Development of sample mismatch detection algorithm in genome sequencing cohorts

Hein Chun

Department of Medical Science

The Graduate School, Yonsei University

Development of sample mismatch detection algorithm in genome sequencing cohorts

Hein Chun

Department of Medical Science

The Graduate School, Yonsei University

Development of sample mismatch detection algorithm in genome sequencing cohorts

Directed by Professor Sangwoo Kim

The Master's Thesis
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Master of Medical Science

Hein Chun

June 2019

This certifies that the Master's Thesis of
Hein Chun is approved.

Thesis Supervisor : Sangwoo Kim

Thesis Committee Member#1 : Hyun Seok Kim

Thesis Committee Member#2 : Yu Rang Park

The Graduate School

Yonsei University

June 2019

ACKNOWLEDGEMENTS

제 자신에 대한 의심과 함께 고민이 끊이지 않았던 석사 생활을 마무리하며 스쳐가는 장면들이 많습니다. 웃음으로 가득했던 TGIL은 제가 질문을 망설이지 않게 해주었고, 여러 선후배님들의 아낌 없는 조언으로 이 연구가 여기까지 올 수 있었습니다. 또한 저의 실수와 고민들을 모두 들어주시고 함께 고민해주셨던 김상우 교수님까지, 이 논문은 석사 기간 동안의 연구 결과뿐 아니라 저의 TGIL에서의 모든 고민과 웃음 그리고 저와 함께 해준 모든 분들의 흔적이 가득한 논문입니다.

석사 기간 동안 마음껏 실수해야 한다며 저의 불안을 배움으로 이끌어주셨던 김상우 교수님. 저의 사수로써, 멘토로써 항상 지켜봐 주었던 우리 미경 언니. 받은 게 너무 많으니 제가 다시 돌려줄 수 있도록 앞으로도 지켜봐 주기를 바래요. 함께 웃었던 기억들만 가득한 우리 소라 언니, 언제나 따뜻하게 온기를 나눠주던 우리 은영 언니, 항상 톰과 제리 같았던 우리 현주 오빠, 질문들에 언제나 열심히 대답해주던 우리 강형, 구박해도 언제나 다정했던 우리 세영쌤. 그리고 언제나 질문을 망설이지 않던 우리 유진이, 흡수력 좋은 우리 범진이, 항상 열심히인 우리 요한이, 제일 웃긴 우리 택쌤 그리고 지금은 TGIL에 있지 않지만 언제나 우리를 먹여 살려주던 우리 건호 오빠, 언제나 어른 같아서 항상 기대고 싶었던 우리 다희쌤까지 모두 감사합니다. 여러분이 주신 의견과 지지, 응원 하나하나가 모여 이 논문이 마무리되었습니다.

치열하게 살지 않아도 된다고 말해주신 분들도, 더 치열하게 살 수 있도록 동기부여가 되어준 분들에게도 모두 감사합니다. 위로 받으면서 조금은 더 열심히 할 수 있었습니다. 그리고 앞으로도 그 따뜻한 마음 안고 더 열심히 앞으로 나아가겠습니다. 마지막으로, 항상 응원을 아끼지 않는 우리 가족들, 든든한 우리 가족이 있어서 제가 저의 있는 그대로 살아갈 수 있었습니다. 앞으로도 잘 부탁합니다.

TABLE OF CONTENTS

ABSTRACT ····· 1
I. INTRODUCTION ····· 3
II. MATERIALS AND METHODS
1. Workflow of a model
A. SNP site selection ····· 6
B. Genotype-based paring ····· 8
C. Pair information obtainment
(A) Name-based pairing ····· 11
(B) List-based pair information ····· 13
D. Decision and report ····· 13
2. Acquisition of datasets ····· 15
3. Comparison with pre-developed tools ····· 16
4. Cancer panel data simulation results with WES ····· 18
5. Accuracy assessment with family dataset ····· 18
6. Accuracy assessment with simulation of shuffled sets ····· 19
7. Evaluation of simulated contamination data ····· 19
III. RESULTS
1. Accuracy comparison with pre-developed tools ····· 21
2. Accuracy of name-based pairing ····· 26
3. Accuracy with family dataset ····· 27
4. Accuracy with simulation of shuffled sets ····· 29
5. Running time comparison with pre-developed tools ····· 29
6. Effect of contamination ····· 30
IV. DISCUSSION ····· 32

V. CONCLUSION ····· 34
REFERENCES ····· 35
ABSTRACT (IN KOREAN) ····· 37
PUBLICATION LIST

LIST OF FIGURES

Figure 1. Overall workflow of BAMixChecker ····· 6
Figure 2. Score distribution of all datasets ····· 9
Figure 3. Score distribution of a panel containing SNPs under 200 having MAF over 0.1 ····· 10
Figure 4. Algorithm for the calculation of file name similarity ····· 12
Figure 5. BAMixChecker report example ····· 14
Figure 6. Accuracy of the four tools ····· 22
Figure 7. Score distribution in panel dataset ····· 25
Figure 8. Accuracy of the file-name-based paring ····· 27
Figure 9. Accuracy with family dataset ····· 28
Figure 10. Mismatch simulation results for WES, KCSG, and KLCC datasets ····· 29
Figure 11. Running times of the four tools ····· 30
Figure 12. Accuracy of contaminated dataset ····· 31

LIST OF TABLES

Table 1. List of test datasets ····· 15
Table 2. Family dataset ····· 16
Table 3. Accuracy of the four tools with sensitivity and specificity ····· 23
Table 4. Number of "INCONCLUSIVE" in BAM-matcher with targeted datasets ····· 24
Table 5. Number of "Require additional investigation" in Conpair with targeted datasets ····· 24

ABSTRACT

Development of sample mismatch detection algorithm
in genome sequencing cohorts

Hein Chun

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sangwoo Kim)

Over the past decade, the incredible development of next generation sequencing (NGS) technology has expanded clinical research using NGS data. NGS enabled large genomic study at low cost. However, expanded use of NGS requires a large number of samples to be processed in a limited time. Therefore, accompanied human errors in sample handling remain constant concerns. Sample mismatch in the process of NGS is one of the frequent problem that can cause an entire genomic analysis to fail. Therefore, a regular cohort-level sample match checkup is needed to ensure that it has not occurred. However, currently developed tools require additional processing besides the main analysis or huge time for a large dataset or show lower performance with targeted sequencing data than other larger target size of sequencing data. We developed a new, automated tool (BAMixChecker) that accurately detects sample mismatches from a given BAM file cohort with minimal user intervention. BAMixChecker compares samples only with 853 well-mappable and frequently mutatable single-nucleotide polymorphisms (SNP) loci for whole genome sequencing (WGS), whole exome sequencing (WES), RNA-seq dataset. BAMixChecker uses a flexible, data-

specific set of SNPs with target region information from BED file for targeted sequencing data. BAMixChecker detects orphan (unpaired) and swapped (mispaired) samples based on genotype-concordance score and entropy-based file name analysis. BAMixChecker shows ~100% accuracy in real WES, RNA-seq, and targeted sequencing data cohorts, even for small panels (<50 genes). BAMixChecker also provides an HTML-style report, with which users can quickly inspect any mismatch events.

Key words : next-generation sequencing, sample mismatch, quality control

Development of sample mismatch detection algorithm
in genome sequencing cohorts

Hein Chun

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sangwoo Kim)

I. INTRODUCTION

Passing by the generation that sequenced one or a few genes, emerging of next-generation sequencing (NGS) made it possible to sequence total human genome. In recent, the technology reached sequencing whole genome sequencing (WGS) for only 1000$[1]. Economical obstacle is dismissing for research using NGS technology. By developing the technology, wider usage in clinical field is fallowing.

Increasing use of NGS in clinical practice requires a large number of samples to be processed in a limited time. While improvements in algorithms have provided more accurate means of detecting genomic variants, human errors in sample handling remain constant concerns. Sample mismatch, in particular, is a frequent occurrence detrimental to sequencing analyses[2]. The more large number of files to handle, the higher the possibility to swap samples exist in the process. Sample mismatch results into misanalysis. Therefore, quality control like sample match check-up is crucial to secure accuracy and reliability of whole analysis.

One of the example is cancer research. To analysis somatic mutation which can cause or relate with cancer, usually tumor and matched normal sample are sequenced for a pair from each patient. Common variation across tumor and normal counted as germline variant which the patient may originally carries and not related with the disease. However, if sample mismatch occurs, germline variants are not eliminated accurately. Therefore, it leads miscalling of somatic variation and affects study result.

