Bioinformatics, 36, 2020, i203–i209
doi: 10.1093/bioinformatics/btaa504
ISMB 2020

MutCombinator: identification of mutated peptides allowing combinatorial mutations using nucleotide-based graph search

Seunghyuk Choi and Eunok Paek*

Department of Computer Science, Hanyang University, Seongdong-gu, Seoul 04763, Republic of Korea

*To whom correspondence should be addressed.

Abstract

Motivation: Proteogenomics has proven its utility by integrating genomics and proteomics. Typical approaches use data from next-generation sequencing to infer proteins expressed. A sample-specific protein sequence database is often adopted to identify novel peptides from matched mass spectrometry-based proteomics; nevertheless, there is no software that can practically identify all possible forms of mutated peptides suggested by various genomic information sources.

Results: We propose MutCombinator, which enables us to practically identify mutated peptides from tandem mass spectra allowing combinatorial mutations during the database search. It uses an upgraded version of a variant graph, keeping track of frame information. The variant graph is indexed by nine nucleotides for fast access. Using MutCombinator, we could identify more mutated peptides than previous methods, because combinations of point mutations are considered and also because it can be practically applied together with a large mutation database such as COSMIC. Furthermore, MutCombinator supports in-frame search for coding regions and three-frame search for non-coding regions.

Availability and implementation: https://prix.hanyang.ac.kr/download/mutcombinator.jsp.

Contact: eunokpaek@hanyang.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the advances in genomics and proteomics technologies such as next-generation sequencing and tandem mass spectrometry, we can better identify sample-specific and/or novel peptides. Construction of protein sequence database plays an important role in identifying novel peptides because reliable peptide identification depends heavily on the database being searched. In terms of mutated peptide identification, many methods have been proposed to construct a proteogenomic database, such as CustomizedDB (Park *et al.*, 2014), CustomProDB (Wang and Zhang, 2013), CanProVar (Zhang *et al.*, 2017) and a variant graph (Woo *et al.*, 2014a). CustomizedDB and CustomProDB assumed that all mutations occur simultaneously in a gene. On the other hand, CanProVar assumed that all digested peptides can have no more than a single mutation. These approaches reduce the search time by avoiding the exhaustive search for all possibilities, but naturally preclude covering all possible mutated peptides at the same time.

Woo and colleagues proposed a 'variant graph', which represents a given transcriptome model as a direct acyclic graph, where each node is a nucleotide sequence representing a part of an exon or a variant call and edges connect neighboring exons indicating splice sites or point mutation occurrences. They also provided a 'variant graph to FASTA' enumeration package because most database search engines only take a FASTA formatted database as an input. The enumerated variant graph can represent almost all possible combinations of mutated peptides depending on a user specific parameter. However, the parameter is not intuitive because it is an internal parameter that controls the algorithm behavior and does not directly describe the desired output. The parameter value has to do with the density of mutation calls, which may vary widely depending on genes, making it very difficult to set the value properly. Furthermore, the variant graph method does not allow a user to set the translation frame: a user may want only in-frame translation, or all three-frame translations.

To overcome such limitations, we developed MutCombinator, which enables us to identify combinatorially mutated peptides by searching a variant graph directly (Fig. 1) without enumerating them into amino acid sequences off-line. A variant graph is built by taking reference genome sequences in FASTA, transcriptome model in GTF and variant calls in VCF format as input. MutCombinator can identify peptides with combinations of maximum n mutations (n is a user-specified parameter) in a peptide. The possible combinations of mutations grow exponentially as n increases; therefore, the search time can grow exponentially. To keep the search time under control, we adopted the two existing techniques: (i) extract short amino acids sequence tags (of length 3) from a spectrum, and search paths containing at least one tag to avoid traversing the whole variant graph [this approach was suggested elsewhere (Mann and Wilm, 1994), but they used FASTA database instead of a variant graph]

