



(51) International Patent Classification:

G16B 20/20 (2019.01) G16B 30/10 (2019.01)
G16B 20/30 (2019.01)

(21) International Application Number:

PCT/EP2024/066271

(22) International Filing Date:

12 June 2024 (12.06.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

2308965.9 15 June 2023 (15.06.2023) GB

(71) Applicant: **ACHILLES THERAPEUTICS UK LIMITED** [GB/GB]; 245 Hammersmith Road, London Greater London W6 8PW (GB).

(72) Inventors: **CHAN, Fong Chun**; c/o Achilles Therapeutics UK Limited, 245 Hammersmith Road, London Greater London W6 8PW (GB). **ROSENTHAL, Rachel**; c/o Achilles Therapeutics UK Limited, 245 Hammersmith Road, London Greater London W6 8PW (GB). **SALM-HORSTMAR, Maximilian Prinz zu**; c/o Achilles Ther-

apeutics UK Limited, 245 Hammersmith Road, London Greater London W6 8PW (GB).

(74) Agent: **MEWBURN ELLIS LLP**; Aurora Building, Counterslip, Bristol BS1 6BX (GB).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE,

(54) Title: IDENTIFICATION OF NEOANTIGEN PEPTIDES

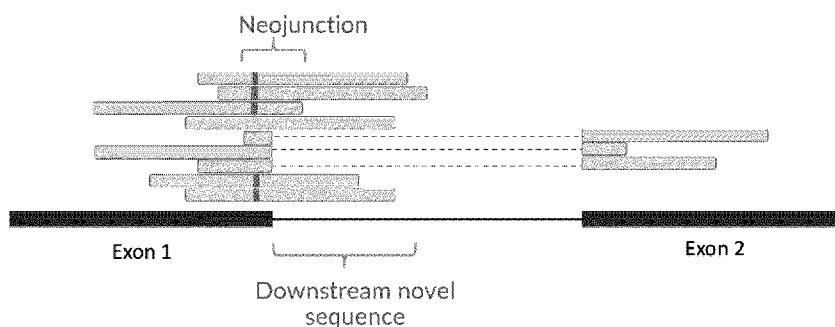


Figure 4

(57) Abstract: Described herein are methods identifying a neoantigen peptide associated with a tumour-specific mutation. The method comprise selecting one or more RNA sequence reads from RNA sequence data from one or more samples, that contain the tumour-specific mutation; assembling a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads; extracting an anchor sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions; identifying a reading frame as the reading frame of a reference transcript: (i) that overlaps the genomic position of the tumour-specific mutation, and (ii) that the anchor sequence maps to or that has a corresponding protein that a translation of the anchor sequence maps. The neoantigen peptide is a translation of at least a part the assembled sequence that uses the identified reading frame. Related methods, systems and products are also disclosed.

SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN,
GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *in black and white; the international application as filed
contained color or greyscale and is available for download
from PATENTSCOPE*

IDENTIFICATION OF NEOANTIGEN PEPTIDES

FIELD OF THE DISCLOSURE

The present disclosure relates to methods for identifying a neoantigen peptide from RNA sequence data. The present disclosure also relates to methods and compositions for the treatment of cancer which make use of or target neoantigens.

BACKGROUND

Cancer neoantigens are antigens that result from the presence of cancer specific variants (also referred to as “tumour-specific mutations”) and are therefore not present on normal cells.

These represent promising therapeutic targets for immunotherapy, and multiple neoantigen vaccine trials are currently underway. The process of designing a personalised neoantigen based immunotherapy typically starts with DNA and/or RNA sequencing of a patient to identify tumour-specific mutations present in the patient. It is then typically necessary to identify the peptides that result from the tumour-specific mutations identified. Indeed, the peptides may be part of the therapy or a necessary reagent to obtain or validate the therapy, or they may at least need to be identified to screen candidate neoantigens for immunogenicity, MHC binding and/or presentation.

Identifying the peptides associated with tumour-specific variants is not a trivial task. Indeed, many factors can impact the peptides obtained, such as the haplotype-specific presence of polymorphisms and the effect of the mutations on the splicing and translation process.

Thus, there is a need for improved methods to identify neoantigen peptides from nucleic acid sequence data.

SUMMARY

Two main types of approaches to identify neoantigen peptides from DNA or RNA sequence data can be distinguished: a first type of approach relies on modifying reference transcripts in the region of tumour-specific mutation by introducing identified mutations and phased polymorphisms, then translating the transcripts using their canonical reading frame. These approaches can make use of DNA sequence data only. A second type of approach is referred to herein as “assembly-based approaches” and assembles transcripts directly from RNA sequence data. While the resulting sequence is expected to more accurately capture the transcripts that are present in the cancer cell, they come with the further challenge of having to identify the peptides that are translated from these transcripts. One approach that has been suggested (called “Isovar”, Rubinsteyn, A. Kodysh J., Aksoy BA. *Hammerlab/isovar: Version*

0.7.0 (2017). doi:10.5281/zenodo.821224) assembles overlapping reads that include a variant allele (resulting in “variant sequence”) then translates the variant sequence using a reading frame for the assembled sequence as that of a reference transcript (“reference context”) identified by matching the variant sequence with reference sequences (i.e. known transcripts at the locus of the variant).

The present inventors have identified that both types of approaches are more problematic in the context of variants that result in tumour-specific splice junctions (also referred to herein as “neojunctions”). In the first type of approaches, when using DNA sequence only, it is necessary to predict the splicing effect of mutations in introns or near canonical splice junctions. This is often inaccurate. Neojunctions can be directly identified from RNA sequence data and spiked into reference transcripts, when available. This is what is performed in the package splice2neo (rdr.io/github/TRON-Bioinformatics/splice2neo/). However, neojunctions can result in a change of the reading frame from the canonical reading frame of the transcript, and therefore using the canonical reading frame of the transcript may result in at least a portion of the translated sequence being erroneous. In the second type of approaches, the present inventors have identified that mutations that result in the inclusion of intronic sequence in the transcript will fail to match to any reference transcript. Another solution is therefore needed.

The present inventors have identified that an assembly-based approach that could deal with neojunctions would be particularly advantageous in that it would naturally capture the presence of polymorphisms in phase with a tumour specific variant, the expected effect of splicing on the resulting peptide, and would also build in an expression filter (as variants that are not expressed simply would not be visible in RNA sequence data). The present inventors have further postulated that all of these benefits could be obtained by providing a novel approach where RNA reads compatible with a variant are assembled then matched with a reference transcript using only those positions of the assembled sequence that are upstream of the mutation and that are exonic. The present inventors showed that this was indeed the case and demonstrated on simulated data, cell line data and patient data that the methods described herein enabled a significant increase in the number of neoantigen peptides that could be identified. This is particularly significant as neojunctions can result in proteins that are very different from their wild-type counterparts and therefore have high immunogenic potential. Thus, the method leads to an improved likelihood of identifying neoantigens that are likely to be successful targets for immunotherapy.

Thus, according to a first aspect, there is provided a method of identifying a neoantigen peptide associated with a tumour-specific mutation, the method comprising: obtaining RNA sequence data from one or more samples comprising tumour genetic material; selecting one

or more RNA sequence reads from the RNA sequence data that contain the tumour-specific mutation; assembling a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads; identifying positions of the assembled sequence as part of intronic or exonic regions; extracting an anchor sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions; identifying a reading frame as the reading frame of a reference transcript: (i) that overlaps the genomic position of the tumour-specific mutation, and (ii) that the anchor sequence maps to or that has a corresponding protein that a translation of the anchor sequence maps; and identifying the neoantigen peptide as a translation of the assembled sequence or a part thereof that uses the identified reading frame.

The method of the present aspect may have one or more of the following features.

Identifying a neoantigen peptide refers to determining the sequence of the neoantigen peptide. This may also be referred to herein as “assembling” a neoantigen peptide.

The use of an anchor sequence that is 5' of the tumour specific mutation but that does not comprise intronic regions ensures that a matching sequence can be found even in cases where a retained intron is present in the assembled sequence.

The use of a sequence assembled only from RNA reads that contain the variant ensures that all polymorphisms present in the vicinity of the variant on the same haplotype will be included in the assembled sequence. In other words, phasing is inherently built into the method. The anchor sequence or translation thereof may be considered to map to a reference transcript or protein corresponding to a reference transcript when the number of mismatched positions between the anchor sequence or anchor sequence translation and the reference sequence or reference sequence protein is at most a predetermined number. The predetermined number may be 0, 1, 2 or 3. The anchor sequence or translation thereof may be considered to map to a reference transcript or protein corresponding to a reference transcript when the length of the matching sequence between the anchor sequence or anchor sequence translation and the reference sequence or reference sequence protein has at least a predetermined length. The predetermined length may be at least 4 amino acids, at least 5 amino acids or at least 6 amino acids, or at least 12, 15, or 18 bases. The anchor sequence or translation thereof may be considered to map to a reference transcript or protein corresponding to a reference transcript when the sequence of the anchor sequence or anchor sequence translation is a perfect match to the reference sequence or reference sequence protein. Alternatively, the anchor sequence or translation thereof may be considered to map to a reference transcript or protein corresponding to a reference transcript when the sequence of the anchor sequence or anchor

sequence translation is a match to the reference sequence or reference sequence protein allowing up to a predetermined number of mismatches. The predetermined number of mismatches may be 1 or 2 bases, or 1 or 2 amino acids.

5 Identifying a reading frame may comprise: translating at least the anchor sequence using all 3 possible reading frames; and identifying a protein corresponding to a reference transcript that a translated anchor sequence maps to, wherein the identified reading frame is the reading frame of the translated anchor that maps to the protein corresponding to a reference transcript. The use of a translated anchor sequence for matching advantageously means that no further
10 translation is needed once a match has been obtained – identifying the anchor translation that maps to a reference protein combines both the matching and translation steps. This makes for a simple and computationally efficient process. Note that when using protein level mapping of the anchor sequence and reference transcripts, no actual transcript sequences need to be analysed: all analyses can be performed at the level of proteins corresponding to reference
15 transcripts. Identifying a reading frame may comprise translating the entire assembled sequence using all 3 possible reading frames.

Identifying a reading frame may comprise identifying a reference transcript that overlaps the genomic position of the tumour-specific mutation and that the anchor sequence maps to, and
20 identifying the reading frame of the identified reference transcript. The use of the anchor sequence and reference transcripts to identify a reading frame advantageously means that the sequences that are matched are longer than when using protein sequences (since each amino acid corresponds to 3 bases in a transcript). This may enable matching of shorter anchor sequences.

25

Identifying the reading frame of the identified reference transcript may comprise obtaining the reading frame from a database of reference transcripts, or determining the reading frame of the identified reference transcript from a corresponding protein. In other words, a reference transcript may be associated with one or both of: a reading frame and a corresponding protein
30 sequence. A reading frame can be identified from a reference transcript and its corresponding protein sequence. Similarly, a corresponding protein sequence can be identified from a reference transcript and its reading frame. Most reference genome databases provide reference transcripts together with corresponding protein sequences (or identifiers from which such sequences can be obtained).

35

The term “reference transcript” may refer to the cDNA sequence of reference transcripts. cDNA sequences of reference transcripts are typically available from reference genome databases, or can easily be obtained from reference transcript mRNA sequences.

5 Assembling a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads may comprise assembling one or more of the selected RNA sequence reads into the longest sequence comprising the tumour-specific mutation compatible with overlapping sequences of the selected one or more RNA sequence reads. The use of an assembled sequence
 10 corresponding to the longest sequence that can be assembled from RNA sequence reads that contain a variant ensures that all polymorphisms are inherently phased with the variant, and will be reflected in the identified protein, while using the most available information about the transcripts comprising the variant.

15 The tumour-specific mutation may be a splice region variant. The tumour-specific mutation may be a 5' splice variant. 5' splice variants can advantageously be analysed using methods as described herein using any sequencing technology including short read sequencing and polyA capture based RNA sequencing. Without wishing to be bound by theory, this is believed to be because such variants do not require knowledge of the length of any retained intron in
 20 order to identify a correct reading frame. Therefore, limited information obtained from short portions of transcripts that may be focussed on exonic sequences is believed to be sufficient to correctly deal with such variants. The tumour-specific mutation may be selected from: a mutation in a 5' exonic splice region, a mutation in a splice donor site, and a mutation in a 5' intronic splice region. In embodiments, the tumour-specific mutation is a 5' splice variant or a
 25 3' splice variant. A 3' splice variant may be selected from: a mutation in a 3' exonic splice region, a mutation in a splice acceptor site, and a mutation in a 3' intronic splice region. In embodiments where the tumour-specific mutation is a 3' splice variant, the RNA sequence data may have been obtained using long read sequencing and/or ribodepleted RNA seq.

30 The RNA sequence data may have been obtained using next generation sequencing. The NGS may be short read sequencing or long read sequencing. The RNA sequence data may have been obtained using polyA capture or ribodepletion RNAseq.

The mutation may be in a gene transcribed from the positive strand and extracting an anchor
 35 sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions may comprise extracting a sequence that is in 5' of the tumour-specific mutation. Alternatively, the mutation

may be in a gene transcribed from the negative strand and extracting an anchor sequence from the assembled sequence may comprise obtaining the reverse complement of the assembled sequence, wherein the anchor sequence is in 5' of the tumour-specific mutation in the reverse complement sequence. Information on whether a gene is transcribed from the negative or the positive strand may be obtained from a genome / transcriptome reference.