For the last few years, several tools have been developed to detect sample matching errors in the regular pipeline of NGS data analysis. The basic approach is observing germline variants which each individual uniquely carries and have no relation with disease. Conpair[3] checks sample swaps for a pair of BAM files based on 7,387 loci with common single nucleotide polymorphism (SNP) sites from 1000 Genomes database[4] by calling variants with GATK HaplotypeCaller[5]. The SNPs are exonic loci with minor allele frequency (MAF) over 40% and linkage disequilibrium (LD) between markers is under 0.8. BAM-matcher[6] uses a similar approach, but allows for faster testing as it only uses 1,500 exotic SNP sites with user-selected tool among GATK, varscan[7], Freebayes[8]. The SNP set is also selected from 1000 Genomes database having 0.45 ~ 0.55 MAF. Even though those tools offer fast result with small set, the tools comparing only a pair of files such as Conpair, BAM-matcher can't compare the relationship among other pairs in dataset. Matched but duplicated pair with different individual is not detectable by only comparing each pair. Analysis in a dataset unit is needed to detect all the possible cases. The recently developed NGSCheckMate[9] works with FASTQ, BAM, or VCF files in a dataset by comparing 21,506 SNPs for BAM, VCF and 11,696 SNPs for FASTQ mode with samtools mpileup[10]. Unlike other two tools, NGSCheckMate uses correlation coefficient to compare samples. The SNP loci to compare are loci having over 0 of median absolute deviation of VAF across samples of 40 WGS stomach cancer patients. Additionally, NGSCheckMate offers different cutoff for family dataset and low coverage

dataset based on observed score distribution for each. NGSCheckMate provides result of matched sample list and all comparison or a tree graph of genotype correlations among given samples. However, when the mislabeling or swap occur, it's hard to catch the unusual result in the tree graph or matched or total pair comparison result at a glance. In addition, all the tools shows lower accuracy for targeted sequencing dataset because they use fixed small SNP set to analysis.

In general, the reported accuracies of these tools are all over 95%. Despite the good accuracy of these tools, we have found areas for improvement in two major features that would allow for more active use in cohort-level checkup. First, the number and the composition of SNP sites for individual matches need to be optimized. These SNP sites should be applicable to various targeted sequencing panels in order to cope with large-scale clinical genomic tests with accurate result. Second, the tool should be fast and automated to minimize intervention from users without pre- and post-processing for conclusion, even with a large number of samples.

Accordingly, we developed BAMixChecker, which facilitates fast and accurate assessment of sample mismatching from WGS/Whole Exome sequencing (WES)/RNA-seq and targeted sequencing panels. BAMixChecker uses 853 SNP sites that are optimized for WGS/WES and RNA-seq data and instantly composes a data-specific SNP list for targeted sequencing data; this provides a reduced running time while maintaining accuracy, even in a small panel. BAMixChecker categorizes orphan and swapped samples with genotype and file-name based pairing or user-given pair information. Overall, the pipeline is fully automated, allowing users to quickly check abnormal events without the need for further intervention to interpret the result. The result is provided with HTML report and heatmap plot to catch the unusual event at a glance as well as with TXT file for non-graphical user interface (GUI) environment.

II. MATERIALS AND METHODS

1. Workflow of a model

BAMixChecker only takes a dataset of BAM files as input with optional genomic region information (BED file) for targeted sequencing and reports mismatched samples and their types (Figure 1). The overall workflow consists of the four major steps. Detailed procedures are described below.

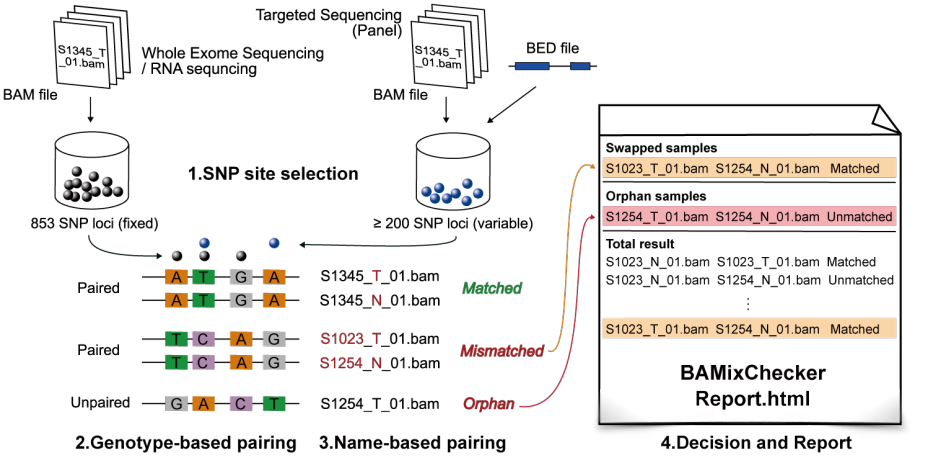

Figure 1. Overall workflow of BAMixChecker.

A. SNP site selection

Initially, 57,582 candidate SNP loci are obtained based on variant/mapping quality and mappability. From the candidate SNP loci, 853 coding SNPs are selected according to global and population-specific MAFs to build a fixed list for WGS/WES and RNA-seq data. For targeted sequencing, BAMixChecker automatically adjusts conditions such that at least 200 SNPs that overlap given genomic regions are selected.

To select only highly informative SNP loci and to reduce ambiguous calls, we considered two criteria: 1) mappability and 2) population allele-frequency. We referred to the gnomAD version r2.0.2 [11] for the criteria. From the database, we collected 57,582 SNPs that passed the following filters, including variant quality, mapping quality, and genomic mappability.

First, the following mandatory filters were applied:

- variant in a coding region,
- variant passes the calling filter,
- variant reported in dbSNP [12],
- variant passes in the random forests classifier of gnomAD, which handles problems with variant quality score recalibration filters applied at the locus level, instead of the allele level, in the ExAC dataset included in gnomAD.
- quality by depth > 2.0, and
- root mean square mapping quality > 50.

Then, we considered the conditions below to extract well mappable SNPs loci:

- variant not in a low complex region,
- variant not in a segment duplicated region, and
- variant not in a simple repeat region.

We selected 853 SNP loci with MAFs over 0.45 and under 0.55 in the overall population and MAFs over 0.35 and under 0.65 within each population as the optimal set for WGS/WES and RNA-seq among 57,582 candidate SNP loci. If the user submits a BED file, BAMixChecker generates dataset-specific SNPs for targeted sequencing data. Given a targeted BED file by a user, BAMixChecker adjusts the MAF conditions such that the list contains at least 200 loci from higher MAFs and creates a list of dataset-specific SNPs within only the target region of the samples. This leads to reduced noise from off-target and possible false positives.

B. Genotype-based paring

For selected SNP sites, BAMixChecker calls genotypes of given samples using GATK HaplotypeCaller with further filtering. Genotype concordance scores are then calculated between all pairs in the cohort. Sample pairs with a concordance >0.7 are considered matched in accordance with large-scale database. Unpaired samples in this step are considered orphans.

BAMixChecker explores genotype on every target SNP locus with a gVCF format of the GATK HaplotypeCaller. Calling with gVCF enables accurate determination of whether a non-called site in regular VCF is the same as the reference or non-sequenced. This helps to maximize usefulness in a small set of loci. By default, unreliable reads are filtered with GATK HaplotypeCaller. After calling variants, BAMixChecker compares the genotype for sites only that have been sequenced, are diallelic, and have a depth greater than five for comparison between two samples.

BAMixChecker matches samples that have a concordance greater than 0.7 in the locus as a pair from the same individual. The cutoff score reflects the score that can divide matched and unmatched pairs in all types of data, even targeted sequencing data with relatively lower MAFs (Figure 2). Targeted sequencing datasets show relatively higher concordance scores in comparison to unmatched pairs, because they compare locus with lower MAFs, which are less possible to be mutated. However, score distributions indicate that the scores of unmatched pairs in targeted sequencing data are still under 0.7. Exceptional case is that the number of dataset-specific targeted SNP is under 200. Score distribution of concordance score in unmatched pair is higher because under 200 dataset-specific targeted SNP means the target region is containing under 200 SNPs even with global MAF > 0.1. So, even the number is close to 200 SNPs consisting the set is less frequently mutated. Therefore, the dataset-specific SNPs having less power for discriminating each individual. In the case, BAMixChecker applies different cutoff score 0.8180. The cutoff is calculated excepting outlier as the middle of

lower whisker of matched and upper whisker of unmatched score in a panel simulation data with 46 genes and containing under 200 SNPs (Figure 3). Even though the final model of BAMixChecker applies different cutoff for the case with SNPs under 200, the cutoff is based on the panel which is one of the simulation datasets. So, we basically illustrated the accuracy with the original cutoff 0.7 to evaluate the performance without over fitting and only tested flexible cutoff for family dataset which is independent with the original WES dataset used to simulate the panel dataset.

