ORIGINAL

Construction of a combinatorial pipeline using two somatic variant calling methods for whole exome sequence data of gastric cancer

Tomohiro Kohmoto[1], Kiyoshi Masuda[1], Takuya Naruto[1], Shoichiro Tange[1], Katsutoshi Shoda[1,2], Junichi Hamada[1,2], Masako Saito[1], Daisuke Ichikawa[2], Atsushi Tajima[1,3], Eigo Otsuji[2], and Issei Imoto[1]

[1]*Department of Human Genetics, Graduate School of Biomedical Sciences, Tokushima University, Tokushima, 770-8503, Japan,* [2]*Division of Digestive Surgery, Department of Surgery, Kyoto Prefectural University of Medicine, Kyoto, 602-8566, Japan,* [3]*Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, Ishikawa 920-8640, Japan*

Abstract : High-throughput next-generation sequencing is a powerful tool to identify the genotypic landscapes of somatic variants and therapeutic targets in various cancers including gastric cancer, forming the basis for personalized medicine in the clinical setting. Although the advent of many computational algorithms leads to higher accuracy in somatic variant calling, no standard method exists due to the limitations of each method. Here, we constructed a new pipeline. We combined two different somatic variant callers with different algorithms, Strelka and VarScan 2, and evaluated performance using whole exome sequencing data obtained from 19 Japanese cases with gastric cancer (GC) ; then, we characterized these tumors based on identified driver molecular alterations. More single nucleotide variants (SNVs) and small insertions/deletions were detected by Strelka and VarScan 2, respectively. SNVs detected by both tools showed higher accuracy for estimating somatic variants compared with those detected by only one of the two tools and accurately showed the mutation signature and mutations of driver genes reported for GC. Our combinatorial pipeline may have an advantage in detection of somatic mutations in GC and may be useful for further genomic characterization of Japanese patients with GC to improve the efficacy of GC treatments. J. Med. Invest. 64 : 233-240, August, 2017

Keywords **:** *Gastric cancer, exome, somatic mutations, variant calling algorithms*

INTRODUCTION

High-throughput next-generation sequencing (NGS), together with the development of powerful computational tools, has transformed biological and biomedical research, particularly cancer research, over the past several years. In a wide variety of tumor types, including gastric cancer (GC), the complex genotypic landscapes of somatic variants have been investigated (1-3). Most significantly, a number of clinically actionable mutations have been identified as therapeutic targets for cancer therapies, narrowing the gap between basic research and clinical application and forming the basis for personalized medicine in the clinical setting (4).

During characterization of cancer genomes, calling somatic variations, mainly single nucleotide variants (SNVs) and small insertions/deletions (indels), by comparing a tumor sample with a matched normal sample is the critical step (5). Although advances in NGS technologies and computational algorithms have led to higher accuracy in somatic variant calling, this step is still difficult due to low allele frequencies, low sample purity, clonal heterogeneity, inadequate sequencing coverage, sequencing errors, and ambiguities in short read mapping (6). To meet the challenges of somatic variant calling, a number of tools with enhanced accuracy have been developed that compare a tumor–normal pair directly at each locus of a possible variant (7). Although each new tool has been compared with some earlier applications (8), the accuracy of the combination methods using multiple tools and their relative advantages in real applications are largely unknown.

GC is a leading cause of global cancer mortality, with high incidence rates in Asia, including Japan (9). Recent genome sequencing studies have provided valuable insights into the key genetic alterations of GC, resulting in identification of its major driver genes (10-14). However, potential genomic alterations among Japanese individuals are not well understood, although several studies identified the potential mutations in specific subtypes of GC, such as diffuse-type (13) and mucinous GCs (14). Recent reports from The Cancer Genome Atlas (TCGA) research network and the Asian Cancer Research Group (ACRG) explored the molecular landscape of GC based on genetic/epigenetic and genetic profiles of GCs, respectively (3, 15), and provided four subtypes in each group. These two classifications showed differences at least partially explained by the difference in ethnic origin of the patients : patients from USA and Western Europe in the TCGA and those from Korea in the ACRG (16). Therefore, further analyses clarifying the molecular landscape of GC in Japanese populations is still needed.

For further detailed exploration of the genetic basis of GC in Japanese individuals, somatic variant detection methods with higher performance compared with frequently used somatic variant calling tools are necessary. Herein, we constructed a new pipeline. We combined two somatic variant callers with different algorithms, Strelka (17) and VarScan 2 (18), and evaluated the performance of this newly constructed method using whole exome sequencing data obtained from 19 Japanese cases with GC ; then, we characterized these tumors based on identified driver molecular alterations.

Received for publication April 28, 2017 ; accepted May 8, 2017.

Address correspondence and reprint requests to Issei Imoto, MD, PhD, Department of Human Genetics, Graduate School of Biomedical Sciences, Tokushima University, Tokushima, 770-8503, Japan and Fax : +81-88-633-7453.

MATERIALS AND METHODS

Patients and DNA samples

Frozen GC samples and paired non-tumorous gastric tissues were obtained from 19 patients with histologically proven primary GC who underwent gastrectomy at the Kyoto Prefectural University of Medicine Hospital (Kyoto, Japan) between 2013 and 2014 (Table 1). None had synchronous or metachronous multiple cancers in other organs. Relevant clinical data were available for all patients. The pathological classification of tumors was determined according to UICC classification (19). Of 19 cases, 12 and 5 cases were differentiated and undifferentiated GCs, respectively. The study was performed according to the Declaration of Helsinki protocols. Formal written consent was obtained from all patients after the local ethics committee (Kyoto Prefectural University of Medicine and Tokushima University) approved all aspects of these studies. Epstein–Barr virus (EBV)-associated GC (EBVaGC) was determined by *in situ* hybridization (ISH) of EBV-encoded small RNAs as described elsewhere (20). Genomic DNA from the cancerous or paired non-tumorous gastric tissues was extracted using the AllPrep DNA/RNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols.

