigprofilermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC genomics*, 20(1):1–12, 2019.

[81] Adi F Gazdar, Venkatesh Kurvari, Arvind Virmani, Lauren Gollahon, Masahiro Sakaguchi, Max Westerfield, Duli Kodagoda, Victor Stasny, H Thomas Cunningham, Ignacio I Wistuba, et al. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International journal of cancer*, 78(6):766–774, 1998.

[82] Junjun Zhang, Rosita Bajari, Dusan Andric, Francois Gerthoffert, Alexandru Lepsa, Hardeep Nahal-Bose, Lincoln D Stein, and Vincent Ferretti. The international cancer genome consortium data portal. *Nature biotechnology*, 37(4):367–369, 2019.

[83] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[84] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.