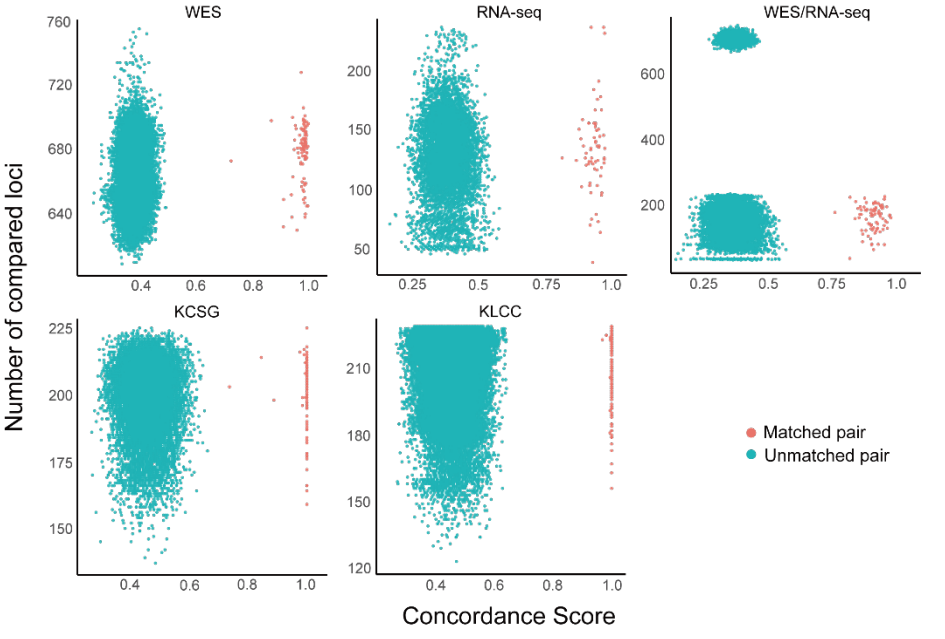

Figure 2. Score distribution of all datasets. Concordance score distributions of BAMixChecker in WES, RNA-seq, WES/RNA-seq, KCSG, and KLCC datasets. Each dot reflects a comparison result between two samples. Red dots indicate unmatched pairs; blue dots are matched pairs.

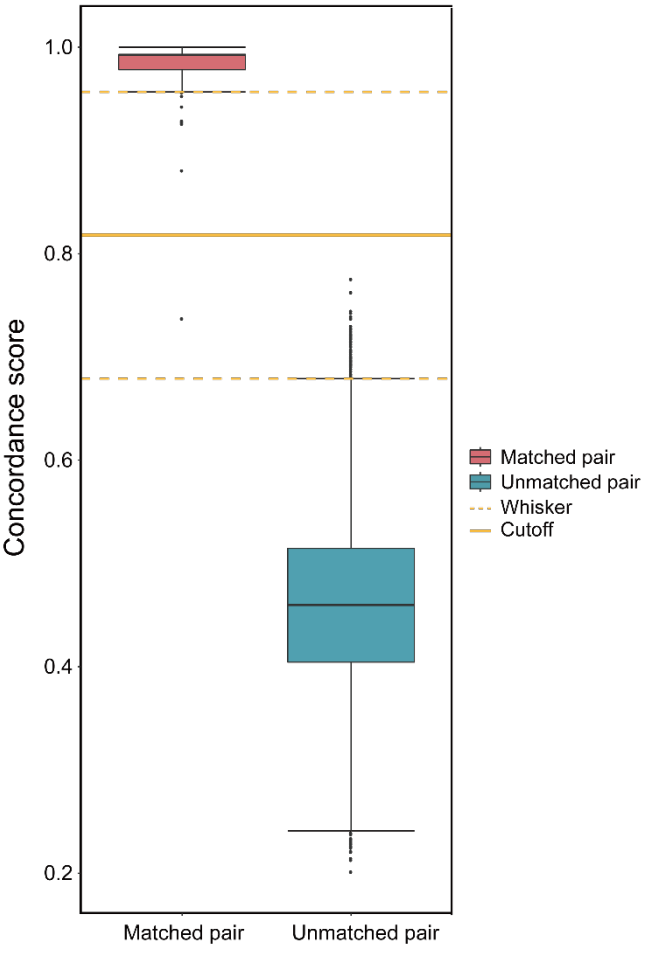

Figure 3. Score distribution of a panel containing SNPs under 200 having MAF over 0.1.

Calling with GATK HaplotypeCaller and concordance scoring are conducted by multiprocessing to allow for reduced running time. The default maximum number of process is four, and the user can adjust the parameter depending on their own computational environment.

C. Pair information obtainment

(A) Name-based pairing

Assuming that file names are rule-based within a cohort, sample relationships can be inferred from the names, just as a human would do. BAMixChecker emulates this using entropy-based file matching. Briefly, the uncertainty of values in the same position of a delimited file name is measured. Positions with high uncertainty tend to represent sample- or individual-specific information (e.g., sample id), while low uncertainty reflects global information (e.g., cohort id). File-name similarity is calculated by adding or subtracting positional entropy for each matched or mismatched value: file names of matched samples only differ in low entropy positions (e.g., T vs. N) and gain a high score in high entropy positions (e.g., sample id), thereby being considered as the best match in the cohort.

After matching highly concordant samples by genotype, BAMixChecker evaluates file name similarity using entropy-based file-name assessment as described below (Figure 4).

1) BAMixChecker detects the regulation of delimiters in file names. It divides the file names by the delimiters and stores each divided part into a vector for each file name.

2) Next, it counts the frequency of each unique value in the part and creates a count vector for each delimited part.

3) After vectors are made, it calculates the entropy score of each delimited part from the count vector among all files.

4) To compare two files by each part divided by delimiters, BAMixChecker scores the entropy score of each part. If the strings of the files in the part are the same, the entropy score is added, and if the values are different, the score is subtracted. Finally, each file is given a similarity score vector against other files.

Based on the similarity score vector, BAMixChecker pairs the best match in the cohort with the maximum similarity.

