Received: 4 August 2020 | Revised: 8 October 2020 | Accepted: 2 November 2020

DOI: 10.1111/cas.14720

ORIGINAL ARTICLE

Cancer Science WILEY

Neoantigen prediction in human breast cancer using RNA sequencing data

Sachie Hashimoto[1] | Emiko Noguchi[2] | Hiroko Bando[3] | Hiroko Miyadera[2] | Wataru Morii[2] | Takako Nakamura[2] | Hisato Hara[3]

[1]Department of Breast and Endocrine Surgery, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Ibaraki, Japan

[2]Department of Medical Genetics, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan

[3]Department of Breast and Endocrine Surgery, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan

Correspondence
Emiko Noguchi, Department of Medical Genetics, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan.
Email: enoguchi@md.tsukuba.ac.jp

Funding information
Japan Society for the Promotion of Science, Grant/Award Number: JP18K08564

Abstract

Neoantigens have attracted attention as biomarkers or therapeutic targets. However, accurate prediction of neoantigens is still challenging, especially in terms of its accuracy and cost. Variant detection using RNA sequencing (RNA-seq) data has been reported to be a low-accuracy but cost-effective tool, but the feasibility of RNA-seq data for neoantigen prediction has not been fully examined. In the present study, we used whole-exome sequencing (WES) and RNA-seq data of tumor and matched normal samples from six breast cancer patients to evaluate the utility of RNA-seq data instead of WES data in variant calling to detect neoantigen candidates. Somatic variants were called in three protocols using: (i) tumor and normal WES data (DNA method, Dm); (ii) tumor and normal RNA-seq data (RNA method, Rm); and (iii) combination of tumor RNA-seq and normal WES data (Combination method, Cm). We found that the Rm had both high false-positive and high false-negative rates because this method depended greatly on the expression status of normal transcripts. When we compared the results of Dm with those of Cm, only 14% of the neoantigen candidates detected in Dm were identified in Cm, but the majority of the missed candidates lacked coverage or variant allele reads in the tumor RNA. In contrast, about 70% of the neoepitope candidates with higher expression and rich mutant transcripts could be detected in Cm. Our results showed that Cm could be an efficient and a cost-effective approach to predict highly expressed neoantigens in tumor samples.

KEYWORDS

breast neoplasms, neoantigen, RNA-seq, sequence analysis, whole-exome sequencing

Abbreviations: BAM, binary alignment map; Cm, Combination method; Dm, DNA method; GATK, Genome Analysis Toolkit; HLA, human leukocyte antigen; Indel, insertion and deletion; IQR, interquartile range; Rm, RNA method; RNA-seq, RNA sequencing; TPM, transcripts per million; VAF, variant allele frequency; VCF, variant call format; VEP, Variant Effect Predictor; WES, whole-exome sequencing.

1 | INTRODUCTION

Neoantigens are tumor-specific antigens derived from somatic variants with amino acid substitutions. A number of studies reported that neoantigens play principal roles in antitumor immune responses.[1,2] It has been reported, through phase I clinical trials of personalized neoantigen vaccines for melanoma and glioblastoma, that vaccination induced neoantigen-specific T-cell responses, which suggested the effectiveness of this therapeutic strategy.[3-5] Other studies reported that the landscape of neoantigens is associated with clinical outcome[6] and with response to immune checkpoint inhibitors.[7,8] Hence, identification of neoantigens is critically important to identify potential biomarkers and immunotherapeutic targets.

Recent improvements in next-generation sequencing have enabled rapid and high throughput prediction of neoantigens. The current standard method for prediction of neoantigens consists of four steps:[9-11] (i) identification of somatic variants through comparison of WES data from tumor and normal materials; (ii) gene expression analysis by RNA-seq of tumor materials; (iii) selection of nonsynonymous variants with expression in tumor samples; and (iv) prediction of binding affinities of a mutated region with the MHC of patients by use of in silico prediction algorithms such as NetMHCpan.[12] Although this approach could comprehensively detect candidate neoantigens and has been used in clinical trials,[3-5] many challenges remain. One of the critical problems of this approach is the cost of obtaining multiple sequencing data, even though the cost of next-generation sequencing is decreasing.

RNA-seq data, which are commonly used for gene expression analysis, gene fusion detection, and identification of splice events, can be used to detect somatic variants. Previous studies showed that variant detection using RNA-seq data is a feasible and cost-effective tool,[13-17] whereas this approach is still challenging in terms of its low accuracy. Compared with the use of WES data, the use of RNA-seq data in variant calling is error-prone owing to several issues, including alignment errors near splice junctions, errors provoked during reverse transcription, and RNA-editing sites.[15] In addition, low coverage and low expression in RNA-seq data are the main causes for missing variants in variant calling with RNA-seq data.[18] However, gene expression and the presence of mutant transcripts are indispensable factors for neoantigen candidates, so the matter of missing variants due to low coverage and expression is considered permissible for prediction of neoantigen candidates. RNA-seq has the potential to enable prediction of neoantigens with increased efficiency and lower cost; however, this possibility has not been fully examined.

In the present study, we evaluated the utility of RNA-seq in the prediction of neoantigens by using breast cancer and matched normal samples from six patients. We established three protocols for prediction of neoantigens: (i) variant calling with tumor and normal WES data (DNA method, Dm); (ii) tumor and normal RNA-seq data (RNA method, Rm); and (iii) combination of tumor RNA-seq data and normal WES data (Combination method, Cm). We examined characteristics such as coverage, gene expression, and type of base substitution for unique or shared neoantigen candidates among each method. We observed that Cm could detect candidate neoantigens, especially those that have higher expression levels and rich variant allele reads in the tumor RNA, and that Cm could be an alternative approach for predicting neoantigens.

2 | MATERIALS AND METHODS

2.1 | Patients and samples

Six patients with breast cancer who underwent surgical resection at University of Tsukuba Hospital were included in this study. All the patients had hormone receptor-positive and human epidermal growth factor receptor 2-negative invasive carcinoma. Median age at diagnosis was 65 years (range, 39-74) and one male patient (BC06) was included. Two patients (BC02 and BC07) underwent neoadjuvant endocrine therapy, and one patient (BC02) developed recurrence in the lung two months after surgery. Clinicopathologic characteristics of the patients are shown in Table 1. The study was approved by the ethics committee of University of Tsukuba Hospital (H29-069), and all the patients provided written informed consent.