Obtaining RNA sequence data from one or more samples comprising tumour genetic material may comprise obtaining RNA sequence data from a plurality of samples comprising tumour genetic material. In such embodiments, selecting one or more RNA sequence reads from the RNA sequence data that contain the tumour-specific mutation may comprise selecting reads across the RNA sequence data from all of the plurality of samples. When data from a plurality of samples is available, the data can be analysed together, combining reads across all regions. This advantageously borrows information across all samples to assemble a sequence, ensuring that the assembled variants are consistent across samples without having to reconcile / merge variant data across samples after protein identification.

The method may comprise determining whether the tumour-specific mutation has support in each of the plurality of samples. The method may comprise determining whether the tumour-specific mutation is ubiquitous in the tumour from which the plurality of samples have been obtained. A tumour-specific mutation that has support in each of the plurality of samples may be considered to be ubiquitous in the tumour.

The one or more samples may be samples from the same subject and/or the same tumour. The one or more samples may be tumour samples. For example, one or more tumour samples may have been obtained from one or more biopsies and/or from a surgically resected tumour. The wording "same tumour" encompasses the same tumour mass as well as a primary tumour and one or more metastases, and/or samples from the same tumour obtained at different times (e.g. samples may be selected from: a diagnostic biopsy, a surgical resection, a metastasis, a relapse sample, etc.)

The method may further comprise identifying the tumour-specific mutation using the RNA sequence data or DNA sequence data from one or more samples comprising tumour genetic material from the same tumour. The method may further comprise obtaining DNA sequence data from one or more samples comprising genetic material from the tumour and identifying the tumour-specific mutation by comparing the DNA sequence data from the one or more samples comprising genetic material from the tumour and a reference genome sequence and/or DNA sequence data from one or more germline samples.

The sequence data may comprise a plurality of sequencing reads. The sequence data may comprise a plurality of aligned sequencing reads, for example in the form of a SAM or BAM file. The sequence data may comprise a plurality of sequencing reads each associated with
5 coordinates in a reference genome or transcriptome to which the sequencing reads align.

The step of obtaining sequence data from one or more samples from the subject may comprise or consist of receiving sequence data from a user (for example through a user interface), from one or more computing device(s), or from one or more data stores or databases. The step of
10 obtaining sequence data (whether genomic sequence data or RNA sequence data or transcriptomic sequence data) may further comprise sequencing (or otherwise determining the sequence composition of genetic material present in a sample) one or more samples from the subject comprising tumour genetic material. The method may further comprise sequencing (or otherwise determining the sequence composition of genomic material present in a sample)
15 one or more germline samples from the subject. The method may further comprise obtaining, from the subject, one or more samples comprising tumour genetic material and optionally one or more germline samples. Genetic material as used herein comprises RNA molecules (e.g. mRNA transcripts), and optionally DNA molecules (e.g. genomic DNA).

Identifying the tumour-specific mutation may comprise identifying genomic coordinates associated with the tumour-specific mutation and the sequence of the mutation. The format of the genomic coordinates may depend on the type of mutation. For example, when the tumour specific mutation is a SNV, the genomic coordinates may be in the form of a single coordinate (e.g. chromosome and location along the chromosome). Then the tumour specific mutation is
20 an indel, the genomic coordinates may be in the form of a pair of coordinates (e.g. start and finish location of the indel) and/or a single coordinate and a length of indel, and optionally the sequence of the mutation.
25

The method may further comprise providing to a user, for example through a user interface,
30 one or more of: the sequence of an identified peptide, information identifying one or more neojunctions in the identified peptide, information characterising one or more tumour-specific mutations present in the identified peptide (such as e.g. genomic location, mutation sequence, ubiquity, likelihood of clonality, etc.), one or more properties of the identified peptide (such as e.g. a predicted binding affinity of the peptide or a part thereof to one or more MHC molecules,
35 a likelihood of the peptide or a part thereof being presented by one or more MHC molecules, a likelihood of the peptide or a part thereof being immunogenic, etc.), and information

identifying one or more amino acids and/or positions in the identified peptide that are not expected to be present in a corresponding normal peptide.

The tumour-specific mutation may be associated with genomic coordinates and selecting one or more RNA sequence reads from the RNA sequence data that contain the tumour-specific mutation may comprise selecting one or more RNA sequence reads that overlap with the genomic coordinates of the tumour-specific mutation and that are compatible with the presence of the tumour-specific mutation.

The tumour-specific mutation may be ubiquitous in the one or more tumour samples. The tumour-specific mutation may be a clonal mutation or a mutation assumed to be clonal in the tumour of the subject. The method may further comprise identifying a plurality of tumour-specific mutations, and for each of the plurality of tumour-specific mutations, determining one or more of: whether the tumour-specific mutations is likely to be clonal in the tumour, and whether the tumour-specific mutation is likely to be expressed in the tumour, and selecting a tumour-specific mutation using the result of the determining. For example, the tumour-specific mutation may be one that has been determined to be likely to be clonal in the tumour and/or likely to be expressed in the tumour. Determining whether a tumour-specific mutation is likely to be clonal in a tumour may be performed using a method as described in WO2022/207925. Determining whether a tumour-specific mutation is likely to be expressed in a tumour may be performed using a method as described in GB 2213928.1.

As the skilled person understands, the complexity of the operations described herein (due at least to the complexity of aligning sequencing data to reference sequences, and the amount of data that is typically generated by sequencing genetic material) are such that they are beyond the reach of a mental activity. Thus, unless context indicates otherwise (e.g. where sample preparation or acquisition steps are described), all steps of the methods described herein are computer implemented.

According to a second aspect, there is provided a method of identifying one or more neoantigens in a subject, the method comprising: identifying a plurality of tumour-specific mutations in the subject (such as for example, using genomic and/or transcriptomic data from one or more tumour samples from said subject); identifying a plurality of candidate neoantigen peptides associated with one or more of the plurality of tumour-specific mutations using the method of any embodiment of the first aspect; and identifying a neoantigen associated with at least one of the one or more of the tumour-specific mutations, wherein a neoantigen is a candidate neoantigen peptide that satisfies one or more predetermined criteria selected from:

the candidate neoantigen peptide being likely to bind to an MHC molecule, the candidate neoantigen peptide being likely to be presented by an MHC molecule, the candidate neoantigen peptide being likely to be immunogenic, the candidate peptide neoantigen peptide being likely to interact with a TCR, the candidate neoantigen peptide being likely to form a stable complex with an MHC molecule, and the candidate neoantigen peptide being likely to be expressed in the tumour.

Also described according to the present aspect is a method of identifying one or more neoantigens in a subject, the method comprising: identifying, by a processor using sequence data from one or more samples from said subject, a plurality of tumour-specific mutations in the subject; identifying, by a processor, a plurality of candidate neoantigen peptides associated with one or more of the plurality of tumour-specific mutations using the method of any embodiment of the first aspect; and selecting, by said processor, one or more of the candidate neoantigen peptides as neoantigens, wherein a neoantigen is candidate neoantigen peptide that satisfies at least one or more predetermined criteria on whether peptide is likely to be expressed and/or likely to be immunogenic. For example, the one or more criteria may be selected from: the candidate neoantigen peptide being likely to bind to an MHC molecule, the candidate neoantigen peptide being likely to be presented by an MHC molecule, the candidate neoantigen peptide being likely to be immunogenic, the candidate peptide neoantigen peptide being likely to interact with a TCR, the candidate neoantigen peptide being likely to form a stable complex with an MHC molecule, and the candidate neoantigen peptide being likely to be expressed in the tumour.

The method of the present aspect may have any one or more of the following features.

A neoantigen may be a peptide that is derived from a tumour-specific mutation that satisfies at least a criterion applied to the likelihood that the tumour-specific mutation is expressed in the tumour in view of the RNA sequence data available from the tumour. Such a criterion can be selected from: having a probability of being expressed above a predetermined threshold, having a probability of being expressed that is above a threshold set adaptively to select a predetermined number of tumour-specific mutations with the highest probabilities of being expressed amongst the tumour-specific mutations for which a probability was determined, having a probability of being expressed that is above a threshold set adaptively to select a predetermined top percentile of tumour-specific mutations amongst the tumour-specific mutations for which a probability was determined, having a power to detect a mutation as expressed that is above a predetermined threshold and a number of RNA reads showing the tumour-specific mutation above a threshold number associated with the power to detect a mutation as being expressed, and having a power to detect a mutation as expressed that is

below a predetermined threshold and a number of RNA reads showing the tumour-specific mutation below a threshold number associated with the power to detect a mutation as being expressed. Methods for calculating each of these are described in GB2213928.1.

The method may further comprise determining whether one or more of the tumour-specific mutations is likely to be clonal in a tumour of the subject, and identifying whether one or more of the tumour-specific mutations is likely to give rise to a clonal neoantigen. A clonal neoantigen may be a tumour-specific mutation that satisfies at least a criterion selected from: having a probability of being clonal above a predetermined threshold, having a probability of being clonal that is above a threshold set adaptively to select a predetermined number of tumour-specific mutations with the highest probabilities of being clonal amongst the tumour-specific mutations for which a probability was determined, and having a probability of being clonal that is above a threshold set adaptively to select a predetermined top percentile of tumour-specific mutations amongst the tumour-specific mutations for which a probability was determined. Thus, the one or more predetermined criteria on whether the tumour-specific mutation is likely to be clonal may be selected from: the mutation having a likelihood of being clonal above a predetermined threshold, the mutation having a likelihood of being clonal that is above a threshold set adaptively to select a predetermined number of tumour-specific mutations with the highest likelihoods of being clonal amongst the tumour-specific mutations for which a likelihood was determined, and having a likelihood of being clonal that is above a threshold set adaptively to select a predetermined top percentile of tumour-specific mutations amongst the tumour-specific mutations for which a likelihood was determined. Methods for calculating each of these are described in WO2022/207925.

According to a third aspect, there is provided a method of determining whether a tumour-specific mutation is likely to give rise to a neoantigen, the method comprising: identifying one or more neoantigen peptides associated with the tumour-specific mutation using the method of any embodiment of the first aspect; and determining whether the tumour-specific mutation satisfies one or more predetermined criteria applying to the identified neoantigen peptide(s), selected from: at least one of the one or more neoantigen peptides being likely to bind to an MHC molecule, at least one of the one or more neoantigen peptides being likely to be presented by an MHC molecule, at least one of the one or more neoantigen peptides being likely to be immunogenic, at least one of the one or more neoantigen peptides being likely to interact with a TCR, and at least one of the one or more neoantigen peptides being likely to form a stable complex with an MHC molecule.

The method of the present aspect may have any one or more of the features of any preceding aspect.

According to a fourth aspect, there is provided a method of providing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising: identifying one or more tumour-specific mutations in the subject; identifying one or more neoantigen peptides associated with one or more tumour-specific mutations using the method of any embodiment of the first aspect; optionally selecting one or more of the neoantigen peptides identified using one or more criteria that apply to the neoantigen peptides; and designing an immunotherapy that targets the one or more neoantigens associated with the, optionally selected, one or more neoantigen peptides.

Selecting one or more of the neoantigen peptides may comprise selecting one or more neoantigen peptides that satisfy one or more predetermined criteria selected from: whether the neoantigen peptide is likely to bind to an MHC molecule, whether the neoantigen peptide is likely to be presented by an MHC molecule, whether the neoantigen peptide is likely to be immunogenic, whether the neoantigen peptide is likely to interact with a TCR, whether the neoantigen peptide is likely to form a stable complex with an MHC molecule, and whether the neoantigen peptide is likely to be expressed in the tumour.

The method may further comprise manufacturing the immunotherapy.

The identity (i.e. sequence) of the neoantigen peptides may be used to manufacture the peptides (e.g. by chemical synthesis) when this forms part of the manufacture (whether as a step to obtain the product or as part of a release assay) and/or testing and development of the immunotherapy. The identity (i.e. sequence) of the neoantigen peptides may be used to select and/or validate candidate neoantigens to be targeted by the immunotherapy, such as e.g. by predicting any property that is directly associated with the peptide and that is indicative of likely efficacy of the immunotherapy (e.g. immunogenicity, MHC binding, etc.)

The present disclosure also relates to immunotherapies that target one or more neoantigen peptides that have been identified using a method as described herein, and to methods for designing and/or providing such immunotherapies.

According to any aspect described herein, an immunotherapy may be an immunogenic composition, a composition comprising immune cells or a therapeutic antibody. The immunogenic composition may comprise one or more neoantigens identified (such as e.g. a neoantigen peptide or protein or a cell displaying the neoantigen), or material sufficient for expression of the one or more neoantigens identified (e.g. a DNA or RNA molecule which encodes the neoantigen(s)). The composition comprising immune cells may comprise T cells, B cells and/or dendritic cells. The composition comprising a therapeutic antibody may comprise one or more antibodies that recognise at least one of the one or more of the neoantigens identified. An antibody may be a monoclonal antibody. For example, the

immunotherapy may be an immunogenic composition comprising the neoantigen peptides or peptide sequences derived therefrom (such as e.g. portions of the neoantigen peptide), or one or more nucleic acids encoding the neoantigen peptides or peptide sequences derived therefrom. The immunotherapy may be an antibody binding to the neoantigen peptides. The immunotherapy may be a cell therapy, such as e.g. T cells that are selected for reactivity to the one or more neoantigens (such as e.g. by selective expansion), or engineered T cells comprising TCR targeting a neoantigen peptide.

