creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:11048-000000147392

이화여자대학교 대학원
2017학년도
석사학위 청구논문

Analysis of inconsistent mutation calls for cell lines in NGS databases

의 과 학 과
Yura Song
2018

Collection @ ewha

Analysis of inconsistent mutation calls for cell lines in NGS databases

이 논문을 석사학위 논문으로 제출함

2017 년 12월

이화여자대학교 대학원

의 과 학 과 Yura Song

dCollection @ ewha

Yura Song 의 석사학위 논문을 인준함

지도교수 김 형 래 _____

심사위원 안 정 혁 _____

김 형 래 _____

홍 경 만 _____

이화여자대학교 대학원

dCollection @ ewha

Table of contents

Collection @ ewha

Collection @ ewha

List of Figures

Collection @ ewha

List of Table

Collection @ ewha

Abbreviation

Allelic Frequency: AF

Antibiotic- Antimycotic: Anti-Anti

Browser Extensible Data: BED

Burrows-Wheeler Aligner: BWA

Cancer Cell Line Encyclopedia: CCLE

Coding DNA Sequence: CDS

Chromatin Immunoprecipitation-sequencing: ChIP-seq

Cancer-Related Analysis of Variants Toolkit: CRAVAT

Fetal Bovine Serum: FBS

False Discovery Rate: FDR

Genome Analysis Toolkit: GATK

Genomics of Drug Sensitivity in Cancer: GDSC

Insertion and Deletion: in/del

Mutation Annotation Format: MAF

Next-Generation Sequencing: NGS

Transversion artifacts on 8-oxoguanine lesions: OXOG

RNA-sequencing: RNA-seq

Real-time Polymerase Chain Reaction: RT-PCR

Single Nucleotide Polymorphism: SNP

Collection @ ewha

The Cancer Genome Atlas: TCGA

Total depth: DP

Variant Calling Format: VCF

Whole Exome Sequencing: WES

Whole Genome Sequencing: WGS

Collection @ ewha

Abstract

Errors in Next-Generation Sequencing (NGS) databases have received much attention due to reports of inconsistent mutation calls from cell-line databases. In order to elucidate the reasons for such inconsistency, we analyzed the mutation calls for 592 cell lines and 897 cancer-driver genes from two databases, namely Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE). Discrepancies in annotation for mutations were found in 7.2% of mutation calls. Even after correction of the discrepancies discovered, significant additional discrepancies (about 33-42%) remained. Most of the mutation calls (98.8%, 162/164) were consistent in our two targeted sequencings for 8 cell lines; 7-8% and 11-13% of mutation calls in CCLE and GDSC, respectively, were not found in targeted sequencing, suggesting that these mutation calls are false positive. Contrary to the generally held notion, however, most (85-86%) of the inconsistent calls might be true mutations, which suggests that the inconsistency in the two databases could be related to false-negative mutations (14% in GDSC, 20% in CCLE). In the course of further analysis of the allele frequencies in the targeted sequencing data, consistent mutant allelic loss (2%, 4/155) or inconsistent allelic loss (4%, 7/155), which can be major sources of inconsistency in the two databases, were found. In conclusion, the mutation databases GDSC and CCLE might contain 7-13% false-positive mutation calls originating from polymerase proofreading

Collection @ ewha

errors. Important reasons for inconsistency and false-negative mutations might be uneven amplification and genetic drift. In order to resolve the problem of inconsistency in mutation databases constructed by NGS, allelic frequency data for all mutations are essential.

Collection @ ewha

I. Introduction

Several pharmacogenomics databases have been built for personalized therapeutic or precision medicine [1-3]. These databases, which have been employed for the development of molecular markers for cancer patients, include information such as mutations and drug responses in cell lines [4]. Two of those databases, Genomics of Drug Sensitivity in Cancer (GDSC) [5], and Cancer Cell Line Encyclopedia (CCLE) [6], include the most extensive data and have been cited by many researchers.

GDSC was built jointly by the Wellcome Trust Sanger Institute (UK) and the Massachusetts General Hospital Cancer Center (USA) in 2013 and updated in 2016. The purpose of this database, which includes the drug sensitivities of cytotoxic and targeted anti-cancer agents as well as the mutation profiles of 19,100 genes and 1,001 human cell lines, is to facilitate the discovery of therapeutic biomarkers that can identify patients most likely to respond to anti-cancer drugs. Additionally, for pharmaco-genetic characterization of large panels of human cancer cells, CCLE, the product of the 2013 collaboration of the Broad Institute and the Novartis Institutes for Biomedical Research, provides genomic data, mRNA expression, and pharmacogenomic features for 947 cell lines.

In the compilation of these and other such databases, Next-Generation Sequencing (NGS) techniques are widely employed [7-9]. NGS techniques have

Collection @ ewha

three major advantages over the traditional Sanger sequencing method: 1) faster data generation, 2) enhanced sensitivity for detection of low-level mutations, and 3) use of DNA samples without further preparation steps such as bacterial cloning [10-12]. Also, due to the current availability of entire-genomic information, NGS facilitates, relative to other methods, obtainment of additional valuable information such as copy-number changes, translocations, repetitive sequences, and insertions of viral genomes [13, 14].

Mutation detection employing NGS techniques includes targeted sequencing, whole exome sequencing (WES), and whole genome sequencing (WGS) [14-17]. In those databases, whole exome sequencing (WES) for GDSC, and targeted sequencing for CCLE were employed. The difference between WES and targeted sequencing is the number of target genes or regions [18]. The targets for WES are all protein-coding regions within the entire genome [19], randomly fragmented genomic sequences are hybridized to exome tiling arrays, and all exome regions are captured prior to the sequencing step [20-22]. Targeted sequencing, by contrast, captures much smaller regions, and the database in CCLE includes exons from 1,651 genes [23, 24].

Although NGS techniques provide much information for molecular changes in cancer cells or tissues, several papers have suggested that NGS analysis can incur sequencing errors [25-27]. Especially, rates of inherent errors in base calling have been reported to be as high as 1% depending on the specific NGS technique [8, 28].

Collection @ ewha

Further, there are slightly different error rates for various types of mutations [15]; these errors require, during bioinformatics analysis, quality recalibration for mutation calls in order to reduce false mutation call rates [29].

A recent report of a consistency rate as low as 57% for missense mutation calls between GDSC and CCLE for cell lines [5, 6] raised a concern regarding mutation call errors by NGS in publicly available mutation databases. Even with such concern about NGS mutation call errors, the extent of and reasons for the discrepancy have not been fully elucidated. In the present study, mutation calls from the GDSC and CCLE databases were re-analyzed, and targeted sequencing was performed two times for 8 cell lines in order to elucidate the reasons for the discrepancy of mutation calls between the databases in NGS analysis.

Collection @ ewha

II. Materials and Methods

A. Retrieval of mutation calls from GDSC and CCLE databases

The mutation calls for cancer cell lines were retrieved from the Genomics of Drug Sensitivity in Cancer (GDSC, http://www.cancerrxgene.org) and the Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle/home) databases. The data from GDSC was an updated version that has been available since July 2016. The mutations included therein were analyzed by whole-exome sequencing (WES) via Illumina HiSeq 2000 [5]. The cell-line sequence-variant data from GDSC were downloaded as an excel worksheet file at the GDSC website. The CCLE data, meanwhile, available since May 2013, was analyzed, and the mutations therein were obtained by targeted sequencing using the Agilent target-enrichment method [6]. The CCLE data were downloaded in the filtered Mutation Annotation Format (MAF) at the CCLE website.

By comparison of cell-line names, 592 common cell lines analyzed in both CCLE and GDSC were found. For comparison of the cell lines, cell-line identities were checked by reference to the information available in ExPASy Cellosaurus [30] to prevent loss or duplication due to the usage of several different names for a specific cell line. For the purposes of the present study, we employed the cell-line names in GDSC.

Collection @ ewha

Mutations from 1,630 genes were compared, as those were the data that were available from both databases. Among those common genes, we filtered rare cancer-driver genes in two databases, The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) and DriverDB v.2 [31]. Among the cancer-driver genes defined in TCGA, those identified by more than three mutation callers were chosen. Finally, 897 cancer-driver genes were selected for the present study.

Collection @ ewha

B. Analysis of consistent and inconsistent mutation calls between GDSC and CCLE

Mutation calls were categorized into six mutation types: 1) missense mutation, 2) nonsense mutation, 3) mutation in termination codon (termination mutation), 4) mutation in splicing site (splicing mutation), 5) inframe insertion/deletion (inframe in/del) mutation, and 6) frameshift insertion/deletion (frameshift in/del) mutation. Mutations located outside of the coding DNA sequences and synonymous mutations in CCLE were filtered out prior to comparison of the mutation calls.