Paired tumor and matched normal breast tissue blocks, collected from all the patients and cut into several slices immediately after surgery, were prepared as fresh-frozen tissues. Peripheral blood samples were obtained from four patients (BC02, 04, 06, and 07) as normal DNA materials. Genomic DNA samples were extracted from the paired tumor and normal fresh-frozen slices by use of a DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany), according to the manufacturer's protocol. Extraction of DNA from peripheral blood samples was carried out using a QuickGene DNA whole blood kit L (Kurabo Industries Ltd., Osaka, Japan) and QuickGene-610L (Fujifilm, Tokyo, Japan). Total RNA was extracted from the paired fresh-frozen slices by use of a TRIzol reagent (Invitrogen, Carlsbad, CA, USA).

2.2 | Whole-exome sequencing and RNA-seq

Twelve genomic DNA samples (six from tumor tissue, two from matched normal breast tissue, and four from peripheral blood) underwent library construction by use of a SureSelectXT Human All Exon V6 (Agilent Technologies, Santa Clara, CA, USA). The captured DNA libraries were sequenced with paired-end reads of 150 bp on a NovaSeq6000 (Illumina, San Diego, CA, USA).

RNA purification was carried out using a NEBNext rRNA Depletion Kit (New England Biolabs, Ipswich, MA, USA) for the normal breast tissue of BC02 and using a NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs) for the other 11 samples according to RNA integrity measured with an Agilent 2100 Bioanalyzer (Agilent Technologies). The RNA libraries were constructed with a NEBNext Ultra Directional RNA Library Prep Kit for

TABLE 1 Clinicopathological characteristics of six patients

Patient	Age (y)	Gender	Neoadjuvant therapy	Histology	Hormone receptor	HER2	Ki-67 LI (%)	Pathological TNM
BC02	64	F	Letrozole 9 months	IDC	Positive	Negative	15	ypT2N1a
BC03	39	F	–	IDC	Positive	Negative	25	pT2N1a
BC04	66	F	–	ILC	Positive	Negative	33	pT4bN2a
BC05	72	F	–	IDC	Positive	Negative	20	pT2N0
BC06	49	M	–	IDC	Positive	Negative	30	pT1cN1a
BC07	74	F	Letrozole 4 months	IDC	Positive	Negative	20	ypT4bN1a

IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; HER2, human epidermal growth factor receptor 2; LI, labeling index.

Illumina (New England Biolabs). These libraries were sequenced with paired-end reads of 36 bp on a NextSeq500 (Illumina).

2.3 | Mapping and data cleanup

We used BWA-MEM (v0.7.17)[19] for WES and STAR aligner (v2.7.3a)[20] for RNA-seq data to align to the reference genome (hg38, https://genome.ucsc.edu/). Alignment and data cleanup were carried out following the workflow of Genome Analysis Toolkit (GATK)[21] Best Practices[22] except for the addition of LeftAlignIndels (GATK v4.1.4.1) at the end of the RNA-seq data procedure to adjust for differences in indel positions between aligners (Doc S1).

2.4 | Gene expression quantification and tumor purity estimation

We used kallisto (v0.46.1)[23] on tumor RNA FASTQ files with Ensembl reference transcriptomes (GRCh38 release 99, http://jan2020.archive.ensembl.org/)[24] and subsequently used sleuth (v0.30.0)[25] for evaluation of gene-level expression. Tumor purity was calculated from the expression data using ESTIMATE (v1.0.13),[26] a package of R (v3.6.3, https://www.R-project.org/).[27]

2.5 | Variant calling and annotation

We tested three methods of variant calling: (i) that with tumor and normal WES data, named the DNA method (Dm); (ii) that with tumor and normal RNA-seq data, named the RNA method (Rm); and (iii) that with tumor RNA-seq data and normal WES data, named the Combination method (Cm).

For variant calling in the Dm, we used Mutect2 (GATK) with the default filtering thresholds.

For Rm and Cm, we used VarScan2 (v2.4.4),[28] which accepts not only DNA data but also RNA-seq data, and applied the parameters based on TCGA-ICGC DREAM-3 SNV Challenge results, adding the estimated tumor purity and adjusting the trimmed read length (Doc S1). To filter potential false-positive variants often observed in variant calling with RNA-seq data,[29-33] we applied four filtering methods of SNPiR[15] on the Rm and Cm: (i) removal of variants located in repetitive regions in RepeatMasker[34] track (hg38); (ii) removal of variants located in homopolymer bases of ≥5 bp; (iii) removal of variants caused by reads mapped to multiple sites using BLAT;[35] and (iv) removal of variants registered as RNA-editing sites in RADAR (version 2).[36]

The VCF files were normalized using vt (v0.5772)[37] and annotated using the Ensembl Variant Effect Predictor (VEP v99).[38] We excluded variants located in immunoglobulin and HLA genes because alignment of these highly polymorphic regions is error-prone and requires specialized analysis tools.[39,40]

2.6 | Prediction of neoantigens

Human leukocyte antigen class I alleles of each patient were determined from normal DNA FASTQ files using HLA-HD (v1.2.0.1).[41] The annotated VCF files were analyzed using pVACseq, a tool of pVACtools (v1.5.9),[42] with the default setting except for turning off the coverage and VAF filters. We used all MHC class I binding algorithms implemented in pVACseq to predict HLA class I (A, B, or C) binding 7- to 11-mer epitopes. The epitopes with a median IC50 binding score ≤500 nM were chosen as neoantigen candidates. Somatic variants that generate neoantigen candidates were manually checked using Integrative Genomics Viewer (v2.8.2).[43]

The general workflow of neoantigen prediction and detailed methods of data processing are described in Figure 1 and Supplementary Methods (Doc S1), respectively.

2.7 | Statistical analysis

The Mann-Whitney *U* test was used to compare coverage and VAF between two groups. For multiple comparison of gene expression levels and variant allele coverage among method unique variants and shared variants, the Kruskal-Wallis test followed by the Mann-Whitney *U* test with Bonferroni adjustment was used. Missing values of coverage, VAF, and gene expression level were excluded from the analysis. The Pearson chi-squared test followed by a Bonferroni post hoc test was used to compare the proportion of coding variants among method unique variants and shared variants. All statistical analyses were carried out using R (v3.6.3).

