Nucleic Acids Research, 2024, **52**, D1276–D1288
https://doi.org/10.1093/nar/gkad922
Advance access publication date: 23 October 2023
Database issue

FusionNeoAntigen: a resource of fusion gene-specific neoantigens

Himansu Kumar [1,†], Ruihan Luo [1,†], Jianguo Wen [1,†], Chengyuan Yang [2], Xiaobo Zhou [1] and Pora Kim [1,*]

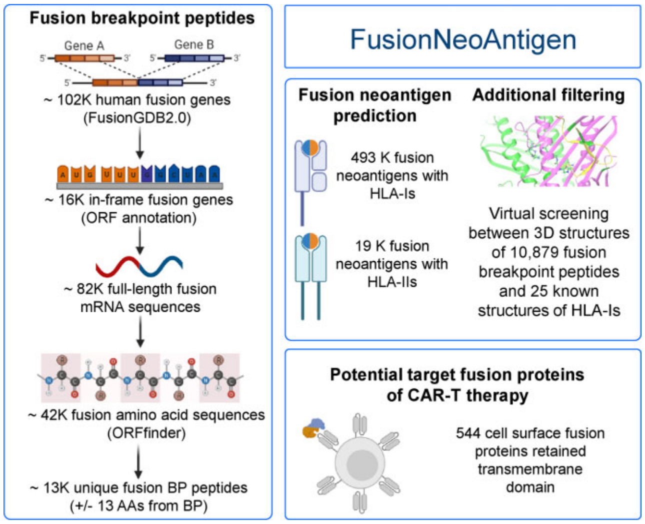
[1] Department of Bioinformatics and Systems Medicine, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2] School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[*] To whom correspondence should be addressed. Tel: +1 713 500 3636; Fax: +1 713 500 3929; Email: Pora.Kim@uth.tmc.edu
[†] The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Abstract

Among the diverse sources of neoantigens (i.e. single-nucleotide variants (SNVs), insertions or deletions (Indels) and fusion genes), fusion gene-derived neoantigens are generally more immunogenic, have multiple targets per mutation and are more widely distributed across various cancer types. Therefore, fusion gene-derived neoantigens are a potential source of highly immunogenic neoantigens and hold great promise for cancer immunotherapy. However, the lack of fusion protein sequence resources and knowledge prevents this application. We introduce 'FusionNeoAntigen', a dedicated resource for fusion-specific neoantigens, accessible at https://compbio.uth.edu/FusionNeoAntigen. In this resource, we provide fusion gene breakpoint crossing neoantigens focused on ∼43K fusion proteins of ∼16K in-frame fusion genes from FusionGDB2.0. FusionNeoAntigen provides fusion gene information, corresponding fusion protein sequences, fusion breakpoint peptide sequences, fusion gene-derived neoantigen prediction, virtual screening between fusion breakpoint peptides having potential fusion neoantigens and human leucocyte antigens (HLAs), fusion breakpoint RNA/protein sequences for developing vaccines, information on samples with fusion-specific neoantigen, potential CAR-T targetable cell-surface fusion proteins and literature curation. FusionNeoAntigen will help to develop fusion gene-based immunotherapies. We will report all potential fusion-specific neoantigens from all possible open reading frames of ∼120K human fusion genes in future versions.

Graphical abstract


Fusion breakpoint peptides
Gene A
Gene B
5'
3'
~ 102K human fusion genes
(FusionGDB2.0)
~ 16K in-frame fusion genes
(ORF annotation)
~ 82K full-length fusion
mRNA sequences
~ 42K fusion amino acid sequences
(ORFfinder)
~ 13K unique fusion BP peptides
(+/- 13 AAs from BP)
FusionNeoAntigen
Fusion neoantigen prediction
Additional filtering
493 K fusion neoantigens with HLA-Is
19 K fusion neoantigens with HLA-IIs
Virtual screening between 3D structures of 10,879 fusion breakpoint peptides and 25 known structures of HLA-Is
Potential target fusion proteins of CAR-T therapy
544 cell surface fusion proteins retained transmembrane domain

Introduction

Tumor neoantigens are products from the cancer-specific mutated genes that are presented by the major histocompatibility complex (MHC) proteins and recognized by CD8+ or CD4+ T cells. Neoantigens derived from single-nucleotide variants (SNVs) and insertions-deletions (Indels) have been the most frequently studied. However, gene fusions are also valuable sources of tumor neoantigens due to their ability to generate new open reading frames (ORFs). More importantly, compared to SNV and Indel neoantigens, fusion gene neoantigens are generally more immunogenic, have multiple targets per mutation and are more widely distributed across various

Received: August 15, 2023. **Revised:** October 2, 2023. **Editorial Decision:** October 8, 2023. **Accepted:** October 9, 2023

cancer types (1,2). As a consequence, they constitute a crucial type of tumor neoantigen and are increasingly being targeted for cancer immunotherapies. In addition, fusion gene neoantigens have multiple applications, including serving as prognostic biomarkers of immune checkpoint blockade and in the development of tumor vaccines, and adoptive cell therapies.

Neoantigen studies have been increasingly reported since 2015. However, so far, there have been only 56 studies reporting on the neoantigens of 33 fusion genes out of 266 manually curated in-frame fusion genes. Information from individual studies is listed in Table 1. Compared to studies on the neoantigens derived from SNVs or Indels, the number of studies on fusion neoantigens remains limited. This is because of the lack of fusion protein sequence resources and limited understanding of this domain. Even though there were several studies/pipelines for the prediction of fusion gene-derived fusion neoantigens in human cancer, those were only focused on specific cancer patients or cell lines. Also, no comprehensive studies were performed on fusion gene-derived neoantigens due to the lack of a reliable source of human fusion protein sequences (3–10). Most importantly, those studies lack functional annotation of the consequences resulting from the inhibition of such fusion gene functions when the immune systems attack those cancer cells with fusion-specific neoantigens (1). This lack of annotation has reduced interest for and prevented consensus in the need for further studies of fusion neoantigens. However, many cancers have a high frequency of fusion gene patients (such as rare childhood cancers) for which there are no efficient targeted therapies developed, and are still awaiting to have individual fusion-specific targeted therapies. Given this context, there is an imperative need to create a comprehensive resource of fusion gene-specific neoantigens for all human cancer types, comprehend the landscape of the fusion neoantigens, understand the functional effect of fusion neoantigens on cancer, to select potential immunotherapeutic targets for the fusion genes lacking targeted therapeutics, and to facilitate the development of effective cancer vaccine.

Table 1. Manual curation of fusion genes that have previous studies on their neoantigens

Hgene	Tgene	# PMIDs	PMIDs
ACTG1	MITF	15	12906767, 27036915, 31883334, 29296522, 35600219, 25477879, 31162138, 27319350, 33438345, 15047891, 28114254, 32640516, 33832648, 29067023, 12362579
BCR	ABL1	7	32117272, 27181332, 32301049, 11697642, 8289483, 22912393, 28153834
SSX1	SYT	5	11559563, 12133991, 15240740, 15647119, 22726592
SSX2	SYT	5	11559563, 12133991, 15240740, 15647119, 22726592
CD74	ROS1	4	30417696, 32793490, 26137589, 23160643
ERG	TMPRSS2	3	34891161, 34527198, 33244314
EWS	FLI1	3	18676758, 34385178, 24963049
TMPRSS2	ERG	3	34891161, 34527198, 33244314
BAP1	PBRM1	2	33823006, 32472114
EGFR	ACADM	2	26216968, 33627408
EML4	ALK	2	32013991, 29732013
EWS	ERG	2	25917459, 24963049
EWS	WT1	2	19206012, 11559563
EWSR1	FLI1	2	34385178, 36900411
NUTM1	BRD4	2	34333275, 36753024
PAX3	FKHR	2	18676758, 16452243
AKAP9	BRAF	1	30776432
BRAF	AKAP9	1	30776432
CBFB	MYH11	1	32831296
CCND1	FGF4	1	35874743
CD46	CD55	1	1371073
ERG	EWSR1	1	36900411
ETV6	ABL1	1	32117272
EWS	ATF1	1	11559563
EWSR1	ERG	1	36900411
EWSR1	FEV	1	36900411
EWSR1	WT1	1	36900411
MYB	NFIB	1	31011208
NDRG1	ERG	1	31475242
NTRK3	ETV6	1	36753024
PML	RARA	1	32117272
PRKAR1A	RET	1	35587601
SS18	SSX2	1	11559563

To fill this gap, we built the FusionNeoAntigen database, which provides potential fusion breakpoint-specific neoantigens throughout ∼102K human fusion genes. From our previous study, FusionGDB2 (11), we successfully translated ∼ 43K fusion protein sequences from ∼82K full-length fusion transcripts based on ∼16K in-frame fusion genes. These fusion protein sequences are currently being imported into UniProt as the reference source of human fusion protein sequences. Furthermore, in FusionPDB, we share the 3D structures of more than 3K fusion proteins. In this study, based on the ∼43K fusion protein sequences, we predicted the fusion gene-derived neoantigens using the HLA binding affinity-based prediction tools (i.e. NetMHCpan and deepHLApan). To provide the fusion-specific neoantigens, we filtered out the neoantigens that are not crossed over the fusion protein breakpoint position. After filtering based on the output values of those tools, we finally identified about 10K and 7K fusion breakpoints that have potential interaction with HLA-Is and HLA-IIs, respectively. In total, we identified ∼52K and ∼19K fusion-specific neoantigens that interact with HLA-Is and HLA-IIs, respectively. Each of these fusion breakpoints in the fusion protein sequences has multiple neoantigens crossing the fusion protein breakpoints. Taking the fusion breakpoint 14 AA sequence (since the minimum protein sequence length for prediction of protein structure is 14 AA in the RoseTTAFold), we predicted the 3D structures of ∼10K fusion breakpoint 14 AA peptides. For these predicted fusion breakpoint peptide structures, we performed virtual screening against 25 different HLAs, which have known 3D structures from PDB. Then, about 85.7% of interactions were still found in the virtual screening results. FusionNeoAntigen will help to develop fusion gene-based immunotherapies.