Because the mutation calls' genomic positions are not provided in GDSC, their cDNA positions were compared instead. The mutation calls found only in GDSC (i.e., the GDSC-only mutation calls) were cross-referenced against the mutation calls in CCLE to remove mutation call duplication errors. The same process was completed for the mutation calls found only in CCLE (i.e., the CCLE-only mutation calls). In the process, annotation discrepancies were found in 1) in/del mutations as well as in 2) mutations in genes with several gene transcripts. For determination of the mutation calls' genomic positions, reference genome version hg19 was employed.

The consistency or inconsistency rate of mutation calls in GDSC was defined as the percentage of consistent or inconsistent mutation calls between GDSC and CCLE divided by the total mutation calls from GDSC. The consistency

Collection @ ewha

or inconsistency rate in CCLE was defined as the percentage of consistent or inconsistent mutation calls divided by the total mutation calls in CCLE.

i) Consistency rate of mutation calls =

$$\frac{n(consistent\ mutation\ calls\ on\ a\ dataset)}{n(total\ mutation\ calls\ on\ two\ databases)} * 100$$

ii) Inconsistency rate of mutation calls =

$$\frac{n(inconsistent\ mutation\ calls\ on\ one\ dataset)}{n(total\ mutation\ calls\ on\ two\ databases)} * 100$$

Collection @ ewha

C. Analysis of mutation calls in CpG island regions

CpG island regions were deduced using the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables) with the reference genome version of the CpG island dataset of GRCh37/hg19 (Feb. 2009). The genomic positions of the CpG islands were included in an output Browser Extensible Data (BED) file. Due to the lack of gene symbols for the positions of the CpG island regions in the BED file, information in Ensembl Gene 90 [32] was employed by using GRCh37 as a reference genome, which is compatible with UCSC genome hg19.

Among the 897 cancer-driver genes, 747 contained CpG island regions in the exons. According to their positions inside or outside of the CpG island regions, the mutation calls in GDSC and CCLE were classified. Using these data, the consistency and inconsistency rates were calculated. In the analysis, inframe and frameshift in/del mutations were excluded, because they showed relatively low consistent mutation call rates between the two databases.

Collection @ ewha

D. Analysis of mutation profiles from cell lines and primary tumors

The mutation prevalence data from 13 types of primary tumors were obtained from the reports of previous studies [33] and modified according to the median mutation prevalence values. We used the R toolkit (Version 3.4.1) to obtain a strip chart for comparison of the mutation distribution patterns for each cancer type with the same tissue origin. Among the 13 tissue-origin- type cell lines, four were not exactly matched, and the primary cancers from ALL, glioblastoma, head and neck, and melanoma cancers in previous papers were compared with the tissue origins of cancer cell lines from haematopoietic and lymphoid tissue, central nervous system, upper aerodigestive tract, and skin, respectively.

For comparison of the mutation profiles between primary tumors and the breast, lung, central nervous system, and colorectal cancer cell lines, two studies for each cancer type were selected [34-41]. The most prevalent 15 cancer-driver mutant genes for each cancer type were selected, and the mutation incidence between two primary tumor studies and cell-line data were compared by Wilcoxon signed-rank test.

Collection @ ewha

E. Targeted sequencing for 8 cell lines

Among the 592 cell lines common cell lines analyzed in both CCLE and GDSC, 8 — A549, Calu-6, HCT116, IGROV-1, KM12, NCI-H460, OVCAR-8, and SK-OV-3 — were selected for Targeted sequencing. These cells were cultured in RPMI 1640 medium (GE Healthcare, Little Chalfont, UK) supplemented with 10% fetal bovine serum (Hyclone, Logan, UT) and 1% Antibiotic-Antimycotic solution (Anti-Anti; Thermo Fisher Scientific, Waltham, MA). Subsequently, all of the cells were incubated at 37℃ and maintained at 5% CO2.

At 80% confluency, the cells were harvested by trypsinization and washed with DPBS (GE Healthcare). The collected cells were lysed with 270μl of ATL buffer (Qiagen, Hilden, Germany) and 20μl of Proteinase K (Qiagen) and incubated at 56℃ for 16 hours. The incubated cells were centrifuged at 13,000 rpm for one minute; 1μl of 10mg/ml RNase A (Qiagen) was added followed by incubation at room temperature for 10 minutes; 200μl of Buffer AL (Qiagen) was next added, followed by incubation at 70℃ for 10 minutes. Then, the samples were treated with 400μl of PCI buffer (Biosesang, Gyeonggi-do, Korea) and mixed by vortexing. All of the samples were centrifuged at 13000 rpm for 10 minutes. The supernatants were transferred to new tubes and centrifuged again. After centrifugation, the supernatants were harvested and treated with 2μl of 3M Sodium Acetate and 800μl of 100% Ethanol (Merck, Kenilworth, NJ) for precipitation of genomic DNAs. The precipitated genomic DNAs were harvested by centrifugation. The pellets were

Collection @ ewha

washed with 75% Ethanol and centrifuged for 1 minute. After removal of the ethanol, the pellets were dissolved in 100μl of Tris-EDTA buffer (pH 8.0; Thermo Fisher Scientific) and incubated at 55℃ for at least 30 minutes. The eluted genomic DNAs were then quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific).

After confirmation of the cell identities by STR markers, targeted sequencing was performed two times for the DNAs using the Axen Cancer Panel (Agilent, US) that can detect mutations in 170 genes, and the depth was x1,000. Analysis with the Panel provides information on mutations for 151 cancer-driver genes, which information is available in both GDSC and CCLE.

Alignment for targeted sequencing reads was performed using the Burrows-Wheeler Aligner (BWA) [42] to the UCSC reference genome hg19, as well as in the mem mode to produce SAM files from fastq files. Duplicated reads were removed by Picard (http://broadinstitute.github.io/picard/), the output file being BAM. The quality of mapping was confirmed, and post-alignment quality control was performed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) based on the phred-scaled score and a cutoff value of 20. Base quality score recalibration was performed with the Genome Analysis Tool Kit (GATK) [43], and the resulting recalibrated BAM files were obtained using dbSNP 147 as a reference for exclusion of Single-Nucleotide Polymorphism (SNP). For mutation calls, GATK MuTect2

Collection @ ewha

was used without any matched normal DNA information, and Annovar was used for the annotation [44].

Collection @ ewha

F. Analysis of allelic ratio in targeted sequencing data

For further analysis of the reasons for the mutation call discrepancies, allelic-ratio data for each of the mutation calls were extracted from Variant Call Format (VCF) files for targeted sequencing of the aforementioned 8 cell lines. For comparison of the allelic-ratio discrepancies, likewise, we extracted the allelic-ratio values for each of the mutation calls from the VCF files.

According to the results of the two targeted sequencings, the allelic fractions of the consistent and inconsistent mutation calls included in the two databases were compared separately in order to detect any differences in the allelic ratios. Additionally, we compared the distributions of allelic ratios for the mutation calls that had been detected only in targeted sequencing but not in GDSC or CCLE. In this case, we divided the mutation calls into three categories: 1) calls detected only in one targeted sequencing result, 2) calls with an allelic ratio (AR) less than 0.02, 3) calls with total read depth (DP) less than 100. Utilizing the R toolkit (Version 3.4.1), we showed all of the cases on one plot for better comparison of the allelic-fraction discrepancies.

Collection @ ewha

II. Results

A. Study scheme

In this study, we obtained mutation datasets from GDSC and CCLE. The GDSC data consists of 1,047 cell lines and 19,100 mutated genes, and the CCLE data, 947 cell lines and 1,651 genes. Upon data filtration, 592 cell lines and 897 cancer-driver genes were found to be common between the two databases. In the analysis of the datasets' combined total of 46,796 mutation sites, 21,414 were consistent. Additionally, we selected, from combined data, 8 cells on which targeted sequencing was performed two times. The overall scheme of the present study is shown in Figure 1.