In any embodiment of any aspect, the cancer may be selected from bladder cancer, gastric cancer, oesophageal cancer, breast cancer, colorectal cancer, cervical cancer, ovarian cancer, endometrial cancer, kidney cancer (renal cell), lung cancer (small cell, non-small cell and mesothelioma), brain cancer (gliomas, astrocytomas, glioblastomas), melanoma, lymphoma, small bowel cancers (duodenal and jejunal), leukemia, pancreatic cancer, hepatobiliary tumours, germ cell cancers, prostate cancer, head and neck cancers, thyroid cancer and sarcomas. The cancer may be lung cancer. The cancer may be melanoma. The cancer may be bladder cancer. The cancer may be head and neck cancer. In any embodiment of any aspect, the subject may be human. Designing an immunotherapy that targets one or more neoantigens identified may comprise designing one or more candidate peptides for each of the one or more neoantigens targeted, each peptide comprising at least a portion of a neoantigen targeted. Designing or providing an immunotherapy may comprise obtaining the one or more candidate peptides. The method may further comprise testing the one or more candidate peptides for one or more properties. Testing may be performed *in vitro* or *in silico*. For example, the one or more peptides may be tested for immunogenicity, propensity to be displayed by MHC molecules (optionally by specific MHC molecule alleles, where the alleles may have been chosen depending on the MHC alleles expressed by the subject), ability to elicit proliferation of a population of immune cells, etc. The method may further comprise producing the immunotherapy. The method may further comprise obtaining a population of dendritic cells that has been pulsed with one or more of the candidate peptides. The immunotherapy may be a composition comprising T cells that recognise at least one of the one or more of the neoantigens identified. The composition may be enriched for T cells that target at least one of the one or more of the neoantigens identified. The method may comprise obtaining a population of T cells and expanding the population of T cells to increase the number or relative proportion of T cells that target at least one of the one or more of the neoantigens identified. The method may further comprise obtaining a T cell population. A T cell population may be isolated from the subject, for example from one or more tumour samples obtained from the subject, or from a peripheral blood sample or a sample from other tissues of the subject. The T cell population may comprise tumour infiltrating lymphocytes. T

cells may be isolated using methods which are well known in the art. For example, T cells may be purified from single cell suspensions generated from samples on the basis of expression of CD3, CD4 or CD8. T cells may be enriched from samples by passage through a Ficoll-opaque gradient. The method may further comprise expanding the T cell population. For example, T cells may be expanded by ex vivo culture in conditions which are known to provide mitogenic stimuli for T cells. By way of example, the T cells may be cultured with cytokines such as IL-2 or with mitogenic antibodies such as anti-CD3 and/or CD28. The T cells may be co-cultured with antigen-presenting cells (APCs), which may have been irradiated. The APCs may be dendritic cells or B cells. The dendritic cells may have been pulsed with peptides containing one or more of the identified neoantigens as single stimulants or as pools of stimulating neoantigen peptides. Expansion of T cells may be performed using methods which are known in the art, including for example the use of artificial antigen presenting cells (aAPCs), which provide additional co-stimulatory signals, and autologous PBMCs which present appropriate peptides. Autologous PBMCs may be pulsed with peptides containing neoantigens as discussed herein as single stimulants, or alternatively as pools of stimulating neoantigens.

According to a further aspect, there is provided a method for expanding a T cell population for use in the treatment of cancer in a subject, the method comprising: identifying one or more neoantigen peptides using a method as described herein; obtaining a T cell population comprising a T cell which is capable of specifically recognising one of the identified neoantigens; and co-culturing the T cell population with a composition comprising the identified neoantigens. The method may have one or more of the following features. The T cell population obtained may be assumed to comprise a T cell capable of specifically recognising one of the identified neoantigens. The method preferably comprises identifying a plurality of neoantigens. The neoantigens may be clonal neoantigens. The T cell population may comprise a plurality of T cells each of which is capable of specifically recognising one of the plurality of identified neoantigens, and co-culturing the T cell population with a composition comprising the plurality of identified neoantigens. The co-culture may result in expansion of the T cell population that specifically recognises the one or more neoantigens. The expansion may be performed by co-culture of a T cell with a neoantigen and an antigen presenting cell. The antigen presenting cell may be a dendritic cell. Thus, the expansion may be a selective expansion of T cells which are specific for the neoantigen. The expansion may further comprise one or more non-selective expansion steps.

According to a further aspect, there is provided a composition comprising a population of T cells obtained or obtainable by a method according to any embodiment of the preceding aspect. According to a further aspect, there is provided a composition comprising a

neoantigen, neoantigen specific immune cell, or an antibody that recognises a neoantigen, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen has been identified as a neoantigen (e.g. identified as being derived from a tumour-specific mutation that is expressed in a tumour of the subject), using the methods described herein. According to a further aspect, there is provided a composition comprising a neoantigen, neoantigen specific immune cell, or an antibody that recognises a neoantigen, wherein said neoantigen has been identified using the methods described herein. According to a further aspect, there is provided a cell or population of cells expressing a neoantigen on its surface, wherein said neoantigen has been identified using the methods described herein. According to a further aspect, there is provided a neoantigen, immune cell which recognises a neoantigen, or antibody which recognises a neoantigen, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen has been identified using the methods described herein. According to a further aspect, there is provided a use of a neoantigen, immune cell which recognises a neoantigen, or antibody which recognises a neoantigen, in the manufacture of a medicament for use in the treatment or prevention of cancer in a subject, wherein said neoantigen has been identified using the methods described herein. According to a further aspect, there is provided a method of treating a subject that has been diagnosed as having cancer, the method comprising administering an immunotherapy that has been provided using the methods described herein, or a composition as described herein.

According to a further aspect, there is provided a method of treating a subject that has been diagnosed as having cancer, the method comprising: identifying one or more neoantigens by: identifying a plurality of tumour-specific mutations in the subject; identifying one or more neoantigen peptides associated with one or more tumour-specific mutations using the method of any embodiment of the first aspect; optionally selecting one or more of the neoantigen peptides using one or more predetermined criteria that apply to the neoantigen peptides; and treating the subject with an immunotherapy that targets a neoantigen associated with one or more of the optionally selected neoantigen peptides.

According to a further aspect, there is provided one or more non-transitory computer readable media comprising instructions that, when executed by one or more processors, cause the one or more processors to perform the steps of any method described herein, such as a method according to any embodiment of the first, second, third or fourth aspects above.

According to a further aspect, there is provided a computer program comprising code which, when the code is executed on a computer, causes the computer to perform the steps of any method described herein, such as a method according to any embodiment of the first, second, third or fourth aspects above.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a flowchart illustrating schematically a method of identifying a neoantigen peptide.

Figure 2 is a flowchart illustrating schematically a method of providing an immunotherapy (A. steps up to and including designing an immunotherapy, B. steps up to and including administering the immunotherapy to the subject).

Figure 3 shows an embodiment of a system for identifying neoantigen peptides and/or for providing an immunotherapy.

Figure 4 illustrates schematically the concept of neojunctions and splice region associated mutations.

Figure 5 illustrates schematically different types of abnormal splicing that can be caused by mutations.

Figure 6 illustrates a problem with obtaining peptide sequences from sequences associated with neojunctions. **A.** Simple sequence with no neojunction. **B.** Sequence with a neojunction.

Figure 7 illustrates different classes of neojunction variants.

Figure 8 shows examples of application of methods according to embodiments of the disclosure to splice donor mutations on the + strand (**A, B**), 5' intronic splice region mutations on the + strand (**C, D**), 5' exonic splice region variants on the + strand (**E, F**), and splice donor mutations on the – strand (**G, H**). “*” in amino acid sequences indicate “stop” codons.

Figure 9 shows results obtained using methods of the disclosure to identify mutant peptides from 5' splice region variants in a synthetic dataset. Left: number of variants in each of the categories identified on Figure 7. Right: number of variants for which a peptide could be assembled using methods of the disclosure.

Figure 10 shows results obtained using methods of the disclosure to identify mutant peptides from splice variants in cell line data (HCC1395). Example of a splice donor variant (chr12:21807476_C>A).

Figure 11 shows results obtained using methods of the disclosure to identify mutant peptides from splice variants in cell line data (HCC1395). Example of a 5' Exonic Splice Region Variant (chr4:119661818_C>T).

Figure 12 illustrates a method to identify neoantigen peptides from sequence data from a plurality of samples (labelled R1, R2). The method considers the union of reads from all samples over a locus when assembling a variant sequence. In the illustrated example, the method also preserves the information about which sample each read came from, which

enables it to determine whether there are reads supporting the mutation in each sample. This can in turn be used to determine whether the mutation is ubiquitous (present in all of the plurality of samples).

Figure 13 shows results obtained using methods of the disclosure to identify mutant peptides from tumour samples. **A.** In the top plot, each bar of the barchart represents a different cancer patient, and the height of the bar indicates the number of splice region associated variants for which peptides could be identified for each patient in each category of variant (in each bar, from top to bottom: exonic variant, splice donor variant, intronic variant). The bottom plot shows the information filtered for variants that are ubiquitously expressed. **B.** Distribution of variant allele fractions (VAF) of splice variants identified across a cohort of patients, by class of splice variant. The diagram below the boxplots illustrates the hypothesis that variants in non-coding regions that are expected due to disrupted splicing should have high VAFs irrespective of tumour purity because transcripts from the normal gene (which do not contain the variant) should not produce reads including the intronic sequence. **C.** Example of a Ubiquitously Expressed Splice Donor Variant identified in a patient (chr5:140041594_T>C).

DETAILED DESCRIPTION

While it is generally accepted that cancer antigens such as those resulting from cancer specific variants (also referred to as “tumour-specific mutations”) represent promising therapeutic targets provided that they are expressed by cancer cells, identifying the peptides that are produced as a result of the presence of these mutations (and which are therefore potential neoantigens) remains difficult.

In the present disclosure, the following terms will be employed, and are intended to be defined as indicated below.

A “sample” as used herein may be a cell or tissue sample, a biological fluid, an extract (e.g. a DNA extract obtained from the subject), from which genomic and/or transcriptomic material can be obtained for genomic and/or transcriptomic analysis, such as genomic sequencing (e.g. whole genome sequencing, whole exome sequencing) or RNA sequencing (also referred to as “RNAseq” or “RNA-seq”). The sample may be a cell, tissue or biological fluid sample obtained from a subject (e.g. a biopsy). Such samples may be referred to as “subject samples”.

In particular, the sample may be a blood sample, or a tumour sample, or a sample derived therefrom. For the purposes of obtaining RNA sequence data, a “sample” as used herein may be a cell or tissue sample, or an extract (e.g. a RNA extract obtained from a subject) from which transcriptomic material can be obtained. For the purposes of obtaining DNA/genomic sequence data, a “sample” as used herein may be a cell or tissue sample, a biological fluid, an extract (e.g. a DNA extract obtained from the subject), from which genomic material can

be obtained for genomic analysis, such as genomic sequencing (e.g. whole genome sequencing, whole exome sequencing). The sample may be one which has been freshly obtained from a subject or may be one which has been processed and/or stored prior to genomic/transcriptomic analysis (e.g. frozen, fixed or subjected to one or more purification, enrichment or extraction steps). The sample may be a cell or tissue culture sample. As such, a sample as described herein may refer to any type of sample comprising cells or genomic and/or transcriptomic material derived therefrom, whether from a biological sample obtained from a subject, or from a sample obtained from e.g. a cell line. In embodiments, the sample is a sample obtained from a subject, such as a human subject. The sample is preferably from a mammalian subject (such as e.g. a mammalian cell sample or a sample from a mammalian subject, such as a cat, dog, horse, donkey, sheep, pig, goat, cow, mouse, rat, rabbit or guinea pig), preferably from a human (such as e.g. a human cell sample or a sample from a human subject). Further, the sample may be transported and/or stored, and collection may take place at a location remote from the sequence data acquisition (e.g. sequencing) location, and/or any computer-implemented method steps described herein may take place at a location remote from the sample collection location and/or remote from the sequence data acquisition (e.g. sequencing) location (e.g. the computer-implemented method steps may be performed by means of a networked computer, such as by means of a “cloud” provider).

The samples used in methods of the present disclosure are typically samples comprising tumour cells (e.g. a tumour sample or sample comprising circulating tumour cells) or genetic material derived from tumour cells (such as e.g. cell free DNA or cell DNA and/or RNA extracted from a sample comprising cells). Such a sample may be a “mixed sample”. A “mixed sample” refers to a sample that is assumed to comprise multiple cell types or genetic material derived from multiple cell types. Within the context of the present disclosure, a mixed sample is typically one that comprises tumour cells or is assumed (expected) to comprise tumour cells, or genetic material derived from tumour cells, and normal cells or genetic material derived from normal cells. Genetic material can comprise genomic material (e.g. DNA) or transcriptomic material (e.g. RNA). Samples obtained from subjects, such as e.g. tumour samples, are typically mixed samples (unless they are subject to one or more purification and/or separation steps). Typically, the sample comprises tumour cells and at least one non-tumour cell type (and/or genetic material derived therefrom). A “tumour sample” refers to a sample derived from or obtained from a tumour. Such samples may comprise tumour cells and normal (non-tumour) cells. The normal cells may comprise immune cells (such as e.g. lymphocytes), and/or other normal (non-tumour) cells (e.g. stromal cells). The lymphocytes in such mixed samples may be referred to as “tumour-infiltrating lymphocytes” (TIL). A tumour may be a solid tumour or a non-solid or haematological tumour. A tumour sample may be a

primary tumour sample, tumour-associated lymph node sample, or a sample from a metastatic site from the subject. A sample comprising tumour cells or genetic material derived from tumour cells may be a bodily fluid sample. Thus, the genetic material derived from tumour cells may be circulating tumour DNA or tumour DNA in exosomes. Instead or in addition to this, the sample may comprise circulating tumour cells. A mixed sample may be a sample of cells, tissue or bodily fluid that has been processed to extract genetic material. Methods for extracting genetic material from biological samples are known in the art. A mixed sample may have been subject to one or more processing steps that may modify the proportion of the multiple cell types or genetic material derived from the multiple cell types in the sample. For example, a mixed sample comprising tumour cells may have been processed to enrich the sample in tumour cells. Thus, a sample of purified tumour cells may be referred to as a “mixed sample” on the basis that small amounts of other types of cells may be present, even if the sample may be assumed, for a particular purpose, to be pure (i.e. to have a tumour fraction of less than 100%).