FusionNeoAntigen is the only resource that provides the potential fusion-specific neoantigens derived from all currently known human cancer fusion genes. FusionNeoAntigen will facilitate the advancement of immunotherapies for fusion-positive patients. We believe that information about cancer fusion neoantigens will grow in both size and importance in the forthcoming years. This resource will be routinely used by diverse cancer and drug research communities. In the future, we will make all potential ORF-based fusion protein sequences including frame-shift fusion genes, and predict the fusion-specific neoantigens from these all potential ORFs.

Materials and methods

Fusion gene information and making fusion protein sequences

We obtained 16146 unique in-frame fusion genes from FusionGDB2.0 (11) (Supplementary Table S1). For these fusion genes, we have the following information: sample ID or expressed sequence tag (EST) ID, 5′-gene, chromosome, breakpoint, strand and 3′-gene, chromosome, breakpoint, strand information and full-length fusion transcript sequences. Out of these, 14 725 in-frame fusion genes could finally make fusion amino acid sequences based on the GRCh37 human reference genome from the result of ORFfinder (12). The methods for using the fusion protein sequences from multiple genomic breakpoints and multiple gene isoforms are described in the FusionGDB2.0 study as follows. Two different genes can form fusion genes with multiple breakpoints based on multiple gene isoforms. Therefore, we considered all gene isoforms at each breakpoint. To help identify and validate fusion genes, we focused on in-frame fusion genes. For more reliable fusion genes, we checked the distance between the two breakpoints in case of intra-chromosomal rearrangements and created fusion sequences when those genes are separated by >100 kb. We also selected fusion genes when both breakpoints aligned at the exon junction. To call each exon sequence of the given breakpoints, transcription start/end sites and CDS start/end sites, we used the nibFrag utility from UCSC Genome Browser based on ENCODE hg19 genome structure. Then, we made 83290 full-length fusion transcript sequences of 16 146 in-frame fusion genes. If these fusion transcript sequences have ORF annotation results from ORFfinder, we selected the longest amino acid sequence as the one for the fusion transcript sequence. As a result, we have 43K fusion protein sequences (level 1) (https://compbio.uth.edu/FusionGDB2/combined_tables/combinedFGDB2genes_chimerkb4_fusiongdb2_AA_seq.txt). Out of these 16K in-frame fusion genes, we decided to focus on the recurrent fusion genes, which were expressed in at least two patients or cell lines. Then, 2300 recurrently expressed fusion genes were used to make 12 354 possible fusion proteins (level 2). To select more reliable fusion genes, we used ChimerKB4.0 from ChimerDB 4.0 fusion gene knowledgebase (13). We obtained 1597 fusion genes with publication support and experimental evidence. Among these, only 266 fusion genes were in-frame and potentially made 1267 fusion proteins. We provide the same annotation result per fusion gene, but we intended to let the user know the importance of individual fusion genes at these levels.

Functional annotation of fusion proteins

To assign functional or genetic categories, we integrated cancer genes, tumor suppressors, epigenetic regulators, DNA damage repair genes, human essential genes, kinases and transcription factors. In each gene group, we checked the retention and ORFs of the main protein functional features. There are 13 features belonging to the ‘region’ category, including ‘calcium binding’, ‘coiled coil’, ‘compositional bias’, ‘DNA binding’, ‘domain’, ‘intramembrane’, ‘motif’, ‘nucleotide binding’, ‘region’, ‘repeat’, ‘topological domain’, ‘transmembrane’ and ‘zinc finger’. To perform the protein functional feature retention search, we first downloaded the GFF (General Feature Format) format protein information of 10 651 UniProt accessions from UniProt (14) for 10 619 genes involved in 15 030 fusion genes. UniProt provides the loci information of 39 protein features, including six molecule processing features, 13 region features, four site features, six amino acid modification features, two natural variation features, five experimental info features, and three secondary structure features. Since such feature loci information was based on amino acid sequence, the genomic breakpoint information was converted into amino acid sequences while considering all UniProt protein accessions, ENST isoforms and multiple breakpoints for each partner. To map each feature to the human genome sequence, we used the GENCODE gene model of human reference genome v19 (15). For the 5′-partner gene, we considered the protein feature to be retained in the fusion gene if the breakpoints occurred on the 3′-end of the protein feature. On the contrary, if a protein domain was not entirely included in the fusion amino acid sequence, we reported that such fusion genes did not retain that protein feature. Similarly, for the 3′-partner gene, we considered the fusion gene to have retained the protein feature if the breakpoints occurred on the 5′-end of the protein feature region.

Identification of the breakpoint in the fusion protein sequences

Even though we can determine fusion protein sequences from the full-length fusion mRNA sequences, we do not know the position of the breakpoint in the fusion protein sequences. To identify the breakpoint of all fusion protein sequences, we ran BLAT (16), a pairwise sequence alignment algorithm with high accuracy for long sequence alignment, by inputting ∼43K fusion protein sequences. Then, we selected the breakpoints from the two major alignment regions that have more than 90 percent identity, two longest matched lengths, and the matched start and end positions are the same as the DNA level fusion breakpoints of our fusion genes.

Prediction of the fusion-specific neoantigens

For 43 464 fusion protein sequences, we made 12 790 unique ±13 AA fusion breakpoint peptide sequences from the breakpoints (Supplementary Table S2) that were identified by running BLAT. We downloaded the information on the HLA-I and HLA-II alleles from The Cancer Immunome Atlas (https://tcia.at/home) and NetMHCIIpan-4.1 server (https://services.healthtech.dtu.dk/services/NetMHCIIpan-4.1/) (17), respectively. Then, 12 790 fusion breakpoint peptide sequences as well as HLA molecules were queried for peptide-MHC complex (pMHC) binding affinity using NetMHCpan 4.1 (https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/) (18) and NetMHCIIpan 4.1. For HLA-I molecules, peptides tagged as strong binders (%rank_EL < 0.5) were retained as the input of deepHLApan (version 1.1) (19). According to the outputs of the deepHLApan model, the candidate neoantigens with immunogenic scores >0.5 were further screened out. For HLA-II alleles, we retained the predicted neoantigens with %rank_EL <0.5. For these predicted fusion neoantigens, we chose the neoantigens that cross the fusion breakpoints so that at least one AA is aligned with one protein and the left length is aligned with the other protein part.

mRNA sequences of the fusion-specific neoantigens

Since we generated the full-length transcript sequence based on the given breakpoint and gene isoforms, we know the exact mRNA sequence of the predicted fusion neoantigen peptide. The table in the section 'Vaccine Design for the FusionNeoAntigens (RNA/protein sequences)' provides the fusion neoantigen sequence and their corresponding RNA sequence. Therefore, our fusion-specific RNA sequence information for the fusion-specific neoantigens will help the researchers to generate the optimal DNA sequence more easily without the codon optimization.

Prediction of the 3D structures of fusion breakpoint 14AA peptides

To check the interaction between the fusion breakpoint peptides and HLAs, we predicted the 3D structures of individual fusion breakpoint peptides through AI-based RosettaFold (20). It is a powerful computational method that combines *ab initio* modeling, comparative modeling and evolutionary information to predict protein structures by taking amino acids as input. RosettaFold has been successful in predicting protein structures, especially for small to medium-sized proteins. Since the minimum length required for the accurate predictions of protein structure through RoseTTAFold, was 14AA, we made the fusion breakpoint 14AA peptide sequences of ∼ 10K fusion breakpoints (Supplementary Table S3) that have the potential fusion neoantigens from running the binding affinity-based prediction tools. We made the library of 3D structures of ∼10K predicted neoantigens for the neoantigen-HLA docking interaction study.

Prediction of the interaction between the fusion breakpoint 14AA peptides and HLAs

We performed molecular docking simulations between the neoantigens library and HLA to obtain the interaction information. Grid size represents the volume of a receptor's active sites where the small molecule can search for binding while docking. The grid around the receptor of HLA-A was generated using the module Receptor Grid Generation (Schrödinger, LLC, New York, NY, 2021). The dimensions of the grid were selected by considering the active site information available in the PDB database. We searched the known structure of HLA-A and bound to it ligands from in the PDB database. While making the grid through the Receptor Grid Generation tool, we consider those residues which are reported in the PDB database as residues that are involved in binding interaction. Virtual screening is a computational technique used in the drug discovery process to select small molecules that are most likely to bind to receptors or target molecules. In this work, we considered around ∼ 10K neoantigen small molecule libraries. These selected libraries were preprocessed by LigPrep (Schrödinger, LLC, New York, NY, 2021) module and made available for virtual screening. Then, finally, we ran the GLIDE (21), a tool for virtual screening provided by Schrödinger (2021) against the known 21 HLA 3D structures (Supplementary Table S4).

Selection of cell surface fusion proteins

We download the cell surface gene list from The Cancer Surfaceome Atlas (22). We have 3557 cell surface protein-coding genes in humans. After overlapping these genes with our ∼14K in-frame fusion genes, there were 4297 fusion genes that have cell surface protein as one of the fusion partners. For these fusion proteins, we investigated the transmembrane domain retention. Then, there were 544 fusion proteins (Supplementary Table S5). To predict the potential of the transmembrane localization, we ran DeepLoc, a multi-label subcellular localization prediction using protein language models (23). Using RoseTTAFold (20), we predicted the 3D structures of these cell surface fusion proteins.