Exome sequencing

A flow chart for exome sequencing and data processing is shown in Figure 1. Exome capture was performed using the Truseq DNA Sample Prep Kit (Illumina, San Diego, CA) or SureSelect XT Human All Exon Kit V5 (Agilent Technologies, Santa Clara, CA). Libraries were sequenced using the HiSeq 1500 or 2500 platform (Illumina) with 101-bp paired-end reads. Image analysis and base calling were performed using HiSeq Control Software v2.2.38 (Illumina), Real Time Analysis v1.18.61 (Illumina), and bcl2fastq Conversion Software v1.8.4 (Illumina). Reads were quality-filtered using a FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and were aligned to the human genome sequence assembly hg19 (GRCh37) using the Burrows-Wheeler Alignment tool v0.7.12 (21). The alignments were converted from a sequence alignment map format to sorted and indexed binary alignment map (BAM) files, and duplicated reads were removed using SAMtools v0.1.19 and v1.2 (22). Local realignments around indels and base quality score recalibration were performed using the Genome Analysis Toolkit version 3.3-0 (23). A summary of exome sequence performance is shown in Table 2. To perform the SNV analysis, pileup files were created from the alignment map files by SAMtools and applied to the two somatic variant calling tools, VarScan 2 v2.3.7 and/or Strelka v1.0.14. SNVs were identified by VarScan 2, filtered with minor allele frequency > 0.05 from the paired non-tumorous tissues, and were considered significant at $P < 0.05$ by Fisher's exact test. Variants that passed the filters were annotated using ANNOVAR ver 2015March (24). To detect somatic copy number variations (CNVs), we applied VarScan 2 to the pileup map files as follows : 1) log coverage ratios were calculated to compare the GC tissues with paired non-tumorous tissues and 2) regions with CNVs were detected using circular binary segmentation with DNAcopy (R /Bioconductor). The relatives were adjusted by the median of each paired sample, and determined as amplification ($\log_2$ ratio > 2) or deletion ($\log_2$ ratio < 0.5).

Global methylation analysis

Bisulfite conversion of DNA was conducted using the EZ DNA Methylation Gold Kit (Zymo Research, Irvine, CA, USA). According to the manufacturer's instructions, HumanMethylation450K BeadChip (Illumina) analysis was performed on 16 cases whose genomic DNA were available for analysis. The default settings of GenomeStudio Software's DNA methylation module (Illumina) were applied to calculate the methylation levels of CpG sites as β-

Table 1. Clinicopathological characteristics of 19 patients with gastric cancer

Sample	Age (yr)	Gender	Main location of Tumor	Borrmann type	Tumor diameter (mm)	Histological predominant type	Histological type	pT	N	Stage	ly	v	EBV
1	65	F	Lower	0-IIc	25	tub1	tub1+2	T1b (SM)	1	IB	1	2	(+)
2	79	F	Lower	2	46	por2	por2>>tub2	T2 (MP)	0	IB	3	0	(-)
3	71	M	Lower	2	80	tub2	tub2>por2	T3 (SS)	0	IIA	3	1	(+)
4	79	F	Upper	1	62	tub1	tub1>tub2	T3 (SS)	1	IIB	0	3	(-)
5	69	M	Lower	2	27	tub1	tub1>tub2	T2 (MP)	2	IIB	1	3	(-)
6	74	M	Upper	2	57	tub2	tub2>por2	T3 (SS)	1	IIB	1	0	(-)
7	84	M	Upper	2	74	por2	por	T2 (MP)	2	IIB	3	1	(+)
8	51	F	Upper	1	32	tub2	tub2>tub1>pap>por2	T2 (MP)	2	IIB	1	2	(-)
9	79	F	Upper	2	82	tub1	tub1>tub2	T3 (SS)	2	IIIA	3	3	(-)
10	74	M	Middle	1	36	tub1	tub1>>tub2>pap<muc	T3 (SS)	2	IIIA	3	1	(-)
11	81	M	Middle	1	51	tub2	tub2>>por2	T3 (SS)	2	IIIA	3	3	(+)
12	66	M	Upper	3	96	tub1	tub1	T3 (SS)	3a	IIIB	3	1	(-)
13	74	M	Upper	3	74	por2	por2>tub2	T3 (SS)	3b	IIIB	3	3	(+)
14	74	F	Lower	1	42	tub1	tub1>pap>muc	T3 (SS)	3a	IIIB	3	0	(-)
15	57	M	Lower	3	75	por2	por2>sig>>tub1	T3 (SS)	3a	IIIB	3	1	(-)
16	65	F	Lower	4	106	sig	sig>por>>tub2	T3 (SS)	3a	IIIB	3	0	(+)
17	76	M	Upper	2	88	por2	tub2>tub1>>por2, tub2>tub1>muc	T4a (SE)	3a	IIIC	3	3	(+)
18	80	F	Middle	4	108	por2	por2>>tub2	T4a (SE)	3b	IIIC	3	0	(-)
19	70	M	Lower	0-IIa+Is	95	tub2	tub2>tub1>por2	T4a (SE)	1	IV	3	0	(+)

The gray shaded area shows samples analyzed using Truseq Exome Kit (Illumina) and Hiseq 1500 sequencer (n=6).