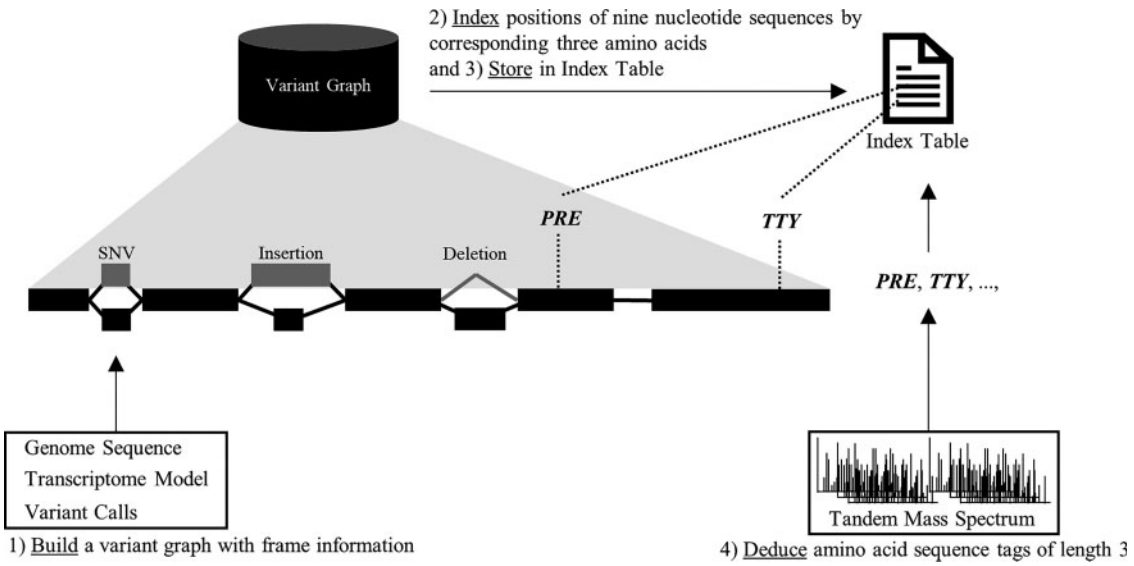


Fig. 1. Overview of MutCombinator. A variant graph is built from the reference genome sequences, a transcriptome model and variant calls. Positions of each nine nucleotide sequence in the variant graph are indexed by corresponding three amino acids and stored in a pre-compiled index table. Positions of sequence tags (e.g. PRE and TTY) deduced from a spectrum are directly recognized by looking up the index table. PSM is processed by traversing the flanking paths of tag positions. The black box represents an exon model in the given transcriptome model. The gray box represents a variant model annotated in the given variant calls

and (ii) indexing variant graph using three amino acids long sequences to directly access nucleotide sequences in the graph. For convenience of a user, MutCombinator supports (i) multi-threading, which enables parallel processing of spectra and (ii) both in-frame coding region search as well as three-frame search that encompasses non-coding regions as well.

We designed a multistage search (Madar *et al.*, 2018) with MutCombinator to effectively identify mutated peptides using proteogenomics data from a previous study (Mun *et al.*, 2019). First, we used unidentified tandem mass (MS/MS) spectra from the previous study as an input for the second stage search using MutCombinator under the conditions: (i) use of 12 688 mutations from sample-specific variant calls and 83 873 mutations from COSMIC database (a total of 96 287 mutations), (ii) 28 843 expressed protein coding transcripts (supported by FPKM > 1) and (iii) allowing up to three mutations per a peptide. As a result, we additionally identified 80 mutated peptides than the previous report. From this result, we could find 10 additional KEGG-pathways and 70 combinations of mutations. Furthermore, we also identified four mutated peptides harboring exclusively expressed mutations.

At the third stage, we further searched unidentified MS/MS spectra from the first and second stage search against the same database with the same search conditions as the second stage except for one thing: use of 12 516 expressed non-coding transcripts (also supported by FPKM > 1). We identified 14 aberrantly translated peptides—five pseudogenes, four frameshifts, two exon extensions, two 5′ UTRs, and one 3′ UTR peptides.

2 Materials and methods

Peptide identification in MutCombinator consists of four parts: (i) construction of a variant graph with frame information, (ii) indexing the variant graph to directly access nucleotide sequences in the graph, (iii) candidate peptides generation by traversing the variant graph from indexed positions and (iv) scoring peptide spectrum matches (PSMs).

2.1 Construction of variant graph with frame-awareness

Originally, a variant graph was not designed to limit the search only to in-frame translation because it assumed that the graph would be built directly from RNA-Seq results. This assumption was made for discovery of potential novel coding regions; however, it is not suitable for identifying mutations in known protein coding regions, because it triples the search space, resulting in increased false positives and execution time. Recent proteogenomic research have focused more on identifying mutated peptides in the coding region because their relation with the disease can be significant (Mertins *et al.*, 2016; Mun *et al.*, 2019; Zhang *et al.*, 2014). To facilitate comprehensive mutation identification in a proteogenomic search, we augmented variant graphs with frame information in each node whenever it is available.

When constructing a variant graph with frame-awareness, each transcript model is initially represented as a linear graph structure (list) where nodes represent nucleotide sequences of exons and edges represent junction sites between two exons (Fig. 2a). When multiple transcripts share a common region (the same genome positions and the same nucleotide sequences, shown in gray in Fig. 2a), it is represented as a single node in the merged transcript as shown in Figure 2b while the remaining parts of the original transcript, i.e. distinct parts, are split from the original node and the split sites are connected by an edge. When splitting nodes, each node inherits its frame information from the original transcript. The frame information is recorded as binary vectors where each row represents each of the three frames and each column represents a transcript of a given gene. For example, if a gene has two transcripts, then the first transcript is represented with a bit value 0b00000001 and second one 0b00000010. Thus, only one of the three rows in the binary vectors has non-zero value in the original transcript, if the transcript model represents a coding sequence (Fig. 2a). We used only seven bits because the sign-bit (the most significant bit) is not suitable for index value. When a gene has more than 7 transcripts, PABPC1 has 18 protein coding transcripts for example and the column size of the binary vectors is determined as $\{1 +$ quotient of dividing the number of transcripts by $7\}$ bytes.