Manual curation of fusion neoantigen literature

By investigating the open reading frames from 1.5K manually curated fusion genes from ChimerKB4 (13), we determined there were only 266 translated fusion genes that could potentially produce fusion proteins. For these 266 fusion genes, we searched the PMC full-text literature that has previous results on neoantigen for individual fusion genes using the search terms, 'BCR and ABL1 and neoantigen' (date: July 2023). For this, we made a Python script to grab the HTML page information of the PMC searching results and parsed the number of studies and PMIDs. Then, we manually checked those papers on the fusion neoantigens. While we searched the web using our script, we considered gene synonyms also. Ultimately, there were only 56 previous studies on fusion neoantigens for 33 fusion genes.

Drug and disease information

For all genes involved in these curated fusion genes, drug–target interactions (DTIs) were extracted from DrugBank (January 2021, version 5.1.8) with the duplicated DTI pairs excluded (24). All drugs were grouped using the Anatomical Therapeutic Chemical (ATC) classification system codes. We also searched PubMed literature evidence on the drugs that were reported as inhibiting or targeting individual fusion genes using this search expression, '((*BCR* [Title/Abstract]) AND *ABL1* [Title/Abstract]) AND fusion [Title/Abstract]) AND drug [Title/Abstract]'. After a manual review of the abstracts, we found 45 fusion genes treated by 34 drugs in 36 types of diseases. We also searched the website of MyCancerGenome of individual 266 fusion genes, we found 34 fusion genes treated by 28 drugs. There were 149 fusion genes expressed in 108 types of diseases. In total, there were 77 fusion genes were targeted by 62 drugs, and 155 fusion genes were reported in 107 diseases.

Database architecture

The FusionNeoAntigen system is based on a three-tier architecture: client, server and database. It includes a user-friendly web interface, Perl's DBI module and MySQL database. This database was developed on MySQL 3.23 with the MyISAM storage engine.

Results

Overview of FusionNeoAntigen

FusionNeoAntigen provides intensive information on the fusion-specific neoantigens that can be potentially derived from 10 320 human fusion genes among ∼14K translatable fusion genes of ∼102K human fusion genes. Figure 1 shows the overview of FusionNeoAntigen annotation. First, we provide ∼13K fusion breakpoint amino acid sequences with ±13 AA length of ∼43K fusion protein sequences as the input for the prediction of the fusion neoantigens

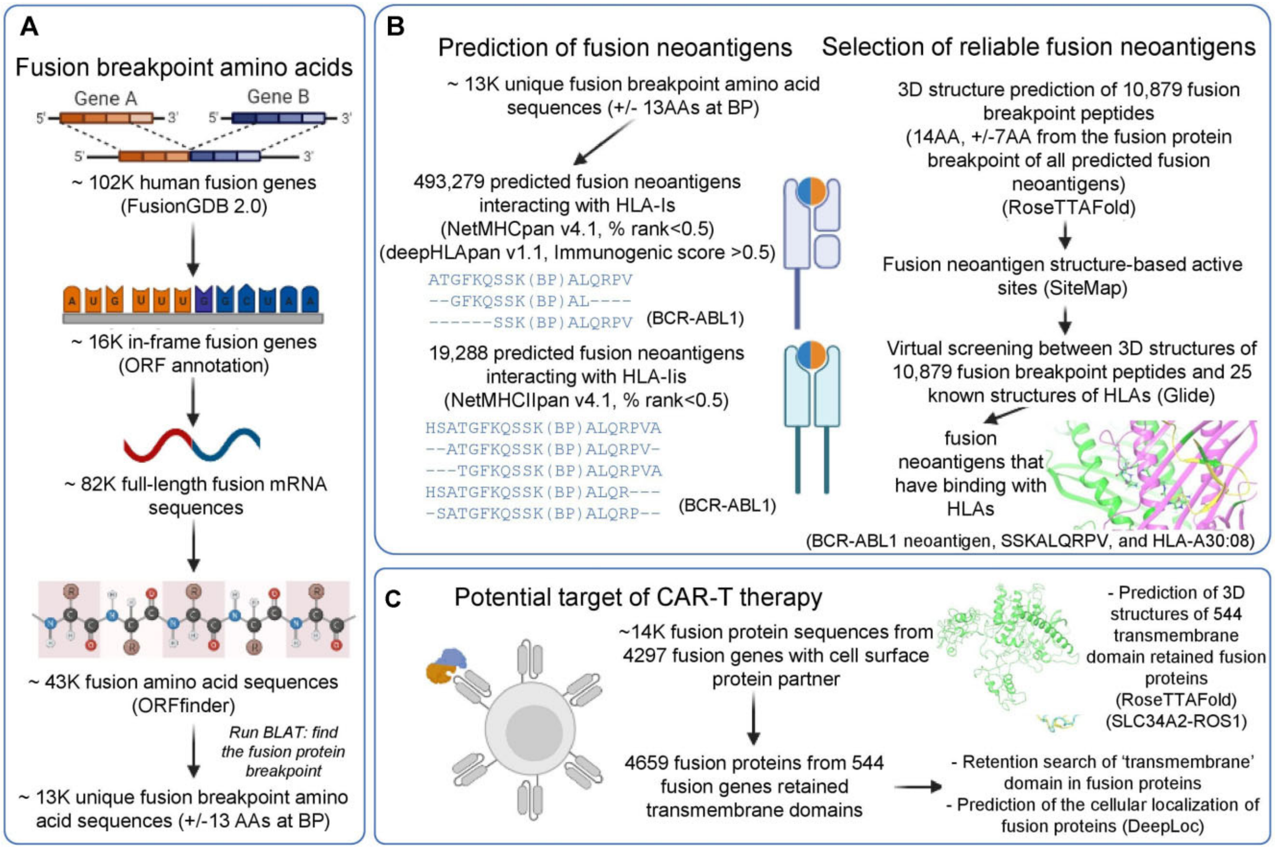

A
Fusion breakpoint amino acids
Gene A
Gene B
5'
3'
~ 102K human fusion genes (FusionGDB 2.0)
A U G U U U G C U A A
~ 16K in-frame fusion genes (ORF annotation)
~ 82K full-length fusion mRNA sequences
~ 43K fusion amino acid sequences (ORFfinder)
Run BLAT: find the fusion protein breakpoint
~ 13K unique fusion breakpoint amino acid sequences (+/-13 AAs at BP)
B
Prediction of fusion neoantigens
~ 13K unique fusion breakpoint amino acid sequences (+/- 13AAs at BP)
493,279 predicted fusion neoantigens interacting with HLA-Is (NetMHCpan v4.1, % rank<0.5) (deepHLApan v1.1, Immunogenic score >0.5)
ATGFKQSSK(BP)ALQRPV
--GFKQSSK(BP)AL----
------SSK(BP)ALQRPV
(BCR-ABL1)
19,288 predicted fusion neoantigens interacting with HLA-IIs (NetMHCIIpan v4.1, % rank<0.5)
HSATGFKQSSK(BP)ALQRPVA
--ATGFKQSSK(BP)ALQRPV-
---TGFKQSSK(BP)ALQRPVA
HSATGFKQSSK(BP)ALQR---
-SATGFKQSSK(BP)ALQRP--
(BCR-ABL1)
Selection of reliable fusion neoantigens
3D structure prediction of 10,879 fusion breakpoint peptides (14AA, +/-7AA from the fusion protein breakpoint of all predicted fusion neoantigens) (RoseTTAFold)
Fusion neoantigen structure-based active sites (SiteMap)
Virtual screening between 3D structures of 10,879 fusion breakpoint peptides and 25 known structures of HLAs (Glide)
fusion neoantigens that have binding with HLAs
(BCR-ABL1 neoantigen, SSKALQRPV, and HLA-A30:08)
C
Potential target of CAR-T therapy
~14K fusion protein sequences from 4297 fusion genes with cell surface protein partner
4659 fusion proteins from 544 fusion genes retained transmembrane domains
- Prediction of 3D structures of 544 transmembrane domain retained fusion proteins (RoseTTAFold) (SLC34A2-ROS1)
- Retention search of 'transmembrane' domain in fusion proteins
- Prediction of the cellular localization of fusion proteins (DeepLoc)

Figure 1. Overview of FusionNeoAntigen. (**A**) Identification of fusion breakpoint amino acid sequences (±13 amino acids (AA) from the breakpoint). FusionNeoAntigen first annotates the open reading frames of ∼ 102K human fusion genes from FusionGDB2.0. For ∼ 16 in-frame fusion genes, we made ∼82K fusion full-length transcript sequences considering multiple breakpoints and gene isoforms. Using ORFfinder, we made ∼43K fusion amino acid sequences. By running the BLAT alignment tool, we identified the breakpoint position of the fusion protein sequences. From these fusion protein breakpoints, we made the ±13 AA fusion breakpoint peptide sequences. There were about 13K of unique fusion protein breakpoints that have these ±13 AA fusion breakpoint peptides. This is the material to predict the fusion-specific neoantigens. (**B**) Prediction of the fusion-specific neoantigens. We input ∼13K fusion breakpoint peptide sequences into NetMHCpan and deepHLApan. After filtering the tools' criteria and checking the position of the neoantigens whether it is across the breakpoint, we identified ∼500K and 20K fusion-specific neoantigens in ∼10K and ∼ 7K unique fusion breakpoints with HLA-Is and HLA-IIs, respectively. To highlight the bindings between the fusion-specific neoantigens and HLAs, we performed the virtual screening between ∼10K predicted fusion neoantigens using RoseTTAFold and 25 HLA-Is that have known 3D structures from the PDB. (**C**) Identification of the potential target of CAR-T therapy. FusionNeoAntigen provides not only the fusion-specific neoantigens but also the potential target of CAR-T therapy in fusion proteins. First, we overlapped ∼14K in-frame fusion genes with 4297 cell surface genes. Then, there were 544 fusion proteins that had a cell surface protein as one of the fusion partner proteins and retained the transmembrane domains. For these, we predicted their potential 3D structures using RoseTTAFold and also predicted the cellular localization of the fusion proteins using DeepLoc.