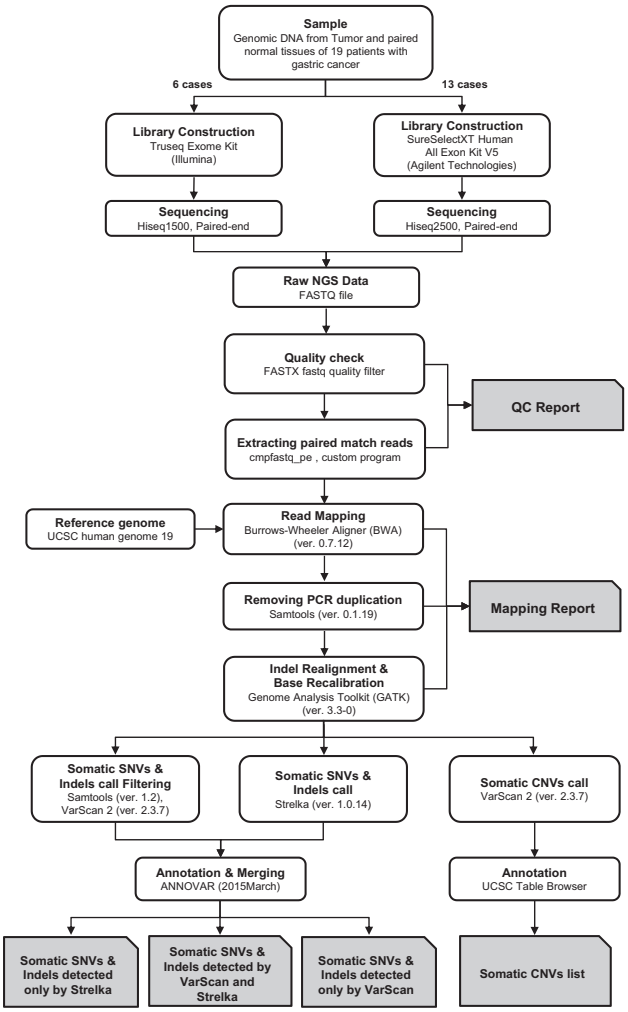

Figure 1. Schematic representation of the overall experimental design
Exome sequencing was performed on paired tumor–normal DNA from 19 GC patients followed by somatic variant calling using two different somatic variant callers.

values [β = intensity (methylated)/intensity (methylated + unmethylated)]. The data were further normalized using a peak correction algorithm embedded in the R-package of Illumina Methylation Analyzer (25). Averaged β-difference in CpG island-based regions of the *MLH1* gene was calculated based on a β-difference matrix in which β-values of paired non-tumorous gastric tissues were subtracted from those of tumors.

RESULTS

Classification and comparison of variants categorized by two different tools

A total 7046 and 4896 SNVs and 795 and 1446 indels were detected by Strelka and VarScan 2, respectively, suggesting that SNVs were better detected by Strelka than VarScan 2, whereas indels were better detected by VarScan 2 than Strelka (Table 3). To analyze common or different variants called by these two tools, we classified detected variants into three categories : (I) detected only by Strelka, (II) detected only by VarScan 2, and (III) detected by both Strelka and VarScan 2. As shown in Figure 2, more than half of variants were commonly detected by both tools : 4177 of 7765 SNVs (53.8%) and 757 of 1484 indels (51.0%). The number of SNVs detected only by Strelka was higher than those detected by VarScan 2 (2869/7765, 36.9% vs. 719/7765, 9.3%), while the number of indels detected only by VerScan 2 was higher than those detected by Strelka (689/1484, 46.4% vs. 38/1484, 2.6%).

Somatic variants detected by each tool

To assess the accuracy of variants called by each tool, we compared detected variants with the pathogenic somatic mutation v70 dataset (the Catalogue of Somatic Mutations in Cancer, COSMIC ; http://cancer.sanger.ac.uk/cosmic) or the genetic variation dataset (dbSNP v138, https://www.ncbi.nlm.nih.gov/projects/SNP/) (Table 4). Accurately called somatic mutations are supposed to overlap mutations in the COSMIC dataset, while inaccurately called germline variants may overlap common variations in the dbSNP dataset. Overlapping rates between called mutations and data in COSMIC were almost the same between Strelka and VarScan 2 (8.4% vs. 8.5%), whereas those between variants called by Strelka and data in dbSNP were smaller than those between variants called by VarScan 2 and data in dbSNP (15.5% vs. 20.6%). Notably higher overlapping rate for COSMIC (9.0%) and lower overlapping rates for dbSNP (14.1%) were obtained using variants in category III (detected by both Strelka and VarScan 2).

Mutation signature of GC

Mutations in human cancer, including GC, are classified into various mutational signatures using base substitution patterns and information of the trinucleotide context of each mutation. There are six classes of base substitutions : C>A, C>G, C>T, T>A, T>C, and T>G. C>T substitution at either NpCpG or TpCpN trinucleotide has been reported as the predominant mutation in GC (26). The cause of increasing C>T substitution is considered to be age-related relatively elevated spontaneous deamination of 5-methylcytosine (NpCpG) or over-activation of the APOBEC family of cytidine deaminases (TpCpN) (27, 28). We classified variants in category III using Maftools (https://github.com/PoisonAlien/maftools), and detected the feature of increasing C>T substitution at NpCpG or TpCpN in our cases of GC (Figure 3).