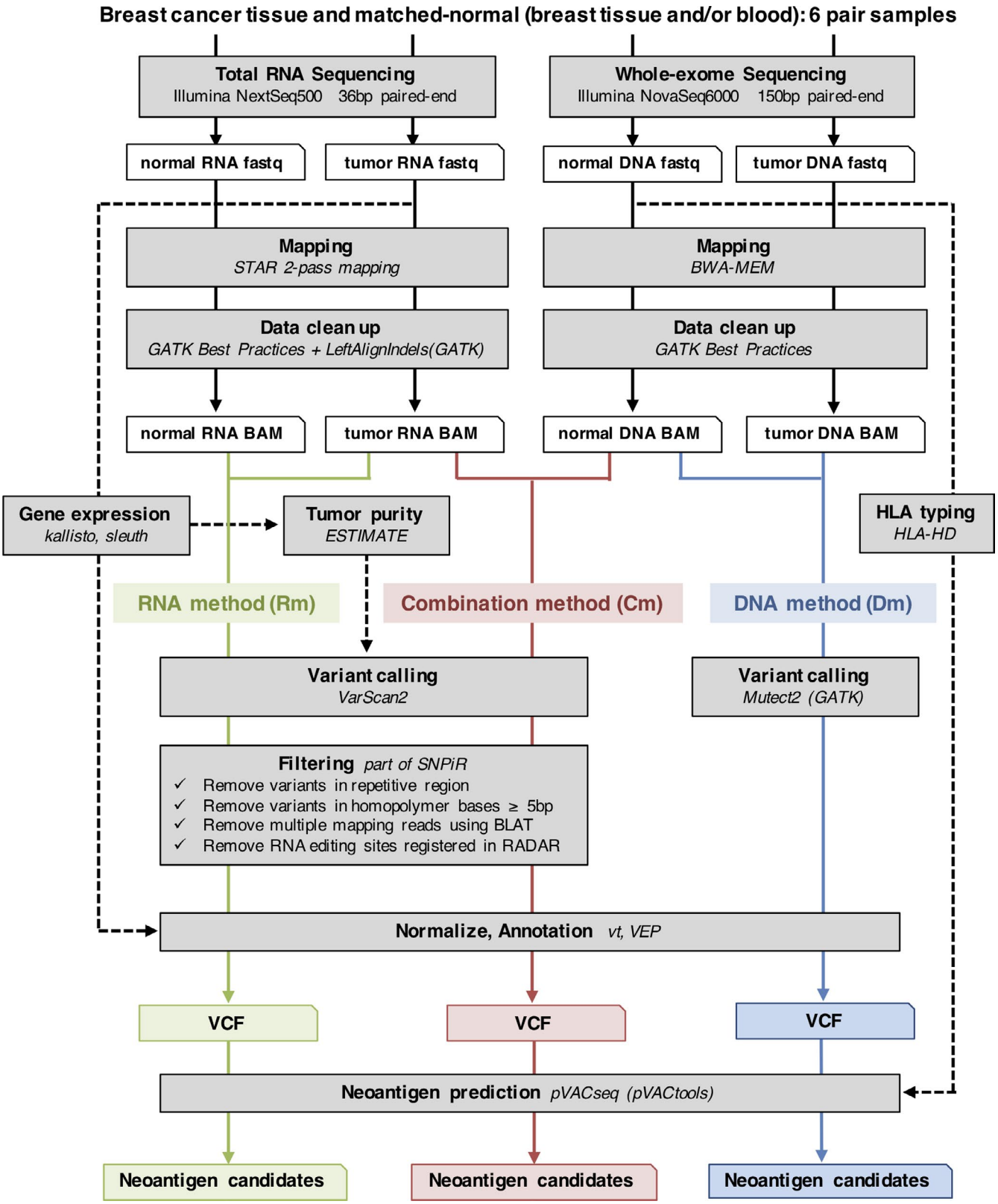

FIGURE 1 Workflow of neoantigen prediction in this study. WES and RNA-seq data were mapped to the reference genome by BWA-MEM and STAR, respectively. After data cleanup, somatic variants were called in three combinations: tumor and normal WES data (Dm); tumor and normal RNA-seq data (Rm); and tumor RNA-seq and normal WES data (Cm). In the Dm, somatic variant calling was carried out by Mutect2 (GATK). In the Cm and Rm, VarScan2 and a part of SNPiR were used for variant detection and subsequent filtering, respectively. After functional annotation, neoantigen prediction was carried out by pVACseq with individual human leukocyte antigen (HLA) class I alleles. BAM, binary alignment map; Cm, Combination method; Dm, DNA method; GATK, Genome Analysis Toolkit; Rm, RNA method; RNA-seq, RNA sequencing; VCF, variant call format; VEP, Variant Effect Predictor; WES, whole-exome sequencing

3 | RESULTS

3.1 | Number of somatic variants detected in each method

Whole-exome sequencing and RNA-seq were carried out on six breast cancer tissues and matched normal samples. Respective median numbers of the total and mapped reads were 158.9 million and 142.9 million for WES data, and 37.9 million and 31.5 million for RNA-seq data, respectively (Table S1). Tumor purities of each sample, calculated from the ESTIMATE scores, were applied to variant calling in Cm and Rm (Table S2). Somatic variants were called by Mutect2 in Dm and by VarScan2 in Cm and Rm, and the respective total numbers of the somatic variants in all the patients detected in Dm, Cm, and Rm were 3443, 401, and 54 (Figure 2). Detailed numbers of the variants for each patient are shown in Table S3. The numbers of variants shared among the methods were small, especially in Rm, with only five common variants being detected in all the methods.

3.2 | Comparison of the RNA method and the Combination method variants

Because the difference between Cm and Rm is based on whether the normal sample is DNA or RNA, we first compared the results of Cm and Rm. A large number of the variants detected in Cm were not detected in Rm because there was either no or low coverage in the normal RNA at the variant sites (Figure 3A). Of the Rm unique variants, 68% (32 of 47) were classified as germline or LOH in Cm because of wild-type allele-specific expression in the normal RNA; 30% (14 of 47) was not detected in Cm, mainly owing to low coverage in the normal DNA; and one was considered an artifactual variant allele in Cm (Figure 3B). Among the 14 Rm unique variants not detected in Cm, all the variants were located in the noncoding region, and 12 variants were deposited in dbSNP (build 146). Of note, most of the Rm unique somatic variants (45 of 47) were included in the dbSNP (build 146) database.