To compact the transcripts into a single merged transcript, nodes in the transcripts are split into common and distinct parts based on both genomic positions and nucleotide sequences. When nodes are split into two or three nodes, each frame information of the split nodes is recalculated in order to keep track of the transcript structures. In the example shown in Figure 2b (focusing on how the frame information of the second transcript changes) frames of the first node do not change. Frames of the remaining nodes are calculated based on their predecessor nodes in a topological order. If the nucleotide sequence length of a predecessor node is a multiple of

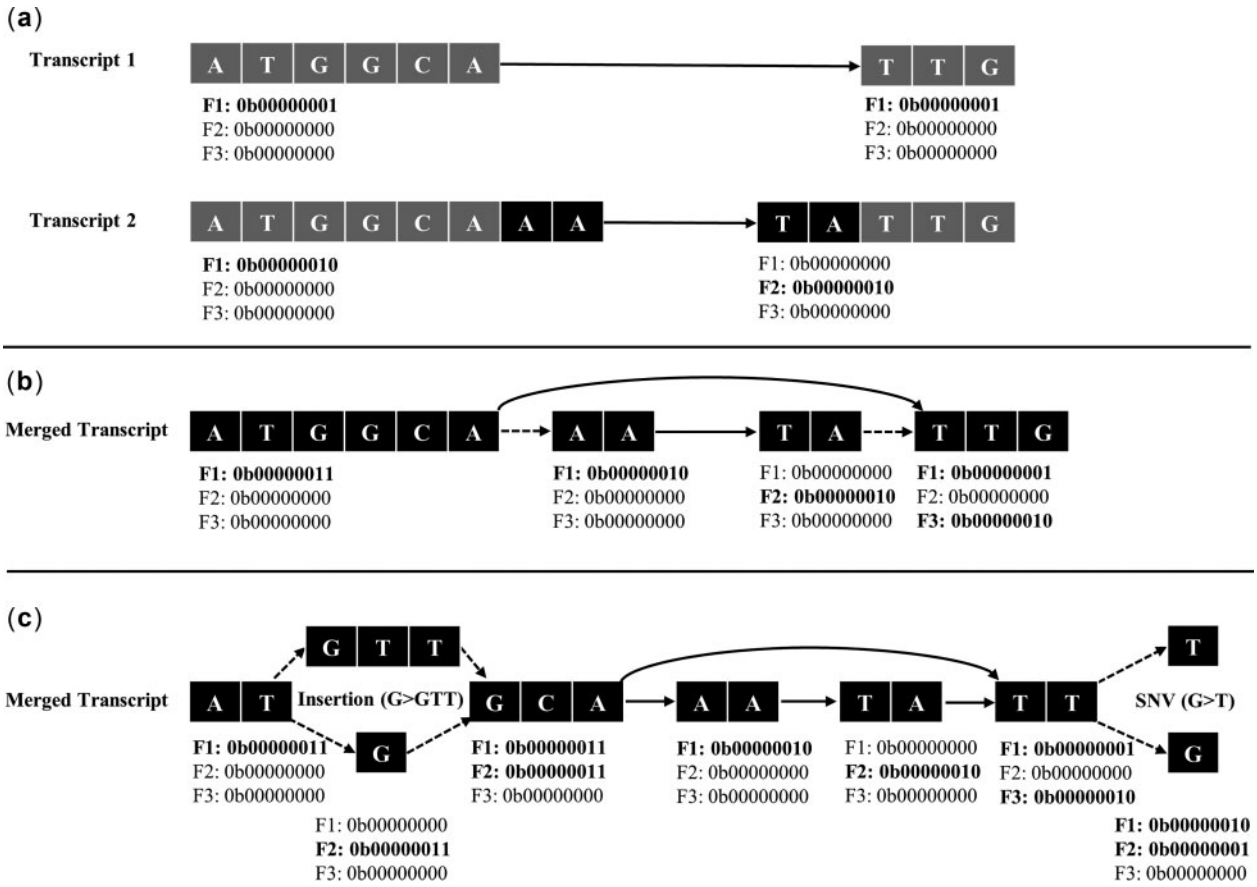


Fig. 2. Construction of a variant graph augmented with frame information. Nodes represent nucleotide sequences and edges connect neighboring nodes. Edges generated by splitting nodes in each step are represented as dotted lines. The letter boxes indicate coding sequences. (**a**) Each transcript is represented as a linear graph structure. Frame information is recorded in each node as binary vectors. Gray nodes represent common regions among multiple transcripts of the same gene. (**b**) The common regions are merged into a single node, and their original frame information is coalesced by bit-OR operation. Each gene is represented as a single directed acyclic graph after this step. (**c**) SNVs and insertions are added, and frames are re-calculated

three, then the current node is assigned the same frame of the predecessor node. Otherwise, the frames of the current node are set as the union of up-rotating each predecessor's frame by the remainder of dividing its nucleotide sequence length by 3. Union operation is actually performed by bit-OR operation. This way, all the transcript models of a gene can be merged into a single variant graph.