Characterization of genomic features in GC

To characterize genomic features of 19 GCs, we first compared the number of somatic SNVs and indels as well as genes with somatic CNVs in each case. Three cases (cases2, 19, and 3) showed higher numbers of SNVs and indels compared with others (Figure 4A), and hypermethylation of *MLH1* was observed in two of those cases (cases 2 and 3 ; Figure 4B). Four cases (cases 17, 13, 10, and 4) showed higher number of genes with CNVs (Figure 4C), and mutations in *TP53* were observed in three of those cases.

The TCGA research network, ACRG, and others reported alterations of various driver genes and therapeutic targets in GC, such as *TP53*, cell cycle mediators, genes related to receptor tyrosine kinases (RTKs), RAS and PI(3)-kinase (RAS-PI3K) signaling, and the DNA mismatch repair (MMR) system (3, 15). Therefore, we next focused on those driver alterations (Figure 5). As with TCGA and ACRG, *TP53* was detected as the most frequently altered gene (13/19, 68%). Amplification of *CCND1* and a loss-of-function mutation of *CDKN2A* in cell cycle mediators were observed. In RTKs and the RAS-PI3K signaling pathway, activating alterations (gain-of-function mutations and amplifications) were observed in *ERBB2*, *PIK3CA*, *KRAS*, *EGFR*, and *FGFR2*, and loss-of-function mutations were observed in *PTEN*.

In the MMR system, at least two *MLH1* methylations (cases 2 and 3 in 16 analyzed cases, Figure 4B) and a loss-of-function mutation of *MSH6* (case 2) were detected. Somatic CNVs were generally lacking in cases 2 and 3 (Figure 4C).

Table 2. Summary of mapped sequencing reads

	Sample	Mapping			Removing PCR dupulication		Depth of coverage	
		Paired reads after QC	Paired mapped reads	Mapping rate	Remained reads	Remained percentage	Mean bases in target region	Percentage of > 15 bases region
1	Non-Tumor	65857120	65614414	99.6%	58724065	89.5%	73.64	97.0%
	Tumor	181603326	180970888	99.7%	131785777	72.8%	158.55	99.4%
2	Non-Tumor	69994510	69749700	99.7%	60962978	87.4%	76.82	97.0%
	Tumor	193041736	192329122	99.6%	145264275	75.5%	178.26	99.4%
3	Non-Tumor	61181016	60957120	99.6%	53734851	88.2%	67.12	96.3%
	Tumor	164285882	163706626	99.6%	119712928	73.1%	147.63	99.4%
4	Non-Tumor	59956756	59747158	99.7%	54074414	90.5%	68.22	96.3%
	Tumor	180965978	180222798	99.6%	129905079	72.1%	157.63	99.3%
5	Non-Tumor	54878494	54686282	99.6%	49419388	90.4%	61.97	95.4%
	Tumor	186128494	185446402	99.6%	127366551	68.7%	156.06	99.4%
6	Non-Tumor	54326490	54135058	99.6%	49366471	91.2%	62.07	95.4%
	Tumor	176612004	175936458	99.6%	131685244	74.8%	159.43	99.5%
7	Non-Tumor	74607464	73883834	99.0%	71542483	96.8%	94.03	97.4%
	Tumor	120809972	120056666	99.4%	112750560	93.9%	142.41	98.8%
8	Non-Tumor	106593550	106086448	99.5%	101611078	95.8%	64.78	85.1%
	Tumor	112144924	111562016	99.5%	106737241	95.7%	66.07	85.4%
9	Non-Tumor	62770174	62556038	99.7%	54126506	86.5%	68.44	96.3%
	Tumor	171822472	171202862	99.6%	121896250	71.2%	148.98	99.2%
10	Non-Tumor	71260430	70997306	99.6%	60971729	85.9%	76.3	97.3%
	Tumor	199795084	199011934	99.6%	141543819	71.1%	174.18	99.4%
11	Non-Tumor	41649774	41499744	99.6%	38145737	91.9%	48.69	92.1%
	Tumor	195045234	194341898	99.6%	138498420	71.3%	171.04	99.5%
12	Non-Tumor	132970538	130783250	98.4%	120431679	92.1%	75.03	87.9%
	Tumor	112515608	110876038	98.5%	101490761	91.5%	60.76	87.0%
13	Non-Tumor	60696194	60482616	99.6%	52415338	86.7%	65.69	96.1%
	Tumor	199245284	198547282	99.6%	136566565	68.8%	163.35	99.5%
14	Non-Tumor	115025312	114532608	99.6%	108029046	94.3%	69.13	85.9%
	Tumor	143450786	142502458	99.3%	136082131	95.5%	75.13	87.4%
15	Non-Tumor	64248108	64038712	99.7%	56975815	89.0%	72.2	96.6%
	Tumor	168567112	167985604	99.7%	129176401	76.9%	158.03	99.4%
16	Non-Tumor	67676754	66517724	98.3%	64714583	97.3%	85.44	96.8%
	Tumor	133779230	132337162	98.9%	125123983	94.5%	166.24	98.8%
17	Non-Tumor	150171702	149139324	99.3%	142595142	95.6%	79.55	87.5%
	Tumor	145581182	144545830	99.3%	137062967	94.8%	85.54	87.3%
18	Non-Tumor	137728078	135804426	98.6%	127916498	94.2%	80.96	88.3%
	Tumor	137414182	135507026	98.6%	127378179	94.0%	80.56	88.3%
19	Non-Tumor	116462760	115036188	98.8%	99865533	86.8%	64.65	86.6%
	Tumor	104988856	103804024	98.9%	95492752	92.0%	58.14	86.8%

The gray shaded area shows samples analyzed using Truseq Exome Kit (Illumina) and Hiseq 1500 sequencer (n = 6).