3.3 | Comparison of the DNA method and the Combination method variants

When comparing Dm and Cm, 3353 somatic variants were detected only in Dm; 311 only in Cm; and 90 in both methods. Features of total coverage and VAF in the tumor DNA for the Dm variants and in the tumor RNA for the Cm variants are shown in Figure 4A. The Cm variants showed a high and wide range of VAF: median 0.40 (IQR, 0.22-0.75), and relatively low coverage: median 8.0 (4.0-17.0). In contrast, the Dm variants showed lower VAF: median 0.13 (IQR, 0.08-0.22), and median coverage was 69.0 (25.0-167.0). Subsequently, we examined the coverage and VAF of the Dm unique variants in the tumor RNA and of the Cm unique variants in the tumor DNA to investigate the reasons for the inconsistency between the results of these methods (Figure 4B). A large number of the Dm unique variants showed low coverage in the tumor RNA and 74% (2471 of 3353) had no reads, whereas the Cm unique variants had a relatively high coverage in the tumor DNA: median 57.0 (IQR, 5.5-168.5). However,

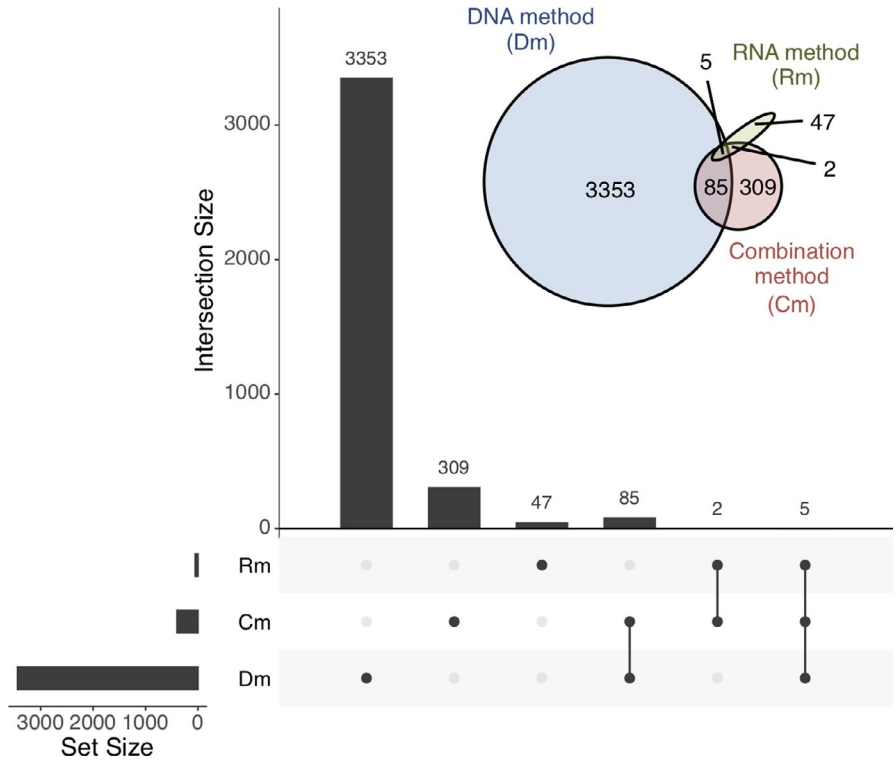

FIGURE 2 Number of intersections of somatic variants called in the three methods (sum of all patients). Cm, Combination method; Dm, DNA method; Rm, RNA method

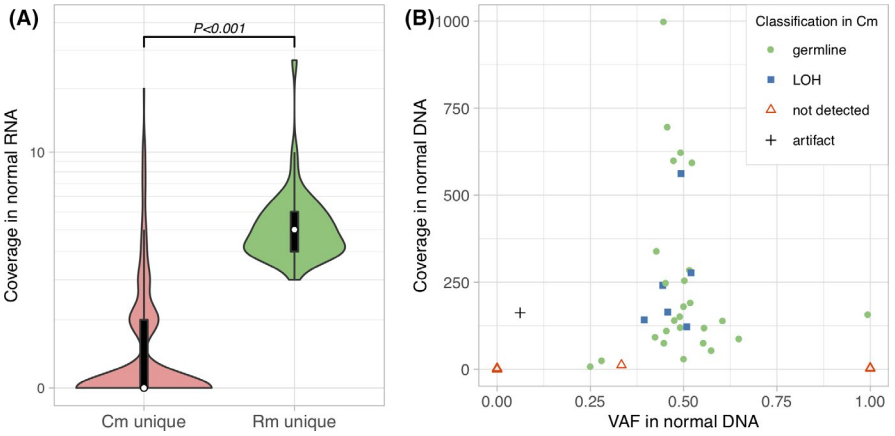

FIGURE 3 Characteristics of the RNA method unique variants compared to the combination method. A, Violin plot of total coverage in the normal RNA for the Cm and Rm unique variants. B, Distribution of VAF (x-axis) and total coverage (y-axis) in the normal DNA for the Rm unique variants. The shape of each plot shows the classifications in the Cm. Cm, Combination method; Rm, RNA method; VAF, variant allele frequency