Collection @ ewha

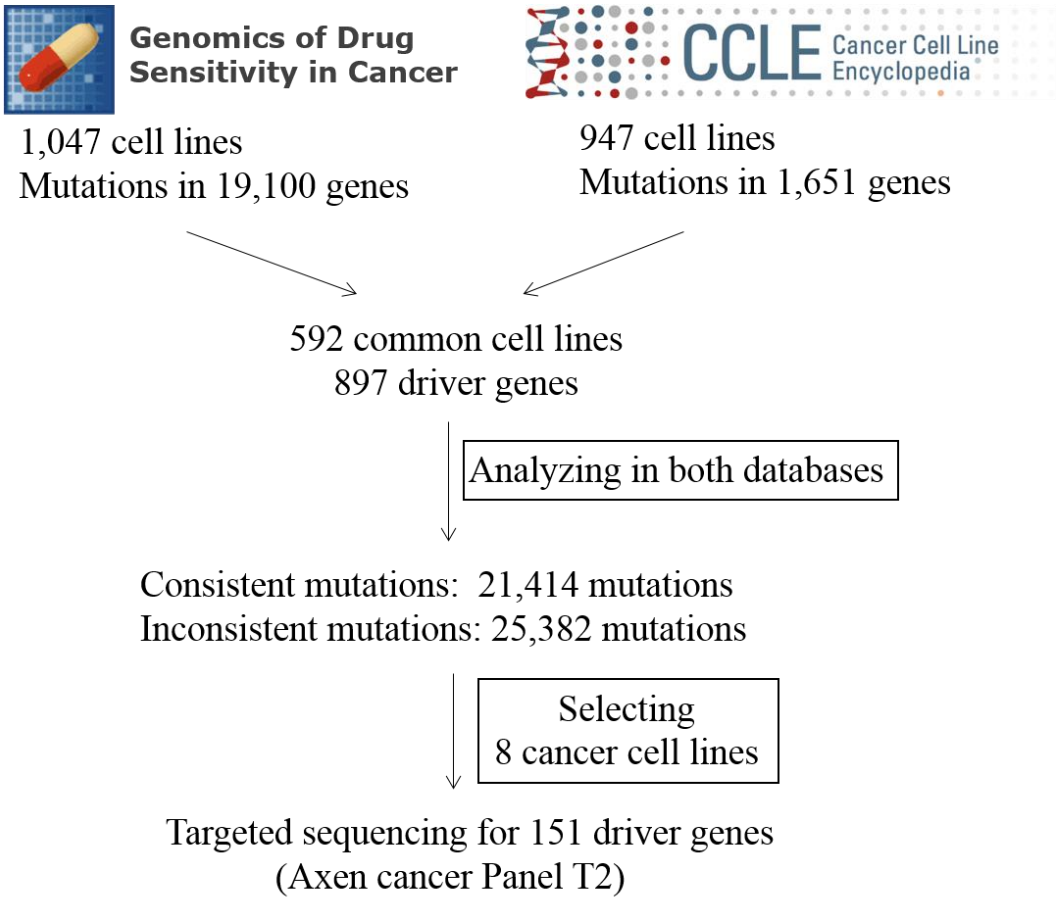

Figure 1. Scheme of present study.

Collection @ ewha

B. Sampling of cell lines

From the GDSC and CCLE, we extracted 592 common cell lines and analyzed the distributions of tissue origins in cancer cell lines. The total number of cell origin types was 24. The tissue of origin with the largest number of cell lines was the lung, with a total of 125 cell lines. By contrast, the salivary gland, biliary tract, prostate and pleura each had less than 10 cell lines for each tissue of origin (Figure 2).

Collection @ ewha

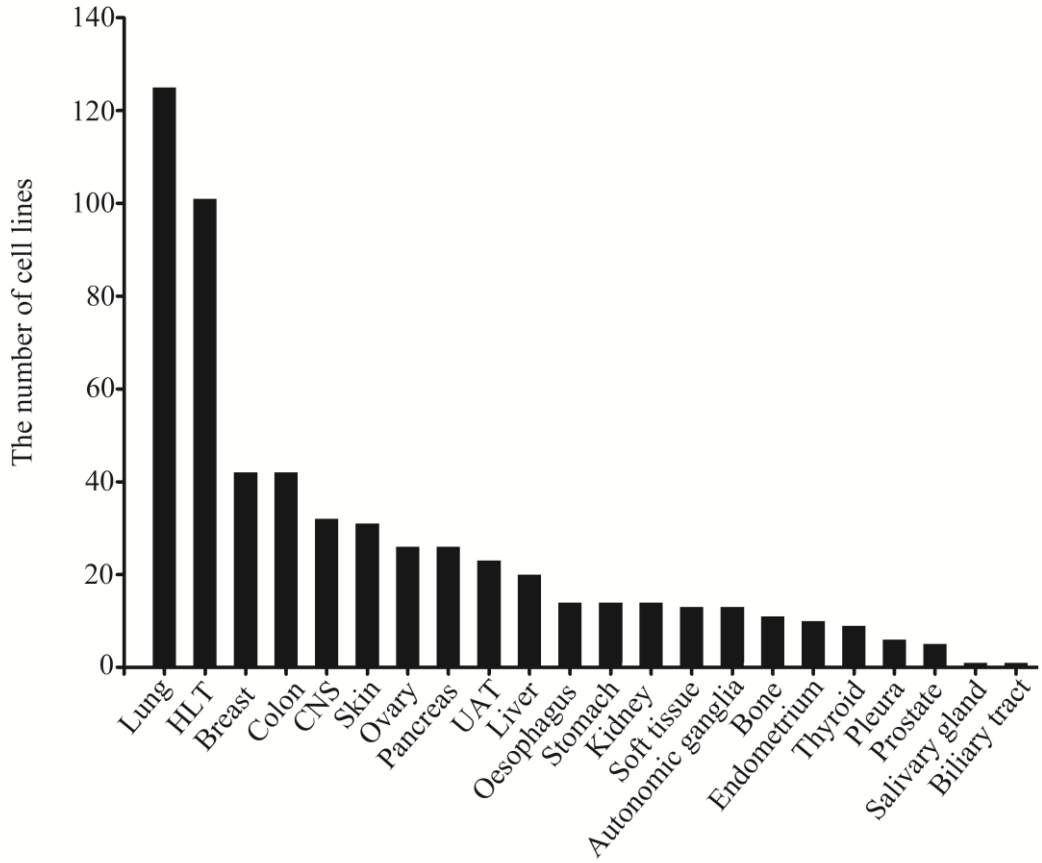

Figure 2. Distribution of tissue origins of 592 cell lines common to GDSC and CCLE databases. X-axis, the tissue origin of cancer cell line. HLT, Haematopoietic and Lymphoid tissue. CNS, Central Nervous System. UAT, Upper Aerodigestive Tract.

Collection @ ewha

C. Annotation discrepancies in GDSC and CCLE

Due to the fact that the genomic coordinates of mutation sites are not available in GDSC, the comparison was conducted based on annotations from the cDNA sequences. During the comparison, inconsistencies in annotations for genes with various transcripts and for in/del mutations in repeated sequence regions were found.

A gene's various transcripts will have various protein products, and its mutations in a specific exon site can provide various mutation products even though they occur at the same genomic site. In our analysis of the CCLE and GDSC databases, 3070 inconsistent mutation calls from 181 genes with various transcripts were found, representing 6.6% of the total (see Figure 3A).

Annotation discrepancies for frameshift and inframe in/del mutations in repeated sequences were found in 312 genes at 718 sites, as representative of 1.5% of the total data. When a unit base or bases in repeated sequences were inserted or deleted, the mutation site could not be specified due to the repeated sequence. For example, an adenosine nucleotide was inserted in a sequence with 7 consecutive adenosines of DYRK3 gene (Figure 3B). Annotations for the mutation in two databases were different: CCLE annotated in front of the first adenosine site; GDSC, however, annotated at the last base at which amino acid change accompanies the insertional site.

Furthermore, although the mutations occurred at a certain site, the annotation tools annotated differently. We plot the annotation results of Oncotator, CRAVAT

Collection @ ewha

(Cancer-Related Analysis of Variants Toolkit) and the Pindel algorithm in Figure 3B.

In addition to annotation discrepancies, there were classification errors in GDSC. A splicing mutation and 2 inframe deletions were mis-classified as frameshift in/del mutations, and 22 frameshift in/del mutations were mis-classified as missense mutations (Table 1). None of the mis-classified mutations in GDSC were detected in CCLE.