The term “tumour fraction” (also sometimes referred to as “tumour purity” or simply “purity”, or aberrant cell fraction (ACF)) refers to the proportion of DNA containing cells within a mixed sample that are tumour cells, or to the equivalent proportion that is assumed to result in a particular mixture of genetic material from tumour and non-tumour cells in a sample. Methods for determining the tumour fraction in a sample are known in the art. For example, in the context of cell or tissue samples, a tumour fraction may be estimated by analysing pathology slides (e.g. hematoxylin and eosin (H&E)-stained slides or other histochemistry or immunohistochemistry slides, by counting tumour cells in one or more representative areas of a sample), or using high throughput assays such as flow cytometry. In the context of samples comprising genomic material, a tumour fraction may be estimated using sequence analysis processes that attempt to deconvolute tumour and germline genomes such as e.g. ASCAT (Van Loo et al., 2010), ABSOLUTE (Carter et al., 2012), or ichorCNA (Adalsteinsson et al., 2017).

A “normal sample”, “healthy sample” or “germline sample” refers to a sample that is assumed not to comprise tumour cells or genetic material derived from tumour cells. A germline sample may be a blood sample, a tissue sample, or a purified sample such as a sample of peripheral blood mononuclear cells from a subject. Similarly, the terms “normal”, “germline” or “wild type” when referring to sequences or genotypes refer to the sequence / genotype of cells other than tumour cells. A germline sample may comprise a small proportion of tumour cells or genetic material derived therefrom, and may nevertheless be assumed, for practical purposes, not to comprise said cells or genetic material. In other words, all cells or genetic material may be

assumed to be normal and/or sequence data that is not compatible with the assumption may be ignored.

The term “sequence data” refers to information that is indicative of the presence of genetic material in a sample that has a particular sequence. Such information may be obtained using sequencing technologies, such as e.g. next generation sequencing (NGS), for example whole exome sequencing (WES), whole genome sequencing (WGS), RNA sequencing or sequencing of captured genomic loci (targeted or panel sequencing), or using array technologies, such as e.g. copy number variation arrays, SNP arrays, expression arrays or other molecular counting assays. In the context of the present disclosure, the sequence data is typically obtained by DNA sequencing, and particularly next generation sequencing. Thus, the sequence data may comprise sequencing reads. Information such as the count of sequencing reads that have a particular sequence or that cover a particular genomic location can be derived from sequencing reads using methods known in the art. When non-digital technologies are used such as array technology, the sequence data may comprise a signal (e.g. an intensity value) that is indicative of the number of sequences in the sample that have a particular sequence, for example by comparison to an appropriate control. Sequence data may be mapped to a reference sequence, for example a reference genome or transcriptome, using methods known in the art (such as e.g. Bowtie (Langmead et al., 2009)). This may result in aligned sequencing reads, for example in the form of a SAM or BAM file. Thus, sequencing reads or equivalent non-digital signals may be associated with a particular genomic location (where the “genomic location” refers to a location in the reference genome to which the sequence data was mapped). Further, a genomic location may contain a mutation compared to the reference sequence at the particular genomic location. The process of identifying the presence of a mutation at a particular location in a sample is referred to as “variant calling” and can be performed using methods known in the art (such as e.g. the GATK HaplotypeCaller, gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller).

The terms “tumour-specific mutation”, “tumour-specific variants”, “somatic mutation” or simply “mutation” or “variant” are used interchangeably and refer to a difference in a nucleotide sequence (e.g. DNA or RNA) in a tumour cell compared to a healthy cell from the same subject. The difference in the nucleotide sequence can result in the expression of a protein which is not expressed by a healthy cell from the same subject. For example, a mutation may be a single nucleotide variant (SNV), multiple nucleotide variant (MNV), a deletion mutation, an insertion mutation, a translocation, a missense mutation, a translocation, a fusion, or any other change in the genetic material of a tumour cell. Mutations may be identified by exome sequencing, RNA-sequencing, whole genome sequencing and/or targeted gene panel sequencing and/or routine Sanger sequencing of single genes, followed by sequence

alignment and comparing the DNA and/or RNA sequence from a tumour sample to DNA and/or RNA from a reference sample or reference sequence (e.g. the germline DNA and/or RNA sequence, or a reference sequence from a database). Suitable methods are known in the art. An "indel mutation" refers to an insertion and/or deletion of bases in a nucleotide
5 sequence (e.g. DNA or RNA) of an organism. Typically, the indel mutation occurs in the DNA, preferably the genomic DNA, of an organism. In embodiments, the indel may be from 1 to 150 bases, for example 1 to 90, 1 to 50, 1 to 23 or 1 to 10 bases.

A "splice region associated mutation", "splice variant", or "splice mutation" refers to a mutation that affects the splicing of a transcript comprising the mutation. These mutations result in the
10 creation of at least one new junction between intronic and/or exonic sequences, also referred to as a "neojunction". An example of this is illustrated in **Figure 4**, which shows a gene model comprising two exons (Exon 1 and Exon 2) separated by an intron, as well as RNA sequencing read data mapped to the gene model. The illustrated sequencing data includes reads that map exclusively to Exons 1 and 2 (illustrated by dashed lines over the intronic sequence where the
15 sequencing reads do not map), as expected for a sequence that is not mutated at the gene locus. The illustrated sequencing data also includes reads that map to Exon 1 and at least part of the following intron, and that include a mutation in the 5' exon near the exon-intron boundary. These reads provide evidence that the mutation affected the splicing of the intron, resulting in novel sequence being included in the mature transcript downstream of the exon.
20 The boundary between the exon sequence (in the illustrated case this is the normal, complete exon sequence) and the retained intronic sequence is a new boundary (or "neojunction") that is not present in normal mature transcripts (where only a boundary between the sequence of exon 1 and the sequence of exon 2 exists). **Figure 4** shows an example of a splice region associated mutation that results in at least partial intron retention. Note that based solely on
25 the read evidence shown it is not possible to determine whether the mutation resulted in partial or full intron retention because there is no read that spans the mutation as well as the next intron-exon junction. Reads that map further down the intron but that do not include the mutation cannot be unambiguously resolved as originating from transcripts that include the mutation.

Any type of mutation above may affect the splicing of a transcript if it is located in a region that is involved in the splicing of the transcript. Regions involved in the splicing of transcripts can include intronic regions, including specifically the splice donor site and splice acceptor site, or in exonic regions, typically near an intron / exon boundary. Such variants are referred to herein as "splice region variants" or "splice region mutations". A splice region variant may be any
35 variant that occurs within a predetermined distance of a splice site, where a splice site may be defined as the location of a splice acceptor site or splice donor site. A predetermined

distance may be set symmetrically for the intron side and the exon side of a splice junction, or may be set differently for the exon side and the intron side. For example, a splice region variant may include any variant that occurs within a predetermined distance of 3 bases of a splice site, in an exon. In other words, a splice region variant may include any variants that affects any one or more of the last 3 bases at the 3' end of an exon, or any one or more of the first 3 bases at the 5' end of an exon. Similarly, a splice region variant may include any variant that occurs within a predetermined distance of 6 bases of a splice site, in an intron. In other words, a splice region variant may include any variants that affects any one or more of the last 8 bases at the 3' end of an intron (i.e. any of: the bases of the splice acceptor site itself -which are the last 2 bases of the intron- and the 6 bases that precede it), or any one or more of the first 8 bases at the 5' end of an intron (i.e. any of: the bases of the splice donor site itself - which are the first two bases of the intron- and the 6 bases that follow it). A splice donor site may be defined as the 2 base region at the 5' end of an intron. A splice acceptor site may be defined as the 2 base region at the 3' end of an intron. A splice donor variant or mutation may be defined as any mutation that occurs within the first or second base at the 5' end of an intron. A splice acceptor variant or mutation may be defined as any mutation that occurs within the last or penultimate base at the 3' end of an intron. These particular definitions are commonly accepted definitions of variant effects as defined by the Sequence Ontology (See www.sequenceontology.org; used by the Ensembl databases - see www.ensembl.org/info/genome/variation/prediction/predicted_data.html). Thus, a splice region associated mutation may be located within or encompass part of an intron, a splice donor site, a splice acceptor site, or part of an exon. A splice donor site is a conserved GU sequence that is normally located at the 5' end of introns. A splice acceptor site is a conserved AG sequence that is normally located at the 3' end of introns. These sites are recognised by the splicing machinery, leading to the intronic sequence between the splice donor site and the splice acceptor site (included) being excised. Mutations can cause the loss of donor or acceptor sites, or the creation of new donor or acceptor sites. These may be referred to as "cryptic splice sites".

Figure 5 illustrates schematically different types of abnormal splicing that can be caused by splice region associated mutations. A pre-mRNA is shown on the top row, including a first exon (Exon 1, 5' exon, upstream exon) an intron and a second exon (Exon 2, 3' exon, downstream exon). This matches a gene model for the gene. The second row shows the corresponding mature mRNA after normal splicing, where the intron is excised, resulting in a junction between Exon 1 and Exon 2. The third row shows a mature mRNA in which a mutation at the 3'end of the 5' exon (i.e. near the splice junction in the 5' exon) has caused splicing of the intron to fail, leading to the intron being retained in the mature mRNA. This results in 2

neojunctions, one between Exon 1 and the retained intron, and one between the retained intron and Exon 2. The fourth row shows a mature mRNA in which a mutation in Exon 2 has caused it to be “skipped”, leading to splicing of the entire sequence between Exon 1 and an exon downstream of Exon 2. This leads to one neojunction between Exon 1 and Exon 3. In the illustrated embodiment the entire Exon 2 is skipped but it is also possible for partial exon skipping to occur. The fifth row shows a mature mRNA in which a mutation in Exon 1 has caused the splicing at the start of the intro to fail, but a cryptic splice donor site in the intron was used instead, leading to retention of part of the intron and creation of two neojunctions (one between Exon 1 and the intron and one between the intronic cryptic site and Exon 2). The sixth row illustrates a mature mRNA in which a mutation in Exon 1 has caused the appearance of a new donor site (Cryptic site) in Exon 1, leading to the sequence of Exon 1 following this site being excised from the mature mRNA and creating a neojunction between the cryptic site in Exon 1 and the start of Exon 2. The last row represents the fact that combinations of these types of abnormal splicing can also occur. For example, a part of an intron can be retained followed by full or partial skipping of an exon. The methods described herein can be used to identify neoantigen peptides associated with any of these types of abnormal splicing.

A “neoantigen” (or “neo-antigen”) is an antigen that arises as a consequence of a mutation within a cancer cell. Thus, a neoantigen is not expressed by normal (i.e. non-tumour) cells. A neoantigen may be processed to generate distinct peptides which can be recognised by T cells when presented in the context of MHC molecules. As described herein, neoantigens may be used as the basis for cancer immunotherapies. References herein to “neoantigens” are intended to include peptides derived from mutated proteins, regardless of their size, provided that they arise as a consequence of a mutation and are immunogenic. Depending on the context, the term may also encompass peptides derived from mutated proteins that are believed to be likely to be immunogenic by virtue of being absent in normal cells. The term “neoantigen” as used herein is intended to encompass any part of a neoantigen that is likely to be immunogenic (whether or not true immunogenicity has been verified). An “antigenic” molecule as referred to herein is a molecule which itself, or a part thereof, is capable of stimulating an immune response, when presented to the immune system or immune cells in an appropriate manner. Immunogenicity is subject-specific and depends on multiple factors. Immunogenicity is believed to require presentation of the neoantigen to an MHC molecule present in the subject, which itself requires binding of the neoantigen to the MHC molecule. The binding of a neoantigen to a particular MHC molecule (encoded by a particular HLA allele) and/or the presentation of the neoantigen by the MHC molecule may be predicted using methods which are known in the art. Examples of methods for predicting MHC binding and/or

presentation include those described by Lundegaard et al., O'Donnel et al. 2020, and Bullik-Sullivan et al. 2018. For example, MHC binding of neoantigens may be predicted using the netMHC-3 (Lundegaard et al.) and netMHCpan4 (Jurtz et al.) algorithms. A neoantigen that has been predicted to bind to or be presented by a particular MHC molecule may be considered to be likely to be presented by said MHC molecule on the cell surface. Immunogenicity is further believed to require interaction between the neoantigen with a T cell receptor (TCR) present in the subject, in the context of an MHC molecule present in the subject. The binding of peptides or peptide-MHC complexes to T cell receptors can be predicted using methods which are known in the art, such as e.g. PMTnet (Lu et al. 2021) and Imrex (Moris et al., 2020) and the method described in application GB2303920.9 filed 17 March 2023. Thus, depending on the context, a "neoantigen" may refer to a peptide that has been predicted as likely to be immunogenic using any of the methods described above for predicting MHC binding, MHC presentation and/or TCR interaction, or that has been verified as immunogenic using methods known in the art such as e.g. ELISpot assays. Neoantigens that have not yet been predicted or verified to be immunogenic may by contrast be referred to as "candidate neoantigens".