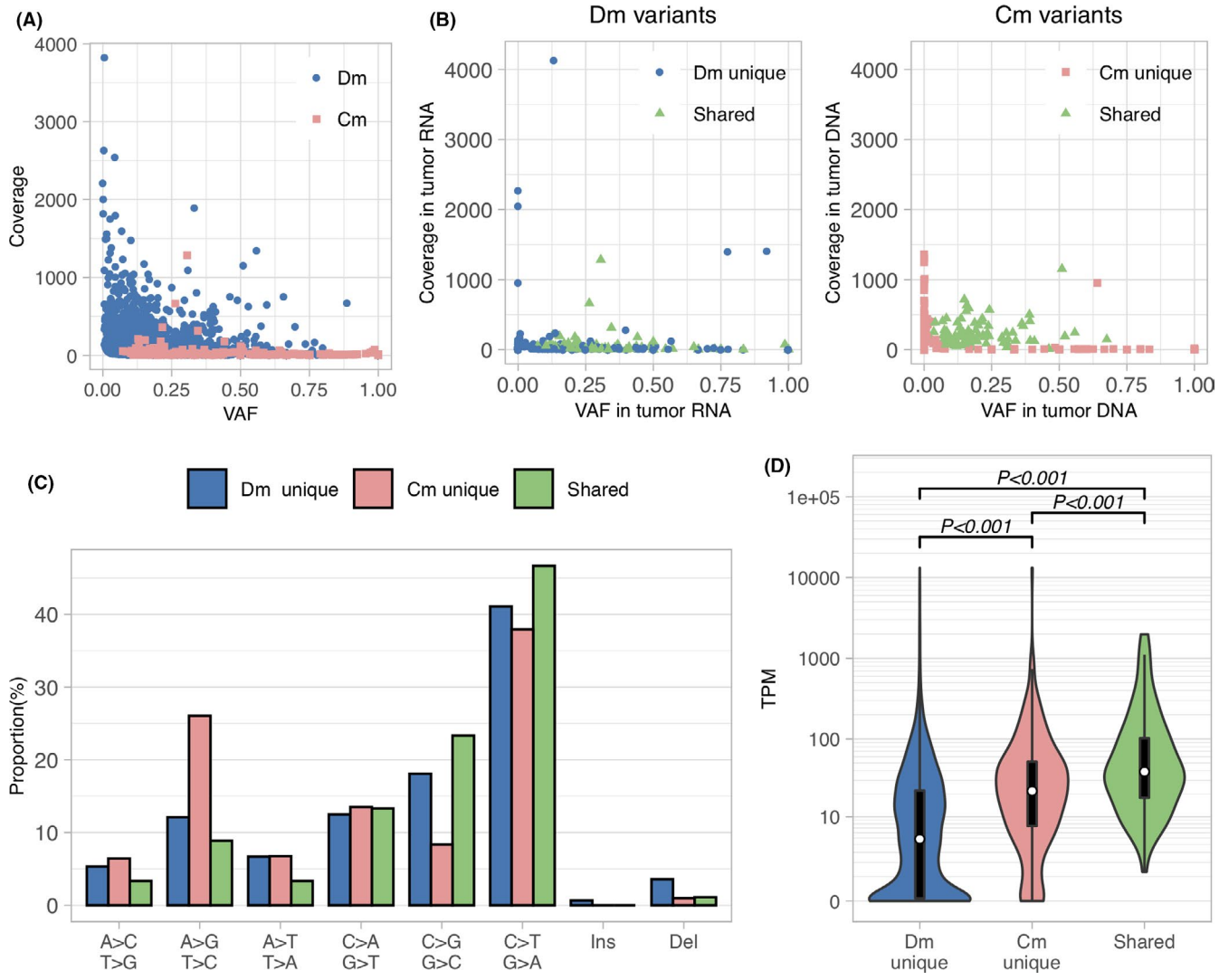

FIGURE 4 Characteristics of somatic variants detected in the DNA method and the Combination method. A, Distribution of VAF (x-axis) and total coverage (y-axis) in the tumor DNA for the Dm variants and in the tumor RNA for the Cm variants. B, Left and right scatterplots show the distribution of VAF (x-axis) and total coverage (y-axis) in the tumor RNA for variants detected in the Dm and in the tumor DNA for variants detected in the Cm, respectively. C, Proportion of base substitution patterns for each method's unique and shared variants. D, Violin plot of TPM for each method's unique and shared variants. Cm, Combination method; Del, deletion; Dm, DNA method; Ins, insertion; TPM, transcripts per million; VAF, variant allele frequency

74% (231 of 311) of the Cm unique variants had no variant allele read in the tumor DNA. The shared variants showed higher coverage and VAF in the tumor DNA than those of the Cm unique variants, and coverage and VAF were also higher in the tumor RNA than those of the Dm unique variants: median coverage of 192.5 (Cm unique variants, median: 57.0, $P < .001$, Mann-Whitney U test) and median VAF of 0.16 (Cm unique variants, median: 0.0, $P < .001$) in the tumor DNA, and median coverage of 16.0 (Dm unique variants, median: 0.0, $P < .001$) and median VAF of 0.31 (Dm unique variants, median: 0.0, $P < .001$) in the tumor RNA. Next, we examined the proportion of base substitution patterns of the Dm unique, Cm unique, and shared variants (Figure 4C). Proportion of A-to-G and T-to-C substitutions, which are known as common substitutions of human RNA editing,[44] were higher in the Cm unique variants than in the Dm unique and shared variants although the 59 Cm unique variants had already been excluded as known RNA-editing sites. Distributions of expression quantification of genes, in which variants were detected, in the tumor samples are shown in Figure 4D. Median transcripts per million (TPM) of the Dm unique, Cm unique, and shared variants were 4.8 (IQR, 0.07-22.3), 21.9 (7.5-51.8), and 38.7 (17.9-102.5), respectively. The Kruskal-Wallis test showed significant differences among these groups ($P < .001$), and multiple comparison tests showed significant differences in gene expression levels among these three groups: that of the shared variants was highest and that of the Dm unique variants was lowest ($P < .001$).

As a result of variant annotation with VEP, the proportions of variants in the protein-coding regions were 23% (787 of 3353) in the Dm unique, 31% (95 of 311) in the Cm unique, and 69% (62 of 90) in the shared variants (Figure S1). Differences among the three groups were significant ($P < .001$, chi-squared test): the shared variants had the highest proportion and the Dm unique variants had the lowest (Table S4).