(Figure 1A). Second, we provide information on the predicted fusion-specific neoantigens of ∼500K interaction pairs between fusion breakpoint-specific neoantigens and HLAs. Third, we provide information on the potential cell surface fusion proteins. The main features of FusionNeoAntigen annotations are summarized below.

i. 'The Fusion Gene and Fusion Protein Summary' category shows a basic fusion gene summary following multiple annotations for each fusion partner gene. It provides the literature curation that has the previous reports on the fusion neoantigens of individual fusion genes. It also provides the functional gene groups to assign potential loss-of-functional effects with ORF annotation and the breakpoints loci on the gene structures of individual partners using the UCSC genome browser. To infer the functional aspect of fusion genes/transcripts/proteins, we provide the link to our FGviewer.

ii. 'The Fusion Amino Acid Sequence' category provides the in-frame fusion gene/protein information, coding potential from our deep learning model and fusion amino acid sequences based on multiple breakpoints and gene isoforms.

iii. 'Fusion Protein Breakpoint Sequence' category provides the ±13 flanking fusion protein breakpoint peptide sequences.

iv. 'Potential FusionNeoAntigens in HLA-I and HLA-II' categories provide the predicted fusion gene-derived fusion neoantigens from the fusion protein breakpoint sequences. If the neoantigen is not aligned over the breakpoint, but aligned on one partner only, then it was filtered out.

v. 'Fusion Breakpoint 14 AA Peptide Structure' category provides the predicted 3D structures of the fusion gene-derived fusion neoantigens.

vi. 'Filtering FusionNeoAntigens Through Checking the Interaction with HLAs in the 3D' category provides the virtual screening results between our predicted 3D structures of fusion neoantigens across 25 HLAs that have known 3D structures.

vii. ‘Vaccine Design of the FusionNeoAntigen’ category provides the multiple-level sequences (fusion breakpoint RNA/protein sequences) of fusion breakpoints that have potential fusion neoantigens.

viii. ‘Potential Target of CAR-T therapy development in Fusion Proteins’ category provides the predicted 3D structures of 544 cell surface fusion proteins and the transmembrane retention information as the potential target of CAR-T therapy development.

ix. ‘Information on the Samples That Have These Potential Fusion Neoantigens’ category provides information on the samples that have potential in-frame fusion genes, fusion proteins and fusion neoantigens.

x. ‘Fusion Protein Targeting Drugs’ and ‘Fusion Protein Related Diseases’ categories provide the manually curated results of used drugs targeting 266 curated fusion genes and reported diseases with these fusion genes from PubMed literature and MyCancerGenome. There were 77 fusion genes targeted by 62 drugs and 155 fusion genes were reported in 107 diseases.

Identified fusion neoantigens (52K and 19K fusion breakpoint-specific neoantigens with HLA-is and HLA-IIs) from ∼43K fusion protein sequences

Table 2 shows the summary statistics of the fusion neoantigen prediction. Out of ∼103K fusion genes in our FusionGDB 2.0 database, there were ∼ 16K in-frame fusion genes. Out of these, there were ∼14.7K fusion genes that can make potential fusion amino acid sequences. Considering the multiple fusion gene breakpoints and multiple gene isoforms of ∼14.7K fusion genes, we could make 83K full-length fusion transcripts. Inputting these 83K full-length fusion transcripts into the ORFfinder tool made by NCBI, we had ∼43K fusion protein sequences. Out of these unique ∼43K fusion protein sequences, we took ±13 amino acids (AAs) from the fusion breakpoint. Because the length of the neoantigens interacting with HLA class I varies as 8–10, 8–11 or 8–13 amino acids, and neoantigens interacting with HLA class II are 12–24 or 13–25 amino acids. Therefore, we used 26 AA (±13 AA from the fusion breakpoint) as the input for the prediction of neoantigens derived from the fusion protein sequences. There were ∼ 13K unique fusion breakpoint peptides of ±13 AA length. After running NetMHCpan and deepHLApan, we identified 10 759 and 5895 fusion genes that have potential interaction with HLA-Is and HLA-IIs, respectively. These fusions have 11 485 and 6184 unique fusion breakpoints, and ∼103K and ∼20K fusion neoantigens, respectively.

Nonetheless, the aim of FusionNeoAntigen is to provide fusion breakpoint-specific neoantigens. Several fusion neoantigens that were reported in the previous studies from well-known fusion genes (i.e. BCR-ABL1 and TMPRSS2-ERG) did not cross the fusion breakpoints but rather aligned to only one gene part, which can occur with wild type genes also. To avoid this situation, we chose neoantigens that crossed the fusion breakpoints so that at least one AA is aligned with one protein and the left length is aligned to the other protein part. Luckily, 93.1% (10 018 out of 10 759) and 97.7% (5762 out of 5895) fusion genes had the potential fusion neoantigens that are crossing the fusion breakpoints. 10 649 and 6039 unique fusion breakpoints had ∼ 52K and ∼ 19K fusion neoantigens that crossed the fusion breakpoints and interacted with the HLA-Is and HLA-IIs, respectively. In total,

Table 2. Statistics of FusionNeoAntigen

Big categories	Categories	Numbers
Fusion gene	# fusion genes	103 344
	# in-frame fusion genes	16 306
	# translated fusion genes	14 725
	# full-length fusion transcripts	83 416
Fusion protein	# fusion proteins	43 464
	# unique fusion breakpoint ±13 AA peptides	12 790
Fusion genes that have potential fusion neoantigens	# fusion genes that interact with HLA-Is	10 759
	# fusion genes that interact with HLA-IIs	5895
	# unique fusion breakpoints that interact with HLA-Is	11 485
	# unique fusion breakpoints that interact with HLA-IIs	6184
	# unique fusion neoantigens that interact with HLA-Is	103 730
	# unique fusion neoantigens that interact with HLA-IIs	20 099
Fusion genes that have potential fusion neoantigens across the breakpoint	# fusion genes that interact with HLA-Is across fusion breakpoint	10 018
	# fusion genes that interact with HLA-IIs across fusion breakpoint	5762
	# unique fusion breakpoints that interact with HLA-Is across fusion breakpoint	10 649
	# unique fusion breakpoints that interact with HLA-IIs across fusion breakpoint	6039
	# unique fusion neoantigens that interact with HLA-Is across fusion breakpoint	51 856
	# unique fusion neoantigens that interact with HLA-IIs across fusion breakpoint	19 288
Fusion neoantigens	# unique fusion neoantigen mRNA sequences that interact with HLA-Is across fusion breakpoint	51 658
	# unique fusion neoantigens mRNA sequences that interact with HLA-IIs across fusion breakpoint	19 240
Fusion breakpoint peptide 3D structures (14AA) that have potential fusion neoantigens	# unique fusion breakpoint 14AA peptides	10 879
	# unique fusion breakpoint 14AA peptides that have predicted 3D structures	10 836
	# unique fusion breakpoint 14 AA peptide that interact with HLA-Is	10 829
	# fusion genes that have predicted 3D structures of fusion breakpoint 14AA peptides	10 017
Potential target of CAR-T	# cell-surface fusion genes from the gene group information	4297
	# cell-surface fusion genes retaining the transmembrane domains	544

there were 1 008 363 interactions with HLA-Is and 452 038 interactions with and HLA-IIs. Furthermore, upon checking the crossing of the breakpoint of all predicted fusion neoantigens, almost half of the fusion neoantigens with potential interaction with HLA-Is were filtered out. As a result, we had 493 279 pairs of interactions between 51 856 fusion breakpoint-specific neoantigen candidates and 351 HLA-I alleles. For the HLA-IIs, there were 432 746 pairs of interactions between 19 223 fusion breakpoint-specific neoantigens candidates and 605 HLA-IIs.

Since the goal of neoantigen prediction is to develop a cancer vaccine, the annotation page of the FusionNeoAntigen website provides the fusion neoantigen mRNA sequences. Since we predicted the fusion neoantigen peptide sequences from fusion transcript sequences, we have the matched fusion mRNA sequence for the fusion neoantigen sequences. In total, we have 51 658 HLA-I interacting and 19 240 HLA-II interacting unique mRNA sequences from the fusion neoantigens. Table 3 shows the most frequent fusion genes that have fusion-specific neoantigens interacting with 25 known HLA structures. Even though TMPRSS2-ERG has many sample frequencies of fusion genes, after investigation of the open reading frames of those fusion gene/transcript sequences, there were not many in-frame and translational fusion genes, indicating a limited number of samples have the fusion-specific neoantigens. However, BCR-ABL1 and EML-ALK had many in-frame fusion genes from multiple breakpoints and had more samples that have potential fusion-specific neoantigens. Out of these top 100 most frequent fusion genes that have fusion-specific neoantigens, only 11 fusion genes had previous records on the studies of fusion neoantigens.

Filtered fusion neoantigens by performing the virtual screening between 25 HLA-is that have known 3D structures and ∼10K fusion protein breakpoint 14AA peptide structures

To further validate the interaction between the potential fusion protein breakpoint-specific neoantigens and provide more reliable fusion breakpoint-specific neoantigens, we used the virtual screening approach between the 3D structures of the fusion neoantigens and the publicly available 3D structures of the HLAs. Between 10 836 3D structures of 10 879 unique ±7 AA (14 AA) from the fusion breakpoints and 21 known 3D structures of the HLAs, we found that 10 017 fusion genes have interaction using the virtual screening approach. We first determined the active site of the predicted structures of 10 879 peptides using SiteMap. Then, we performed the virtual screening of individual pairs using Glide. This yielded a total of 422872 pairs of interactions between 10 879 peptides and 21 HLA known molecules in 10 017 fusion genes (10 649 fusion breakpoints). Overall, 85.7% (422 872 out of 493 279) pairs of fusion-specific neoantigens and HLAs had virtual screening evidence. Figure 2A shows the example of nine frequent and well-known fusion genes' complexes in 3D between the HLAs and individual fusion neoantigens. Figure 2B shows their information with the interaction scores.