Collection @ ewha

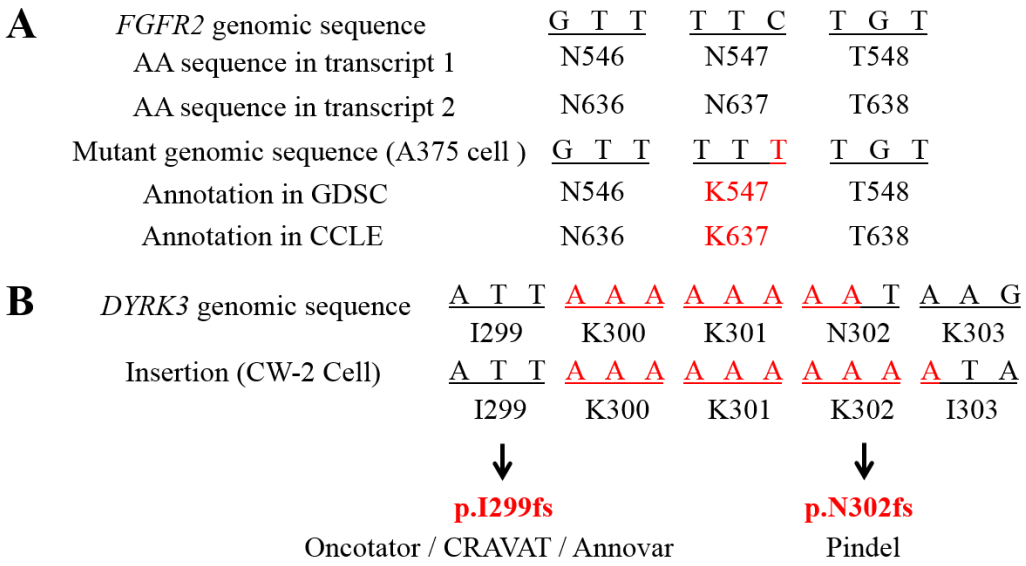

Figure 3. Discrepancies due to different annotation calls from GDSC and CCLE. (A) Annotation discrepancy due to various transcripts. *FGFR2* has several transcripts. GDSC and CCLE mutation calls are based on different transcripts, and the annotations were different for the same genomic mutation. Annotation discrepancy due to the usage of various transcripts was found in 3,070 mutation calls from 181 genes. (B) Annotation discrepancy due to inconsistent mutation calls for same in/del changes in repeated sequences. *DYRK3* has 8 adenosines in the sequence, but the annotations were different due to incorrect mutation call site in CCLE. Annotation discrepancies due to inconsistent mutation calls were found in 718 mutation calls from 312 genes.

Collection @ ewha

Table 1. Errors in classification of GDSC mutation calls

Gene	Cell line	CDS change	Amino acid change	GDSC	Classification
MLH1	SNU-1040	c.1731G>A	p.S556fs*14	Frameshift	ess_splicing
BRAF	NCI-H2405	c.1454_1469_16>A	p.L485_P490>Y	Frameshift	Inframe_deletion
FIP1L1	Mewo	c.961_989_29>CT	p.I321_R330>L	Frameshift	Inframe_deletion
FMN2	NCI-H1048	c.1105_1106insAGC	p.Q368_L369insQ	Missense	Inframe_insertion
FMN2	A704	c.1105_1106insAGC	p.Q368_L369insQ	Missense	Inframe_insertion
PPM1E	MPP-89	c.131_132insACCCGAACCCGA	p.E44_S45insPEPE	Missense	Inframe_insertion
PIK3R1	OVTOKO	c.1356_1357insAAC	p.N453_T454insN	Missense	Inframe_insertion
PIK3R1	Hs-578-T	c.1354_1355insATA	p.N453_T454insN	Missense	Inframe_insertion
FURIN	EC-GI-10	c.1467_1468insCGGCTCACC	p.R497_R498insLTLSYNR	Missense	Inframe_insertion
BRD2	OVK-18	c.1509_1510insGAA	p.E506_S507insE	Missense	Inframe_insertion
FLT3	MOLM-13	c.1774_1775insTTGATTTC	p.E598_Y599insFDFREYE	Missense	Inframe_insertion
FLT3	MV-4-11	c.1771_1772insACGTTGATTTCA	p.D600_L601insHVDFREYEYD	Missense	Inframe_insertion
IRS1	SUP-HD1	c.2036_2037insCAG	p.S686_N687insS	Missense	Inframe_insertion
TBX18	ECC10	c.255_256insACGTCTGGGCCG	p.P85_A86insTSGP	Missense	Inframe_insertion
IRS1	MOLM-16	c.2626_2627insAGC	p.Q882_P883insQ	Missense	Inframe_insertion
MYH9	CAL-33	c.2787_2788insGAG	p.E929_R930insE	Missense	Inframe_insertion
CSF1R	BFTC-905	c.2799_2800insGGCAGC	p.S937_S938insGS	Missense	Inframe_insertion
MAP3K1	EB2	c.2821_2822insCAA	p.T949_E950insT	Missense	Inframe_insertion
COL18A1	HCC-15	c.3518_3519insCGGCCCCCC	p.P1176_S1177insGPP	Missense	Inframe_insertion
PCDH15	BL-41	c.4305_4306insCCG	p.P1443_G1444insP	Missense	Inframe_insertion
PCDH15	KMS-12-BM	c.4305_4306insCCGCCG	p.P1443_G1444insPP	Missense	Inframe_insertion
CLTCL1	MOLM-13	c.4413_4414insCTG	p.L1471_T1472insL	Missense	Inframe_insertion
CASC5	TOV-21G	c.5387_5388insAGA	p.E1795_D1796insE	Missense	Inframe_insertion
MCL1	EN	c.55_56insGGG	p.G18_A19insG	Missense	Inframe_insertion
DOM3Z	SW900	c.924_925insATGAAG	p.M311_F312insKM	Missense	Inframe_insertion

dCollection @ ewha

D. Consistency or inconsistency rates of mutation calls between GDSC and CCLE

In the analysis of the consistency rates among the six mutation types, consistency was higher in point mutations including termination (76-87%), nonsense (77-79%), and missense mutations (69-76%). However, the rate was much lower in frameshift (46-57%) and inframe (3-69%) in/del mutations, especially in the CCLE database, suggesting that mutation calls for in/del mutations in various databases might not be consistent or reliable. Also, the inframe in/del mutations were not included in the present consistency and inconsistency rate analyses. Accordingly, the consistency rate of mutation calls between GDSC and CCLE ranged from 67 to 69% (Figures 4B and 4C).

A previous report [45] compared missense mutations from two databases, and concluded, after calculating all of the consistent missense mutation calls divided by all of the missense mutation calls from the two databases, that the consistency rate was about 57.38%. Our rate, as calculated with the same formula for only missense mutations, was 56.67% (Figure 4A), quite similar to the previous result.

Collection @ ewha

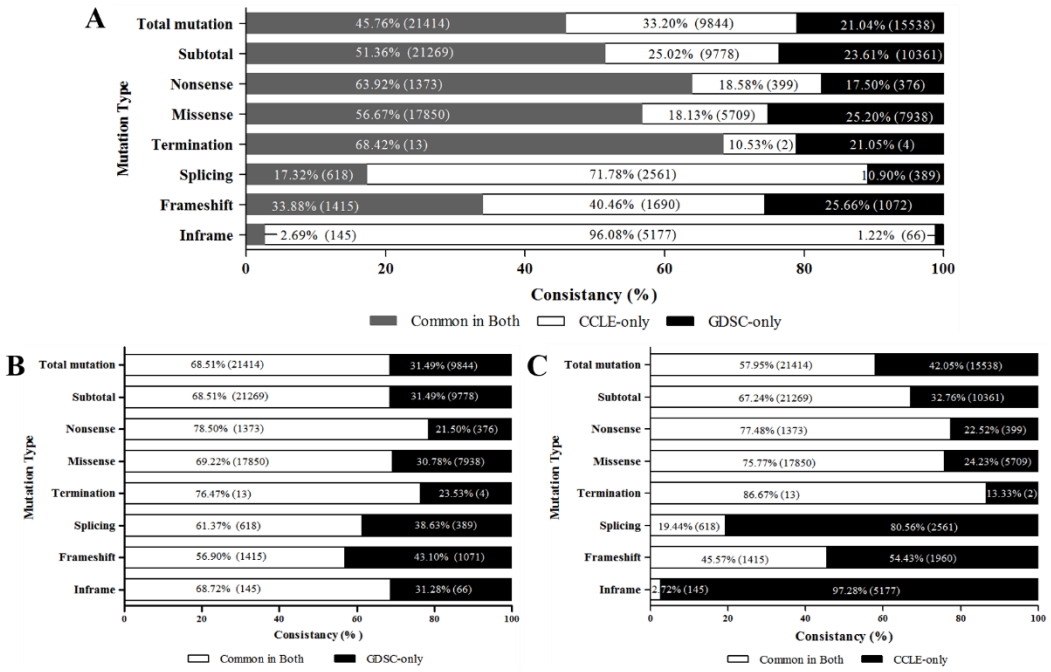

Figure 4. Consistent and inconsistent mutation calls from GDSC and CCLE databases. (A) Consistent and inconsistent mutation calls using total number of all mutant variants found in GDSC or CCLE as denominator. (B) Consistent and inconsistent mutation calls in GDSC using total number of mutation calls in GDSC as denominator. (C) Consistent and inconsistent mutation calls in CCLE using total number of mutation calls in CCLE as denominator. *In the subtotal rates, the inframe in/del mutations were excluded.