Table 3. Number of SNVs and Indels detected in each case

Sample ID	SNVs			Indels		
	Strelka alone[a]	VarScan alone[b]	VarScan and Strelka[c]	Strelka alone[a]	VarScan alone[b]	VarScan and Strelka[c]
1	199	44	53	3	6	1
2	755	59	1923	12	347	475
3	1014	42	341	12	127	53
4	31	33	173	1	7	8
5	75	43	58	3	7	1
6	168	26	81	0	5	1
7	99	21	16	2	4	0
8	31	44	78	2	6	2
9	42	32	117	0	3	6
10	21	58	79	0	7	6
11	48	37	112	0	8	5
12	45	37	6	0	7	0
13	46	27	117	0	7	4
14	80	38	88	0	6	1
15	33	18	178	1	2	10
16	71	22	77	0	8	2
17	15	54	79	1	9	7
18	13	10	2	0	5	0
19	83	74	599	1	119	175
Total	2869	719	4177	38	690	757

[a]Number of SNVs or Indels detected only by Strelka
[b]Number of SNVs or Indels detected only by VarScan
[c]Number of SNVs or Indels detected by Strelka and VarScan

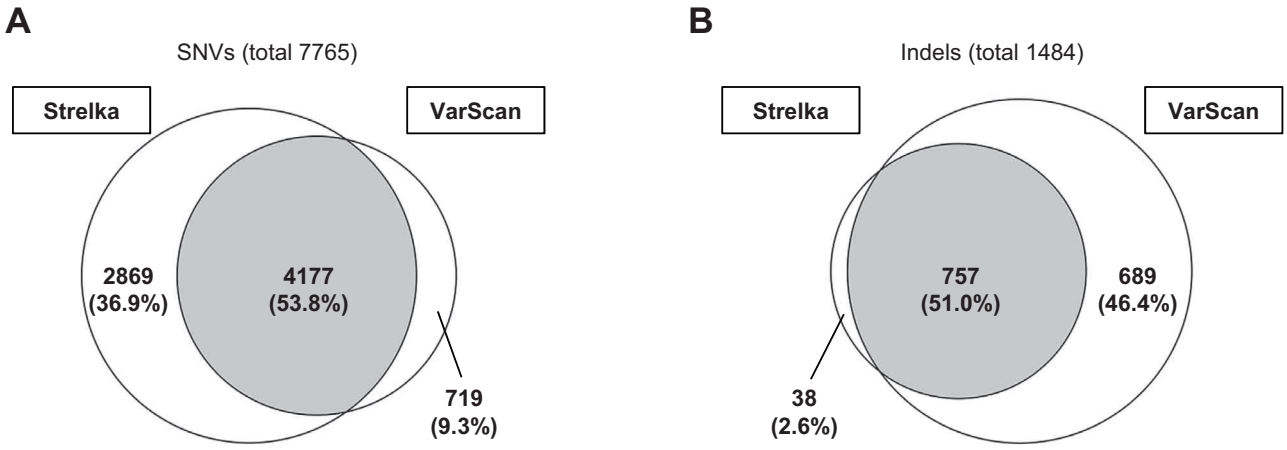

Figure 2. Venn diagrams to summarize the somatic variants called by Strelka or VarScan 2 in the exome sequencing data of 19 GC samples SNV (A) and indel (B) calls are quantified for each caller and the combination.

Table 4. Comparison of the concordance rate between SNVs detected by each analysis and data from COSMIC or dbSNP

			COSMIC			dbSNP		
		Total SNVs	Matched	Non-matched	Matching rate	Matched	Non-matched	Matching rate
Combinatorial analysis	Strelka alone	2869	211	2658	7.4%	505	2364	17.6%
	VarScan alone	719	40	679	5.6%	419	300	58.3%
	both Strelka and VarScan	4177	378	3799	9.0%	589	3588	14.1%
Conventional analysis	Strelka	7046	589	6457	8.4%	1094	5952	15.5%
	VarScan	4896	418	4478	8.5%	1008	3888	20.6%