In case there are SNVs and insertions given as input (in VCF format), a node containing mutation site is split based on its mutation position. A new node representing the mutation is created and added to the graph, and their frames must be recalculated. The recalculation method is the same as above. In the example shown in Figure 2c, there are two mutations: insertion 'G > GTT' and SNV 'G > T'. While there is a single predecessor node for the last node SNV 'G > T' (the split node and new node), the last node 'GCA', which is caused by the insertion, has two predecessors such as node 'G' and node 'GTT'. In this case, we coalesce both up-rotated frame information by bit-OR operation, and apply the coalesced frame information as that of the last node 'GCA'. In the case of deletion, we simply make a new transcript model in the gene before the merge step.

As an concrete example, the first node in Figure 2b, representing a nucleotide sequence 'ATGGCA', is split into three nodes such as 'AT', 'G' and 'GCA' shown in Figure 2c. Node 'AT' inherited the same frame as the original node 'ATGGCA' because it is the first node among the three. The successor of node 'AT' is node 'G'; therefore, the frames of node 'G' is set as up-rotating that of node 'AT' twice. Similarly, the frame information of the last node 'GCA' is set as bit-OR operation of both up-rotating that of node 'G' once and up-rotating that of node 'GTT' three times (thus no rotation operation). Notice that node 'GTT' as the insertion 'G > GTT' just copies the same frame information with node 'G' because they have the same predecessor node 'AT'.

2.2 Indexing variant graph

In a typical database search approach to peptide identification, each spectrum is compared with candidate peptide sequences in a sequence database. There have been three major methods to avoid searching the whole database: (i) limit candidate peptides only to those that match the precursor mass of a given spectrum, (ii) select candidate peptides using fragment ion matches (Kong *et al.*, 2017) and (iii) select candidate peptides that contain sequence tags derived by applying de novo sequencing to a given spectrum (Mann and Wilm, 1994; Na *et al.*, 2012; Tabb *et al.*, 2003). The first and second approaches enumerate all possible enzymatic peptides and find the best match for each spectrum. Enumerating all possible peptides of a given variant graph is not practical once we start to consider combinatorial mutation. MutCombinator adopted the third approach: extracting three amino acids sequence tag from an input spectrum and directly accessing the tag positions in variant graph by pre-compiled index.

Each index is generated by traversing the whole variant graph and recording the following information: (i) a start node, (ii) an end node, (it must be noted that a sequence tag may span over multiple nodes), (iii) a tag start position within the start node, (iv) a tag end position within the end node, (v) a gene id, (vi) the number of mutation sites included in the nine nucleotide sequence of a tag and (vii) the frame information of a tag obtained by bit-AND operation of all the nodes in a path spanning the nine nucleotide sequence. During index generation, there are two cases when an index of tag should

be discarded: (i) the number of mutation occurrences exceeds the maximum allowable mutations n, specified as a user parameter, or (ii) all three frame information is 0b00000000, meaning that there is no proper path denoted by the tag.

2.3 Generating candidate peptides

Sequence tags are inferred from a spectrum by *de novo* sequencing. The positions of these tag occurrences in a variant graph can be retrieved by looking up the pre-compiled index table. Candidate peptides, the masses of which match to a precursor mass of the spectrum, are generated by traversing the flanking nodes neighboring the tag positions. When traversing the flanking nodes, nucleotide sequences in a node is virtually translated to amino acids for all the valid frames and the peptide mass is calculated while extending the tag sequence into the neighboring nodes until the peptide mass just exceeds the precursor mass given the tolerance. The frame information in the tag is updated during the traversal by bit-AND operation among the visited nodes to confirm the validity of a path, i.e. a variant graph merged all the transcript models of a gene into a single graph, thus an integrity check is necessary to confirm that a path actually corresponds to some transcript model (Supplementary Fig. S1). The traversal stops extending into the neighboring nodes whenever the frame information becomes 0b00000000 (i.e. there is no valid transcript that matches the nucleotide sequence of the current path), or when a path contains more mutations than the number of maximum allowable mutation n.

2.4 Evaluating PSM quality

MODa (Na *et al.*, 2012) evaluated PSMs using a logistic regression of four component scores such as: (i) prefix residue mass (PRM) score, (ii) mass error of matched fragment ions, (iii) the fractions of b and y ions found and (iv) the propensity to a particular ion type. We used the same scoring method to evaluate PSMs. Briefly, an experimental spectrum is first converted into a PRM spectrum, and the PRM spectrum is used to match against the candidate peptides using an alignment based on dynamic programming.

3 Results

3.1 Multistage search to further identify mutated peptides

A huge number of disease-related mutations could be obtained by several resources such as ClinVar (Landrum *et al.*, 2018) and COSMIC databases (Tate *et al.*, 2019). To use these resources for proteogenomic study, we usually make a mutated peptide sequence database by considering all possible combinations of mutations because we cannot be sure which mutations might be observed in our samples.