A "clonal neoantigen" (also sometimes referred to as "truncal neoantigen") is a neoantigen that results from a mutation that is present in essentially every tumour cell in one or more samples from a subject (or that can be assumed to be present in essentially every tumour cell from which the tumour genetic material in the sample(s) is derived). Similarly, a "clonal mutation" (sometimes referred to as "truncal mutation") is a mutation that is present in essentially every tumour cell in one or more samples from a subject (or that can be assumed to be present in essentially every tumour cell from which the tumour genetic material in the sample(s) is derived). Thus, a clonal mutation may be a mutation that is present in every tumour cell in one or more samples from a subject. A "sub-clonal" neoantigen is a neoantigen that results from a mutation that is present in a subset or a proportion of cells in one or more tumour samples from a subject (or that can be assumed to be present in a subset of the tumour cells from which the tumour genetic material in the sample(s) is derived). Similarly, a "sub-clonal" mutation is a mutation that is present in a subset or a proportion of cells in one or more tumour samples from a subject (or that can be assumed to be present in a subset of the tumour cells from which the tumour genetic material in the sample(s) is derived). A neoantigen or mutation may be clonal in the context of one or more samples from a subject while not being truly clonal in the context of the entirety of the population of tumour cells that may be present in a subject (e.g. including all regions of a primary tumour and metastasis). Thus, a clonal mutation may be "truly clonal" in the sense that it is a mutation that is present in essentially every tumour cell (i.e. in all tumour cells) in the subject. This is because the one or more

samples may not be representative of each and every subset of cells present in the subject. Thus, within the context of the present disclosure, a “clonal neoantigen” or “clonal mutation” may also be referred to as a “ubiquitous neoantigen” or “ubiquitous mutation”, to indicate that the neoantigen is present in essentially all tumour cells or all tumour samples that have been analysed, but may not be present in all tumour cells that may exist in the subject. The terms “clonal” and “ubiquitous” are used interchangeably unless context indicates that reference to “true clonality” was intended. The wording “essentially every tumour cell” in relation to one or more samples or a subject may refer to at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94% at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% of the tumour cells in the one or more samples or the subject. Nevertheless, a neoantigen/mutation that is identified as likely to be clonal (or “ubiquitous”) may be considered likely to be truly clonal, or at least more likely to be truly clonal than a neoantigen/mutation that is identified as unlikely to be clonal. Further, the confidence in the probability that a clonal neoantigen/mutation identified in a subject is truly clonal increases when the sample(s) used to identify the clonal neoantigen/mutation capture a more complete picture of the genetic diversity of the tumour (e.g. by including a plurality of samples from the subject, such as e.g. samples from different regions of the tumour, and/or by including samples that inherently capture a diversity of tumour cells such as e.g. ctDNA samples). Conversely, a neoantigen/mutation that is identified as unlikely to be clonal is unlikely to be truly clonal, because the identification that the neoantigen/mutation is unlikely to be clonal indicates that even in the restricted view afforded by the sampling process, there is evidence that the neoantigen/mutation is not present in all tumour cells. Thus, the process of identifying clonal neoantigens/mutations may be seen as prioritising which candidate neoantigens/mutations are most likely to be clonal, based on the restricted view of the clonal structure of the subject’s tumour available from the one or more samples.

Methods to identify clonal neoantigens are known in the art and include the methods described in WO 2022/207925, WO 2016/16174085, Landau et al. (2013), Roth et al. (2014), McGranahan et al. (2016). Clonal neoantigens may in particular be identified using a method as described in WO 2022/207925.

The term “cancer cell fraction” (or “CCF”) refers to the proportion of tumour cells that contain a mutation, such as e.g. a mutation that results in a particular neoantigen. A cancer cell fraction may be estimated based on one or more samples, and as such may not be equal to the true cancer cell fraction in the subject (as explained above). Nevertheless, the cancer cell fraction estimated based on one or more samples may provide a useful indication of the likely true cancer cell fraction. Further, as explained above, the accuracy of such an estimate may increase when the sample(s) used to estimate the cancer cell fraction capture a more

complete picture of the genetic diversity of the tumour. Additional sources of noise and confounding factors in genomic data mean that a cancer cell fraction determined from one or more samples represents an estimate. As such, although a truly clonal mutation/neoantigen should have a CCF=1, in practice mutations/neoantigens that are more likely to be clonal are expected to be associated with a higher CCF estimate (which may not be equal to 1) than mutations that are less likely to be clonal, which are expected to be associated with a lower CCF estimate. For example, a cancer cell fraction estimate may be obtained by integrating variant allele frequencies with copy numbers and purity estimates as described by Landau et al. (2013). Such a CCF estimate can also be used to identify mutations that are likely to be clonal. For example, a clonal mutation may be defined as a mutation which has an estimated cancer cell fraction (CCF) ≥ 0.75 , such as a CCF $\geq 0.80, 0.85, 0.90, 0.95$ or 1.0 . A subclonal mutation may be defined as a mutation which has a CCF $< 0.95, 0.90, 0.85, 0.80$, or 0.75 . Further, a CCF estimate may be associated with (e.g. derived from) a distribution associating a probability with each of a plurality of possible values of CCF between 0 and 1, from which statistical estimates of confidence may be obtained. For example, a mutation may be defined as likely to be a clonal mutation if the 95% CCF confidence interval is ≥ 0.75 , i.e. the upper bound of the 95% confidence interval of the estimated CCF is greater than or equal to 0.75. In other words, a mutation may be defined as likely to be a clonal mutation if there is an interval of CCF with lower bound L and upper bound H that is such that $P(L < \text{CCF} < H) = 95\%$ with $H \geq 0.75$. Alternatively, a mutation may be identified as clonal if there is more than a 50% chance or probability that its cancer cell fraction (CCF) reaches or exceeds the required value as defined above, for example 0.75 or 0.95, such as a chance or probability of 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95% or more. In other words, a mutation may be identified as clonal if $P(\text{CCF} > 0.75) \geq 0.5$. For example, mutations may be classified as likely clonal or subclonal based on whether the posterior probability that their CCF exceeds 0.95 (or 0.75, or any other chosen threshold) is greater or lesser than 0.5, respectively. In this context, as will be explained further below, a mutation may be identified as likely to be clonal if $P(\text{CCF}=1)$ exceeds a threshold. The threshold may be fixed. For example, a mutation may be identified as likely to be clonal if $P(\text{CCF}=1) > 0.05$. Alternatively, the threshold may be determined for a particular set of mutations that are investigated. In embodiments, the threshold may be set based on a benchmarking data set with known clonal / non-clonal status, to reach a predetermined precision and/or recall. A benchmarking data set may be obtained using synthetic data and/or using a data set obtained from a population with known clonality structure (for example a cell line mixture data). For example, a mutation may be identified as likely clonal if $P(\text{CCF}=1) > t$ where t is the maximum value that is such that 95% (or any other value such as e.g. 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%) of true clonal mutations in a benchmarking dataset are identified (i.e. a false negative rate of at

most 5%). As another example, a mutation may be identified as likely clonal if $P(\text{CCF}=1) > t$ where t is the minimum value that is such that at least 50% (or any other value such as e.g. 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%) of the mutations that exceed the threshold in a benchmarking dataset are true clonal mutations (i.e. a true positive rate of at least 50%).

Alternatively, the threshold may be set such that any mutation (or a certain % of mutations) that is associated with an estimated CCF that has a confidence interval meeting the criteria described above (e.g. it is such that the upper bound of the 95% confidence interval of the estimated CCF is greater than or equal to 0.75) is selected as likely to be clonal. Alternatively, the threshold may be set such that any mutation (or a certain % of mutations) that is associated with an estimated CCF that has a posterior probability distribution meeting the criteria described above (e.g. a posterior probability that their CCF exceeds 0.95 (or 0.75, or any other chosen threshold) is greater than 0.5) is selected as likely to be clonal.

A cancer immunotherapy (or simply "immunotherapy") refers to a therapeutic approach comprising administration of an immunogenic composition (e.g. a vaccine), a composition comprising immune cells, or an immunoactive drug, such as e.g. a therapeutic antibody, to a subject. The term "immunotherapy" may also refer to the therapeutic compositions themselves. In the context of the present disclosure, the immunotherapy typically targets a neoantigen. For example, an immunogenic composition or vaccine may comprise a neoantigen, neoantigen presenting cell or material necessary for the expression of the neoantigen. As another example, a composition comprising immune cells may comprise T and/or B cells that recognise a neoantigen. The immune cells may be isolated from tumours or other tissues (including but not limited to lymph node, blood or ascites), expanded ex vivo or in vitro and re-administered to a subject (a process referred to as "adoptive cell therapy"). Instead or in addition to this, T cells can be isolated from a subject and engineered to target a neoantigen (e.g. by insertion of a chimeric antigen receptor that binds to the neoantigen) and re-administered to the subject. As another example, a therapeutic antibody may be an antibody which recognises a neoantigen. One skilled in the art will appreciate that if the neoantigen is a cell surface antigen, an antibody as referred to herein will recognise the neoantigen. Where the neoantigen is an intracellular antigen, the antibody will recognise the neoantigen peptide-MHC complex. As referred to herein, an antibody which "recognises" a neoantigen encompasses both of these possibilities. Further, an immunotherapy may target a plurality of neoantigens. For example, an immunogenic composition may comprise a plurality of neoantigens, cells presenting a plurality of neoantigens or the material necessary for the expression of the plurality of neoantigens. As another example, a composition may comprise immune cells that recognise a plurality of neoantigens. Similarly, a composition may comprise a plurality of immune cells that recognise the same neoantigen. As another example, a

composition may comprise a plurality of therapeutic antibodies that recognise a plurality of neoantigens. Similarly, a composition may comprise a plurality of therapeutic antibodies that recognise the same neoantigen.

A composition as described herein may be a pharmaceutical composition which additionally
5 comprises a pharmaceutically acceptable carrier, diluent or excipient. The pharmaceutical composition may optionally comprise one or more further pharmaceutically active polypeptides and/or compounds. Such a formulation may, for example, be in a form suitable for intravenous infusion.

References to "an immune cell" are intended to encompass cells of the immune system, for
10 example T cells, NK cells, NKT cells, B cells and dendritic cells. In a preferred embodiment, the immune cell is a T cell. An immune cell that recognises a neoantigen may be an engineered T cell. A neoantigen specific T cell may express a chimeric antigen receptor (CAR) or a T cell receptor (TCR) which specifically binds a neoantigen, or an affinity-enhanced T cell receptor (TCR) which specifically binds a neoantigen (as discussed further hereinbelow). For
15 example, the T cell may express a chimeric antigen receptor (CAR) or a T cell receptor (TCR) which specifically binds to a neoantigen (for example an affinity enhanced T cell receptor (TCR) which specifically binds to a neo-antigen or a neo-antigen peptide). Alternatively, a population of immune cells that recognise a neoantigen may be a population of T cell isolated from a subject with a tumour. For example, the T cell population may be generated from T
20 cells in a sample isolated from the subject, such as e.g. a tumour sample, a peripheral blood sample or a sample from other tissues of the subject. The T cell population may be generated from a sample from the tumour in which the neoantigen is identified. In other words, the T cell population may be isolated from a sample derived from the tumour of a patient to be treated, where the neoantigen was also identified from a sample from said tumour. The T cell
25 population may comprise tumour infiltrating lymphocytes (TIL).

The term "Antibody" (Ab) includes monoclonal antibodies, polyclonal antibodies, multispecific antibodies (e.g., bispecific antibodies), and antibody fragments that exhibit the desired biological activity. The term "immunoglobulin" (Ig) may be used interchangeably with "antibody". Once a suitable neoantigen has been identified, for example by a method
30 according to the disclosure, methods known in the art can be used to generate an antibody.

An "immunogenic composition" is a composition that is capable of inducing an immune response in a subject. The term is used interchangeably with the term "vaccine". The immunogenic composition or vaccine described herein may lead to generation of an immune response in the subject. An "immune response" which may be generated may be humoral
35 and/or cell-mediated immunity, for example the stimulation of antibody production, or the

stimulation of cytotoxic or killer cells, which may recognise and destroy (or otherwise eliminate) cells expressing antigens corresponding to the antigens in the vaccine on their surface. The immunogenic composition may comprise one or more neoantigens, or the material necessary for the expression of one or more neoantigens. In addition, a neoantigen may be delivered in the form of a cell, such as an antigen presenting cell, for example a dendritic cell. The antigen presenting cell such as a dendritic cell may be pulsed or loaded with the neo-antigen or neo-antigen peptide or genetically modified (via DNA or RNA transfer) to express one, two or more neo-antigens or neoantigen peptides, for example 2, 3, 4, 5, 6, 7, 8, 9 or 10 neo-antigens or neo-antigen peptides. Methods of preparing dendritic cell immunogenic compositions or vaccines are known in the art.