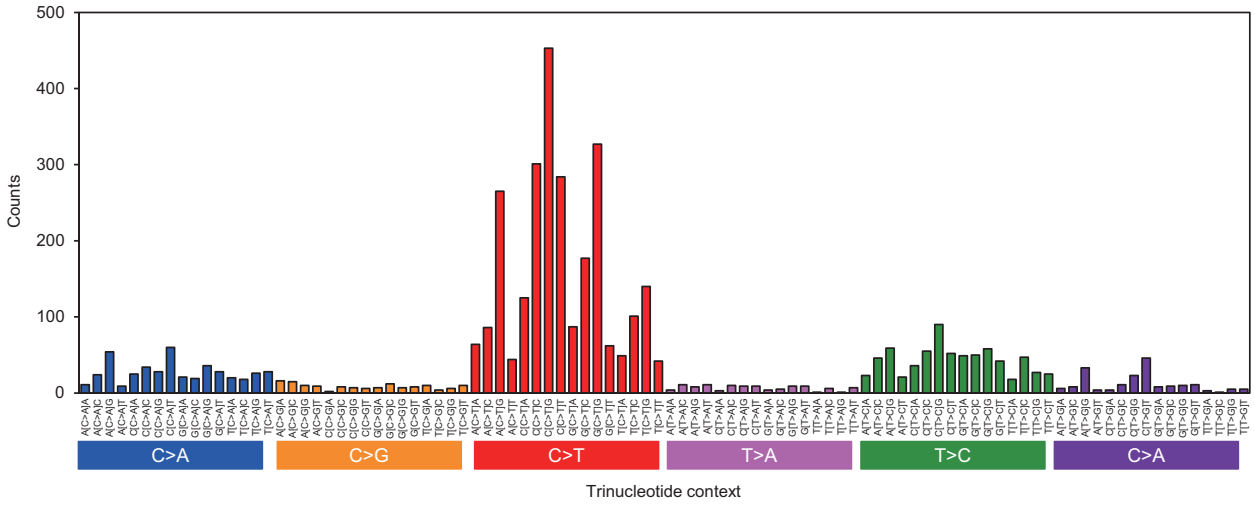

Figure 3. Mutational signatures using SNVs identified by both Strelka and VarScan 2
Vertical axis depicts the number of mutations attributed to a specific mutation type.

DISCUSSION

It is challenging for a somatic variant calling tool to balance between detecting true low-allelic somatic variants and reducing the number of false positive calls. Sensitivity, specificity, and accuracy have been discussed for various somatic variant calling tools (6, 29-31). The calling algorithms of Strelka and VarScan 2 used in this research are different from each other, with each tool having

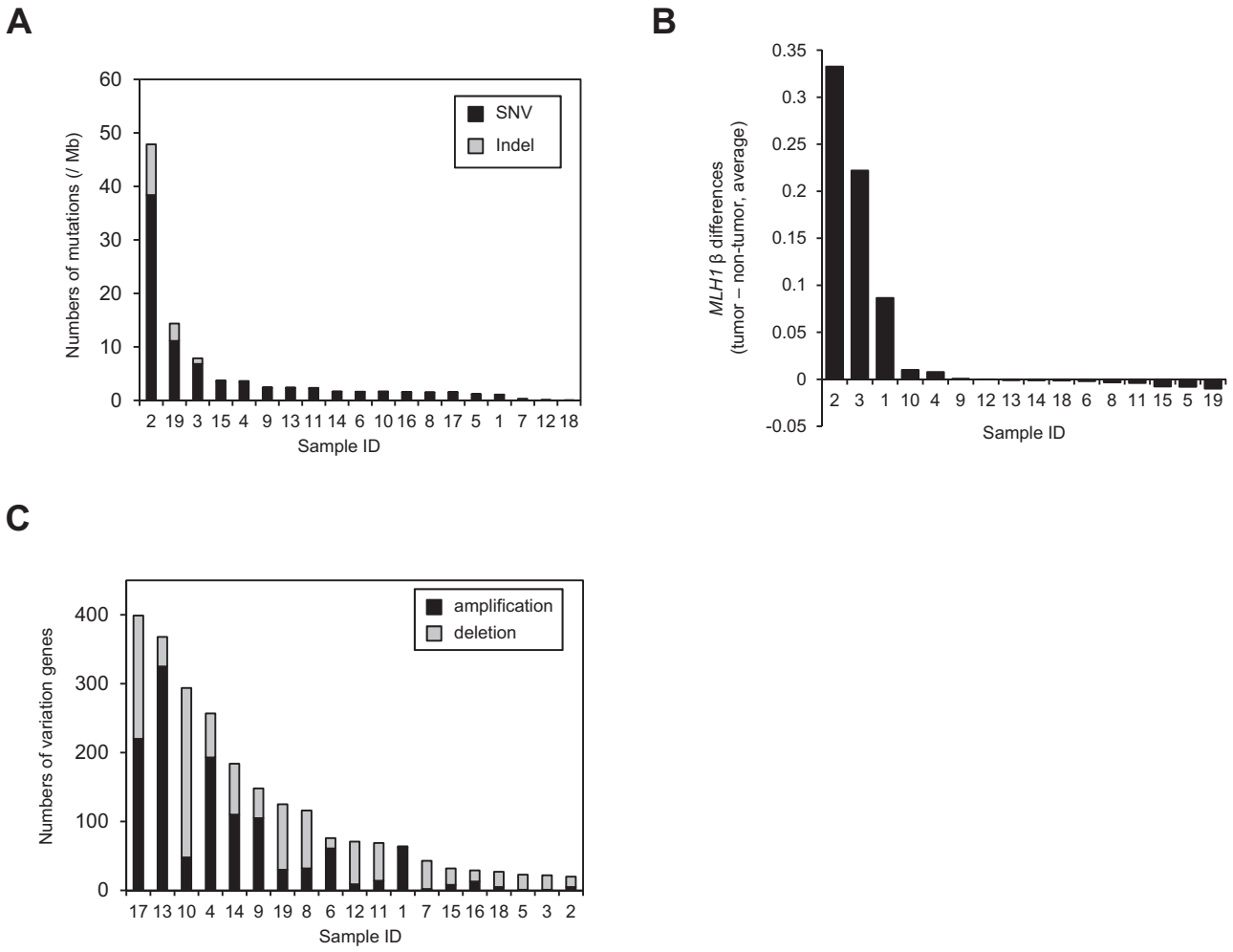

Figure 4. The prevalence of somatic variants, DNA methylations in CpG islands of *MLH1*, and genes with copy number alterations in GC samples

(A) The number of somatic variants per Mb of DNA in 19 GC samples. SNVs (black) and indels (gray) were called by both Strelka and VarScan 2. (B) The average difference in β-value (methylation level) of *MLH1* CpG islands between tumor and non-tumor tissues in 16 GC cases, as determined using an Illumina HumanMethylation450K BeadChip. (C) The number of genes with copy number alterations in 19 GC samples. Focal amplification (black) and deletion (gray) were called by VarScan 2 and DNACopy.