Collection @ ewha

E. Inconsistencies of mutation calls in CpG islands between GDSC and CCLE

The reason for the inconsistency in mutation calls between GDSC and CCLE was investigated in an earlier study [45], which found that an important cause of inconsistency was the lower sequencing depth due to poor amplification in CpG island regions. In the present study, the inconsistency rates of mutation calls inside and outside of CpG island regions were compared. The inconsistent mutation calls in CpG island regions were much higher (about 1.5 times higher) than those in non-CpG island regions from the same gene list or in genes without CpG islands, which results confirmed the previous report's conclusion that mutation calls in CpG island regions are an important cause of inconsistency. However, the mutations in CpG island regions represent only about 5% of total mutation calls (Table 2). As such, the major reasons for inconsistent mutation calls outside of CpG island regions remain unclear.

Collection @ ewha

Table 2. Consistent and inconsistent mutation calls in CpG island regions or in non-CpG island regions

	GDSC		CCLE	
	Consistent calls	**Inconsistent calls**	**Consistent calls**	**Inconsistent calls**
CpG island Region	649 (55.7%)	517 (44.3%)	649 (51.3%)	617 (48.7%)
Out of CpG Island	15818 (70.6%)	6595 (29.4%)	15818 (68.9%)	7149 (31.1%)

The difference in consistency between the regions inside and outside of CpG islands was significant ($P < 0.001$ for both GDSC and CCLE, by Chi-square test).

Collection @ ewha

F. Comparison between mutation profiles of cell lines and primary tumors

Between the modified reference data and the mutation distribution of the cell lines, the mutation prevalence for 13 cell types from a previous study showed no similarity for either consistent or inconsistent mutation calls from the two databases (Figure 5). However, between the mutation incidences of consistent and inconsistent calls, the mutation distribution for each of the cell types showed a similar tendency.

With the Wilcoxon-signed rank test, the statistical significance was calculated for the comparison of the mutation prevalence of primary tumor with that of the cell-line data from GDSC and CCLE (Table 3). The comparison between the two datasets from primary tumor studies showed significant correlations for all four types of cells ($P < 0.05$). On the other hand, in our comparison between each primary tumor study and the cell-line data, we did not found any significant pairing between the primary tumor study and the cell-line datasets (Table 4).

Collection @ ewha

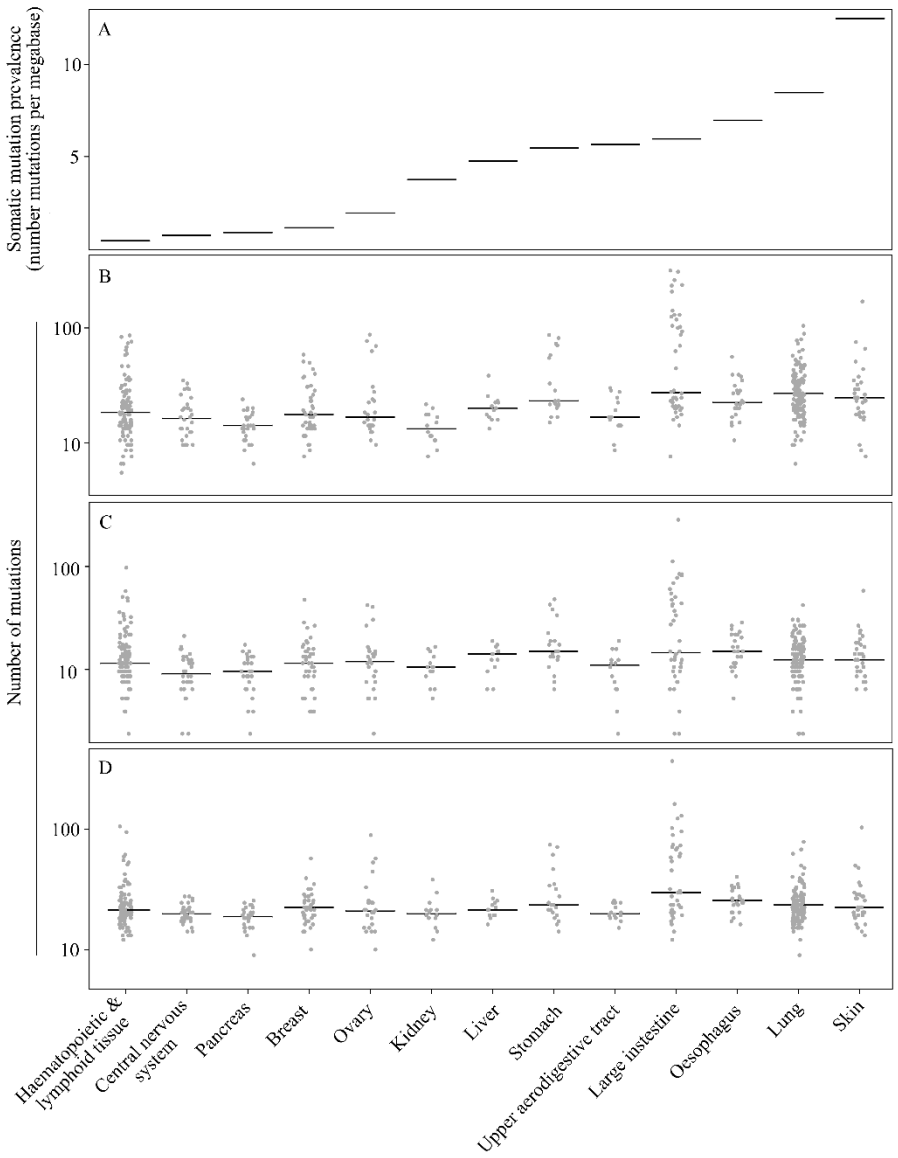

Figure 5. Distributions of mutation rate in various tissue origins of cancers. (A) Mutation prevalence in primary cancers modified from published data (Alexandrov, L.B., et al, 2013) by extraction of mean values. (B) Distributions of mutation calls in both GDSC and CCLE. (C) Distribution of GDSC-only inconsistent mutations. (D) Distribution of CCLE-only inconsistent mutations

Collection @ ewha

Table 3. Preparation of input for Wilcoxon-signed rank test

Colon				Breast			
Gene	**Study I**	**Study II**	**Cell line**	**Gene**	**Study I**	**Study II**	**Cell line**
APC	1	1	1	PIK3CA	1	1	2
TP53	2	2	4	TP53	2	3	1
KRAS	3	3	3	GATA3	3	5	
TTN	4		2	KMT12C	4		
PIK3CA	5	4	10	TBX3	5		
FBXW7	6	6	12	ARID1A	6	12	
SMAD4	7	5		CBFB	7	13	
NRAS	8			AKT1	8	14	
TCF7L2	9			MAP2K4	9	8	
FAM123B	10			NCOR1	10		
SMAD2	11			NF1	11		4
CTNNB1	12	15		PTEN	12	11	
KIAA1804	13			SF3B1	13		
SOX9	14	9		RUNX1	14		
ACVR1B	15	16		CTCF	15		
PTEN		7		CCND1		2	
BRAF		8		MYC		4	
GNAS		10		ERBB2		6	
ATM		11		MAP3K1		7	
ERBB2		12		CDH1		9	5
MLL2		13		ZNF217		10	
LRP1B		14	6	MLLT4		15	
MLL3			5	TTN			3
BMPR2			7	PKHD1			6
HERC2			8	BIRC6			7
UNC13C			9	BRCA2			8
DOCK3			11	CSMD3			9
BRCA2			13	FBN2			10
EP300			14	MLL3			11
LRP2			15	OBSCN			12
				PCDH15			13
				MYH9			14
				HERC2			15

dCollection @ ewha

Table 3. Preparation of input for Wilcoxon-signed rank test (continued)

Central Nervous System				Lung			
Gene	**Study I**	**Study II**	**Cell line**	**Gene**	**Study I**	**Study II**	**Cell line**
TP53	1	2	1	TP53	1	1	1
IDH1	2			KRAS	2	2	7
ATRX	3	9		STK11	3	4	
PTEN	4	1	3	EGFR	4	5	
CDKN2A	5			LRP1B	5		3
EGFR	6	3		NF1	6	6	
MET	7			ATM	7		
H3F3A	8			APC	8		
NF2	9			EPHA3	9		
DDX3X	10			PTPRD	10		
MYCN	11			CDKN2A	11		
PIK3CA	12	5		ERBB4	12		
RB1	13	8	4	KDR	13		
CIC	14			FGFR4	14		
IDH2	15	12		NTRK3	15		
PIK3R1		4		KEAP1		3	
NF1		6	5	BRAF		7	
SPTA1		7		SETD2		8	
TCHH		10		RBM10		9	
KEL		11		MGA		10	
SEMA3C		13		MET		11	
SCN9A		14		ARID1A		12	
PDGFRA		15		PIK3CA		13	
TTN			2	SMARCA4		14	
DNAH8			6	RB1		15	6
PKHD1			7	CSMD3			2
OBSCN			8	TTN			4
MGA			9	NAV3			5
BRAF			10	LRP2			8
HERC2			11	PCDH15			9
CHL1			12	NLRP3			10
				FBN2			11
				DNAH8			12
				FBN1			13
				DST			14
				CTNNA2			15

Collection @ ewha

Table 4. Comparison of mutation incidence in most prevalent 15 cancer-driver genes between primary tumors and cell lines by Wilcoxon-signed rank test

Comparison	Study I Study II	Study I Cell line	Study II Cell line
Colon	0.0006	0.0514	0.0514
Breast	0.0481	0.5000	0.2083
CNS	0.0002	0.1667	0.2083
Lung	0.0083	0.5000	0.5000

The Wilcoxon-signed rank test was used to calculate the correlation significance. *P*-values less than 0.05 were considered significant. In this table, study I and study II mean the two individual studies dealing with the mutation incidence of primary tumors.