Neoantigen peptides may be synthesised using methods which are known in the art. The term "peptide" is used in the normal sense to mean a series of residues, typically L-amino acids, connected one to the other typically by peptide bonds between the α -amino and carboxyl groups of adjacent amino acids. The term includes modified peptides and synthetic peptide analogues. The neoantigen peptide may comprise the cancer cell specific mutation (e.g. the non-silent amino acid substitution encoded by a single nucleotide variant (SNV)) at any residue position within the peptide. By way of example, a peptide which is capable of binding to an MHC class I molecule is typically 7 to 13 amino acids in length. In embodiments, neoantigen peptides (such as e.g. for use in an antigen-specific expansion step, inclusion in an immunotherapeutic composition, etc.) may be from 7 to 15, such as 8 to 13 amino acids in length. The peptides may be 7, 8, 9, 10, 11, 12, 13, 14 or 15 amino acids long. In embodiments, 15 amino acids long peptides are designed as a set of overlapping sequences with 11 amino acid overlaps (i.e. a "jump" of 4 amino acids from the start of one sequence to the start of the next) each including at least one amino acid that results from the presence of a tumour specific mutation, to provide an overlapping peptide pool. In embodiments, longer peptides, for example 15-31-mers, may be used. In such embodiments, the mutation (i.e. any one or more amino acids resulting from the presence of a tumour specific mutation) may be at any position, for example at the centre of the peptide, e.g. at positions 7, 8, 9, 10, 11, 12, 13, 14, 15 or 16. Such peptides can also be used to stimulate both CD4 and CD8 cells to recognise neoantigens. For examples, longer peptides, such as peptides that are 27, 28, 29, 30 or 31 amino acids long, may be used to stimulate both CD4+ and CD8+ cells.

The mutation may be present at any residue position(s) within the peptide. Thus, the peptide may comprise one or more amino acids that are not present in a corresponding native peptide expressed by a healthy cell at any residue position(s) within the peptide. For example, the mutation, for example an amino acid substitution(s) or any other new sequence resulting from the presence of a tumour specific mutation, may be present (or may start) at any one or more

of positions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or 13 in a peptide comprising 13 amino acids. As another example, a mutation comprising a plurality of amino acids may be present at any one or more of positions 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or 13 in a peptide comprising 13 amino acids. Reference to the mutation being at a particular position within the peptide refers to the residue position of the mutation relative to the amino acid sequence of the corresponding native peptide expressed by a healthy cell. In embodiments, the mutation is at or near the centre of the peptide. For example, the mutation (i.e. any one or more amino acids resulting from the presence of a tumour specific mutation) may be present (or may start) at any one or more of positions 4, 5 or 6 in a peptide comprising 9 amino acids. As another example, the mutation i.e. any one or more amino acids resulting from the presence of a tumour specific mutation) may be present (or may start) at any one or more of positions 6, 7, 8 or 9 in a peptide comprising 13 amino acids. As a further example, the mutation i.e. any one or more amino acids resulting from the presence of a tumour specific mutation) may be present (or may start) at any one or more of positions 12, 13, 14, 15 or 16 in a peptide comprising 29 amino acids. A mutation comprising a plurality of amino acids may start at a predetermined position within the peptide. For example, a predetermined position may be selected from position 1, position 1+j, position 1+2j, position 1+3j, etc. where "j" is a predetermined "jump" value, such as e.g. 3, 4 or 5 amino acids. A mutation may result in a sequence of amino acids that are not present in a corresponding peptide expressed by a healthy cell, the sequence being longer than the length of the peptide. Thus, the peptide may consist of amino acids that result from the mutation. Further, the sequence of amino acids that result from the mutation may not be fully included in the peptide, which may instead comprise a portion of said sequence.

As used herein "treatment" refers to reducing, alleviating or eliminating one or more symptoms of the disease which is being treated, relative to the symptoms prior to treatment. "Prevention" (or prophylaxis) refers to delaying or preventing the onset of the symptoms of the disease. Prevention may be absolute (such that no disease occurs) or may be effective only in some individuals or for a limited amount of time.

As used herein, the terms "computer system" includes the hardware, software and data storage devices for embodying a system or carrying out a method according to the above described embodiments. For example, a computer system may comprise a central processing unit (CPU) and/or a graphic processing unit (GPU), input means, output means and data storage, which may be embodied as one or more connected computing devices. A computer system may comprise a display or comprises a computing device that has a display to provide a visual output display. The data storage may comprise RAM, disk drives or other non-transitory computer readable media. The computer system may include a plurality of

computing devices connected by a network and able to communicate with each other over that network. It is explicitly envisaged that computer system may consist of or comprise a cloud computer.

As used herein, the term “computer readable media” includes, without limitation, any non-transitory medium or media which can be read and accessed directly by a computer or computer system. The media can include, but are not limited to, magnetic storage media such as floppy discs, hard disc storage media and magnetic tape; optical storage media such as optical discs or CD-ROMs; electrical storage media such as memory, including RAM, ROM and flash memory; and hybrids and combinations of the above such as magnetic/optical storage media.

Identification of neoantigen peptides

The present disclosure provides methods for identifying a neoantigen peptide associated with a tumour-specific mutation, using RNA sequence data from one or more samples comprising tumour genetic material. An illustrative method will be described by reference to **Figure 1**.

The method may comprise optional step 10 of obtaining one or more samples comprising tumour genetic material (such as e.g. one or more tumour samples). The sample(s) may be mixed samples comprising genomic material from multiple cell types including tumour cells and non-tumour cells (also referred to as “reference”, “healthy”, “normal” or “germline” cells). One or more germline samples may also be obtained, which do not comprise tumour genetic material. Germline samples may be matched germline samples, obtained from the same subject as the subject from which the one or more tumour samples are obtained. The use of a matched germline sample improves the accuracy of calling of somatic (tumour-specific) mutations, as any variant position identified in a tumour sample can be compared to variant positions in a matched germline sample to exclude germline variants. The same matched germline sample may be used to analyse a plurality of tumour samples from a subject. Further, the matched germline sample and one or more tumour samples may have been obtained at different times. For example, a first tumour sample and matched germline sample may have been obtained at the time of diagnosis or resection of a tumour, and a further tumour sample may be obtained and analysed together with the initial matched germline sample at a later time point. When a matched germline sample is not available, a reference sample or genome including common germline variants may be used. Alternatively, a process-matched normal sample may be used, which may not have been obtained from the same subject, or may have been obtained from a pool of subjects. Optionally, the samples may be sequenced at step 12, to obtain at least RNA sequence data, and also optionally DNA sequence data. The RNA sequence data may be obtained by RNA sequencing. The DNA sequence data may be

obtained using one of whole exome sequencing, or whole genome sequencing. Alternatively, the sequence data may have been previously obtained and may be received from a user interface, computing device or database, At optional step 14, the sequence data may be analysed to identify one or more mutations that are likely to be present in the tumour cells but not in non-cancerous cells. These represent tumour-specific mutations. These may be used as candidate neoantigens as explained further below. Step 14 may comprise the steps of aligning the sequences from the one or more samples (i.e. the mixed sample(s) and the germline sample(s), if available) to a reference such as e.g. a reference genome or transcriptome, and identifying genomic locations (also referred to as "positions" or "loci") where the sequence of the tumour differs from the germline sequence or can be assumed to differ from the germline sequence (e.g. if a germline sequence for the subject is not available). This analysis may be performed using RNA sequence data, DNA sequence data, or both. For example, DNA sequence data may be used to identify tumour-specific mutations that are single nucleotide variants, multiple nucleotide variants or indels, and RNA sequence data may be used to identify tumour-specific mutations that are gene fusions or splicing variants. Thus, the tumour-specific mutations analysed may be somatic mutations present in a tumour of the subject from which the samples have been obtained. Any one or more of the tumour-specific mutations identified (or otherwise selected for example by a user through a user interface, or obtained from a computing device or database), may then be analysed to identify peptides associated with the mutations. The subsequent steps will be described in relation to a single tumour-specific mutation, and can be repeated for each of a plurality of tumour-specific mutations identified. The tumour-specific mutation may be a tumour-specific mutation that has been identified as a splice region variant. The tumour-specific mutation may be a tumour-specific mutation that has been identified as a 5' splice variant. These types of variants may be identified based on the position of the mutation relative to annotated splice sites in a reference genome or transcriptome (sometimes also referred to as a gene model). The tumour-specific mutation may be a tumour that has been filtered or annotated for whether the tumour-specific mutation has RNA sequence read support in each of a plurality of tumour samples from a patient. Such mutations may be considered ubiquitous in the tumour. The tumour-specific mutation may be a tumour that has been filtered or annotated for whether the mutation is likely to be clonal in a tumour. The tumour-specific mutation may be a tumour that has been filtered or annotated for whether the mutation is likely to be expressed in a tumour.

At step 16, one or more RNA sequences reads that contain the tumour specific mutations are selected from the RNA sequence data. Reads may be reads that have been mapped to a reference genome or transcriptome and therefore reads may be selected by selecting reads that overlap the genomic position of the tumour-specific mutation and where the sequence of

the read is compatible with the presence of the mutation. The reads may be selected from sequence data from a single sample or from a plurality of samples. In other words, selecting one or more RNA sequence reads is performed across the RNA sequence data from all of a plurality of samples (e.g. a plurality of samples from different tumour regions of a patient).

- 5 At step 18, the reads selected at step 16 are used to assemble a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads. This may comprise assembling one or more of the selected RNA sequence reads into the longest sequence comprising the tumour-specific mutation compatible with overlapping sequences of the selected one or more RNA sequence reads.
- 10 At step 20, an anchor sequence is extracted from the assembled sequence as a subsequence (i.e. a part of a sequence) that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions. Thus, positions of the assembled sequence may be identified (e.g. annotated) as intronic or exonic regions, by reference to a reference genome / transcriptome in which exonic and intronic regions are annotated. Extracting an anchor
15 sequence may comprise determining whether the mutation is in a gene transcribed from the positive strand or the negative strand. Information on whether a gene is transcribed from the negative or the positive strand may be obtained from a genome / transcriptome reference. A mutation located in a genomic region annotated in a genome / transcriptome reference as located in a gene transcribed from the positive strand may be considered a mutation in a gene
20 transcribed from the positive strand, and similarly for genes annotated on the negative strand. When the mutation is in a gene transcribed from the positive strand, extracting an anchor sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions may comprise extracting a sequence that is in 5' of the tumour-specific mutation. When the mutation is in a
25 gene transcribed from the negative strand, extracting an anchor sequence from the assembled sequence may comprise obtaining the reverse complement of the assembled sequence, wherein the anchor sequence is in 5' of the tumour-specific mutation in the reverse complement sequence.

- At step 22, a reading frame is identified for the assembled sequence. The identified reading
30 frame is the reading frame of a transcript that satisfies at least the following conditions: (i) it overlaps the genomic position of the tumour-specific mutation, and (ii) the anchor sequence maps to the reference transcript or the reference transcript has a corresponding protein that a translation of the anchor sequence maps to. An anchor sequence may be considered to map to a reference sequence if it has a matching sequence of a minimum length and/or a number
35 of mismatches below a predetermined number. Identifying a reading frame may comprise

translating at least the anchor sequence using all 3 possible reading frames and identifying a protein corresponding to a reference transcript that a translated anchor sequence maps to, wherein the identified reading frame is the reading frame of the translated anchor that maps to the protein corresponding to a reference transcript. Alternatively, identifying a reading frame may comprise identifying a reference transcript that overlaps the genomic position of the tumour-specific mutation and that the anchor sequence maps to, and identifying the reading frame of the identified reference transcript. In other words, the anchor sequence may be matched to reference transcripts at the protein level (i.e. matching translated anchor sequences to protein sequences associated with reference transcripts) or at the cDNA level (i.e. matching cDNA anchor sequences to reference transcripts sequences). In the former case, the identified reading frame is the reading frame of the mapping anchor sequence, and no separate further translation step may be needed. In the latter case, the identified reading frame is the reading frame of the peptide corresponding to the reference transcript and this can be used to translate the complete assembled sequence.

At step 24, a neoantigen peptide that corresponds to the tumour-specific mutation is identified as a translation of the assembled sequence or a part of this translation, where the translation uses the reading frame identified at step 22. In other words, the assembled sequence or a part thereof (which contains the mutation or contains a sequence that results in a peptide that differs from the wild type sequence as a result of the presence of the mutation) are translated using the identified reading frame and the resulting peptide is considered as a candidate neoantigen.

At optional step 26, the results of one or more of the preceding steps are provided to a user, for example through a user interface.

Applications

The above methods find applications in the context of designing immunotherapies, particularly immunotherapies that use peptides or sequences encoding peptides associated with tumour-specific mutations to generate or promote an immune response. Indeed, such methods typically require the identification of peptides associated with the mutations for the purpose of manufacturing the immunotherapy and/or for the purpose of selecting peptides for inclusion in the immunotherapy based on any criteria that apply to the peptide sequence (such as e.g. predicted immunogenicity). Indeed, peptides that are predicted to be likely to be immunogenic are more promising candidates for inclusion in an immunotherapy, and for it is therefore beneficial to identify neoantigen peptide sequences in order to test this even when the immunotherapy uses sequences encoding the peptides selected. In particular, the above methods may be used to provide cancer immunotherapies that target cancer-specific antigens

(also referred to herein as “cancer neoantigens”, or simply “neoantigens”. The cancer neoantigens may be clonal neoantigens. Thus, also described herein are methods of providing an immunotherapy for a subject, the method comprising identifying one or more peptides associated with one or more tumour specific mutations, wherein the identifying is based on data from one or more samples from the subject and further is performed using a method as describe herein. An example of such a method will be described by reference to **Figure 2**.