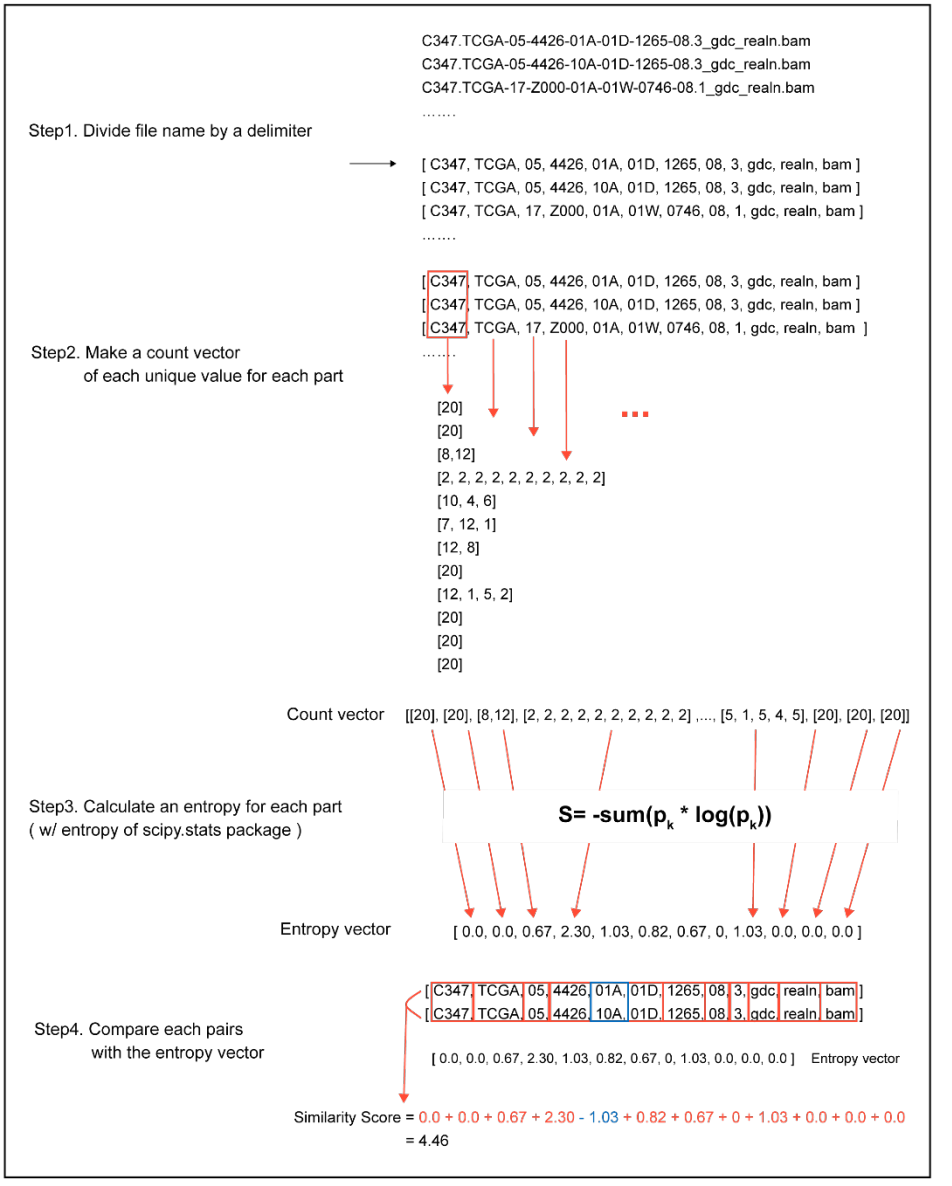

Figure 4. **Algorithm for the calculation of file name similarity.** An example of file name comparison to extract pair information with file name similarity. P_k is the probability of each unique value.

(B) List-based pair information

To extract pair information from file names, the number of files must be greater than 6, and the file names should have common regulation when divided by the delimiters. Otherwise, the user can offer pair information by submitting a list of tab-divided samples of an individual on each line as an input instead of whole list of samples on each line or a directory path to files. In the case of containing two or more sample of an individual, BAMixChecker can be applied with the list.

D. Decision and report

After genotype-based paring and file-name-based paring or user-given matched file name obtaining, BAMixChecker categorizes all given samples into three classes: matched (match for genotype and file-name), swapped (genotype match that is not file-name matched, or vice versa), and orphan (no genotype match found). Final results are reported in an HTML document (Figure 5).

BAMixChecker investigates all pairs in regards to whether they have high genomic concordance and paired file names. BAMixChecker classifies samples into three groups as a ‘matched sample’, ‘swapped sample’, or ‘orphan sample’. Matched samples are samples having a highly concordant pair by genotype and a best match in the cohort by file name or user inputted match. Swapped samples are samples having a pair by genotype concordance, but not matched by its file name, or vice versa. If a sample has no pair by genotype or concordance, it is reported as an orphan sample. Mixed samples are reported as ‘mismatched samples’ in an HTML file and TXT files for a non-GUI environment; the files also include the total concordance analysis results for all samples.

Sample Mix-up analysis result by BAMixChecker

Mismatched samples

1. Swapped samples
- Matched samples only by *genotype or file name* but not by both

Sample1	Sample2	Concordance rate	Conclusion
S1254_N.bam	S1254_T.bam	0.38	Unmatched
S1254_N.bam	S1345_T.bam	0.97	Matched
S1345_N.bam	S1254_T.bam	0.95	Matched
S1345_N.bam	S1345_T.bam	0.36	Unmatched

2. Orphan samples
- Samples matched with nothing by *genotype*

Orphan sample	Best match by file name	Concordance rate	Conclusion
S1983_N.bam	S1983_T.bam	0.37	Unmatched
S1983_T.bam	S1983_N.bam	0.37	Unmatched

Matched samples
- Matched samples by *file name* and *genotype*

Sample1	Sample2	Concordance rate	Conclusion
S1023_N.bam	S1023_T.bam	0.92	Matched

Total result

Sample1	Sample2	Concordance rate	Conclusion
S1023_N.bam	S1023_T.bam	0.92	Matched
S1023_N.bam	S1254_N.bam	0.37	Unmatched
S1023_N.bam	S1254_T.bam	0.34	Unmatched
S1023_N.bam	S1345_N.bam	0.34	Unmatched
S1023_N.bam	S1345_T.bam	0.36	Unmatched
S1023_N.bam	S1983_N.bam	0.36	Unmatched
S1023_N.bam	S1983_T.bam	0.35	Unmatched
S1023_T.bam	S1254_N.bam	0.35	Unmatched
S1023_T.bam	S1254_T.bam	0.32	Unmatched
S1023_T.bam	S1345_N.bam	0.33	Unmatched
S1023_T.bam	S1345_T.bam	0.35	Unmatched
S1023_T.bam	S1983_N.bam	0.34	Unmatched
S1023_T.bam	S1983_T.bam	0.35	Unmatched
S1254_N.bam	S1254_T.bam	0.38	Unmatched
S1254_N.bam	S1345_N.bam	0.37	Unmatched
S1254_N.bam	S1345_T.bam	0.97	Matched
S1254_N.bam	S1983_N.bam	0.36	Unmatched
S1254_N.bam	S1983_T.bam	0.33	Unmatched
S1254_T.bam	S1345_N.bam	0.95	Matched
S1254_T.bam	S1345_T.bam	0.37	Unmatched
S1254_T.bam	S1983_N.bam	0.39	Unmatched
S1254_T.bam	S1983_T.bam	0.35	Unmatched
S1345_N.bam	S1345_T.bam	0.36	Unmatched
S1345_N.bam	S1983_N.bam	0.38	Unmatched
S1345_N.bam	S1983_T.bam	0.34	Unmatched
S1345_T.bam	S1983_N.bam	0.36	Unmatched
S1345_T.bam	S1983_T.bam	0.32	Unmatched
S1983_N.bam	S1983_T.bam	0.37	Unmatched

Figure 5. BAMixChecker report example. HTML style report from BAMixChecker when ‘S1254_T.bam’ and ‘S1345_T.bam’ are swapped with each other and ‘S1983_N.bam’ and ‘S1983_T.bam’ samples are not from the same patient.

2. Acquisition of datasets

We evaluated the performance of BAMixChecker in five public and private cohorts with accompanying genomic analysis and patient information (Table 1). All datasets consisted of matched tumor-normal pairs. A WES dataset from TCGA lung adenocarcinoma patients (202 samples) was downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/). RNA sequencing data (RNA-seq) covered samples from cholangiocarcinoma (CHOL), liver hepatocellular carcinoma (LIHC), and pancreatic adenocarcinoma (PAAD) patients in TCGA. A WES/RNA-seq dataset was obtained from WES and RNA-seq pairs from the same TCGA LIHC patient. Performance for targeted sequencing data was assessed in two previously published or on-going umbrella trials with genomic data: the Korean Cancer Study Group (KCSG) cohort contains 192 tumor-normal matched samples from 96 head and neck esophageal squamous cell carcinoma patients reflected in a custom Illumina sequencing panel (244 genes)[13], and the Korean Lung Cancer Consortium (KLCC) contains 402 tumor-normal matched samples from 201 non-small cell lung adenocarcinoma patients reflected in the same sequencing panel.

Table 1. List of test datasets

Dataset	Data type	#Sample	#Individual	Sequencing Depth
WES	Whole exome sequencing	202	101	40 ~ 160X
RNA-seq	RNA sequencing	130	65	60 ~ 260X
WES/ RNA-seq[1]	Whole exome sequencing / RNA sequencing	168	84	30 ~ 160X / 60 ~ 270X
KCSG	Targeted sequencing	192	96	120 ~ 2310X
KLCC	Targeted sequencing	402	201	70 ~ 1810X

[1]WES - RNA-seq pair

Additionally, we tested a trio dataset with 105 WES of ADHD patient and their parent[14] to evaluate discriminant ability of BAMixChecker when the individuals share partial of genetic information (Table 2). The trio dataset is single samples from each individual including probands and their parents without a pair of samples from an individual.