Collection @ ewha

G. Comparison of mutation calls by targeted sequencing

Mutation calls from GDSC and CCLE were compared by targeted sequencing twice. The consistency rate of the two targeted sequencing results was 98% (162/165), with discrepancies only in 3 mutation calls (Table 5), which suggested that most of the targeted sequencing results are reliable.

In the comparison of mutation calls from databases with targeted sequencing, 90 or 91% (89-90/99) of the consistent mutation calls were found in targeted sequencing (Table 5), suggesting that only 9-10% of consistent mutations are false-positive mutations. Among the inconsistent mutation calls, 82 or 85% (54-56/66) were found in targeted sequencing, suggesting that although the errors in inconsistent mutation calls might be higher than those in consistent calls, most of the identified mutation calls from the two NGS analyses might be true mutations.

We next analyzed all of the mutation calls from the two targeted sequencings (Table 6). We made mutation calls first with MuTect2, and analyzed again by using three filters: allelic frequency, total depth, and transversion artifacts. The consistency rate of mutation calls between the two targeted sequencings was 90.24 - 93.48% in the first targeted sequencing, and 93.77 - 96.80% in the second.

Collection @ ewha

Table 5. Validation of mutation calls from GDSC or CCLE with targeted sequencing in 8 cancer cell lines

Database vs.	**Mutation calls in targeted sequencing* (Confirming rate in targeted sequencing)****	
	Consistent calls	**Inconsistent calls**
1st targeted sequencing	89/99 (89.9%)	54/66 (81.8%)
2nd targeted sequencing	90/99 (90.9%)	56/66 (84.8%)

*Only mutations found in two databases were analyzed.

** The number of mutations found in targeted sequencing was divided by the total number of mutations from the two databases.

Collection @ ewha

Table 6. Comparison of total mutation calls between two targeted sequencings

Filtration	Consistent mutations	Total mutations in 1st targeted sequencing	Total mutations in 2nd targeted sequencing	Consistent rate for 1st targeted sequencing	Consistent rate for 2nd targeted sequencing
No filtration	17445	19332	18604	90.24% (17445/19332)	93.77% (17445/18604)
*AF >= 0.02	17060	18250	17339	93.48% (17060/18250)	96.17% (17060/17339)
AF >= 0.02 **DP > 100	12722	13824	13143	92.03% (12722/13824)	96.80% (12722/13143)
AF >= 0.02 DP > 100 ***OXOG filter	903	979	948	92.24% (903/979)	95.25% (903/948)

*AF, Allelic Frequency; **DP, Total depth; ***OXOG, transversion artifacts on 8-oxoguanine lesions

Collection @ ewha

H. Analysis of allelic ratio in targeted sequencing

In analysis of allelic frequencies, sequencing depth is important; however, we found that most of the alleles analyzed had enough sequencing depth for the analysis of allelic frequency: in fact, only 16 among the 1022 mutation calls had a sequencing depth less than 100. In the comparison of the mutant allelic ratios from the two targeted sequencing results (Figures 6 and 7), severe discrepancies due to imbalance or loss of either allele were found in 4% of the mutation calls (7/165; see the ovals in Figure 6). These can be related to uneven amplification of alleles, indicating that uneven amplification leads to the loss of mutation information, which results in inconsistencies in mutation calls between GDSC and CCLE.

In 2.4% of the mutation calls (4/165), two targeted sequencings showed no or low (less than 2%)-level mutant allele consistently (see the dotted circle in Figure 6), suggesting that the cells used in the present study might contain no or least level mutant variants due to genetic drift. In 2 other cases, the allelic frequency of the mutant allele was consistently quite low (about 10%), as indicative of the presence of rare cell variants among the cells used in the present study.

We also analyzed mutant allelic frequencies that were not detected in GDSC or CCLE. In most of the mutation calls, the allelic frequencies were linearly correlated in the two targeted sequencing results, though imbalances and allele losses were still shown (Figure 7). Most of the imbalances of allelic frequencies were observed when the mutant allelic frequency was 0.0 - 0.1. Among them, 3.7% (38/1022) of

Collection @ ewha

the mutant alleles were detected in either of the targeted sequencings. Also, 3.2% (33/1022) of the mutant alleles showed skewed allelic frequency, and the allelic ratios from the two targeted sequencings were quite different. These mutant alleles showing allelic imbalances or skewed allelic frequency also seem to be related to uneven amplification during the PCR reaction for the targeted sequencing procedures.

Collection @ ewha

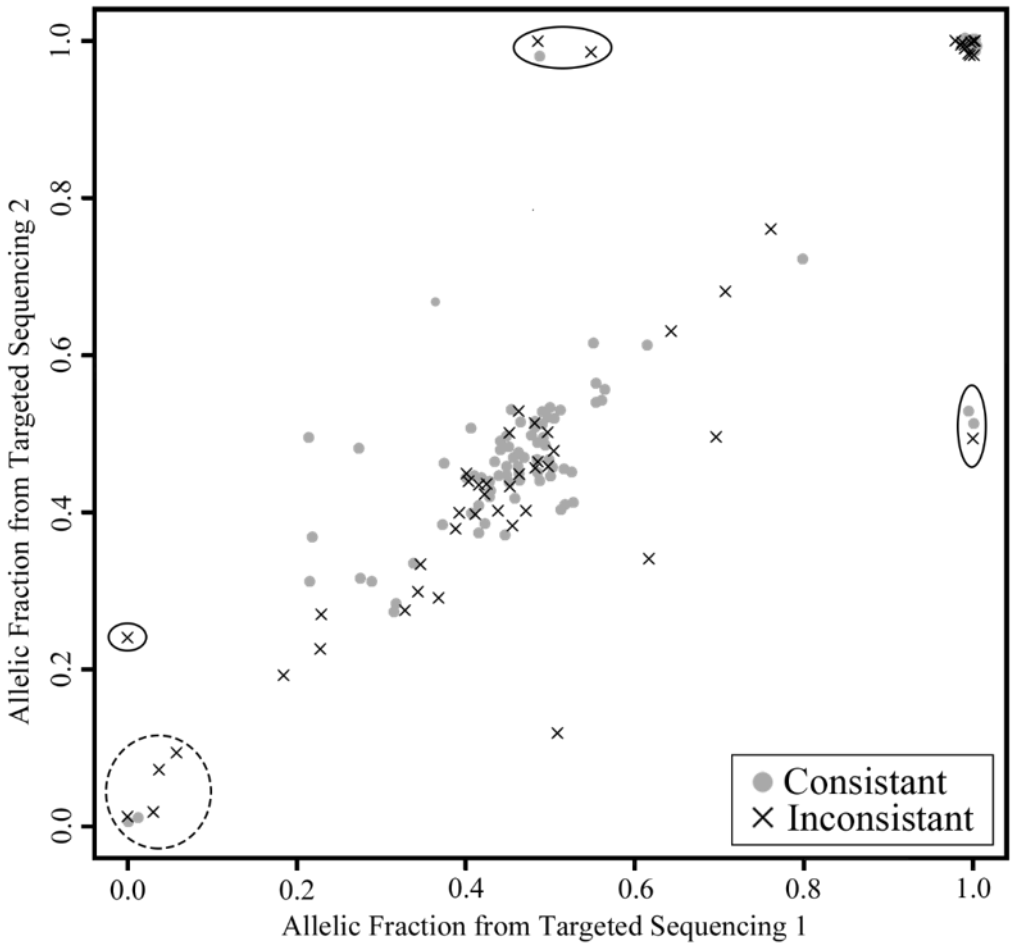

Figure 6. Discrepancy in allelic ratios from two targeted sequencings. The sites on the plot were positive in both GDSC and CCLE. The discrepancies in the allelic frequency of mutation calls in the black ovals might be due to uneven amplification. The discrepancies in the allelic frequency of mutation calls in the dotted circle might be due to either uneven amplification or population drift.