FusionNeoAntigen annotation page

Figure 3 shows the major annotation categories in the FusionNeoAntigen website. First, we show the fusion protein sequences of each fusion gene (Figure 3A). Considering multiple breakpoints and gene isoforms, each fusion gene can be expressed as multiple fusion protein sequences. From the given DNA breakpoints of individual fusion genes, we made full-length fusion mRNA sequences and checked the open reading frames. Then, for the in-frame fusion mRNA sequences, we input to the ORFfinder and chose the longest ORF's amino acid sequences. For the well-known fusion genes, we checked their DNA breakpoints and their fusion amino acid sequences whether our annotation is right or not. From these fusion protein sequences, we made ±13 AA long peptides from the breakpoints at the top of Figure 3B. This is the starting material to predict the fusion breakpoint-specific neoantigens. Then, we show the information of the predicted fusion-specific neoantigens in the table including the breakpoints, fusion neoantigen peptide sequence, interacting HLA allele name, binding affinity scores and immunogenic scores. We also provide the multiple sequence alignment of all predicted fusion neoantigens crossing the breakpoint per fusion breakpoint. To have more reliable fusion neoantigen candidates, we predicted the 3D structures of the peptide of ±7 AA length from the breakpoints and checked the binding between these structures and known 3D structures of 21 HLAs. We show these 3D structures and their docking scores as shown at the bottom of Figure 3B. In FusionNeoAntigen, we also provide information on the potential cell surface fusion proteins as the potential source of CAR-T therapy. As shown in Figure 3C, we show their potential 3D structure, transmembrane domain retention information and subcellular localization prediction result of the given cell surface fusion protein sequence.

Clinical usage of FusionNeoAntigen

Current research primarily focuses on neoantigens resulting from single-nucleotide variants (SNVs) and insertions or deletions (Indels) (25). However, the scope of tumor neoantigens is limited in tumors with high burdens of SNVs or Indels. Therefore, expanding the range of tumor neoantigens is crucial to extend the applicability of neoantigen-based immunotherapies to a broader population of cancer patients. Gene fusions represent a significant form of genetic alteration in cancer, playing a crucial role in early tumorigenesis and accounting for approximately 20% of global cancer cases (26). These fusions occur due to structural variations in chromosomes (SV) and can result in the creation of new open reading frames (ORFs) (27). Peptides derived from the regions where two genes fuse differ from self-antigens and serve as excellent sources of neoantigens. Patients' T cells can recognize these neoantigens, as observed in cases like BCR-ABL in chronic myelogenous leukemia and SYT-SSX1 in synovial cell sarcoma (28). Consequently, neoantigens derived from gene fusions have the potential to expand the existing repertoire of tumor neoantigens and offer promising prospects for cancer immunotherapy. Notably, even in tumors with low tumor mutational burden (TMB) and limited immune cell infiltration, neoantigens generated by gene fusions can still activate cytotoxic T cells (27–29). A study examining fusion neoantigens from the TCGA data found that fusion genes have the potential to produce a significantly higher number of novel open reading frames (ORFs) compared to single-nucleotide variants (SNVs) and insertions/deletions (INDELs), resulting in a six-fold increase in neoantigens and an eleven-fold increase in specific candidate neoantigens. Fusion neoantigens are more

Table 3. Number of samples that have fusion neoantigens

Fusion genes		# samples with fusion genes	# samples with in-frame fusion genes	# samples with neoantigen interacting with HLA-Is	# cancer types with neoantigen interacting with HLA-Is	# samples with neoantigen interacting with HLA-IIs	# cancer types with neoantigen interacting with HLA-IIs	PMIDs of previous studies
BCR	ABL1	194	16	9	3	8	3	32117272, 27181332, 32301049, 11697642, 8289483, 22912393, 28153834
EML4	ALK	78	21	8	5	7	5	32013991, 29732013
RPS6KB1	VMP1	98	14	7	5	6	4	
CCDC6	ANK3	27	11	6	3	5	2	
CBFB	MYH11	57	10	6	3	1	1	32831296
ARL8B	ITPR1	9	6	5	4	3	3	
RUNX1	RUNX1T1	99	8	5	2	2	1	
EWSR1	FLI1	90	4	5	3	2	2	34385178, 36900411
REPS2	TXLNG	13	7	5	4	2	2	
NCOR2	SCARB1	28	10	4	4	4	4	
FLNB	SLMAP	14	6	4	4	4	4	
PUM1	NKAIN1	13	13	4	4	4	4	
KIF5B	RET	24	4	4	2	3	1	
MPP5	GPHN	20	5	4	3	3	2	
EWSR1	ATF1	16	3	4	3	3	3	
SMARCA4	DNM2	11	6	4	4	3	3	
ERC1	RET	10	5	4	3	3	2	
SND1	BRAF	10	3	4	3	3	3	
ASCC1	MICU1	11	8	4	2	2	2	
PLG	LPA	16	4	4	1	1	1	
WNK2	CENPP	5	4	4	3	1	1	
LAPTM4B	MTDH	8	5	3	3	4	4	
ETV6	NTRK3	36	7	3	3	3	3	36753024
CAMKK2	KDM2B	15	8	3	1	3	1	
PRRC2B	NUP214	14	7	3	3	3	3	
LTBP1	BIRC6	13	6	3	3	3	3	
TRIM27	RET	7	3	3	3	3	3	
AP2A2	TALDO1	6	5	3	3	3	3	
CABLES1	RBBP8	6	3	3	1	3	1	
IKBKB	ANK1	6	4	3	3	3	3	
TMPRSS2	ERG	975	18	3	1	2	1	34891161, 34527198, 33244314
EWSR1	ERG	16	5	3	2	2	2	36900411
KAT6A	ADAM2	14	6	3	1	2	1	
BCAS3	VMP1	13	7	3	3	2	2	
CD44	PDHX	12	8	3	3	2	2	
RFTN1	DAZL	5	4	3	2	2	2	
WWOX	VAT1L	4	3	3	3	2	2	
ZC3H7A	GSPT1	3	3	3	3	2	2	
AFF1	KMT2A	45	3	3	2	1	1	
PPP1CB	PLB1	20	7	3	2	1	1	
ERG	TMPRSS2	18	10	3	1	1	1	34891161, 34527198, 33244314
BPTF	PITPNC1	13	5	3	2	1	1	
COMMD10	AP3S1	10	4	3	3	1	1	
KDM5A	NINJ2	9	9	3	3	1	1	
LMBR1	PTPRN2	7	4	3	3	1	1	
ZMYND8	EYA2	7	5	3	2	1	1	
DLG1	BDH1	6	4	3	3	1	1	
ZMYND8	SULF2	5	5	3	2	1	1	
LTBP1	TTC27	4	3	3	2	1	1	
SNTB2	TANGO6	4	3	3	3	1	1	
BAZ2A	PRIM1	3	3	3	3	1	1	
PML	RARA	102	13	2	2	2	1	32117272
TMPRSS2	ETV4	49	7	2	1	2	1	