Table 2. Family dataset

Dataset	Data type	#Sample	#Individual	#Genetically related pair[1]	Sequencing Depth
Family	Whole exome sequencing	105	105	70	40 ~ 160X

[1]Genetically related pair: mother-proband / father-proband

3. Comparison with pre-developed tools

The performance of BAMixChecker was demonstrated in comparison to pre-developed tools that can be used in the regular NGS data analysis pipeline: NGSCheckMate, BAM-matcher, and Conpair. The only tool for multiple samples of NGS data without additional pre-processing is NGSCheckMate, which offers VCF, BAM, and FASTQ modes. The VCF mode requires germline variant calling. However, germline variant calling is optional in cancer studies, depending on the objective of the study. Therefore, we only tested BAM and FASTQ modes for NGSCheckMate. BAM-matcher offers three VCF files for target SNP region information for 1,511 to 7,550 SNPs, which the user can select depending on the target size of the sample. Therefore, we tested WES, RNA, and WES/RNA datasets with a list of the 1,511 SNPs mentioned by Wang, et al in the paper. For the targeted sequencing dataset, we used a VCF with 7,550 SNPs because some of the samples would not have contained enough SNPs with which to compare to the other two VCFs. BAM-matcher also offers a calling step with three kinds of

variant callers, all of which we tested: the three callers showed the same accuracy. Also, we recorded running time with GATK as it is considered the fastest caller among the three callers in the implementation. Conpair recommends the option to use only homozygous markers in supplementary material, although the function is not set as a default. Also, the option can be used to compare a pair of tumor and normal samples, not between tumor samples or between normal samples. Therefore, we tested Conpair without the option to see the concordance with all files, even between tumor samples and normal samples. Additionally, Conpair only offers the concordance score without concluding whether the samples are matched, although a recommendation is given in supplementary materials. Accordingly, we followed this standard to interpret results.

The TCGA WES BAM file was mapped with GRCh38/hg38; however, NGSCheckMate and BAM-matcher did not offer GRCh38/hg38 SNPs. Thus, we converted the GRCh37/hg19 SNPs to GRCh38/hg38 using the LiftOver utility provided by the USCS genome browser[15]. All parameters for all tools were default, except the number of threads in the NGSCheckMate FASTQ mode and BAMixChecker for running time comparison: BAMixChecker uses four processes as a default, while the NGSCheckMate FASTQ mode uses one thread as a default. We evaluated both tools with processes or threads set to both one and four. Only for the targeted sequencing and panel data, we also used a non-default option of NGSCheckMate which uses non-zero mean depth of targeted loci for VAF correlation cutoff determination as in the original paper.

Additionally, we included both conclusive and inconclusive data in the accuracy evaluation. A result deemed ‘INCONCLUSIVE’ in BAM-matcher and data with a score in the range requiring ‘additional investigation’ were also included as miscalls, because they also reflect a failure to obtain information expected by a user.

4. Cancer panel data simulation results with WES

To evaluate the performance with targeted sequencing data, we performed additional evaluation with popular commercial cancer panels: Ion AmpliSeq Comprehensive Cancer Panel (Ion-CCP, 409 genes), Foundation One (FONE, 315 genes), xGen Pan-Cancer Panel (xGen-PCP, 127 genes), and Comprehensive Common Cancer Panel (CCCP, 46 genes). We extracted the reads in the target gene region of four commercial cancer panels from the 202 WES data from TCGA and created new BAM files for each panel. The BAM files are similar, with a lower depth of targeted sequencing, not including reads for off-target regions. A lower depth than normal for targeted sequencing data does not make a difference in analysis because all tools show almost perfect accuracy for WES data with the depth. However, by excluding off-target reads from real data, it is possible to affect the performance of the tools, except for BAMixChecker. Using a targeted BED file by the user, BAMixChecker explores dataset-specific SNP loci only in targeted sequenced regions. Thus, we expect that BAMixChecker would not be affected by reads on off-target regions, which can lower accuracy, and still perform well with real cancer panel data, unlike other tools.

5. Accuracy assessment with family dataset

We tested trio dataset with 105 WES of ADHD patient and their parent to evaluate discriminant ability of BAMixChecker when the individuals are related. Even though copy number variation in the probands are reported in the paper, we confirmed that all targeted loci are not contain CNV reported position. Because the dataset consist of a sample from each individual, we applied an option '—OFFFileNameMatching' and evaluated only with genotype-based score result. Due to absence of matched samples (same individual), the accuracy is same as specificity which is the ratio whether different individual is called as 'Unmatched' in all comparison. Additional to normal accuracy, we calculated accuracy only in

comparison between one of parent and proband to reduce the impact of weakness of some tools in small panel dataset.

6. Accuracy assessment with simulation of shuffled sets

We tested the sample mix-up detecting ability of BAMixChecker with artificially shuffled datasets by random shuffling the file name and data. We assumed three cases:

1. 10% of swapped samples,
2. 10% of swapped samples + 10% of orphan samples, and
3. 10% of orphan samples.

Swapped samples were samples that showed high concordance by genotype, but not by file name, or vice versa. Orphan samples included samples with no genotype concordance for any sample. We made new datasets by switching the file names to simulate sample mix-up. Each case tested 100 times of random selection of 100 samples from each dataset. Then, we renamed 10% of the samples for each condition. We changed the file names in the portion for ‘swapped samples’. To make ‘orphan samples’, we copied one of the tumor or normal file names on a sample from a different patient. Thus, the orphan samples had a pair according to file name but not by genotype. We demonstrated the mismatched sample detecting ability of BAMixChecker with these simulated datasets.

7. Evaluation of simulated contamination data

The contaminated data was created in silico with WES dataset. Randomly selected 100 samples were assumed cross-contaminated in specific ratio. We tested with 0.01, 0.05, 0.1, 0.25% of contamination condition. Selected samples were down sampled into 0.99, 0.95, 0.90, 0.75 % and also randomly selected

contaminant samples were down sample into 0.01, 0.05, 0.1, 0.25 % with picard DownsampleSam. And we merged two samples and made in silico contaminated dataset with samtools merge. The contaminated panel datasets were also made with the same approach using the dataset created for panel simulation dataset from the WES dataset.

III. RESULTS

1. Accuracy comparison with pre-developed tools

We evaluated the accuracy of BAMixChecker in comparison to previously reported tools (NGSCheckMate, BAM-matcher and Conpair) in five real NGS cohorts: (1) TCGA WES cohort (*n*=202), (2) TCGA RNA-seq cohort (*n*=130), (3) TCGA WES/RNA-seq cohort (*n*=168), (4) KCSG panel sequencing cohort (*n*=192), and (5) KLCC panel sequencing cohort (*n*=402).

For all cohorts, only BAMixChecker showed perfect accuracy (Figure 6A). For TCGA WES and RNA-seq, all tools exhibited good accuracy, except for a few miscalls by BAM-matcher and Conpair. However, there was a noticeable drop in accuracy for the targeted sequencing cohorts (KCSG and KLCC in Figure 6A) with Conpair. Other tools also showed some miss calls in targeted sequencing cohorts (Table 3). BAM-matcher and Conpair had lower specificity because of inconclusive result. If target loci is under 100 and the concordance score is under 0.9 and above 0.6 or under 20 in the comparison between two samples, BAM-matcher determine it as “INCONCLUSIVE” case which means fail to determinate whether the samples are from same individual (Table 4). Conpair also recommend if the score ranged from 50 to 80, it “Requires additional investigation” (Table 5). The inconclusive results in both tools lower the specificity in targeted sequencing. Because target SNP set is too small, targeted sequencing can’t contain enough SNP loci to compare. Even though NGSCheckeMate targets 21,506 SNPs for BAM, VCF and 11,696 SNPs for FASTQ mode, NGSCheckeMate also missed some calls in targeted sequencing cohorts. It’s because the cutoff to determine whether the samples are from same individual was based on the observed score distribution of a WGS dataset. The cutoff doesn’t reflect score distribution in targeted sequencing cohort.

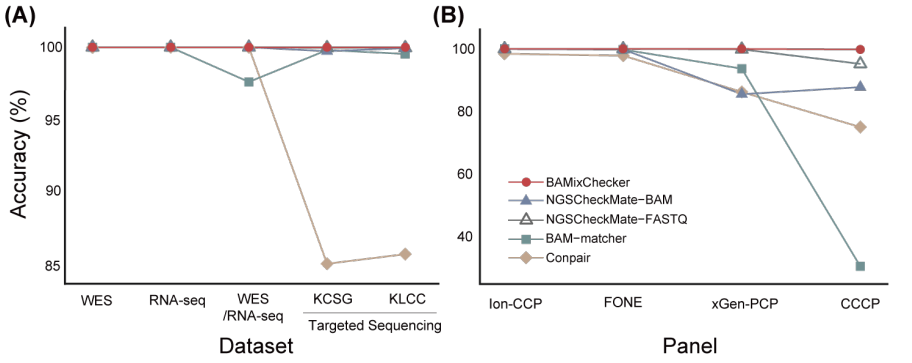

Figure 6. **Accuracy of the four tools. (A)** Accuracies of the four tools in five cohorts. NGSCheckMate contains two different modes (BAM and FASTQ input). WES/RNA-seq represents a WES-RNA-seq pair. **(B)** Accuracy of the four tools in downsampled cohorts. Ion AmpliSeq Comprehensive Cancer Panel (Ion-CCP, 409 genes), Foundation One (FONE, 315 genes), xGen Pan-Cancer Panel (xGen-PCP, 127 genes), and Comprehensive Common Cancer Panel (CCCP, 46 genes).