Collection @ ewha

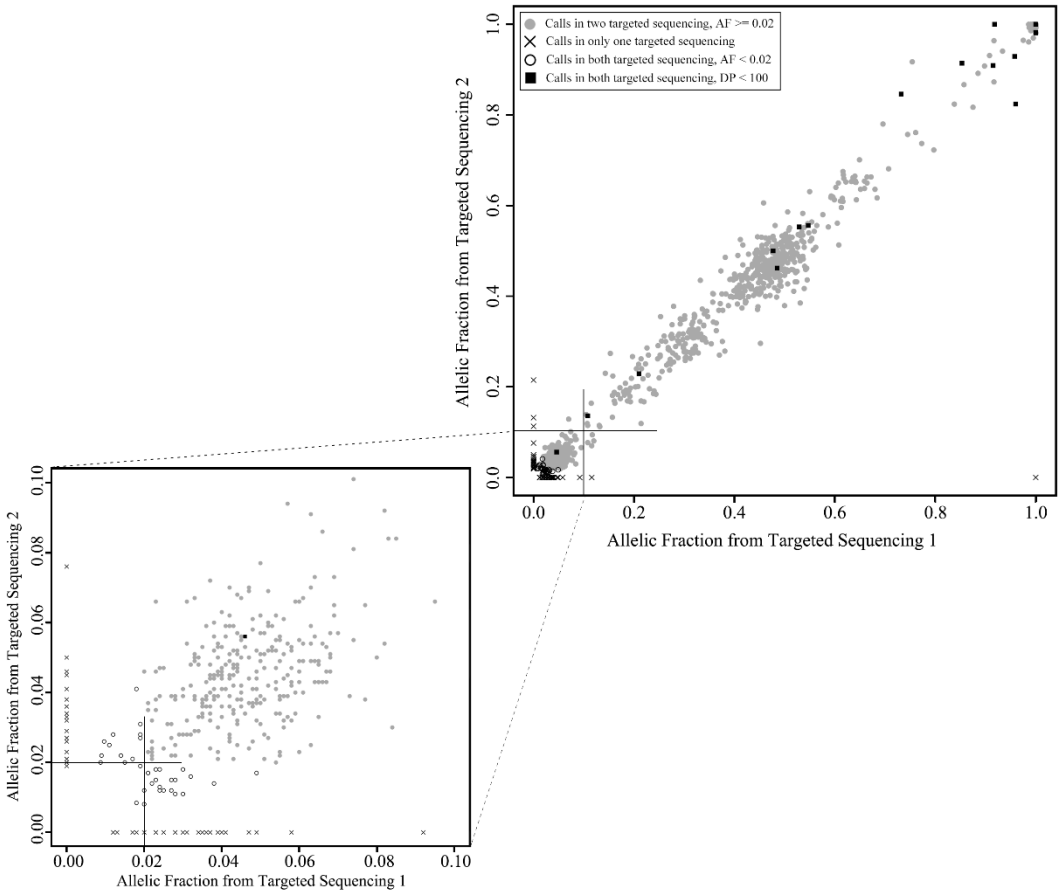

Figure 7. Discrepancies of two targeted sequencing results in allelic ratios for mutation calls not found in GDSC or CCLE. On the plot, the mutation calls that are positive only on two times of targeted sequencing result. We enlarged the plot where the allelic fraction is between 0.0 and 0.1.

Collection @ ewha

IV. Discussion

In the analysis of 897 cancer-driver mutations from the 592 cell lines common to the two cell-line databases, GDSC and CCLE, inconsistencies were found in about 30% of mutation calls. In the comparison of our targeted sequencing results with mutation calls from GDSC and CCLE, 12-13% of mutation calls might be false mutations that had been introduced by proofreading errors from polymerases during target amplification. In the analysis of the allele frequencies among the two independent targeted sequencing results for 8 cell lines, allelic losses and imbalances were found in about 6.7% (11/165) of mutation calls, which can also contribute to the other inconsistencies in mutations calls between the two databases.

In the annotation comparisons between GDSC and CCLE, we found annotation discrepancies in mutation calls, and those were related to 1) splicing variants, and 2) in/del mutation calls in repeated sequences. Many genes have splicing variants [46], and their annotation can be inconsistent when different transcripts are used as the reference sequence. Because the mRNA sequence information is not available for most cancer cell types or cell lines, correct annotations for mutations in genes with multiple transcripts usually are not possible. Therefore, determination of the means of selecting a reference transcript among multiple transcripts for a specific gene is required. For instance, using the longest transcript as a reference can reduce annotation discrepancies for mutations in genes with various transcripts. In addition

Collection @ ewha

to splicing variants, there are discrepancies in in/del mutation calls in repeated sequences. The current annotation tools such as Oncotator [47], Annovar [44] or CRAVAT [48] should consider and correct these annotation discrepancies in mutation calls for genes with various splicing variants, and in in/del mutations in repeated sequences.

The inconsistency rate of point mutation calls from GDSC and CCLE in the present study was about 30%. In a previous report of a study that analyzed only missense mutations [45], the consistency rate was about 57.38%. In that study, all of the mutation calls from the two databases were used for the denominator in the calculation. In the present study, when calculated with the same formula, the discrepancy rate for missense mutations was quite similar (56.67%), even after correction of annotation discrepancies. However, using all mutations identified from several studies as a denominator might exaggerate the discrepancy rates in repeated experiments. Therefore, the present study, using all of the mutations found in a specific study as a denominator, found that the consistency rates for missense mutations were 77.5 and 78.5%, and the inconsistency rates 21.5 and 22.5%, for CCLE and GDSC respectively.

The reasons for inconsistencies in mutation calls have not been clear. One report suggests that the low reading depth in CpG island regions is an important reason [45], and our analysis also showed that the discrepancy was significantly higher in CpG island regions than in non-CpG island regions in the same gene list. However,

Collection @ ewha

mutations in CpG island regions comprise only 5% of all mutations, and the reasons for inconsistency in most other regions are not clear. In the comparison with targeted sequencing results in the present study, we estimated that up to 13% of mutation calls in GDSC and CCLE are false-positive mutations related to polymerase proofreading errors. In the current analysis of allelic ratios from the two targeted sequencings, severe allelic imbalances and losses were found in 4.2% of mutations, which was indicative of their significant contribution to the inconsistency of mutation calls. Uneven or biased amplification has already been recognized as a cause of allelic imbalance during amplification [49], but the effect on inconsistent mutation calls has not yet been evaluated. In our results, uneven and biased amplification can be another important reason for inconsistency of mutation calls in NGS data.

In our analysis of targeted sequencing, only a limited number of cell lines and genes were analyzed. Therefore, more extensive targeted sequencing studies for cell lines are warranted for confirmation of mutation status in cancer cell lines. Additionally, the current study did not estimate the false-negative rates of mutation calls; future targeted sequencing studies will reveal more accurate mutation profiles by simultaneous analysis of allelic ratios for all mutation calls. Although the present study did not explore the influence of using various SNP databases as a polymorphism reference on mutation calls, inconsistent mutation calls, in fact, might increase abruptly when various SNP databases are employed. Therefore,

Collection @ ewha

reference SNPs, at least in cancer-driver genes among various ethnic populations, should be defined more precisely.

In conclusion, most of the inconsistent mutation calls in GDSC and CCLE might be true mutations, and uneven or biased amplification and genetic drift can be major reasons for discrepant mutation calls in GDSC and CCLE cell-line data. In future studies, other reasons for inconsistent mutation calls due to employment of various SNP databases, as well as for false-negative mutation calls, should be evaluated further.