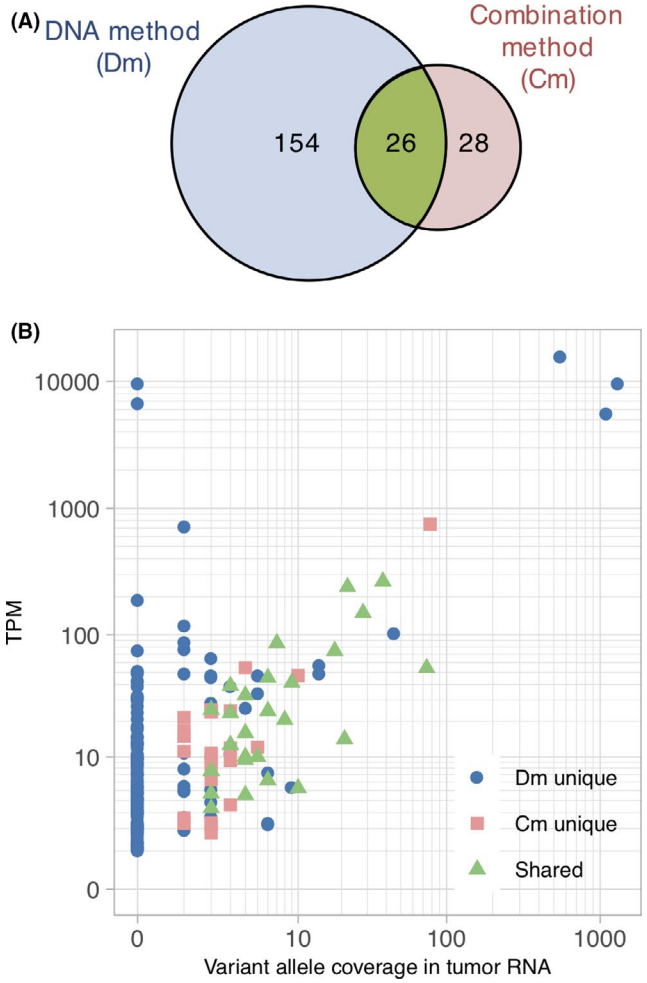

FIGURE 5 Number and features of neoantigen candidates detected in the DNA method and the Combination method. A, Venn diagram of neoepitope candidates detected in the Dm and Cm. B, Scatterplot of variant allele coverage in the tumor RNA (x-axis) and TPM (y-axis) for the Dm unique, Cm unique, and shared neoepitope candidates. Cm, Combination method; Dm, DNA method; TPM, transcripts per million

3.4 | Neoantigen candidates detected in the DNA method and in the Combination method

With in silico prediction of neoantigen, we used multiple prediction algorithms on somatic variants detected in Dm and Cm with patients' individual HLA-A, -B, and -C alleles determined by HLA-HD. There were 154 epitope candidates detected in the Dm only, 28 candidates detected in the Cm only, and 26 candidates shared by both methods (Figure 5A). Detailed information on these neoantigen candidates is shown in Table S5.

First, we examined why the Dm unique epitope candidates were missed in the Cm, by manually checking with Integrative Genomics Viewer and also by checking the filters in VarScan2 to determine if the variant allele was present in the tumor RNA. A large proportion of the Dm unique candidates (135 of 154) was not present in the output files of VarScan2, mainly owing to no coverage (n = 29) and to no or low variant allele reads (n = 97) in the tumor RNA. There were 12 insertion/deletion (two insertions and 10 deletions) that generated the Dm unique candidates, and eight of the 10 deletions were misidentified as soft clipped reads: bases in the 5′ or 3′ ends of the reads that are not aligned to the reference sequence (Table 2).

Second, we examined the characteristics of the 28 Cm unique epitope candidates. Two candidates were generated from variants in the variant-clustered reads and excluded because these were considered false-positive owing to mapping error. Of the 26 candidates, one had six variant allele reads at the DNA level (VAF of 0.042) but was not detected in Dm, and five of the remaining 25 candidates that had no variant allele reads in the tumor DNA were A-to-G or T-to-C substitutions (Table 3).

Third, we compared gene expression levels of the Dm unique epitope candidates, Cm unique ones (excluding two candidates considered false-positive), and the shared ones. Median TPM of the Dm unique, Cm unique, and shared candidates were 6.0 (IQR, 2.7-16.7), 11.6 (6.7-23.4), and 22.0 (9.6-44.3), respectively. There were significant differences among these three groups ($P < .001$, Kruskal-Wallis

TABLE 2 Reasons why the DNA method unique neoantigen candidates were not detected in the Combination method

Total number	154
Not detected in VarScan2	135 (88%)
No coverage in the tumor RNA	29
Low variant allele coverage (≤3)	97
Misidentified indels	8
Low mapping quality	1
Filtered by VarScan2 filters	14 (9%)
Low VAF (<0.1)	2
Short read length (<33)	8
Low variant base quality (<30)	6
Difference of mapping quality between REF and VAR (>10)	1
Low variant allele coverage (<3)	1
Filtered by SNPiR filters	5 (3%)
Not uniquely mapped reads (BLAT)	4
Repetitive region in RepeatMasker	1

REF, reference; VAR, variant.

TABLE 3 Characteristics of the Combination method unique neoantigen candidates

Total number	28
Likely false positive due to mapping error	2 (7%)
No variant allele reads in the tumor DNA	25 (89%)
A-to-G and T-to-C substitutions	5
Existing variant allele reads in the tumor DNA	1 (4%)

test). The shared ones had higher expression levels than those of the Dm unique ones ($P < .001$, Mann-Whitney U test with the Bonferroni adjustment), but no significant differences were observed between the Cm unique ones and the Dm unique ones ($P = .29$) and the shared ones ($P = .23$) (Figure S2A).

Theoretically, not only high expression levels in the tumor but also actual expression of the mutant alleles, rich variant allele reads in the tumor RNA, is desirable as neoantigen candidates, especially for neoantigen-targeted therapies. The means of variant allele coverage in the tumor RNA were 0 (IQR, 0-1.0), 2.0 (1.3-3.0), and 5.5 (3.3-9.8) in the Dm unique, Cm unique, and shared epitope candidates, respectively. The Kruskal-Wallis test ($P < .001$) followed by a post hoc test showed significant differences among the three groups (Figure S2B). Finally, we assessed the number of mutant allele reads in the tumor RNA in conjunction with the gene expression levels (Figure 5B). All the Dm unique candidates with >1000 TPM (n = 5) were derived from mitochondrial DNA, and three of the five candidates, which also had abundant variant allele reads in the tumor RNA, were filtered in Cm by BLAT filter because of mapping to multiple sites. Of the neoepitope candidates with >10 TPM and variant allele reads in the tumor RNA >5, 68% (13 of 19) could be identified in the Cm. These results showed that neoepitope candidates detected in both methods had higher gene expression levels and a rich amount of mutant transcripts.