Under the assumption that some of these mutated peptides, derived from the public mutation resources, could also be observed by MS/MS spectra for the sample of our interest, we designed multistage search (Madar *et al.*, 2018) using MutCombinator (Fig. 3). We chose N33T34 dataset, obtained from a tissue sample of a microsatellite instability (MSI) high cancer patient from a previous proteogenomics study on EOGC (early onset gastric cancer) (Mun *et al.*, 2019). N33T34 dataset included three types of data—4 215 882 MS/MS spectra [the spectra were processed by PE-MMR (Shin *et al.*, 2008)] labeled with 4-plex iTRAQ, mRNA-seq and whole exome-seq. We used a preprocessed dataset provided by Mun and colleagues. There were a total of 41 359 expressed transcripts annotated in Ensembl transcriptome model v71. Among them, 28 843 and 12 516 transcripts were protein coding and non-coding, respectively. As for mutations, there were a total of 12 688 mutations matched to the expressed transcripts.

Among 12 688 sample-specific mutations, only 254 mutations (247 SNVs and 7 insertions) were found among 83 873 stomach cancer-related mutations of COSMIC database (version 87) with the following conditions: (i) available genomic positions and mutated nucleotide sequences, (ii) categorized as SNV, insertion, or deletion and (iii) matched to the expressed transcripts. Assuming that the two mutation sources can be complementary to each other, we constructed CnSSVG (Cosmic and sample-specific variant graph) using 96 307 unified mutations (83 873 stomach cancer-related mutations as well as 12 688 sample-specific mutations).

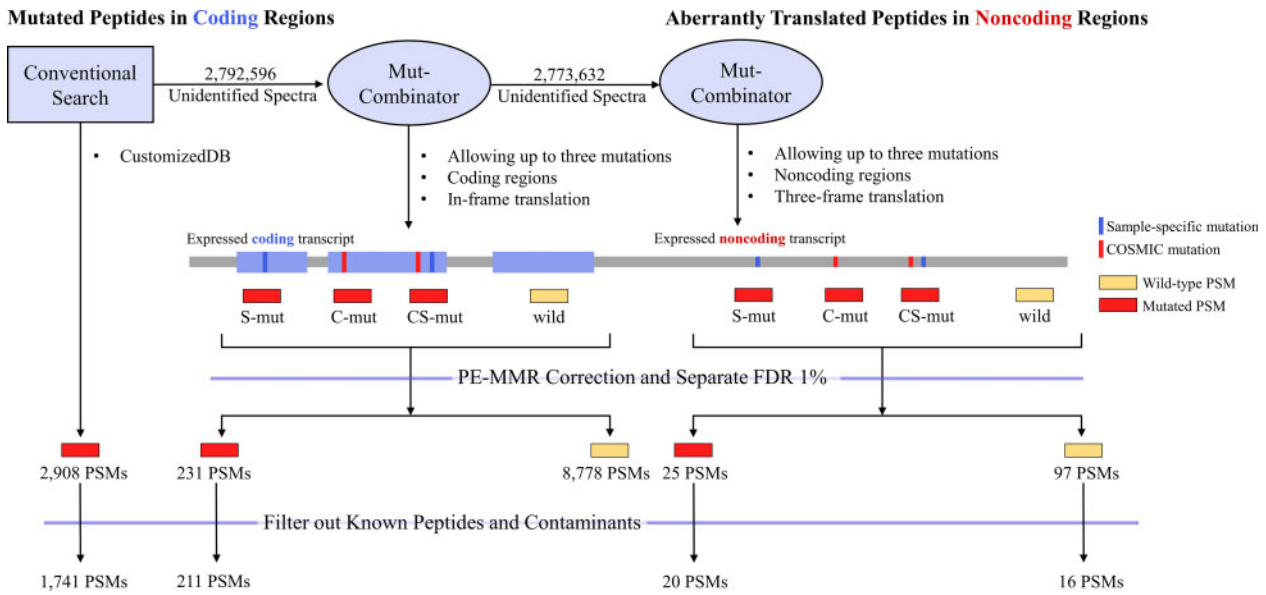


Fig. 3. Multistage search using MutCombinator. Unidentified MS/MS spectra from the previous result (EOGC second-stage dataset) are subjected to MutCombinator as an input. MutCombinator identifies mutated peptides considering combinations of mutations of both sample-specific and COSMIC mutations in the expressed coding transcripts. Unidentified MS/MS spectra from the expressed coding transcripts are subjected to identification of aberrantly translated peptides in the expressed non-coding transcripts. The identified PSMs are filtered out if there is the same sequence in UniProt proteome or contaminants. Note that the result of conventional search was provided by Mun and colleagues

EOGC group identified 588 483 PSMs by searching N33T34 spectra against CustomizedDB (Park *et al.*, 2014) using MS-GF+ search (Kim and Pevzner, 2014) and we denoted this search strategy as a conventional search in Figure 3. To further identify mutated peptides considering combinations of mutations of both sample-specific and COSMIC mutations, we used 2 792 596 unidentified MS/MS spectra. Note that PE-MMR generates multiple spectra per scan by correcting precursor m/z and charge state; therefore, we filtered out 1 423 286 MS/MS spectra corresponding to 588 483 identified MS/MS scans in the previous result.