Collection @ ewha

V. References

1. Bean, L.J. and M.R. Hegde, Gene Variant Databases and Sharing: Creating a Global Genomic Variant Database for Personalized Medicine. *Hum Mutat*, 2016. **37**(6): p. 559-63.
2. Dong, L., et al., Clinical Next Generation Sequencing for Precision Medicine in Cancer. *Curr Genomics*, 2015. **16**(4): p. 253-63.
3. Xue, Y. and W.R. Wilcox, Changing paradigm of cancer therapy: precision medicine by next-generation sequencing. *Cancer Biol Med*, 2016. **13**(1): p. 12-8.
4. Sim, S.C., R.B. Altman, and M. Ingelman-Sundberg, Databases in the area of pharmacogenetics. *Hum Mutat*, 2011. **32**(5): p. 526-31.
5. Yang, W., et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, 2013. **41**(Database issue): p. D955-61.
6. Barretina, J., et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 2012. **483**(7391): p. 603-7.
7. Mardis, E.R., Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 2008. **9**: p. 387-402.
8. Liu, L., et al., Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012. **2012**: p. 251364.
9. Ansorge, W.J., Next-generation DNA sequencing techniques. *N Biotechnol*, 2009. **25**(4): p. 195-203.
10. Quail, M.A., et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 2012. **13**: p. 341.
11. Grotta, S., et al., Advantages of a next generation sequencing targeted approach for the molecular diagnosis of retinoblastoma. *BMC Cancer*, 2015. **15**: p. 841.
12. van Dijk, E.L., et al., Ten years of next-generation sequencing technology. *Trends Genet*, 2014. **30**(9): p. 418-26.
13. Reis-Filho, J.S., Next-generation sequencing. *Breast Cancer Res*, 2009. **11 Suppl 3**: p. S12.
14. Metzker, M.L., Sequencing technologies - the next generation. *Nat Rev*

Collection @ ewha

Genet, 2010. **11**(1): p. 31-46.
15. Shendure, J. and H. Ji, Next-generation DNA sequencing. *Nat Biotechnol*, 2008. **26**(10): p. 1135-45.
16. Davey, J.W., et al., Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, 2011. **12**(7): p. 499-510.
17. Simon, R. and S. Roychowdhury, Implementing personalized cancer genomics in clinical trials. *Nat Rev Drug Discov*, 2013. **12**(5): p. 358-69.
18. Bashiardes, S., et al., Direct genomic selection. *Nat Methods*, 2005. **2**(1): p. 63-9.
19. Ng, S.B., et al., Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009. **461**(7261): p. 272-6.
20. Turner, E.H., et al., Methods for genomic partitioning. *Annu Rev Genomics Hum Genet*, 2009. **10**: p. 263-84.
21. Mertes, F., et al., Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics*, 2011. **10**(6): p. 374-86.
22. Hodges, E., et al., Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, 2007. **39**(12): p. 1522-7.
23. Bodi, K., et al., Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech*, 2013. **24**(2): p. 73-86.
24. Mamanova, L., et al., Target-enrichment strategies for next-generation sequencing. *Nat Methods*, 2010. **7**(2): p. 111-8.
25. Meacham, F., et al., Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 2011. **12**: p. 451.
26. Dohm, J.C., et al., Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*, 2008. **36**(16): p. e105.
27. Junemann, S., et al., Updating benchtop sequencing performance comparison. *Nat Biotechnol*, 2013. **31**(4): p. 294-6.
28. Fox, E.J., et al., Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl*, 2014. **1**.
29. DePristo, M.A., et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 2011. **43**(5): p. 491-8.
30. Artimo, P., et al., ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*, 2012. **40**(Web Server issue): p. W597-603.

dCollection @ ewha

31. Chung, I.F., et al., DriverDBv2: a database for human cancer driver gene research. *Nucleic Acids Res*, 2016. **44**(D1): p. D975-9.
32. Spudich, G.M. and X.M. Fernandez-Suarez, Touring Ensembl: a practical guide to genome browsing. *BMC Genomics*, 2010. **11**: p. 295.
33. Alexandrov, L.B., et al., Signatures of mutational processes in human cancer. *Nature*, 2013. **500**(7463): p. 415-21.
34. Cancer Genome Atlas, N., Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 2012. **487**(7407): p. 330-7.
35. Schrock, A.B., et al., Genomic Profiling of Small-Bowel Adenocarcinoma. *JAMA Oncol*, 2017.
36. Ding, L., et al., Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 2008. **455**(7216): p. 1069-75.
37. Cancer Genome Atlas Research, N., Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 2014. **511**(7511): p. 543-50.
38. Pereira, B., et al., Erratum: The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*, 2016. **7**: p. 11908.
39. Nik-Zainal, S., et al., Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 2016. **534**(7605): p. 47-54.
40. Nikiforova, M.N., et al., Targeted next-generation sequencing panel (GlioSeq) provides comprehensive genetic profiling of central nervous system tumors. *Neuro Oncol*, 2016. **18**(3): p. 379-87.
41. Brennan, C.W., et al., The somatic genomic landscape of glioblastoma. *Cell*, 2013. **155**(2): p. 462-77.
42. Li, H. and R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009. **25**(14): p. 1754-60.
43. McKenna, A., et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010. **20**(9): p. 1297-303.
44. Wang, K., M. Li, and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 2010. **38**(16): p. e164.
45. Hudson, A.M., et al., Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery. *Cancer Res*, 2014. **74**(22): p. 6390-6396.
46. Modrek, B. and C. Lee, A genomic view of alternative splicing. *Nat Genet*, 2002. **30**(1): p. 13-9.

Collection @ ewha

47. Ramos, A.H., et al., Oncotator: cancer variant annotation tool. *Hum Mutat*, 2015. **36**(4): p. E2423-9.
48. Douville, C., et al., CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics*, 2013. **29**(5): p. 647-8.
49. Kebschull, J.M. and A.M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res*, 2015. **43**(21): p. e143.

Collection @ ewha

VII. Abstract in Korean

송유라

의과학과 생화학전공

이화여자대학교 대학원

동일한 세포주에 대한 돌연변이가 보고된 데이터베이스들에서 일치하지 않는다는 것이 보고되면서, 차세대 염기서열 분석기술을 적용한 데이터베이스가 가질 수 있는 오류에 대한 문제점이 제기되었다. 이러한 돌연변이 결과의 불일치에 대한 원인을 찾기 위해, GDSC 와 CCLE 와 같은 두 데이터베이스에서 공통으로 분석하였던 592 개 세포주의 897 가지 유전자에 나타난 돌연변이를 분석하였다. 두 데이터베이스의 돌연변이를 비교하는 과정에서, 전체 돌연변이의 7.2%가 동일한 변이를 다른 돌연변이로 명명한 것을 확인할 수 있었다. 그런데 이러한 돌연변이의 명명의 불일치를 교정하였음에도 불구하고, 여전히 두 데이터베이스 사이에 불일치를 보이는 돌연변이가 전체 돌연변이 중 33-42% 정도라는 것을 확인하였다. 이러한 불일치의 원인을 분석하기 위해 8 개의 세포주에 대해 표적 서열분석을 두 번 시행하였다. 두 번의 표적서열분석 결과는 두 데이터베이스에서 보고하였던 돌연변이 중 98.8% (162/164)가 일관된 결과를 보였으나, CCLE 에서 발견된 7-8% 돌연변이와, GDSC 에서 발견된 11-13% 돌연변이가 표적 서열분석 결과에서는 검출되지 않았다. 이러한 CCLE 혹은 GDSC 에서만 발견되고 표적서열분석에서는 발견되지 않은 돌연변이들은 위양성일 가능성이 크다고 할 수 있다. 그렇지만 통상적으로 두 데이터베이스간 불일치한 돌연변이들이 모두 거짓일 수 있다는 가정과는 다르게, 두 데이터베이스에서 일치하지 않았던 돌연변이라도 이중 85-86%는 진짜 존재하는

Collection @ ewha

돌연변이라는 것을 시사하는 결과이다. 이렇게 두 데이터베이스에서 불일치 하였지만 표적열분석에서 검출된 돌연변이는 GDSC 의 돌연변이 중 14%와 CCLE 돌연변이 중 20%를 차지하였다. 이러한 결과는 두 데이터베이스에서 돌연변이가 불일치한 이유로 데이터베이스에 존재하는 위음성과 연관이 있다는 것을 시사한다. 이러한 불일치의 원인을 더 분석하기 위하여 표적서열분석 결과로부터 대립유전자 빈도를 추가로 분석하였더니, 두 데이터베이스의 공통 돌연변이 중 돌연변이유전자가 표적서열분석결과 전혀 발견되지 않은 경우는 2% (4/155)였고, 한 개의 데이터베이스에서만 나타난 불일치 돌연변이가 표적서열분석에서 발견되지 않은 경우는 4% (7/155)를 차지하였다. 이와 같은 결과는, 데이터베이스간의 불일치와 차세대염기서열분석으로 인한 돌연변이 검출 실패가 관련이 있다는 것을 시사하는 것이다. 이상과 같은 결과를 종합해 보면, 두 데이터베이스간 돌연변이결과 비교 연구를 통해, GDSC 와 CCLE 와 같은 데이터베이스는 전체 돌연변이 중 7-13%의 위양성 돌연변이이며, 이는 중합효소 교정 오류(polymerase proofreading error)에 의해 발생하는 것이다. 또한 불규칙한 증폭(uneven amplification)과 유전적 부동(genetic drift)은 돌연변이의 불일치와 위음성 돌연변이의 중요한 원인이 될 수 있다는 사실을 알게 되었다. 이상의 연구결과는 차세대 염기서열 분석 기술을 기반으로 하여 만들어진 데이터베이스가 갖고 있는 돌연변이 분석결과의 불일치 문제의 원인을 이해하고 불일치를 해소할 수 있는 중요한 열쇠가 될 수 있을 것이다.

dCollection @ ewha