4 | DISCUSSION

In the present study, we investigated whether we could detect neoantigen candidates efficiently using RNA-seq data instead of WES data in the variant calling step. Our results showed that the method using RNA-seq data for both tumor and normal tissues (RNA method, Rm) had both high false-positive and high false-negative rates and is not suitable for neoantigen prediction, whereas the method combining tumor RNA-seq data and normal WES data (Combination method, Cm) may be an efficient neoantigen prediction method because this method could detect neoepitope candidates that showed high expression levels and abundant variant allele reads in the tumor RNA. Although Cm requires normal WES data, we propose that Cm could be a cost-effective alternative strategy to the conventional method (DNA method, Dm) because it omits the most costly tumor WES, which generally demands higher coverage than the normal WES owing to tumor heterogeneity.[45]

The main causes for the high false-positive and false-negative rates in Rm were the misidentification of germline variants as somatic ones when there were no variant alleles in the normal RNA, or the overlooking of somatic variants when there were no or few transcripts in the normal sample (Figure 3). Although the accuracy of variant calling in the Rm depended greatly on the expression status of the normal transcripts, Cm could avoid these types of errors caused by transcripts from normal samples. As a result of comparing the somatic variants between Dm and Cm, a majority of variants detected in Dm were missed in Cm, mainly owing to low coverage in the tumor RNA (Figure 4B). The detection rate was improved at the level of neoantigen candidates (Figure 5A), but still 86% (154 of 180) of candidates detected in the Dm were not detected in the Cm, mainly owing to no or few variant allele reads in the tumor RNA (Table 2). Previous studies regarding somatic variant detection using RNA-seq data have also reported a small overlap between DNA and RNA, with recall rates of approximately 10%-20% in the exonic regions, and the fundamental factor of the missing variants at the RNA level was low expression.[14,16-18] However, not only the gene expression but also the assured existence of mutant transcripts in tumor tissue is a crucial factor for precise identification of true neoantigens;[46] therefore, neoepitopes missed for this reason could be permissible in terms of neoantigen prediction. Actually, about 70% of neoepitope candidates with higher expression and abundant variant allele reads in the tumor RNA could be detected in the Cm (Figure 5B).

In contrast, a large proportion of the Cm unique variants had no variant alleles in the tumor DNA (Figure 4B). A possible reason for this could be RNA editing. RNA editing is the post-transcriptional process leading to nucleotide substitution in mRNA, and the major types of human RNA editing are A-to-I substitution catalyzed by Adenosine Deaminase Acting on RNA (ADAR).[47] Aberrant regulation

of RNA editing is reported in multiple cancer types,[48-50] and a recent study showed that several epitopes generated from RNA editing have a function to elicit the immune response as cancer antigens.[51] Although we excluded variants known as RNA-editing sites in the present study to verify the accuracy of the variant calling step among different methods, the proportion of A-to-G and T-to-C substitutions was high in Cm (Figure 4C). The Cm has a potential to find novel RNA editing-derived epitopes.

Another possible cause of the Cm unique variants is technical artifacts or true variants not found in the tumor DNA. Variant calling with RNA-seq data is known to have a high false-positive rate due to several reasons such as alignment or sequencing errors around the splice junction or repetitive regions, and misalignment to paralogous regions.[29-32] Although we reduced a large amount of false-positive variants by adding several filters of SNPiR,[15] some of the Cm unique epitope candidates were derived from the variant-clustered reads that considered mapping error. It is probably effective to add a filter to remove clustered variants for improving the accuracy of the Cm. As reported previously, variant calling with RNA-seq data may find true mutations that were missed at the DNA level or a novel RNA-editing site.[14,16] In the present study, we detected one epitope candidate missed in the Dm and 25 Cm unique candidates or novel RNA-editing sites (Table 3). Hence, modifying the filters and manual checking are necessary not only to reduce false-positive variants but also to detect true neoantigen candidates identified only at the RNA level.

We acknowledge that there are several limitations to the present study. First, this study showed a small overlap between Dm and Cm even in the coding region and one that was slightly inferior to those shown in previous studies.[14,16-18] Possible reasons are the low number of RNA-seq reads and short read length (36 bp). The majority of RNA-seq data used in previous studies were with reads of ≥50 million and read lengths of ≥50 bp. In the present study, the number of total reads in the tumor RNA-seq data was <40 million, whereas it has been shown that total reads ≥50 million are required to improve the detection accuracy of mutant mRNA.[46] In addition, a short read length has the risk of missing indels[52] and multiple mapping, especially around the splice junction. In fact, a majority of indels could not be detected correctly in the Cm in the present study and some neoantigen candidates were filtered by multiple mapping (Table 2). Neoantigens derived from indels are known to be highly immunogenic,[53] and accurate identification of indels is essential for neoantigen candidate prediction. It has been reported that STAR aligner misidentified indels as soft clips at read lengths ≤50 bp, but the detection rate of indels improved at read lengths of 100 bp.[54] As several software programs have been developed for indel-sensitive detection with RNA-seq data,[55,56] the detection rate of indels could be improved using longer read length and appropriate tools. Moreover, longer read length is known to improve the detection of the splice junction.[57] Further investigation regarding appropriate library size and read length in RNA-seq is necessary to optimize neoantigen prediction using RNA-seq data.

Second, although we applied the GATK Best Practices pipeline, one of the most popular workflows for processing next-generation sequencing data, and major variant calling software (Mutect2 and VarScan2), it has been reported that the results of somatic calls differ depending on the variant calling algorithms,[58] and the choice of alignment tools also affects the results of variant detection with RNA-seq data.[59] Somatic variant calling using RNA-seq data is still challenging and a standardized analytical pipeline has not yet been established, although several algorithms were developed for RNA-seq data analysis.[15-17] Sophistication of the analysis pipeline for RNA-seq data is needed to improve the accuracy of variant and neoantigen candidate discovery in Cm.

Third, neoantigen prediction using in silico MHC-binding predictors is not sufficient to determine the immunogenicity of the neoantigens. There are many factors associated with immunogenicity of neoepitopes besides MHC peptide-binding affinity, such as T-cell recognition, processing in tumor, T-cell repertoires of individual patients, and clonality of the neoantigen.[60] Experimental validations are necessary to detect truly immunogenic neoepitopes.

In summary, the present study showed that the neoantigen prediction method using tumor RNA-seq data and normal WES data could detect neoantigen candidates that have higher expression and rich variant transcripts and also has a potential to find novel neoantigen candidates that were not detected using the conventional strategy. The Cm has potential clinical applications, including vaccine target detection and prediction of therapeutic effects of immune checkpoint inhibitors that have been reported to be associated with the degree of tumor mutation burdens.[7,8] Although we focused on SNV in the present study, several software programs for predicting neoantigens derived from fusion genes[61] and alternative splice events[62,63] have been developed using RNA-seq data. These variants are known to be an important source of potentially immunogenic neoantigens;[64-66] therefore, further utilization of RNA-seq data in addition to Cm may provide a comprehensive landscape of immunogenic neoantigens.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP18K08564. We are also grateful to Flaminia Miyamasu for suggestions that greatly improved the manuscript.