We searched 2 792 596 MS/MS spectra using MutCombinator against CnSSVG. The search parameters were set as follows: 10 ppm for precursor tolerance, 0.025 Da for fragment tolerance, three fixed modifications (carbamidomethylation at cysteine and iTRAQ label at peptide N-terminal and lysine), semi-tryptic for enzyme specificity allowing up to two miscleavages and eight for minimum peptide length. We also set n to three, allowing up to three mutations per peptide. After the search, the same scans could appear more than one time in the PSM list because the spectra were processed by PE-MMR; therefore, we selected a PSM having the highest score among PSMs with the same scan number. And then, we applied separate false discovery rate (FDR) strategy (Woo *et al.*, 2014b) so that mutated peptides and wild-type peptides could fairly compete with each other. We divided the search results into two: (i) PSMs with wild-type peptide match including both target or decoy and (ii) PSMs with mutated peptide match including both target or decoy. If a sequence of mutated peptide is equivalent to a sequence of wild-type peptide, we assigned the PSM as a wild-type PSM. The result was estimated at 1% local-FDR at PSM level. We identified 8778 wild-type PSMs and 231 mutated PSMs. From 231 mutated PSMs, we filtered out those sequences of which were found in UniProt proteome (release 2019-11) or contaminants. Repeating a similar workflow, we further identified aberrantly translated peptides from 12 516 non-coding transcripts using non-coding search mode in MutCombinator as the last stage. We also compared MutCombinator with a conventional search (MS-GF+ applied against CustomizedDB including mutation) when executed in a single stage search mode (Supplementary Fig. S2). It must be noted that the search space of the two can be tremendously different.

3.2 Mutated peptides in coding regions

With a multistage search using MutCombinator, we further identified 211 mutated PSMs in the coding regions (Fig. 4a). This result amounts to additional identification of 80 mutated peptides, 52 genes and 70 combinations of mutations. We compared two KEGG-pathways (P-value < 0.05) from (i) genes from the conventional search and (ii) genes from the conventional search together with MutCombinator, using DAVID (Huang da *et al.*, 2009) to see whether the additional gain in peptide identification could lead to different interpretation in terms of pathways (Fig. 4b). In the original conventional search results, two pathways were strongly enriched in ECM-receptor interaction and focal adhesion, showing significantly negative mRNA-survival correlation (Mun *et al.*, 2019). Our approach resulted in 10 additional significantly enriched pathways. To make sure that such additional enriched pathways are not random, possibly due to high proportions of such genes in CnSSVG, we further calculated P-values using Fisher exact test. We used all genes harboring mutations in CnSSVG, as a background population and then calculated P-value of each pathway using the genes found by MutCombinator only. All of the pathways showed P-value below 0.05, showing that the pathways are significantly enriched in the search results (details in Supplementary Table S1). Proteoglycans in cancer, one of the additional pathways, also showed significant negative mRNA-survival correlation in the previous report. Intriguingly, inflammation related pathways such as phagosome, leukocyte transendothelial migration, bacterial invasion of epithelial cells and viral myocarditis were enriched and this result is consistent with the already known relationship between inflammation and cancers (Coussens and Werb, 2002).

Owing to MutCombinator, we can further identify 70 combinations of mutations including 60 SNVs, 6 INDELs, 3 double SNVs and 1 double INDEL (Fig. 4c). Nine SNVs were derived from the

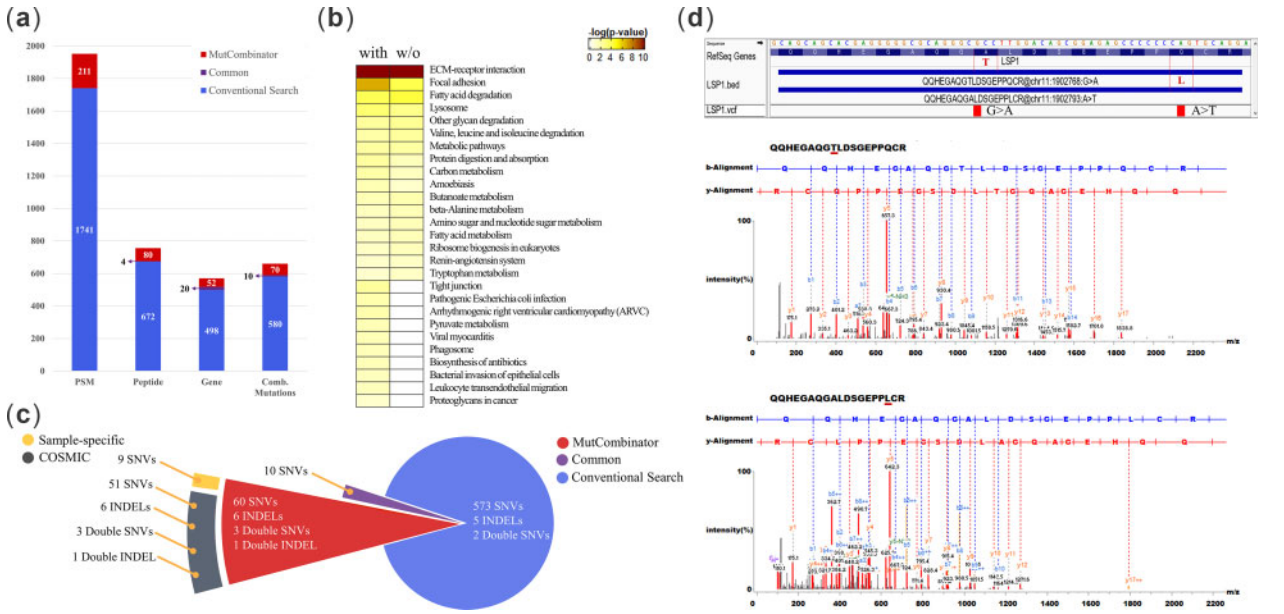
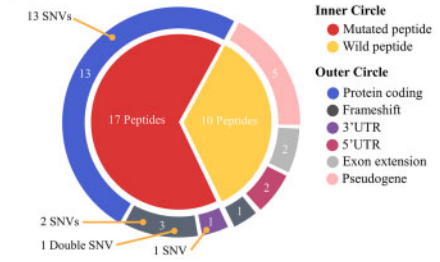