Table 3. Accuracy of the four tools with sensitivity and specificity

		BAMixChecker	NGSCheckMate-BAM	NGSCheckMate-FASTQ	BAM-matcher	Conpair
WES	*Accuracy (%)*	100	100	100	100	99.99
	Sensitivity (%)	100	100	100	100	97.33
	Specificity (%)	100	100	100	100	100
RNA-seq	*Accuracy (%)*	100	100	100	100	100
	Sensitivity (%)	100	100	100	100	100
	Specificity (%)	100	100	100	100	100
WES /RNA-seq	*Accuracy (%)*	100	100	100	99.99	100
	Sensitivity (%)	100	100	100	97.62	100
	Specificity (%)	100	100	100	100	100
KCSG	*Accuracy (%)*	100	99.73	99.96	99.81	85.07
	Sensitivity (%)	100	48.96	91.67	100	100
	Specificity (%)	100	100	100	99.81	84.99
KLCC	*Accuracy (%)*	100	99.94	99.96	99.54	85.73
	Sensitivity (%)	100	87.50	96.02	100	100
	Specificity (%)	100	99.97	99.97	99.54	85.69

Table 4. Number of "INCONCLUSIVE" in BAM-matcher with targeted datasets

	"INCONCLUSIVE"
KCSG	44 / 19,900
KLCC	370 / 80,599

Table 5. Number of "Require additional investigation" in Conpair with targeted datasets

	"Requires additional investigation"
KCSG	2,737 / 19,900
KLCC	11,501 / 80,599

For evaluation of smaller panels, TCGA WES data were downsampled to gene lists for four popular commercial panels: Ion AmpliSeq Comprehensive Cancer Panel (Ion-CCP, 409 genes), Foundation One (FONE, 315 genes), xGen Pan-Cancer Panel (xGen-PCP, 127 genes), and Comprehensive Common Cancer Panel (CCCP, 46 genes). We found BAMixChecker showed almost perfect accuracy in all panels (>99.8%), while the other tools showed lower accuracy in smaller panels (Figure 6B). The score distribution shows clear discrimination between matched pair and unmatched pair in BAMixChecker even with the small size of panel (Figure 7). However, NGSCheckMate and Conpair showed almost mixed distribution for CCCP and close distribution in larger panel. Even though BAM-matcher shows clear separate distribution, the number of dot is smaller by decreasing of panel size. The same tendency is also appeared in Conpair. With smaller panel, the possibility of inconclusive result is increasing. So, the accuracy is lower in small panel.

The cutoff score of NGSCheckMate is based on the score distribution in WGS dataset. However, we observed higher score distribution of targeted

sequencing dataset in unmatched pair than other dataset and it isn't reflected in the cutoff score. On the other hands, the cutoff score of BAMixChecker was considered with the common cutoff in all kinds of datasets. So the almost perfect accuracy is observed.

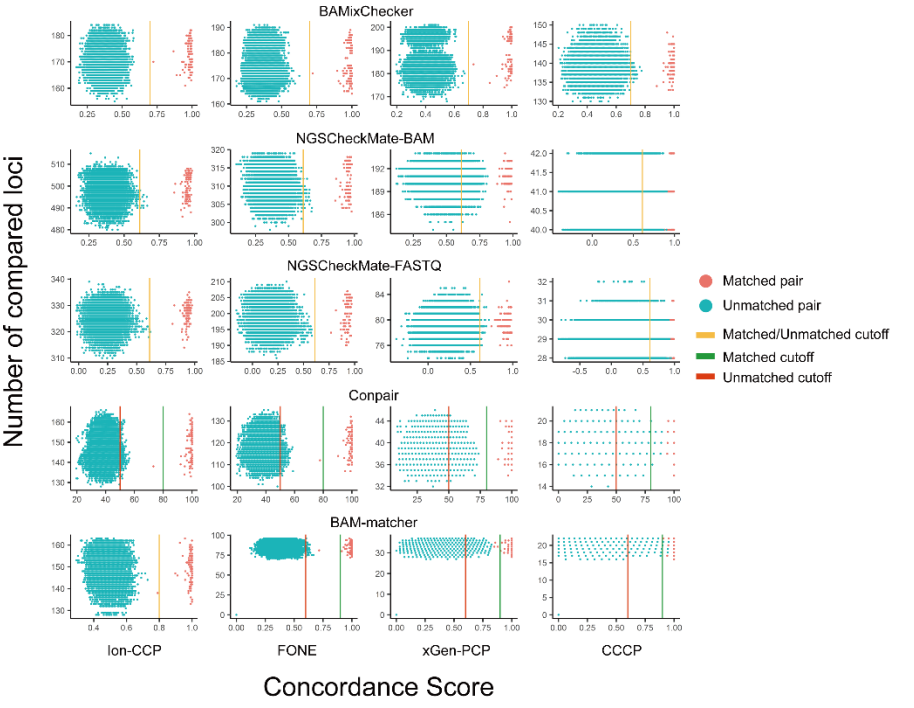

Figure 7. Score distribution in panel datasets. Ion-CCP : Ion AmpliSeq Comprehensive Cancer Panel (409 genes), FONE : Foundation One (315 genes), xGen-PCP : xGen Pan-Cancer Panel (127 genes), CCCP : Comprehensive Common Cancer Panel (46 genes)

Additionally, even though BAMixChecker shows almost perfect accuracy in all panel dataset, we identified the cutoff 0.7 is not equally applicable to small panel including ~200 SNPs as it can see in CCCP result. So, we applied flexible cutoff in a final model for the case which have ~200 SNPs even with MAF over

0.1 condition. The cutoff is based on the concordance score distribution of matched and unmatched pair in CCCP dataset. It is calculated the middle of upper whisker of unmatched pair score and lower whisker of matched pair score to except outlier (See Method).

2. Accuracy of name-based pairing

To evaluate the accuracy of name-based pairing, we paired samples with the algorithm and recorded the resultant accuracy (Figure 8). We tested the accuracy of BAMixChecker in four datasets in which we did not modify the file name regulation: WES comprised an original bam file name downloaded from TCGA. The file names in RNA-seq were modified with sample IDs from TCGA in another study by our laboratory independent of the present study. KCSG and KLCC file names were original file names from each study. Among the datasets, we could not test the WES/RNA-seq dataset as it contains files with different numbers of delimited parts. Overall, our accuracy analysis indicated that pairs were perfectly detected by the file-name-based pairing algorithm in all four datasets.

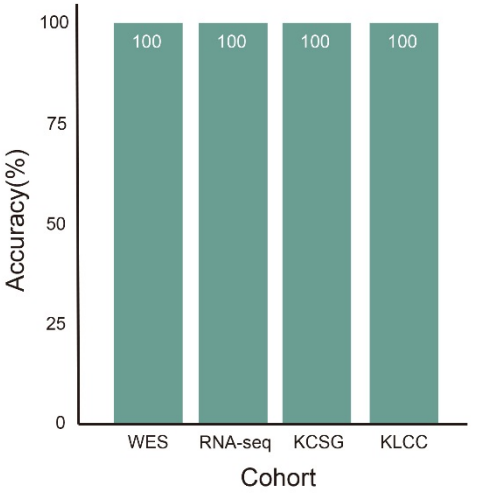

Figure 8. Accuracy of the file-name-based paring. WES is the case with the original bam file name downloaded from TCGA. The file names in RNA-seq was modified with the sample ID in TCGA in other study of our laboratory which is independent with this study. KCSG and KLCC file names is original file name from each study.

3. Accuracy with family dataset

There is a concern that the genotype-based approach is not able to apply to family dataset because of genetic similarity. So we evaluated the performance of BAMixChecker whether each individual is discriminated from their family member by BAMixChecker. We tested with 70 WES of ADHD patient and their parent trio dataset (Figure 9A). All tested tools showed robust accuracy for the WES dataset as for the WES dataset without related individuals. However, for the panel simulation dataset only BAMixChecker and NGSCheckMate which applies more strict cutoff score for family dataset showed almost perfect accuracy while other tools missing some correct calls (Figure 9B). The accuracy whether each tool determine the family members are separate individual makes it clear the discriminant ability of each tool without weakness of some tools for small size of

data (Figure 9C). To observe the effect of the flexible cutoff, we evaluated accuracy both original and flexible cutoff in the family dataset. Observed result demonstrates that BAMixChecker performs with best accuracy for all datasets even with related individuals even with the original model cutoff 0.7. With the flexible cutoff in final model, it discriminates each individual successfully. At the same time, it demonstrate that the lower accuracy of NGSCheckMate in panel dataset is related with normal cutoff score which is not reflected panel dataset score distribution.