DISCLOSURE STATEMENT

The authors have no conflicts of interest.

ORCID

Sachie Hashimoto https://orcid.org/0000-0001-5973-9051
Emiko Noguchi https://orcid.org/0000-0001-9319-1763
Hiroko Bando https://orcid.org/0000-0002-7361-3647
Hiroko Miyadera https://orcid.org/0000-0002-6805-414X
Wataru Morii https://orcid.org/0000-0002-8907-5169
Hisato Hara https://orcid.org/0000-0002-2764-0353

REFERENCES

1. Heemskerk B, Kvistborg P, Schumacher TN. The cancer antigenome. *EMBO J*. 2013;32:194-203.
2. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348:69-74.
3. Ott PA, Hu Z, Keskin DB, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*. 2017;547:217-221.
4. Sahin U, Derhovanessian E, Miller M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*. 2017;547:222-226.
5. Keskin DB, Anandappa AJ, Sun J, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature*. 2019;565:234-239.
6. Brown SD, Warren RL, Gibb EA, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res*. 2014;24:743-750.
7. Snyder A, Makarov V, Merghoub T, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med*. 2014;371:2189-2199.
8. Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348:124-128.
9. Sahin U, Tureci O. Personalized vaccines for cancer immunotherapy. *Science*. 2018;359:1355-1360.
10. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat Rev Immunol*. 2018;18:168-182.
11. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Med*. 2019;11:56.
12. Hoof I, Peters B, Sidney J, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. 2009;61:1-13.
13. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*. 2009;37:e106.
14. Coudray A, Battenhouse AM, Bucher P, Iyer VR. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ*. 2018;6:e5362.
15. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet*. 2013;93:641-651.
16. Neums L, Suenaga S, Beyerlein P, et al. VaDiR: an integrated approach to Variant Detection in RNA. *Gigascience*. 2018;7:1–13.
17. Sheng Q, Zhao S, Li CI, Shyr Y, Guo Y. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics*. 2016;107:163-169.
18. O'Brien TD, Jia P, Xia J, et al. Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: a case study in lung cancer. *Methods*. 2015;83:118-127.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
20. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15-21.
21. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297-1303.
22. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491-498.
23. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34:525-527.
24. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. *Nucleic Acids Res*. 2020;48:D682-D688.
25. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods*. 2017;14:687-690.
26. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.
27. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2020. https://www.R-project.org/. Accessed May 2, 2020
28. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568-576.
29. Schrider DR, Gout JF, Hahn MW. Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One*. 2011;6:e25842.
30. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*. 2012;335:1302. author reply.
31. Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*. 2012;335:1302.
32. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*. 2012;335:1302.
33. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods*. 2012;9:579-581.
34. Smit A, Hubley R, Green PR. Open-4.0, 2013–2015 http://www.repeatmasker.org. Accessed May 27, 2020.
35. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res*. 2002;12:656-664.
36. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res*. 2014;42:D109-D113.
37. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics*. 2015;31:2202-2204.
38. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17:122.
39. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res*. 2013;41:W34-W40.
40. Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol*. 2015;33:1152-1158.
41. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat*. 2017;38:788-797.
42. Hundal J, Kiwala S, McMichael J, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. Cancer. *Immunol Res*. 2020;8:409–420.
43. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29:24-26.
44. Ramaswami G, Zhang R, Piskol R, et al. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods*. 2013;10:128-132.
45. Griffith M, Miller C, Griffith O, et al. Optimizing cancer genome sequencing and analysis. *Cell Syst*. 2015;1:210-223.
46. Karasaki T, Nagayama K, Kuwano H, et al. Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer Sci*. 2017;108:170-177.
47. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem*. 2010;79:321-349.
48. Fumagalli D, Gacquer D, Rothé F, et al. Principles governing A-to-I RNA editing in the breast cancer transcriptome. *Cell Rep*. 2015;13:277-289.
49. Han L, Diao L, Yu S, et al. The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell*. 2015;28:515-528.
50. Paz-Yaacov N, Bazak L, Buchumenski I, et al. Elevated RNA editing activity is a major contributor to transcriptomic diversity in tumors. *Cell Rep*. 2015;13:267-276.

51. Zhang M, Fritsche J, Roszik J, et al. RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat Commun*. 2018;9:3919.
52. Neuman JA, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform*. 2013;14:46-55.
53. Turajlic S, Litchfield K, Xu H, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol*. 2017;18:1009-1021.
54. Sun Z, Bhagwate A, Prodduturi N, Yang P, Kocher JA. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief Bioinform*. 2017;18:973-983.
55. Yang R, Van Etten JL, Dehm SM. Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics*. 2018;19:270.
56. Mose LE, Perou CM, Parker JS. Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics*. 2019;35:2966-2973.
57. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biol*. 2015;16:131.
58. 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*. 2016;11:e0151664.
59. Hong JH, Ko YH, Kang K. RNA variant identification discrepancy among splice-aware alignment algorithms. *PLoS One*. 2018;13:e0201822.
60. Vitiello A, Zanetti M. Neoantigen prediction and the need for validation. *Nat Biotechnol*. 2017;35:815-817.
61. Zhang J, Mardis ER, Maher CA. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*. 2017;33:555-557.
62. Zhang Z, Zhou C, Tang L, et al. ASNEO: Identification of personalized alternative splicing based neoantigens with RNA-seq. *Aging (Albany NY)*. 2020;12:14633-14648.
63. Smart AC, Margolis CA, Pimentel H, et al. Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol*. 2018;36:1056-1058.
64. Yang W, Lee K-W, Srivastava RM, et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat Med*. 2019;25:767-775.
65. Kahles A, Lehmann KV, Toussaint NC, et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*. 2018;34:211-224. e6.
66. Vauchy C, Gamonet C, Ferrand C, et al. CD20 alternative splicing isoform generates immunogenic CD4 helper T epitopes. *Int J Cancer*. 2015;137:116-126.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.