Fig. 4. The identification of mutated peptides in coding regions. (**a**) Peptides, genes and combinations of mutations corresponding to a total of 211 mutated PSMs are described. (**b**) KEGG-pathways of two gene groups—results of conventional search with/without MutCombinator analysis—show different patterns. Pathways with P-value < 0.05 are used. (**c**) Combinations of mutations observed in MS/MS are categorized into three groups—conventional search, MutCombinator and commonly observed by both. The combinations of mutations in MutCombinator group are further classified into sample-specific and COSMIC mutations. (**d**) Mutated peptides harboring exclusively expressed mutations in LSP1 protein. Identified peptides and corresponding gene model are shown, and amino acid changes are indicated by red underline

(a)



(b)

Novel Peptide	Loci (GRCh37)	Gene	Type	Mutations
QVIILNHPGQISTGYAPVLDCHTAHIACK	chr7:22550507-22550593 or chr2:107314063-107314149	EEF1A1P6 or EEF1A1P12	Pseudogene	None
SELSSGTR	chr20:34148374-34148397	FER1L4	Pseudogene	None
QAVCGAYQVLASKLCPIK	chr1:43417381-43417434	SLC2A1	Pseudogene	None
AFLADPSAFVAAAPVAAATTAAPAAAAAAPAK	chr11:61404950-61405045	RPLP0P2	Pseudogene	None
LVKPGNQNTQVTQAWNK	chr3:156528006-156528056	PA2G4L1	Pseudogene	None
SHCIENSK	chr11:43352069-43352092	API5	Frameshift	chr11:43352073:G>A and chr11:43352088:A>G
RMILENER	chr20:57561152-57561175	NELFCD	Frameshift	chr20:57561169:G>T
VANTLSPK	chr3:184024180-184024203	PSMD2	Frameshift	chr3:184024196:C>G
VFVSSTPEKADQVAR	chr5:179048327-179048371	HNRNPH1	Frameshift	None
QESTLHLVLR	chr2:55461964-55461993	RPS27A	Exon-extension	None
TILSSEVVCAPPTAYIDFAR	chr12:6978012-6978071	TPI1	Exon-extension	None
IISAPTMATFVELSTK	chr7:134212646-134212693	AKR1B10	5'UTR	None
SAVEMAVWEAEQLGGLQR	chr20:62203697-62203750	HELZ2	5'UTR	None
TEGLGEAR	chr17:3799917-3799940	P2RX1	3'UTR	chr17:3799918:G>C

Fig. 5. Aberrantly translated peptides supported by MS/MS assay. (**a**) Mutated and wild peptides are categorized into variant types such as (1) protein coding, (2) frameshift, (3) UTRs, (4) exon extension and (5) pseudogene. (**b**) Details of novel peptides supported by MS/MS assay. The first peptide in the list is matched to two genes

sample-specific mutations, and the rest were derived from the COSMIC mutations.

On the other hand, MutCombinator enables considering combinations of mutations allowing up to three mutations per peptide. We could identify four mutated peptides harboring exclusively expressed mutations in RHOA, RRBP1, HIST1H3H and LSP1. For example, we identified two mutated peptides resided in a genomic region from 1 902 744 to 1 902 800 in chromosome 11 (Fig. 4d). One of them had Alanine changed into Threonine at position 9 because of SNV (G > A) at locus chr11:1 902 768. The other peptide had Glutamine changed into Leucine at position 17 because of SNV (A > T) at locus chr11:1 902 793. Although these mutations originated from the same sample-specific mutations, they were expressed exclusively at the protein level. Next-generation sequencing analyses bulk of cells simultaneously, thus the actual combinations of mutations are not distinguishable at the genomic level. Certain conventional proteogenomic analyses could have missed identifying these two exclusively mutated peptides, but our approach could successfully resolve mutational ambiguities at the protein level by considering mutations combinatorially during the second stage search.