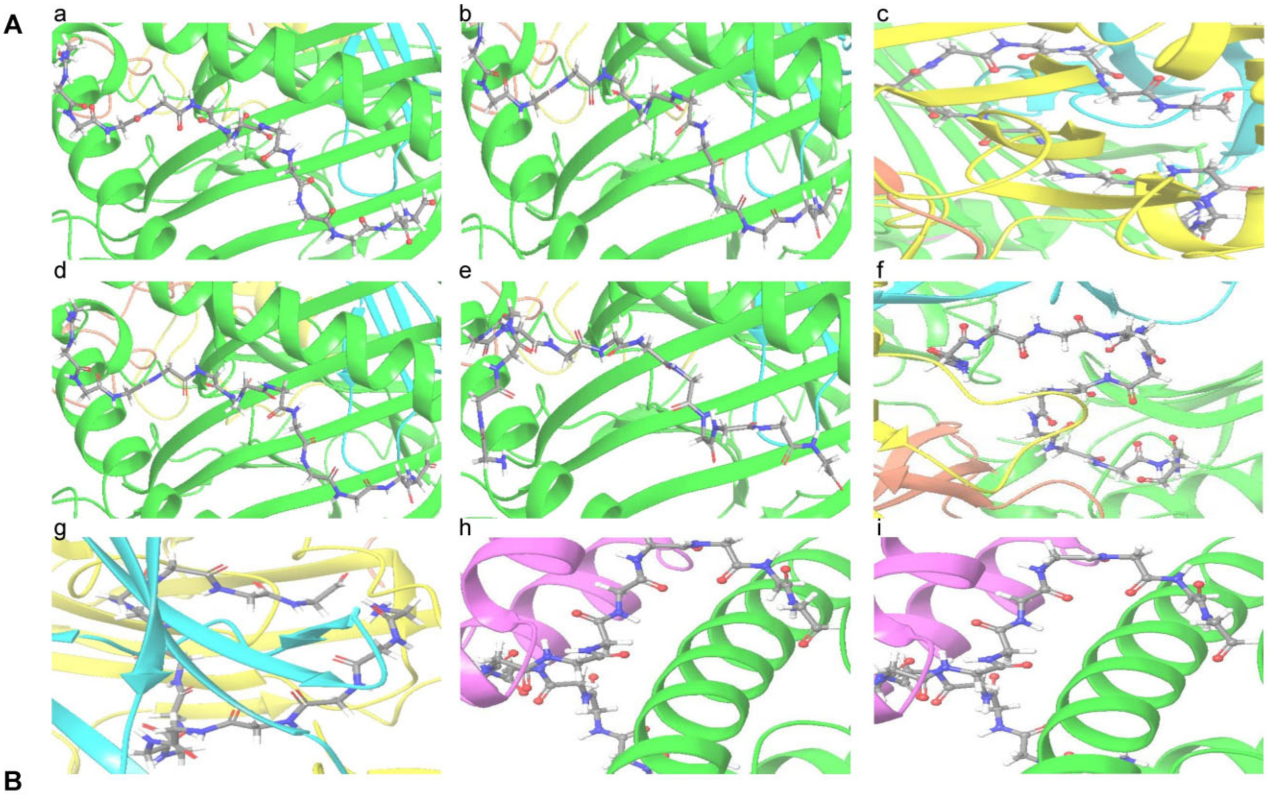
Table 3. Continued

Fusion genes		# samples with fusion genes	# samples with in-frame fusion genes	# samples with neoantigen interacting with HLA-Is	# cancer types with neoantigen interacting with HLA-Is	# samples with neoantigen interacting with HLA-IIs	# cancer types with neoantigen interacting with HLA-IIs	PMIDs of previous studies
KMT2A	MLLT3	32	3	2	2	2	2	
MYB	NFIB	18	4	2	2	2	2	31011208
KDM2A	RHOD	16	3	2	2	2	2	
TPX2	HM13	16	2	2	2	2	2	
PPP6R3	LRP5	15	5	2	2	2	2	
KDM2A	C11orf80	14	4	2	2	2	2	
QKI	PACRG	14	5	2	2	2	2	
PRCC	TFE3	13	3	2	2	2	2	
VTI1A	TCF7L2	13	2	2	2	2	2	
FUS	ERG	12	3	2	2	2	2	
SMCHD1	NDC80	12	3	2	2	2	2	
PTPN12	CCDC146	11	4	2	2	2	2	
ERBB2	PSMB3	10	4	2	1	2	1	
TFG	ALK	10	3	2	2	2	2	
ACO2	ZC3H7B	9	2	2	2	2	2	
FGFR2	BICC1	9	7	2	2	2	2	
FGFR2	ATE1	9	3	2	2	2	2	
SDC4	ROS1	9	3	2	2	2	2	
TBCD	FOXK2	9	5	2	2	2	2	
UBTF	MAML3	9	4	2	1	2	1	
ERC1	WNK1	8	5	2	2	2	2	
EWSR1	WT1	8	2	2	3	2	2	36900411
STRN	ALK	8	5	2	2	2	2	
TC2N	TRIP11	8	3	2	2	2	2	
UBE2K	LIAS	8	2	2	2	2	2	
ESR1	MTHFD1L	7	2	2	2	2	2	
KMT2A	EPS15	7	1	2	2	2	2	
AGBL4	FAF1	6	2	2	2	2	2	
BCR	JAK2	6	1	2	3	2	3	
EFHD1	EIF4E2	6	4	2	1	2	1	
KMT2A	AFF3	6	1	2	2	2	2	
PPFIA1	SLC39A11	6	3	2	1	2	1	
SMARCAD1	GRID2	6	5	2	2	2	2	
TAF15	AP2B1	6	3	2	2	2	2	
WHSC1L1	ADAM32	6	6	2	1	2	1	
MSH2	TTC7A	5	3	2	2	2	2	
NEMF	C14orf182	5	2	2	2	2	2	
PTPN11	EPYC	5	4	2	1	2	1	
RAB3GAP1	ACMSD	5	3	2	2	2	2	
SEC31A	ALK	5	1	2	2	2	2	
UBR1	TGM5	5	5	2	2	2	2	
ATXN2	TRIM37	4	3	2	1	2	1	
CTPS1	NFYC	4	3	2	1	2	1	
CUX1	RET	4	1	2	2	2	2	
DDX10	NUP98	4	2	2	1	2	1	
EHMT1	GRIN1	4	3	2	2	2	2	

likely to elicit a robust immune response compared to neoantigens generated by SNVs and INDELs. Similar to the burden of candidate neoantigens derived from SNVs and INDELs, the burden of fusion neoantigens is closely associated with the frequency of fusion mutations, particularly in microsatellite-stable tumors with a higher fusion mutation burden (1,30). In an expanded study across 30 different tumor types, it was observed that 24% of fusion protein-expressing cancers harbored neoepitopes resulting from such fusion events, and these neoantigens were predicted to bind to patient-specific MHC-I molecules (27,31). In this context, the information provided in FusionNeoAntigen is useful to develop new immunotherapeutics for multiple types of human cancer.

Discussion

In FusionNeoAntigen, we provide comprehensive information on ∼500K and 19k fusion breakpoint-specific neoantigens that interact with HLA-Is and HLA-IIs in ∼ 10K and 5.7K human fusion genes in cancer, respectively. Further virtual screening between the fusion breakpoint-specific peptides of 14 AA length and the known 3D structures of 21 HLAs identified about 85.7% of interacting pairs kept from almost all fusion neoantigens that were predicted as interacting with HLA-Is. We also provide information on the potential cell surface fusion proteins. To predict the structures of the fusion breakpoint peptides and cell surface fusion proteins, we used



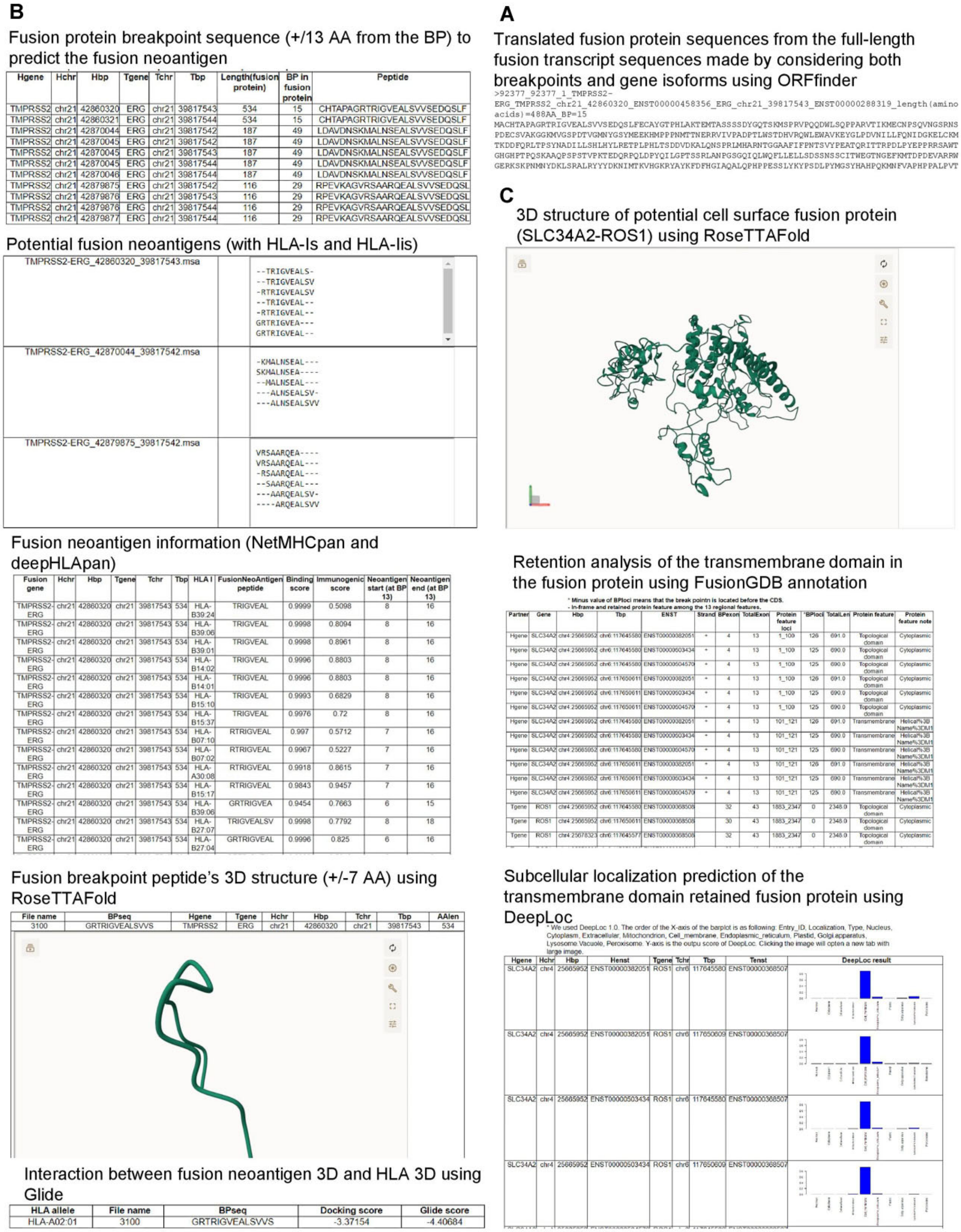
B

ID in panel A	5'-gene	3'-gene	5'-chr	5'-breakpoint	3'-chr	3'-breakpoint	HLA allele	HLA PDB ID		14 AA fusion breakpoint peptide sequence that have fusion neoantigens	Docking score	Glide score
a	BCR	ABL1	chr22	23613779	chr9	133729451	HLA-A24:02	5HGA	4161	KDPTTKNSLEKALQ	-7.52403	-7.63743
b	CBFB	MYH11	chr16	67116211	chr16	15820911	HLA-A24:02	5HGA	8628	SHREEMEAKANLDK	-7.52403	-7.63743
c	CCDC6	RET	chr10	61554231	chr10	43612032	HLA-B14:02	3BVN	3178	GTSHGITEDPKWEF	-7.4838	-7.5956
d	EML4	ALK	chr2	42552694	chr2	29446394	HLA-A24:02	5HGA	1487	DYEILYLYRRKHQE	-7.52403	-7.63743
e	EWSR1	FLI1	chr22	29683123	chr11	128675261	HLA-A24:02	5HGA	9079	SSYGQQNPSYDSVR	-8.26666	-9.30196
f	PML	RARA	chr15	74325754	chr17	38504567	HLA-B52:01	3W39	590	ASGAGEAAIETQSS	-7.72469	-7.73669
g	RPS6KB1	VMP1	chr17	57992064	chr17	57886156	HLA-B14:02	3BVN	430	AMKVLKKDFASRAK	-7.9962	-8.1096
h	RUNX1	RUNX1T1	chr21	36231771	chr8	92972745	HLA-A11:01	4UQ2	6102	NCRLSLLLVCVCVC	-10.0067	-10.1201
i	TMPRSS2	ERG	chr21	42860320	chr21	39817543	HLA-A11:01	4UQ2	3100	GRTRIGVEALSVVS	-10.0067	-10.1201