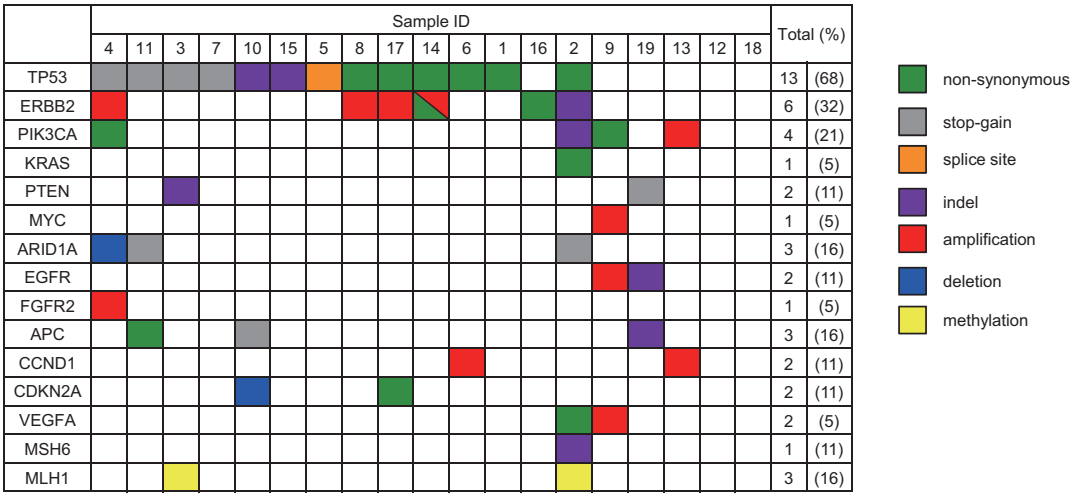

Figure 5. Landscape of genetic and epigenetic changes observed in possible driver genes

The matrix displays individual somatic alterations in each case. Color indicates the class of alteration. The percentage of samples with somatic alterations in each gene is shown on the right.

unique features. Strelka uses a complex set of calculations based on a Bayesian approach, wherein the tumor and normal allele frequencies from realigned BAM files are treated as continuous values (17). VarScan 2 applies Fisher's exact test to the tumor and normal allele frequencies obtained from a pileup file (18). Strelka identifies low-allelic-fraction candidate mutations with high sensitivity, whereas VarScan 2 detects little low-allelic-fraction candidates (6, 29). Therefore, tumor purity has a relatively higher impact on the numbers of variants detected by VarScan 2. In the present study, SNVs were better detected by Strelka than VarScan 2 (Figure 2 A), suggesting that our GC tumor samples had a relatively low degree of purity. Notably, more indels were detected by VarScan 2 (Figure 2B) because Strelka filtered out indels in microsatellites and tandem repeats (17). This effect showed higher impact on MSI samples (Table 3). Therefore, SNVs and indels are preferably evaluated by Strelka and VarScan 2, respectively, for the genetic characterization of GC. Conversely, variants in category III (detected by both Strelka and VarScan 2) showed a higher overlapping rate with COSMIC and a lower overlapping rate with dbSNP, compared with other categories. Therefore, the highest accuracy for estimating variants can be obtained using both Strelka and VarScan 2.

Using variants in category III, we found that mutation signatures in Japanese GC cases were the same as the previously reported signatures of GC using the data from TCGA and the International Cancer Genome Consortium, which contains Japanese cases (26). This result indicates that (i) variants in category III of our approach have accuracy sufficient to categorize mutation signatures in GC and (ii) C>T substitution at NpCpG or TpCpN trinucleotide is the predominant mutation in GC regardless of ethnicity or race.

The TCGA research network reported that GC could be classified into four molecular subtypes : EBV-positive (EBV), microsatellite instability (MSI), genomically stable (GS), and chromosomal instability (CIN) (3). In the EBV-positive subtype, frequent *PIK3CA* mutation and DNA promoter hypermethylation were reported. In our cases, however, neither *PIK3CA* mutation (Figure 5) nor DNA hypermethylation pattern (data not shown) was observed. ACRG provided four different subtypes : tumors with microsatellite instability (MSI), tumors with epithelial-mesenchymal transition (EMT), tumors with microsatellite stability and p53 activity (MSS/TP53$^+$), and tumors with microsatellite stability and loss of p53 activity (MSS/TP53$^-$) (15). In those subtypes, MSS/TP53$^+$ showed the highest frequency of EBV positivity. However, 5 of 13 cases with the *TP53* mutation and 3 of 6 cases without the *TP53* mutation showed EBV positivity. These results suggested that genomic features of GC may be different between Japanese patients and other patients from different ethnic origins, such as the USA and Western Europe in the TCGA and Korea in the ACRG. Thus, further characterization of GC in Japanese patients remains to be performed for identifying bona fide molecular targets and developing solid therapeutic approaches.

In conclusion, we constructed a combinatorial pipeline using two different somatic variant calling methods, which may be useful for accurately detecting mutations in GC. Personalized medicine for Japanese patients with GC needs accurate and detailed molecular characteristics of this disease to provide tailored patient treatments. Genomic characterization through application of our pipeline to larger cohorts of Japanese patients is expected be useful to improve the efficacy of GC treatments.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

ACKNOWLEDGEMENT

This study was supported in part by a Grant-in-Aid for Scientific Research (KAKENHI) Grant Numbers 26293304 (I.I.), 16K 15618 (I.I.), 24590943 (K.M), 15K19620 (T.N.), and the Tailor-Made Medical Treatment with the BioBank Japan Project (BBJ) (I.I.) from Japan Agency for Medical Research and development (AMED).