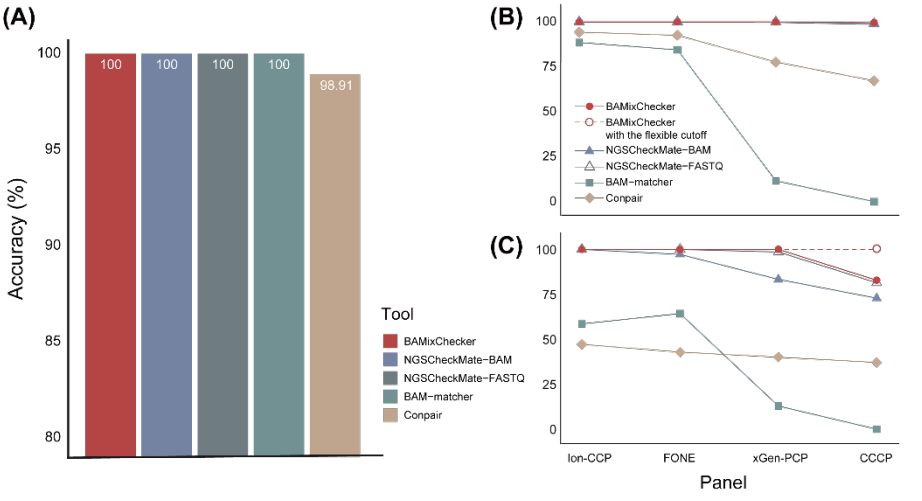

Figure 9. Accuracy with family dataset. (A) Accuracy for the WES family dataset among all comparisons. (B) Accuracy for panel simulation dataset from WES among all comparisons. (C) Accuracy for panel simulation dataset among comparisons with family members. All samples in the dataset are from different individuals. The accuracy in (A) and (B) is the rate of correct determination as 'unmatched' for all comparisons. BAMixChecker evaluated with both original cutoff 0.7 and flexible cutoff 0.8180 for a panel with ~200 SNPs having MAF over 0.1 in the family dataset.

4. Accuracy with simulation of shuffled sets

Additionally, we generated artificial mismatches (10% swaps and orphans) in all of the cohorts by altering file names to confirm that BAMixChecker correctly detects all mismatches and their types with 100% accuracy in three different datasets (Figure 10).

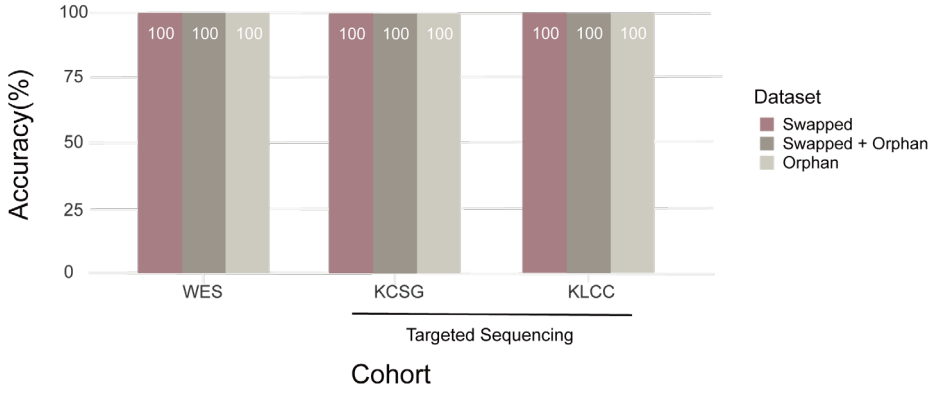

Figure 10. Mismatch simulation results for WES, KCSG, and KLCC datasets. BAMixChecker results with artificially mixed up datasets generated by changing file names.

5. Running time comparison with pre-developed tools

Finally, running times were assessed for all tools. We runned 10 times with 30 samples of each dataset on Intel® Core™ i7-4790 CPU 3.60GHz with quad cores and 32 GB memory (Figure 11). BAMixChecker compares only 853 loci for WES data. And it can calls multiple samples at the same time with a maximum number of processes which the user set. BAMixChecker returns of the result for 30 WES samples about 5 minutes with the default 4 processes set. BAMixChecker exhibited comparable or faster speed than BAM-matcher and Conpair, and was remarkably faster (~18x) than NGSCheckMate. Considering

the reduced need for intervention from users and automated output, we expect that the practical hands-on time would be much shorter with BAMixChecker.

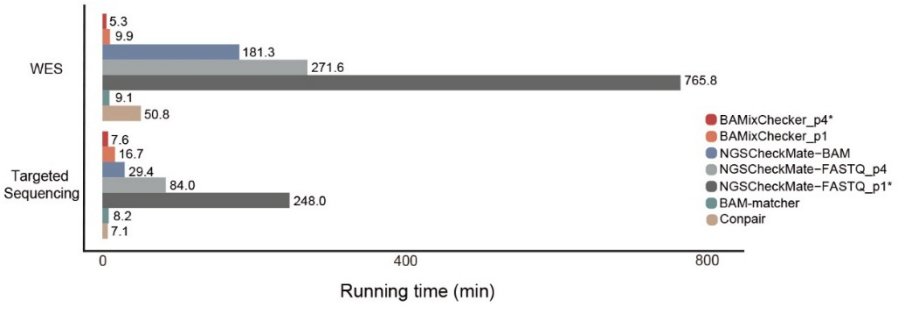

Figure 11. Running times of the four tools. The running times of BAMixChecker and NGSCheckMate were measured in two different modes (p1: single-thread, p4: multi-thread with four processors). * : default.

6. Effect of contamination

Contamination is a factor that can affect the result of genotype-based pairing. So the researcher who deal with samples expected contamination or copy number variation have to consider the effect on the pairing result. To observe an effect of contamination on the individual identification process based on genotype, we simulated contamination dataset in silico with the ratio of 0.01, 0.05, 0.1, 0.25%. The result shows almost perfect accuracy in 0.01, 0.05, 0.1 % of cross-contamination for WES (Figure 12A). However, 0.25 % of contamination result showed slightly lower accuracy. This tendency is also observed in panel datasets (Figure 12B). As the contamination rate increases and becomes closer to the cutoff score, and the accuracy of the result decreases as the interval covering the disconcordant information decrease. And the effect is sharp as smaller the panel size. Therefore, datasets that can be expected to be contaminated needs to remove of contamination risk or interpret pair-match analysis result by genotype-based pairing.

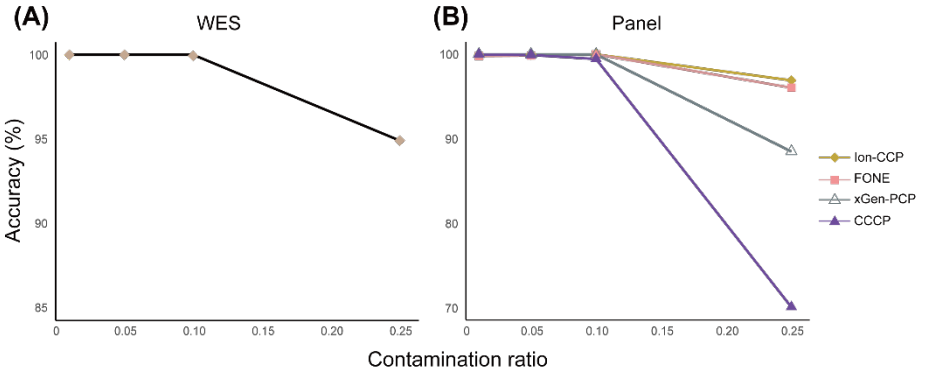

Figure 12. Accuracy of contaminated dataset. (A) Contaminated WES dataset (B) Contaminated panel datasets. Ion-CCP: Ion AmpliSeq Comprehensive Cancer Panel (409 genes), FONE: Foundation One (315 genes), xGen-PCP: xGen Pan-Cancer Panel (127 genes), CCCP: Comprehensive Common Cancer Panel (46 genes)

IV. DISCUSSION

The accuracy result in five real dataset with pre-developed tools demonstrated more accurate analysis of BAMixChecker even with small size of panel. The problem of lower accuracy in pre-developed tools is based on the scoring method or cutoff score or target SNP set. NGSCheckMate uses VAF correlation coefficient than genotype. However, the score distribution in panel datasets shows the score is not distinguishable than other tools including BAMixChecker which use genotype-based pairing.