Figure 2. The 3D complexes of the top nine most frequent fusion genes' neoantigens and HLA-Is. We investigated the interaction between the 3D structures of ∼ 10K fusion protein breakpoint peptide (14 AA length), that have the potential fusion breakpoint-specific neoantigens, and known 3D structures of 21 HLAs. (**A**) Fusion protein breakpoint peptides and HLA complex in 3D – a, BCR-ABL1; b, CBFB-MYH11; c, CCDC6-RET; d, EML4-ALK; e, EWSR1-FLI1; f, PML-RARA; g, RPS6KB1-VMP1; h, RUNX1-RUNX1T1; I, TMPRSS2-ERG. (**B**) Detailed information on the interaction between fusion protein breakpoint peptides and HLAs.

an AI-based protein structure prediction tool, RosettaFold. However, since RoseTTAFold relies on a database of known protein structures to make predictions and neoantigens derived from the fusion breakpoints are unique. As they may not have structurally similar counterparts in the database, we admit that a lack of training data for fusion proteins can affect the accuracy of the fusion breakpoint peptide structure predictions. Even so, about 85.7% of the binding affinity-based predicted interactions between fusion breakpoint peptides of 14 AA and 21 HLA-Is recurred in the virtual screening results. These fusion-specific neoantigen candidates can be novel resources for new therapy development in cancer.

In this study, we predicted the fusion-specific neoantigens that are crossing the breakpoints of all human in-frame fusion genes since our study material, which is the fusion breakpoint peptide sequences was made based on our previous study, FusionGDB2.0. In FusionGDB2.0, we focused on the in-frame fusion genes only and did not generate frame-shift fusion protein sequences. This was due to the significance of the reliably translatable fusion genes and the huge data size of all open reading frames' full-length fusion transcript sequences considering multiple breakpoints and multiple gene isoforms of each fusion gene. However, when we checked the reported fusion neoantigens from previous studies, all the fusion neoantigens regardless of crossing the breakpoints or not, the fusion neoantigen sequences of the well-known fusion genes (BCR-ABL1, EWSR1-FLI1, TMPRSS2-ERG and CBFB-MYH11) were all from the in-frame fusion genes that were included in our prediction. Also, we provide the multiple sequence alignment of all predicted neoantigens at each fusion breakpoint if the neoantigen crosses the breakpoint at least one residue. Even though focused on the in-frame fusion genes, we have ∼500K fusion neoantigens from ∼ 14K in-frame fusion genes, which means there are about 30 fusion neoantigens predicted per fusion gene. Therefore, at this stage, we think in-frame fusion gene-derived neoantigens still provide useful information from the unique combination of two different genes' degraded peptides. In the future, we will



B

Fusion protein breakpoint sequence (+/13 AA from the BP) to predict the fusion neoantigen

Hgene	Hchr	Hbp	Tgene	Tchr	Tbp	Length(fusion protein)	BP in fusion protein	Peptide
TMPRSS2	chr21	42860320	ERG	chr21	39817543	534	15	CHTAPAGRTRIGVEALSVVSEDQSLF
TMPRSS2	chr21	42860321	ERG	chr21	39817544	534	15	CHTAPAGRTRIGVEALSVVSEDQSLF
TMPRSS2	chr21	42870044	ERG	chr21	39817542	187	49	LDAVDNSKMALNSEALSVVSEDQSLF
TMPRSS2	chr21	42870045	ERG	chr21	39817542	187	49	LDAVDNSKMALNSEALSVVSEDQSLF
TMPRSS2	chr21	42870045	ERG	chr21	39817543	187	49	LDAVDNSKMALNSEALSVVSEDQSLF
TMPRSS2	chr21	42870045	ERG	chr21	39817544	187	49	LDAVDNSKMALNSEALSVVSEDQSLF
TMPRSS2	chr21	42870046	ERG	chr21	39817544	187	49	LDAVDNSKMALNSEALSVVSEDQSLF
TMPRSS2	chr21	42879875	ERG	chr21	39817542	116	29	RPEVKAGVRSAARQEALSVVSEDQSL
TMPRSS2	chr21	42879876	ERG	chr21	39817543	116	29	RPEVKAGVRSAARQEALSVVSEDQSL
TMPRSS2	chr21	42879876	ERG	chr21	39817544	116	29	RPEVKAGVRSAARQEALSVVSEDQSL
TMPRSS2	chr21	42879877	ERG	chr21	39817544	116	29	RPEVKAGVRSAARQEALSVVSEDQSL

Potential fusion neoantigens (with HLA-Is and HLA-Iis)

TMPRSS2-ERG_42860320_39817543.msa	--TRIGVEALS- --TRIGVEALSV -RTRIGVEALSV --TRIGVEAL-- -RTRIGVEAL-- GRTRIGVEA--- GRTRIGVEAL--
TMPRSS2-ERG_42870044_39817542.msa	-KMALNSEAL--- SKMALNSEA---- --MALNSEAL--- ---ALNSEALSV- ---ALNSEALSVV
TMPRSS2-ERG_42879875_39817542.msa	VRSAARQEA---- VRSAARQEAL--- -RSAARQEAL--- --SAARQEAL--- ---AARQEALSV- ----ARQEALSVV

Fusion neoantigen information (NetMHCpan and deepHLApan)

Fusion gene	Hchr	Hbp	Tgene	Tchr	Tbp	HLA I	FusionNeoAntigen peptide	Binding score	Immunogenic score	Neoantigen start (at BP 13)	Neoantigen end (at BP 13)
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B39:24	TRIGVEAL	0.9999	0.5098	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B39:06	TRIGVEAL	0.9998	0.8094	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B39:01	TRIGVEAL	0.9998	0.8961	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B14:02	TRIGVEAL	0.9996	0.8803	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B14:01	TRIGVEAL	0.9996	0.8803	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B15:10	TRIGVEAL	0.9993	0.6829	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B15:37	TRIGVEAL	0.9976	0.72	8	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B07:10	RTRIGVEAL	0.997	0.5712	7	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B07:02	RTRIGVEAL	0.9967	0.5227	7	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-A30:08	RTRIGVEAL	0.9918	0.8615	7	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B15:17	RTRIGVEAL	0.9843	0.9457	7	16
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B39:06	GRTRIGVEA	0.9454	0.7663	6	15
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B27:07	TRIGVEALSV	0.9998	0.7792	8	18
TMPRSS2-ERG	chr21	42860320	chr21	39817543	534	HLA-B27:04	GRTRIGVEAL	0.9996	0.825	6	16

Fusion breakpoint peptide's 3D structure (+/-7 AA) using RoseTTAFold

File name	BPseq	Hgene	Tgene	Hchr	Hbp	Tchr	Tbp	AAlen
3100	GRTRIGVEALSVVS	TMPRSS2	ERG	chr21	42860320	chr21	39817543	534

Interaction between fusion neoantigen 3D and HLA 3D using Glide

HLA allele	File name	BPseq	Docking score	Glide score
HLA-A02:01	3100	GRTRIGVEALSVVS	-3.37154	-4.40684

A

Translated fusion protein sequences from the full-length fusion transcript sequences made by considering both breakpoints and gene isoforms using ORFfinder

```
>92377_92377_1_TMPRSS2-
ERG_TMPRSS2_chr21_42860320_ENST00000458356_ERG_chr21_39817543_ENST00000288319_length(amino
acids)=488AA_BP=15
MACHTAPAGRTRIGVEALSVVSEDQSLFECAYGTPHLAKTEMTASSSSDYGQTSKMSPRVPQQDWLSQPPARVTIKMECNPSQVNGSRNS
PDECSVAKGGKMVGSPDTVGMNYGSYMEEKHMPPPNMTTNERRVIVPADPTLWSTDHVRQWLEWAVKEYGLPDVNILLFQNIDGKELCKM
TKDDFQRLTPSYNADILLSHLHYLRETPLPHLTSDDVDKALQNSPRLMHARNTGGAAFIFPNTSVYPEATQRITTRPDLPYEPPRRSAWT
GHGHPTPQSKAAQPSPSTVPKTEDQRPQLDPYQILGPTSSRLANPGSGQIQLWQFLLELLSDSSNSSCITWEGTNGEFKMTDPDEVARRW
GERKSKPNMNYDKLSRALRYYYDKNIMTKVHGKRYAYKFDFHGIAQALQPHPPESSLYKYPSDLPYMGSYHAHPQKMNFVAPHPPALPVT
```

C

3D structure of potential cell surface fusion protein (SLC34A2-ROS1) using RoseTTAFold

Retention analysis of the transmembrane domain in the fusion protein using FusionGDB annotation

* Minus value of BPloci means that the break pointn is located before the CDS.
- In-frame and retained protein feature among the 13 regional features.