REFERENCES

1. Cancer Genome Atlas Network : Comprehensive molecular characterization of human colon and rectal cancer. Nature 487 : 330-7, 2012
2. Cancer Genome Atlas Network : Comprehensive molecular portraits of human breast tumours. Nature 490 : 61-70, 2012
3. Cancer Genome Atlas Research Network : Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513 : 202-9, 2014
4. Jackson SE, Chester JD : Personalised cancer medicine. Int J Cancer 137 : 262-6, 2015
5. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G : Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31 : 213-9, 2013
6. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z : Detecting somatic point mutations in cancer genome sequencing data : a comparison of mutation callers. Genome Med 5 : 91, 2013
7. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, Marra MA, Aparicio S, Shah SP : JointSNVMix : a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 28 : 907-13, 2012
8. Cai L, Yuan W, Zhang Z, He L, Chou KC : In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. Sci Rep 6 : 36540, 2016
9. Jemal A, Center MM, DeSantis C, Ward EM : Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev 19 : 1893-907, 2010
10. Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, Chan TL, Kan Z, Chan AS, Tsui WY, Lee SP, Ho SL, Chan AK, Cheng GH, Roberts PC, Rejto PA, Gibson NW, Pocalyko DJ, Mao M, Xu J, Leung SY : Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat Genet 43 : 1219-23, 2011
11. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, Rajasegaran V, Heng HL, Deng N, Gan A, Lim KH, Ong CK, Huang D, Chin SY, Tan IB, Ng CC, Yu W, Wu Y, Lee M, Wu J, Poh D, Wan WK, Rha SY, So J, Salto-Tellez M, Yeoh KG, Wong WK, Zhu YJ, Futreal PA, Pang B, Ruan Y, Hillmer AM, Bertrand D, Nagarajan N, Rozen S, Teh BT, Tan P : Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet 44 : 570-4, 2012
12. Wang K, Yuen ST, Xu J, Lee SP, Yan HH, Shi ST, Siu HC, Deng S, Chu KM, Law S, Chan KH, Chan AS, Tsui WY, Ho SL, Chan AK, Man JL, Foglizzo V, Ng MK, Chan AS, Ching YP, Cheng GH, Xie T, Fernandez J, Li VS, Clevers H, Rejto PA, Mao M, Leung SY : Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet 46 : 573-82, 2014
13. Kakiuchi M, Nishizawa T, Ueda H, Gotoh K, Tanaka A,

Hayashi A, Yamamoto S, Tatsuno K, Katoh H, Watanabe Y, Ichimura T, Ushiku T, Funahashi S, Tateishi K, Wada I, Shimizu N, Nomura S, Koike K, Seto Y, Fukayama M, Aburatani H, Ishikawa S : Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat Genet 46 : 583-7, 2014

14. Rokutan H, Hosoda F, Hama N, Nakamura H, Totoki Y, Furukawa E, Arakawa E, Ohashi S, Urushidate T, Satoh H, Shimizu H, Igarashi K, Yachida S, Katai H, Taniguchi H, Fukayama M, Shibata T : Comprehensive mutation profiling of mucinous gastric carcinoma. J Pathol 240 : 137-48, 2016
15. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, Ye XS, Do IG, Liu S, Gong L, Fu J, Jin JG, Choi MG, Sohn TS, Lee JH, Bae JM, Kim ST, Park SH, Sohn I, Jung SH, Tan P, Chen R, Hardwick J, Kang WK, Ayers M, Hongyue D, Reinhard C, Loboda A, Kim S, Aggarwal A : Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 21 : 449-56, 2015
16. Corso S, Giordano S : How Can Gastric Cancer Molecular Profiling Guide Future Therapies? Trends Mol Med 22 : 534-44, 2016
17. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK : Strelka : accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28 : 1811-7, 2012
18. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK : VarScan 2 : somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22 : 568-76, 2012
19. Sobin LH WC, Gospodarowicz M, editors, TNM Classification of Malignant Tumors, 7th Edition. Wiley-Blackwell : 2011.
20. Shoda K, Ichikawa D, Fujita Y, Masuda K, Hiramoto H, Hamada J, Arita T, Konishi H, Kosuga T, Komatsu S, Shiozaki A, Okamoto K, Imoto I, Otsuji E : Clinical utility of circulating cell-free Epstein-Barr virus DNA in patients with gastric cancer. Oncotarget 8 : 28796-804, 2017
21. Li H, Durbin R : Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 : 1754-60, 2009
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S : The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 : 2078-9, 2009
23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA : The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 : 1297-303, 2010
24. Wang K, Li M, Hakonarson H : ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38 : e164, 2010
25. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S : IMA : an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28 : 729-30, 2012
26. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR : Signatures of mutational processes in human cancer. Nature 500 : 415-21, 2013
27. Pfeifer GP : Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol 301 : 259-81, 2006
28. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR, Breast Cancer Working Group of the International Cancer Genome C : Mutational processes molding the genomes of 21 breast cancers. Cell 149 : 979-93, 2012
29. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y : Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics 15 : 244, 2014
30. Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, Scott HS, Glonek G, Adelson DL : A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics 29 : 2223-30, 2013
31. Kroigard AB, Thomassen M, Laenkholm AV, Kruse TA, Larsen MJ : Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLoS One 11 : e0151664, 2016