The cutoff score of NGSCheckMate is based on the score distribution in WGS dataset. However, we observed higher score distribution of targeted sequencing dataset in unmatched pair than other dataset and it isn't reflected in the cutoff score. On the other hands, the cutoff score of BAMixChecker was considered with the common cutoff in all kinds of datasets even available in family dataset. Nevertheless, BAMixChecker still miss some calls when the target size is too small like CCCP with under 50 of genes. Even though the accuracy is still higher than other tools, careful usage is required to interpret the result with small panel data. When target size is small and contain less target SNP loci, higher concordance score shows in different individual (unmatched samples) while almost same score in same individual (matched samples). If the score of reported mismatched samples is slightly higher than cutoff and score of other unmatched samples in the dataset are close to the score than matched samples, it's possible to false call. In that case with small panel, user can consider a distance with scores of matched samples or scores of unmatched samples.

Target SNPs set is also crucial factor for accurate analysis. NGSCheckMate targets 21,506 SNPs for BAM, VCF and 11,696 SNPs for FASTQ mode. On the other hands, BAM-matcher and Conpair target 7,387, 1,500 SNPs for each. The small set of BAM-matcher and Conpair is the crucial factor that relatively lower accuracy in targeted sequencing datasets. Even though NGSCheckMate targets

more site and compare largest number of loci, it can give false positive information without proper depth and quality filtering for off-target region in panel dataset. Additional to depth filtering by in-house process and quality filtering by caller, extraction target loci of BAMixChecker on only target cite helps the negative effect of off-target region. At the same time, BAMixChecker can secure enough number of target cite by adjusting AF condition. However, it also still not enough for small panel depends on genes the panel contain**.** The case needs additional inspection with different approach.

There are factors that have an effect to lower concordance in matched samples like contamination, copy number variation. Contamination simulation result shows the weakness of genotype-based pairing approach. Smaller panel is more vulnerable from the effect of contamination. Therefore, contamination or copy number variation expected dataset is also needed to cautious analysis. Advanced approach to overcome contamination and copy number variation is required to develop for more precise result with samples in various condition.

V. CONCLUSION

We developed the fast and efficient to use for regular sample match check-up tools for NGS dataset. Based on the basic principle of experiment procedure with variant mappability and quality, the optimal SNP set is possible to offer more correct result to user not only fast output. The small but enough informative SNP set may be applicable to construct model with FASTQ. FASTQ file contains reads information before mapping. So the process to search reads containing each SNP takes huge time with currently developed SNP sets. However, 853 loci for WGS/WES/RNA-seq and at least 200 loci for targeted sequencing dataset may reduce total analysis time.

BAMixChecker is constructed to check sample match check-up for pairs or more sample from each individual. However, a usage of BAMixChecker can be expanded to check duplicated sample in a dataset that consist one sample of an individual with option that turns off file name matching and only compares samples by genotype-based matching. Furthermore, this approach can be applicable for similar purpose in non-human dataset with appropriate SNP information.

Additional to accurate genotype-based pairing, file-name-based paring by human behave-based algorithm serves utility to user. The higher utility increases the possibility to placing of sample match check-up into regular quality control procedure.

Even though the pair match or patient identification is crucial issue, the appropriate process is not conveyed in plenty study. The biggest obstacle is an absence of proper tool to give fast and accurate result with no intervention for user. BAMixChecker helps user to obtain the accurate sample match check-up result in proper time without additional pre- or post-processing in regular NGS data analysis procedure. Moreover, the accurate sample match results across various sequencing type may lead to precise and meaningful research outcome.

REFERENCES

1. Plothner M, Frank M, von der Schulenburg JG. Cost analysis of whole genome sequencing in German clinical practice. Eur J Health Econ. 2017;18(5):623-33.
2. Westra HJ, Jansen RC, Fehrmann RS, te Meerman GJ, van Heel D, Wijmenga C, et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27(15):2104-11.
3. Bergmann EA, Chen BJ, Arora K, Vacic V, Zody MC. Conpair: concordance and contamination estimator for matched tumor-normal pairs. Bioinformatics. 2016;32(20):3196-8.
4. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
6. Wang PP, Parker WT, Branford S, Schreiber AW. BAM-matcher: a tool for rapid NGS sample matching. Bioinformatics. 2016;32(17):2699-701.
7. Koboldt DC, Zhang QY, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;22(3):568-76.
8. Garrison E MG. Haplotype-based variant detection from short-read sequencing. arXiv 2012;1207(3907).
9. Lee S, Lee S, Ouellette S, Park WY, Lee EA, Park PJ. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. 2017;45(11):e103.
10. Li H. A statistical framework for SNP calling, mutation discovery,

association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

11. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308-11.
13. Lim SM, Cho SH, Hwang IG, Choi JW, Chang H, Ahn MJ, et al. Investigating the Feasibility of Targeted Next-Generation Sequencing to Guide the Treatment of Head and Neck Squamous Cell Carcinoma. Cancer Res Treat. 2018.
14. Lima Lde A, Feio-dos-Santos AC, Belangero SI, Gadelha A, Bressan RA, Salum GA, et al. An integrative approach to investigate the respective roles of single-nucleotide variants and copy-number variants in Attention-Deficit/Hyperactivity Disorder. Sci Rep. 2016;6:22851.
15. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.

ABSTRACT (IN KOREAN)

유전체 데이터 코호트 내 시료 불일치 검출 방법 개발

<지도교수 김 상 우>

연세대학교 대학원 의과학과

전 혜 인

Next-generation sequencing (NGS)를 이용한 기술이 발전함 따라 여러 연구에서 다수의 NGS 데이터를 이용한 연구가 확대되고 있다. 이러한 NGS 데이터를 생산하는 과정에서 발생하는 동일 개체 유래 시료의 불일치는 전체 유전체 분석 결과에 영향을 줄 수 있는 문제 중 하나이다. 시료 수집부터 실험 과정에서나 bioinformatics 분석 과정에서 다수의 실험자 혹은 연구자들을 거치게 되면서 각 개인에서 유래한 시료들이 섞이거나 mislabeling 될 가능성이 있다. 그러므로 코호트 단위의 전체적인 시료의 일치 여부를 확인하는 것이 필요하다. 그러나 현재 개발되어 있는 프로그램은 일반적인 NGS 데이터 분석 과정에 추가적인 전처리 혹은 해석을 위한 후처리 등의 추가적인 과정을 필요로 하거나, 다수의 데이터 분석에 많은 시간이 걸리거나, targeted sequencing 데이터 등 target 영역이 작은 데이터에서 낮은 정확도를 보이는 등의 실제 분석 단계에서 활용에의 어려움이 있다. 본 연구에서는 사용자가 기본적인 NGS 분석 과정에서 사용할 수 있는 BAM file 코호트를 이용하여 정확하고 빠르게 시료의 불일치

를 검출해낼 수 있는 자동화된 프로그램인 BAMixChecker를 개발하였다. BAMixChecker는 오직 853개의 mapping이 잘 되며 자주 mutation이 되는 single-nucleotide polymorphisms (SNP) 영역을 비교하여 whole genome sequencing (WGS), whole exome sequencing (WES), RNA-seq 코호트 시료를 비교한다. 그리고 targeted sequecing의 경우 BED file을 추가적으로 입력 받아 해당 타겟 영역에 맞는, 개별 코호트 특이적인 SNP set을 구성한다. 이렇게 정해진 비교 영역에서의 체세포 변이 정보와 함께 파일명 유사도 분석 알고리즘 기반 혹은 사용자 입력 기반 시료 쌍 정보를 바탕으로 시료의 불일치 여부를 분석한다. 불일치 시료는 경우에 따라, 어떤 시료와도 유전 정보로 일치하지 않는 시료의 경우 "Orphan", 유전 정보와 파일명 기반 혹은 사용자 입력 정보 기반 시료 쌍 정보가 일치하지 않는 경우 "Swapped" 시료로 분류하여 보고 된다. 분석 결과는 한 눈에 결과를 확인할 수 있는 HTML 파일과 함께 리눅스 환경에서 결과를 바로 확인할 수 있는 TXT 파일을 통해 제공된다. 해당 알고리즘은 50개 이하의 panel data를 포함한 실제 WES, RNA-seq, and targeted sequencing 코호트에서 ~100%의 정확도를 보였다.

핵심되는 말 : 차세대 유전체 시퀀싱, 시료 불일치, 퀄리티 컨트롤

PUBLICATION LIST

1. Chun H, Kim S. BAMixChecker: an automated checkup tool for matched sample pairs in NGS cohort. Bioinformatics. 2019.