At optional step 210, one or more samples comprising tumour genetic material and one or more germline samples are obtained from a subject. The subject may be a subject that has been diagnosed as having cancer, and may be (but does not need to be) the same subject for which the immunotherapy is provided. At step 212, a list of tumour-specific variants is obtained. This may comprise step 212' of obtaining a list of tumour-specific variants from genomic sequence data and/or step 212'' of obtaining a list of tumour-specific variants from RNA sequence data. Sequences comprising tumour-specific variants that are believed to potentially lead to the expression of a protein that is not present in a normal cell may be referred to as “candidate neoantigens”. In the context of the present disclosure, tumour-specific variants may be filtered for variants that are likely to result in the expression of a protein that is not present in a normal cell, and therefore reference to tumour-specific variants may equally refer to candidate neoantigens unless context indicates otherwise. At optional step 212', a list of candidate neoantigens may be obtained from genomic sequence data from the sample(s) using methods known in the art, for example as described in WO 2016/16174085, Landau et al. (2013), Lu et al. (2018), Leko et al. (2019), Hundal et al. (2019), and others. The list of candidate neoantigens may comprise a single neoantigen, or a plurality of neoantigens. Preferably, the list comprises a plurality of neoantigens. The neoantigens may be clonal neoantigens. Methods to identify clonal neoantigens are known in the art and include the methods described in WO 2016/16174085, Landau et al. (2013), Roth et al. (2014), McGranahan et al. (2016), and in WO 2022/207925. Instead or in addition to step 212', one or more candidate neoantigens may be identified from RNA sequence data from the sample(s) at step 212'', for example by identifying one or more RNA sequence reads that include a variant. Variants can be identified by comparison with an expected healthy sequence such as a reference genome or transcriptome or RNA/DNA sequence from a normal sample of the subject. Thus, step 212'' may comprise optional step 212a'', where the RNA sequence content of the one or more samples comprising tumour genetic material and optionally the matched germline sample may be determined, for example by sequencing the RNA (or mRNA) in the sample using RNA sequencing. Alternative methods such as e.g. expression arrays may be used, although sequencing methods are preferred since they generate a digital output representative of the number of each particular sequence in a sample. Step 212'' may further

comprise optional step 212b'' of analysing the RNA sequence data to identify one or more mutations that are likely to be present in the tumour cells but not in non-cancerous cells. These represent tumour-specific mutations and may be used as candidate neoantigens. This may comprise the steps of aligning the RNA sequences from the one or more samples (i.e. the sample(s) comprising tumour genetic material and the germline sample(s), if available). This may further comprise identifying locations where the RNA sequence of the tumour differs from the germline sequence or can be assumed to differ from the germline sequence (e.g. if a germline sequence for the subject is not available), for example based on a reference genome or transcriptome. Instead or in addition to this, step 212b'' may comprise aligning the RNA sequences from the one or more samples to a reference genome or transcriptome and identifying sequences that are not expected to be present in such a reference (e.g. novel transcripts, splicing variants, fusions, single nucleotide variants, indels). For example, a fusion or splicing variant may be identified if one or more reads align to non-contiguous sections of the reference transcriptome or genome, or fail to align to the reference genome or transcriptome.

Step 212' may comprise optional step 212a', where the sequence content of the one or more samples comprising tumour genetic material and optionally the matched germline sample may be determined, for example by sequencing the genomic material in the sample using one of whole exome sequencing or whole genome sequencing. At optional step 212b', the sequence data may be analysed to identify one or more mutations that are likely to be present in the tumour cells but not in non-cancerous cells. These represent tumour-specific mutations and may be used as candidate neoantigens. This may comprise the steps of aligning the sequences from the one or more samples (i.e. the mixed sample(s) and the germline sample(s), if available), and identifying genomic locations where the sequence of the tumour differs from the germline sequence or can be assumed to differ from the germline sequence (e.g. if a germline sequence for the subject is not available). At optional step 212c', genomic sequence data for the sample comprising tumour genetic material may be used to determine the probability of the tumour-specific mutations being clonal, for example as described in WO 2022/207925.

At step 214, one or more candidate neoantigen peptides corresponding to the (optionally clonal) candidate neoantigens identified at step 212 are obtained, using a method as described herein. This may comprise, for each tumour-specific mutation identified using genomic sequence data (steps 212'), analysing RNA sequence data obtained at step 212a'' to identify one or more candidate neoantigen peptides for each tumour-specific variant (optionally restricted to clonal variants) that has support in the RNA sequence data (e.g. for which at least 1 RNA sequencing read containing the variant can be identified). Where such

candidate neoantigen peptides can be obtained, the corresponding tumour-specific variants are by definition candidate neoantigens as they lead to the expression of a protein or peptide not present in a normal cell, due to the presence of the variant.

At step 216, an immunotherapy that targets at least one (and optionally a plurality) of the candidate (optionally clonal) neoantigens is designed. Designing such an immunotherapy comprises identifying one or more candidate peptides for each of the candidate clonal neoantigens (step 216A). For example, a plurality of peptides may be designed for at least one of the candidate clonal neoantigens, which differ in their lengths and/or the location of a sequence variation that characterises the neoantigen compared to the corresponding germline peptide. At step 216B, the one or more peptides identified are analysed to determine whether they have one or more properties, such as being likely to be immunogenic (e.g. as described in GB 2303920.9), being expressed in the subject's tumour (e.g. as described in GB2213928.1), being expressed in reference samples or datasets, having a low similarity to a corresponding normal peptide, having a high likelihood of manufacturability (e.g. as described in application PCT/EP2023/055383), etc. At step 216C, one or more of the peptides are selected for production based on at least some of the results of step 216B.

At step 218, the selected peptides (or sequences encoding said peptides) may be obtained. Peptides with selected sequences may be obtained using any method known in the art but they are preferably obtained using chemical synthesis. Methods for obtaining sequences that encode peptides of interest are known. For example, tandem minigenes may be obtained which encode the selected one or more peptide. At step 220, an immunotherapy may be produced using at least some of the one or more peptides or sequences encoding said peptides produced at step 218. The immunotherapy may comprise the one or more peptides (e.g. in the case of an immunogenic composition such as a synthetic long peptide vaccine), sequences encoding said peptides (e.g. in the case of a DNA or RNA vaccine) or may comprise molecules or cells that have been obtained using the selected peptides (e.g. in the case of therapeutic antibodies that selectively bind the candidate peptides, or immune cells that specifically recognise the candidate peptides). In the illustrated embodiment, the immunotherapy comprises cells that have been obtained using the selected peptides. Methods of producing an immunotherapy comprising cells that have been obtained using neoantigen peptides are known in the art, for example as described in WO 2022/207925, WO 2016/16174085, McGranahan et al. (2016), Lu et al. (2018), and Leko et al. (2019). At optional step 222, the immunotherapy may be administered to a subject, which is preferably the subject from which the samples used to identify the neoantigens have been obtained. An example of producing an immunotherapy comprising a T cell population selectively enriched with T cells that recognise one or more neoantigens, preferably clonal neoantigens, will be described. At

step 220A, a population of T cells may be obtained. The T cells may be obtained from the subject to be treated, but do not need to be. The T cells may be obtained from a tumour sample, from a blood sample, or from any other tissue sample. At step 220B, a population of antigen presenting cells (e.g. dendritic cells) may be obtained. For example, a population of dendritic cells may be derived from mononuclear cells (e.g. peripheral blood mononuclear cells, PBMCs) from the subject to be treated. At step 220C, the population of dendritic cells may be pulsed with the selected peptides. At step 220D, the T cell population may be selectively expanded using the population of pulsed dendritic cells. Additional expansion factors such as e.g. cytokines or stimulating antibodies may be used.

Thus, the disclosure provides a method of providing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising: identifying one or more neoantigen peptides associated with tumour-specific mutations present in the subject using a method described herein, and designing an immunotherapy that targets one or more of the candidate neoantigens. The method may have any one or more of the following features. The immunotherapy that targets the one or more of the neoantigens may be an immunogenic composition, a composition comprising immune cells or a therapeutic antibody. The immunogenic composition may comprise one or more of the neoantigen peptides (such as e.g. a neoantigen peptide or protein or a cell displaying the neoantigen). The composition comprising immune cells may comprise T cells, B cells and/or dendritic cells. The composition comprising a therapeutic antibody may comprise one or more antibodies that recognise at least one of the one or more of the neoantigen peptides. An antibody may be a monoclonal antibody. The immunogenic composition may comprise one or more nucleic acids encoding the one or more peptides, or a construct comprising such a nucleic acid.

Designing an immunotherapy that targets one or more of the neoantigen peptides identified may comprise identifying one or more candidate peptides for each of the one or more neoantigens peptides targeted, each peptide comprising at least a portion of a neoantigen peptide targeted. The method may further comprise obtaining the one or more candidate peptides. The method may further comprise testing the one or more candidate peptides for one or more further properties. Further testing may be performed *in vitro* or *in silico*. For example, the one or more peptides may be tested for immunogenicity, propensity to be displayed by MHC molecules (optionally by specific MHC molecule alleles, where the alleles may have been chosen depending on the MHC alleles expressed by the subject), ability to elicit proliferation of a population of immune cells, etc.

A method of designing, or providing an immunotherapy may further comprise producing the immunotherapy. The method may further comprise obtaining a population of dendritic cells

that has been pulsed with one or more of the candidate peptides. The immunotherapy may be a composition comprising T cells that recognise at least one of the one or more of the neoantigens identified. The composition may be enriched for T cells that target at least one of the one or more of the neoantigens identified. The method may comprise obtaining a population of T cells and expanding the population of T cells to increase the number or relative proportion of T cells that target at least one of the one or more of the neoantigens identified. The method may further comprise obtaining a T cell population. A T cell population may be isolated from the subject, for example from one or more tumour samples obtained from the subject, or from a peripheral blood sample or a sample from other tissues of the subject. The T cell population may comprise tumour infiltrating lymphocytes. T cells may be isolated using methods which are well known in the art. For example, T cells may be purified from single cell suspensions generated from samples on the basis of expression of CD3, CD4 or CD8. T cells may be enriched from samples by passage through a Ficoll-paque gradient. The method may further comprise expanding the T cell population. For example, T cells may be expanded by ex vivo culture in conditions which are known to provide mitogenic stimuli for T cells. By way of example, the T cells may be cultured with cytokines such as IL-2 or with mitogenic antibodies such as anti-CD3 and/or CD28. The T cells may be co-cultured with antigen-presenting cells (APCs), which may have been irradiated. The APCs may be dendritic cells or B cells. The dendritic cells may have been pulsed with the candidate peptides (containing one or more of the identified neoantigens) as single stimulants or as pools of stimulating neoantigen peptides. Expansion of T cells may be performed using methods which are known in the art, including for example the use of artificial antigen presenting cells (aAPCs), which provide additional co-stimulatory signals, and autologous PBMCs which present appropriate peptides. Autologous PBMCs may be pulsed with peptides containing neoantigens as discussed herein as single stimulants, or alternatively as pools of stimulating neoantigens.

Also described herein is a method for expanding a T cell population for use in the treatment of cancer in a subject, the method comprising: identifying one or more neoantigen peptides using a method as described herein; obtaining a T cell population comprising a T cell which is capable of specifically recognising one of the neoantigen peptides; and co-culturing the T cell population with a composition comprising the neoantigen peptide. The method may have one or more of the following features. The T cell population obtained may be assumed to comprise a T cell capable of specifically recognising one of the neoantigen peptides. The method preferably comprises identifying a plurality of neoantigen peptides. The neoantigen peptides may comprise one or more clonal neoantigens. The T cell population may comprise a plurality of T cells each of which is capable of specifically recognising one of the plurality of neoantigen peptides, and co-culturing the T cell population with a composition comprising the plurality of

neoantigen peptides. The co-culture may result in expansion of the T cell population that specifically recognises one or more of the neoantigen peptides. The expansion may be performed by co-culture of a T cell with the one or more neoantigen peptides and an antigen presenting cell. The antigen presenting cell may be a dendritic cell. Thus, the expansion may be a selective expansion of T cells which are specific for the neoantigen peptides. The expansion may further comprise one or more non-selective expansion steps. Thus, also described herein is a composition comprising a population of T cells obtained or obtainable by a method as described above.

Thus, the disclosure also provides a T cell composition comprising a T cell population selectively enriched with T cells that recognise one or more neoantigens, preferably clonal neoantigens, wherein the T cell population has been selectively enriched using peptides that have been identified using any of the methods described herein.

In a T cell composition as described herein the expanded population of neoantigen-reactive T cells may have a higher activity than the population of T cells which have not been expanded, as measured by the response of the T cell population to restimulation with a neoantigen peptide. Activity may be measured by cytokine production, and wherein a higher activity is a 5-10 fold or greater increase in activity.

References to a plurality of neoantigens may refer to a plurality of peptides or proteins each associated with a different tumour-specific mutation that gives rise to a neoantigen. Said plurality may be from 2 to 250, from 3 to 200, from 4 to 150, or from 5 to 100 tumour-specific mutations, for example from 5 to 75 or from 10 to 50 tumour-specific mutations. Each tumour-specific mutation may be represented by one or more neoantigen peptides. In other words, a plurality of neoantigens may comprise a plurality of different peptides, some of which comprise a sequence that includes the same tumour-specific mutation (for example at different positions within the sequence of the peptide, or within peptides of varying lengths). Thus, the one or more selected peptides obtained at step 218 (or any method comprising selecting peptides using the methods described herein) may comprise from 2 to multiple hundred peptides, such as e.g. between 2 and 400 peptides, between 2 and 300 peptides, between 2 and 250 peptides, between 2 and 200 peptides, between 10 and 400 peptides, between 10 and 300 peptides, between 10 and 250 peptides, between 10 and 200 peptides, between 50 and 400 peptides, between 50 and 300 peptides, between 50 and 250 peptides, or between 50 and 200 peptides. In particular, the one or more selected peptides may comprise up to a maximum number of peptides that is set by the capacity of a synthesis process or a step thereof, such as for example the number of wells in a reaction plate used for a single synthesis run or a multiple thereof. For example, when 96 wells plates are used the number of selected peptides

may be set to a maximum of 96, 192, 288, or 384. Instead or in addition to this, in the context of peptides comprising a different tumour-specific mutation that gives rise to a neoantigen, the number of peptides selected may be set to a maximum corresponding to the number of tumour-specific mutations that give rise to a neoantigen identified in a subject, or to the number of different peptides of a predetermined length that comprise said tumour-specific mutations. For example, as many as 1,000 to 10,000 peptides comprising one or more coding mutations may be identified using the methods described herein. A T cell population that is produced in accordance with the present disclosure will have an increased number or proportion of T cells that target one or more neoantigens that are represented in peptides identified using the methods described herein. That is to say, the composition of the T cell population will differ from that of a "native" T cell population (i.e. a population that has not undergone the expansion steps discussed herein), in that the percentage or proportion of T cells that target a neoantigen that is identified and produced as described herein will be increased. The T cell population according to the disclosure may have at least about 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or 100% T cells that target a neoantigen for which a peptide is identified and produced as described herein.