Partner	Gene	Hbp	Tbp	ENST	Strand	BPexon	TotalExon	Protein feature loci	*BPloci	TotalLen	Protein feature	Protein feature note
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000382051	+	4	13	1_100	126	691.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000503434	+	4	13	1_100	125	690.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000504570	+	4	13	1_100	125	690.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000382051	+	4	13	1_100	126	691.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000503434	+	4	13	1_100	125	690.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000504570	+	4	13	1_100	125	690.0	Topological domain	Cytoplasmic
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000382051	+	4	13	101_121	126	691.0	Transmembrane	Helical%3B Name%3DM1
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000503434	+	4	13	101_121	125	690.0	Transmembrane	Helical%3B Name%3DM1
Hgene	SLC34A2	chr4:25665952	chr6:117645580	ENST00000504570	+	4	13	101_121	125	690.0	Transmembrane	Helical%3B Name%3DM1
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000382051	+	4	13	101_121	126	691.0	Transmembrane	Helical%3B Name%3DM1
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000503434	+	4	13	101_121	125	690.0	Transmembrane	Helical%3B Name%3DM1
Hgene	SLC34A2	chr4:25665952	chr6:117650611	ENST00000504570	+	4	13	101_121	125	690.0	Transmembrane	Helical%3B Name%3DM1
Tgene	ROS1	chr4:25665952	chr6:117645580	ENST00000368508		32	43	1883_2347	0	2348.0	Topological domain	Cytoplasmic
Tgene	ROS1	chr4:25665952	chr6:117650611	ENST00000368508		30	43	1883_2347	0	2348.0	Topological domain	Cytoplasmic
Tgene	ROS1	chr4:25678323	chr6:117645577	ENST00000368508		32	43	1883_2347	0	2348.0	Topological domain	Cytoplasmic

Subcellular localization prediction of the transmembrane domain retained fusion protein using DeepLoc

* We used DeepLoc 1.0. The order of the X-axis of the barplot is as following: Entry_ID, Localization, Type, Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell_membrane, Endoplasmic_reticulum, Plastid, Golgi apparatus, Lysosome Vacuole, Peroxisome. Y-axis is the outpu score of DeepLoc. Clicking the image will opten a new tab with large image.

Hgene	Hchr	Hbp	Henst	Tgene	Tchr	Tbp	Tenst	DeepLoc result
SLC34A2	chr4	25665952	ENST00000382051	ROS1	chr6	117645580	ENST00000368507	
SLC34A2	chr4	25665952	ENST00000382051	ROS1	chr6	117650609	ENST00000368507	
SLC34A2	chr4	25665952	ENST00000503434	ROS1	chr6	117645580	ENST00000368507	
SLC34A2	chr4	25665952	ENST00000503434	ROS1	chr6	117650609	ENST00000368507	

Figure 3. Representative annotation categories in FusionNeoAntigen. (**A**) Fusion protein sequences are provided from the ORF annotation and coding potential investigation. (**B**) From the fusion breakpoint sequence of ±13 AA length from the breakpoint, we provide the fusion neoantigen candidate information such as ±13 AA peptide sequence, binding affinity-based prediction scores, multiple sequence alignment of multiple fusion neoantigens per fusion breakpoint, the 3D structure of the fusion breakpoint peptide of ±7 AA length and the docking score between the fusion breakpoint peptide 3D structure and known 3D structure of HLAs. (**C**) We also provide information on the potential cell surface-located fusion proteins as the potential target of CAR-T therapy. For the cell-surface located fusion proteins, we provide the predicted 3D structure, transmembrane domain retention result in the fusion protein and the predicted location for the given fusion protein sequences.

work to derive fusion protein sequences from all open reading frames (including frame-shift ORF) and predict all potential fusion-specific neoantigens in the next version.

We hope FusionNeoAntigen can provide helpful knowledge on the development of cancer type-specific fusion-specific targeted immune therapies so that fusion gene-induced cancer patients can have the right personalized therapeutics. We believe FusionNeoAntigen will be routinely used in diverse research communities, including cancer research, genomic study, pharmacology study, etc. We will provide guidelines to improve the findability, accessibility, interoperability and reuse of digital assets (FAIR).

Data availability

All annotation results are available from the FusionNeoAntigen website (https://compbio.uth.edu/FusionNeoAntigen). Further information and requests should be directed to Dr. Pora Kim (Pora.kim@uth.tmc.edu).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank the fusion gene researchers, the reviewers of this study, and the members of the Center for Computational Systems Medicine for valuable questions, discussions and suggestions. We also thank Marielle Holmblad and Grant K. Wang for assistance in improving the English of the manuscript and website.

Funding

National Institutes of Health grants [R35GM138184 to P.K.]; Funding for open access charge: Startup Fund to Dr Kim from the University of Texas Health Science Center at Houston. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement

None declared.

References

1. Wei,Z., Zhou,C., Zhang,Z., Guan,M., Zhang,C., Liu,Z. and Liu,Q. (2019) The landscape of tumor fusion neoantigens: a Pan-Cancer analysis. *iScience*, **21**, 249–260.
2. Wang,Y., Shi,T., Song,X., Liu,B. and Wei,J. (2021) Gene fusion neoantigens: emerging targets for cancer immunotherapy. *Cancer Lett.*, **506**, 45–54.
3. Haas,B.J., Dobin,A., Li,B., Stransky,N., Pochet,N. and Regev,A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**, 213.
4. Zhang,J., Mardis,E.R. and Maher,C.A. (2017) INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics*, **33**, 555–557.
5. Hundal,J., Kiwala,S., McMichael,J., Miller,C.A., Xia,H., Wollam,A.T., Liu,C.J., Zhao,S., Feng,Y.Y., Graubert,A.P., *et al.* (2020) pVACtools: a Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol. Res.*, **8**, 409–420.
6. Chang,T.C., Carter,R.A., Li,Y., Li,Y., Wang,H., Edmonson,M.N., Chen,X., Arnold,P., Geiger,T.L., Wu,G., *et al.* (2017) The neoepitope landscape in pediatric cancers. *Genome Med.*, **9**, 78.
7. Rubinsteyn,A., Kodysh,J., Hodes,I., Mondet,S., Aksoy,B.A., Finnigan,J.P., Bhardwaj,N. and Hammerbacher,J. (2017) Computational Pipeline for the PGV-001 Neoantigen Vaccine Trial. *Front. Immunol.*, **8**, 1807.
8. Rech,A.J., Balli,D., Mantero,A., Ishwaran,H., Nathanson,K.L., Stanger,B.Z. and Vonderheide,R.H. (2018) Tumor Immunity and Survival as a Function of Alternative Neopeptides in Human Cancer. *Cancer Immunol. Res.*, **6**, 276–287.
9. Weber,D., Ibn-Salem,J., Sorn,P., Suchan,M., Holtstrater,C., Lahrmann,U., Vogler,I., Schmoldt,K., Lang,F., Schrors,B., *et al.* (2022) Accurate detection of tumor-specific gene fusions reveals strongly immunogenic personal neo-antigens. *Nat. Biotechnol.*, **40**, 1276–1284.
10. Luo,R., Chyr,J., Wen,J., Wang,Y., Zhao,W. and Zhou,X. (2023) A novel integrated approach to predicting cancer immunotherapy efficacy. *Oncogene*, **42**, 1913–1925.
11. Kim,P., Tan,H., Liu,J., Lee,H., Jung,H., Kumar,H. and Zhou,X. (2022) FusionGDB 2.0: fusion gene annotation updates aided by deep learning. *Nucleic Acids Res.*, **50**, D1221–D1230.
12. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S., *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
13. Jang,Y.E., Jang,I., Kim,S., Cho,S., Kim,D., Kim,K., Kim,J., Hwang,J., Kim,S., Kim,J., *et al.* (2020) ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res.*, **48**, D817–D824.
14. UniProt,C. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
15. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
16. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
17. Reynisson,B., Barra,C., Kaabinejadian,S., Hildebrand,W.H., Peters,B. and Nielsen,M. (2020) Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data. *J. Proteome Res.*, **19**, 2304–2315.
18. Reynisson,B., Alvarez,B., Paul,S., Peters,B. and Nielsen,M. (2020) NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.*, **48**, W449–W454.
19. Wu,J., Wang,W., Zhang,J., Zhou,B., Zhao,W., Su,Z., Gu,X., Wu,J., Zhou,Z. and Chen,S. (2019) DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. *Front. Immunol.*, **10**, 2559.
20. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D., *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
21. Friesner,R.A., Murphy,R.B., Repasky,M.P., Frye,L.L., Greenwood,J.R., Halgren,T.A., Sanschagrin,P.C. and Mainz,D.T. (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.*, **49**, 6177–6196.
22. Hu,Z., Yuan,J., Long,M., Jiang,J., Zhang,Y., Zhang,T., Xu,M., Fan,Y., Tanyi,J.L., Montone,K.T., *et al.* (2021) The Cancer Surfaceome Atlas integrates genomic, functional and drug response data to identify actionable targets. *Nat Cancer*, **2**, 1406–1422.
23. Thumuluri,V., Almagro Armenteros,J.J., Johansen,A.R., Nielsen,H. and Winther,O. (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.*, **50**, W228–W234.

24. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z., *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
25. Gubin,M.M., Artyomov,M.N., Mardis,E.R. and Schreiber,R.D. (2015) Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.*, **125**, 3413–3421.
26. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
27. Yang,W., Lee,K.-W., Srivastava,R.M., Kuo,F., Krishna,C., Chowell,D., Makarov,V., Hoen,D., Dalin,M.G. and Wexler,L. (2019) Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat. Med.*, **25**, 767–775.
28. Leko,V. and Rosenberg,S.A. (2020) Identifying and targeting human tumor antigens for T cell-based immunotherapy of solid tumors. *Cancer Cell*, **38**, 454–472.
29. Hindson,J. (2019) Gene-fusion neoantigens stimulate T cells. *Nat. Rev. Cancer*, **19**, 364–364.
30. Wang,L., Shamardani,K., Babikir,H., Catalan,F., Nejo,T., Chang,S., Phillips,J.J., Okada,H. and Diaz,A.A. (2021) The evolution of alternative splicing in glioblastoma under therapy. *Genome Biol.*, **22**, 48.
31. Frankiw,L., Baltimore,D. and Li,G. (2019) Alternative mRNA splicing in cancer immunotherapy. *Nat. Rev. Immunol.*, **19**, 675–687.

Received: August 15, 2023. **Revised:** October 2, 2023. **Editorial Decision:** October 8, 2023. **Accepted:** October 9, 2023