3.3 Aberrantly translated peptides in non-coding regions

Proteogenomic approach can be useful in identifying peptides deduced not only from mutations but also from aberrant expression of non-coding RNAs and pseudogenes (Kim *et al.*, 2014; Nesvizhskii, 2014). Protein sequence database built from three frame translation of genes of interest such as non-coding RNAs and/or pseudogenes is used to identify aberrantly expressed peptides from MS/MS spectra. Such an approach could be useful to correct gene annotation or, perhaps, analyze disease-specific patterns (Stewart *et al.*, 2019). We added three frame translation mode for non-coding RNAs and pseudogenes to MutCombinator so that a user can easily identify aberrantly expressed peptides with/without combinatorial mutations.

We applied third stage search to identify aberrantly expressed peptides, after the two-stage search in the coding regions. We searched 2 773 632 unidentified spectra against 12 516 expressed non-coding/pseudogene transcripts, considering sample-specific and COSMIC mutations. We estimated at 1% local FDR and identified 122 PSMs. We removed 86 PSMs, peptide sequences of which exactly match UniProt (release 2019-11) or common contaminant sequences. We obtained genomic loci of 27 peptides corresponding to the remaining 36 PSMs by applying ACTG tool (Choi *et al.*, 2017). Thirteen mutated peptides could be matched to coding regions in Ensembl v71 so we further discarded the mutated PSMs in the identifications of aberrantly translated peptides. The summary of non-coding search result is described in Figure 5.

4 Discussion

Proteogenomics has improved understandings of biology, via integration of genomics and proteomics. The baseline results of the integration depend on identifications of expressed and mutated peptides; however, there is no practically available software tool to identify mutated peptides considering all possible combinations of mutations in coding regions. We designed MutCombinator so that it can be applied to identify mutated peptides allowing combinatorial mutations using a reasonable amount of computational resources. A total of 2 792 596 spectra were processed using 75 GB of memory and 16 threads, taking 42 h on workstation computers.

We demonstrated the usefulness of MutCombinator in two aspects: (i) identifications of mutated peptides with combinatorial mutations and (ii) incorporation of large-scale mutation database such as COSMIC. By considering combinations of mutations, MutCombinator facilitates identifying mutated peptides regardless of where the mutations really come from. In other words, we now can decode the combinations of mutations even when mutations from different sources are aggregated in a single database.

Funding

This work was supported by the National Research Foundation of Korea [NRF-2017M3C9A5031597, NRF-2017R1E1A1A01077412 and NRF-2019M3E5D3073568]; and the BK21 plus program through the National Research Foundation funded by the Ministry of Education of Korea.

Conflict of Interest: none declared.

References

Choi,S. *et al.* (2017) ACTG: novel peptide mapping onto gene models. *Bioinformatics*, **33**, 1218–1220.

Coussens,L.M. and Werb,Z. (2002) Inflammation and cancer. *Nature*, **420**, 860–867.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Kim,M.S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.

Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.

Kong,A.T. *et al.* (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods*, **14**, 513–520.

Landrum,M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.

Mann,M. and Wilm,M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, **66**, 4390–4399.

Madar,I.H. *et al.* (2018) Comprehensive and sensitive proteogenomics data analysis strategy based on complementary multi-stage database search. *Int. J. Mass Spectrom.*, **427**, 11–19.

Mertins,P. *et al.*, NCI CPTAC. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, **534**, 55–62.

Mun,D.G. *et al.* (2019) Proteogenomic characterization of human early-onset gastric cancer. *Cancer Cell*, **35**, 111–124 e110.

Na,S. *et al.* (2012) Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteomics*, **11**, M111.010199.

Nesvizhskii,A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*, **11**, 1114–1125.

Park,H. *et al.* (2014) Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics*, **14**, 2742–2749.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Shin,B. *et al.* (2008) Post-experiment monoisotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometric data increases accuracy of peptide identification in LC/MS/MS. *Mol. Cell. Proteomics*, **7**, 1124–1134.

Stewart,G.L. *et al.* (2019) Aberrant expression of pseudogene-derived lncRNAs as an alternative mechanism of cancer gene regulation in lung adenocarcinoma. *Front. Genet.*, **10**, 138.

Tabb,D.L. *et al.* (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.*, **75**, 6415–6421.

Tate,J.G. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.

Wang,X. and Zhang,B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics*, **29**, 3235–3237.

Woo,S. *et al.* (2014a) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.*, **13**, 21–28.

Woo,S. *et al.* (2014b) Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, **14**, 2719–2730.

Zhang,B. *et al.*, the NCI CPTAC. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.

Zhang,M. *et al.* (2017) CanProVar 2.0: an updated database of human cancer proteome variation. *J. Proteome Res.*, **16**, 421–432.