The immunotherapies described herein may be used in the treatment of cancer. Thus, the disclosure also provides a method of treating cancer in a subject comprising administering an immunotherapeutic composition as described herein to the subject.

Suitably, in any embodiment of any aspect described herein, the cancer may be ovarian cancer, breast cancer, endometrial cancer, kidney cancer (renal cell), lung cancer (small cell, non-small cell and mesothelioma), bladder cancer, gastric cancer, oesophageal cancer, colorectal cancer, cervical cancer, endometrial cancer, brain cancer (gliomas, astrocytomas, glioblastomas), melanoma, merkel cell carcinoma, clear cell renal cell carcinoma (ccRCC), lymphoma, small bowel cancers (duodenal and jejunal), leukemia, pancreatic cancer, hepatobiliary tumours, germ cell cancers, prostate cancer, head and neck cancers, thyroid cancer and sarcomas. For example, the cancer may be lung cancer, such as lung adenocarcinoma or lung squamous-cell carcinoma. As another example, the cancer may be melanoma. The cancer may be bladder cancer. The cancer may be head and neck cancer. In embodiments, the cancer may be selected from melanoma, merkel cell carcinoma, renal cancer, non-small cell lung cancer (NSCLC), urothelial carcinoma of the bladder (BLAC) and head and neck squamous cell carcinoma (HNSC) and microsatellite instability (MSI)-high cancers. In some embodiments, the cancer is non-small cell lung cancer (NSCLC). In any embodiment of any aspect, the subject may be human.

Treatment using the compositions and methods of the present disclosure may also encompass targeting circulating tumour cells and/or metastases derived from the tumour. Treatment according to the present disclosure targeting one or more neoantigens, preferably clonal neoantigens, may help prevent the evolution of therapy resistant tumour cells which may occur with standard approaches such as chemotherapy, radiotherapy, or non-specific immunotherapy. The methods and uses for treating cancer described herein may be performed in combination with additional cancer therapies. In particular, the immunotherapies (including but not limited to T cell compositions) described herein may be administered in combination with immune checkpoint intervention, co-stimulatory antibodies, chemotherapy and/or radiotherapy, targeted therapy or monoclonal antibody therapy. Further, any immunotherapy (including but not limited to T cell compositions) described herein may be administered in combination with another immunotherapy described herein. For example, a T cell composition as described herein may be used in combination with an immunogenic composition as described herein. 'In combination' may refer to administration of the additional therapy (whether it is an immunotherapy or otherwise) before, at the same time as or after administration of the immunotherapy (e.g. T cell composition) as described herein.

The invention also provides a method for producing an immunotherapeutic composition, the method comprising identifying one or more candidate neoantigen peptides associated with a tumour-specific mutation, predicting whether one or more candidate neoantigen peptides are likely to be immunogenic (based on e.g. likely MHC binding, likely MHC presentation, likely TCR binding, etc.), selecting one or more peptides from the candidate neoantigen peptides based on the predicting, and producing an immunotherapeutic composition that targets the selected neoantigens.

Also described herein is a composition comprising a neoantigen peptide, neoantigen peptide specific immune cell, or an antibody that recognises a neoantigen peptide, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been identified using the methods described herein. Also described herein is a composition comprising a neoantigen peptide, neoantigen peptide specific immune cell, or an antibody that recognises a neoantigen peptide, wherein said neoantigen peptide has been produced using the methods described herein. Also described herein is a neoantigen peptide, immune cell which recognises a neoantigen peptide, or antibody which recognises a neoantigen peptide, for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been produced using the methods described herein. Also described herein is the use of a neoantigen peptide, immune cell which recognises a neoantigen peptide, or antibody which recognises a neoantigen peptide, in the manufacture of a medicament for use in the treatment or prevention of cancer in a subject, wherein said neoantigen peptide has been identified and

optionally produced using the methods described herein. Also described herein is a method of treating a subject that has been diagnosed as having cancer, the method comprising administering an immunotherapy that has been provided using the methods described herein, or a composition as described herein.

5 Systems

Figure 3 shows an embodiment of a system for identifying a neoantigen peptide associate with a tumour-specific mutation, and/or for providing an immunotherapy based at least in part on the identified neoantigen peptides, according to the present disclosure. The system comprises a computing device 1, which comprises a processor 101 and computer readable
10 memory 102. In the embodiment shown, the computing device 1 also comprises a user interface 103, which is illustrated as a screen but may include any other means of conveying information to a user such as e.g. through audible or visual signals. The computing device 1 is communicably connected, such as e.g. through a network 6, to sequence data acquisition means 3, such as a sequencing machine, and/or to one or more databases 2 storing sequence
15 data. The one or more databases may additionally store other types of information that may be used by the computing device 1, such as e.g. reference sequences, parameters, etc. The computing device may be a smartphone, tablet, personal computer or other computing device. The computing device is configured to implement a method for identifying a neoantigen peptide and/or providing an immunotherapy, as described herein. In alternative embodiments,
20 the computing device 1 is configured to communicate with a remote computing device (not shown), which is itself configured to implement a method of identifying a neoantigen peptide and/or providing an immunotherapy, as described herein. In such cases, the remote computing device may also be configured to send the result of the method to the computing device. Communication between the computing device 1 and the remote computing device
25 may be through a wired or wireless connection, and may occur over a local or public network such as e.g. over the public internet or over WiFi. The sequence data acquisition 3 means may be in wired connection with the computing device 1 and/or the database 2, or may be able to communicate through a wireless connection, such as e.g. through a network 6, as illustrated. The connection between the computing device 1 and the sequence data acquisition
30 means 3 may be direct or indirect (such as e.g. through a remote computer). The sequence data acquisition means 3 are configured to acquire sequence data from nucleic acid samples, for example RNA samples and also optionally genomic DNA samples extracted from cells and/or tissue samples. In some embodiments, the sample may have been subject to one or more preprocessing steps such as RNA purification, fragmentation, library preparation, target
35 sequence capture (such as e.g. exon capture and/or panel sequence capture). The sequence data acquisition means is preferably a next generation sequencer. The sequence data

acquisition means 3 may be in direct or indirect connection with one or more databases 2, on which sequence data (raw or partially processed) may be stored.

The following is presented by way of example and is not to be construed as a limitation to the scope of the claims.

5 EXAMPLES

Introduction

In this work, the inventors present and exemplify a new method to identify neoantigen peptides, which is applicable to all types of mutations but is particularly useful in the context of neojunctions (mutations that result in the presence of new splicing junctions that are not
10 present in the corresponding wild-type transcript).

Figure 6 illustrates a problem with obtaining peptide sequences from sequences associated with neojunctions. Figure 6A illustrates a simple sequence with no neojunction. In this case, a variant sequence can be assembled from RNA sequence reads that are compatible with the variant (illustrated as “Assembled Variant Sequence”). This may comprise the variant itself
15 (labelled as sSNV – single nucleotide variant) and a single nucleotide polymorphism (SNP). The assembled variant sequence can be matched with a reference transcript cDNA, provided that one or two mismatches are allowed in a successful match. The protein sequence corresponding to the matching codons of the reference transcript can be used as the sequence
20 of the neoantigen peptide, by replacing the amino acid for any codon with a coding variant mutation by its corresponding modified amino acid. Figure 6B illustrates the same process applied to a sequence with a neojunction. In this case, a variant is present in the intron sequence between exon E1 and exon E2. This is visible in the RNA sequence data (labelled “Read support”) because it causes some of the intron sequence to be retained. In this case
25 the reference transcript cDNA only contains the sequence for E1 and E2, and therefore the entire sequence of the retained intron will not match to the reference transcript and no matching reference transcript will be identified from which a protein sequence can be derived.

The methods described herein provide a solution to this problem.

Example 1 – A new method to assemble peptides from splice region associated mutations

Figure 7 illustrates different classes of neojunction variants considered in the methods of the present disclosure. These include: 5' exonic splice region variants, splice donor variants, 5' intronic splice region variants, 3' intronic splice region variants, splice acceptor variants and 3' exonic splice region variants. The figure shows two exons, labelled E1 (5' exon) and E2 (3' exon), and intronic sequence between the two exons. The dashed lines indicate the normal
30

splicing event that would occur in the wild-type sequence, which would result in a mature transcript comprising only the sequence of E1 and E2.

5' exonic splice region variants are variants that occur in the 5' exon (E1) and that affect the splicing event between exons E1 and E2. Splice donor variants are variants that occur at the splice donor site. The splice donor site is typically a GU sequence forming the first two bases at the 5' end of the intron. 5' intronic splice region variants are variants that occur in the intron, near the 5' junction (i.e. near the boundary between E1 and the intron) and that affect the splicing event. 3' intronic splice region variants are variants that occur in the intron, near the 3' junction (i.e. near the boundary between E2 and the intron) and that affect the splicing event. Splice acceptor variants are variants that occur at the splice acceptor site. The splice acceptor site is typically an AG sequence forming the last two bases at the 3' end of the intron. 3' exonic splice region variants are variants that occur in the 3' exon (E2) and that affect the splicing event between exons E1 and E2.

In these examples, variants were categorised based on the Ensembl Variant Effect Prediction definitions, i.e.:

- 5' exonic splice region variants are variants that change any of the last 3 bases of an exon (i.e. bases 1 to 3 located 5' of a splice donor site),
- splice donor variants are variants that change the 2 base region at the 5' end of an intron,
- 5' intronic splice region variants are variants that change any of bases 3 to 8 counting from the 5' end of an intron (i.e. bases within 6 bases of a splice donor site, in the intron),
- 3' intronic splice region variants are variants that change any of bases 3 to 8 counting from the 3' end of an intron (i.e. bases within 6 bases of a splice acceptor site, in the intron),
- splice acceptor variants are variants that change the 2 base region at the 3' end of an intron, and
- 3' exonic splice region variants are variants that change any of the first 3 bases of an exon (i.e. bases 1 to 3 located 3' of a splice acceptor site).

The methods of the present disclosure will be illustrated below by reference to each of these classes, on both the + strand and the - strand.

In general, the method assembles a sequence that contains the variant of interest from RNA sequence reads, then translates this into all 3 possible reading frames, identifies an anchor sequence that comprises the coding part of the assembled sequence that precedes the variants (or its corresponding translated sequences), finds reference transcripts that overlap

the assembled sequence and their protein sequences, and uses the reading frame of the reference transcript (or corresponding protein) that best matches the anchor (or conversely, the translated anchor that best matches the protein sequence of a reference transcript). In the examples below, the anchor sequences are obtained as protein sequences (3 possible anchors per assembled sequence, one for each possible reading frame) and compared to protein sequences corresponding to reference transcripts. In either case, the method captures any polymorphisms that are part of the assembled sequence because the sequence is translated directly from the assembled sequence.

Figures 8A, 8B illustrates the use of a method of the disclosure to assemble a peptide associated with a variant of interest that is a splice donor mutation on the + strand. The method starts from aligned RNA sequences, illustrated here as a BAM file, and information identifying one or more variants comprising for each variant: the variant sequence (e.g. in the illustrated embodiment the variant sequence is T>G or G) and the variant genomic coordinates in the same reference genome used to create the aligned RNA sequences. All RNA sequence reads comprising a variant of interest (illustrated as a star in the reads aligned to the gene model on Figure 8A) are extracted from the BAM file and assembled into an assembled sequence. These are sequence reads that overlap with the genomic location of the variant and that include the variant sequence. The assembled sequence is then translated into all 3 possible frames (labelled as Frame 1, Frame 2, Frame 3) to obtain 3 translated sequences. A corresponding set of anchor sequences are obtained by trimming each of the 3 translated sequences to exclude all amino acids from the mutated codon (amino acid that is underlined in the translated sequences) onwards (in the 5'-3' direction), and any remaining intronic amino acids between the mutated codon and the nearest upstream exon (i.e. intronic amino acids that are 5' of the mutated position). In this particular case it is not necessary to remove additional intronic amino acids because the mutation occurs in the first codon of the retained intron. The variant information is also used to identify overlapping reference transcripts, i.e. all transcripts from a reference transcriptome that overlap with the genomic coordinates of the variant of interest. These reference transcripts are associated with corresponding reference protein sequences. The set of 3 anchors are compared with the reference protein sequences to identify the anchor that best matches to any of the reference protein sequences. A best match is identified as that with the highest sequence identity, allowing for a predetermined number of mismatches to allow for polymorphisms. The full translated sequence that corresponds to the best matching anchor sequence is identified as the neoantigen peptide associated with the variant of interest. In the illustrated embodiment it can be seen that this neoantigen includes a series of intronic amino acids that would not be present in the wild-type protein.

Figures 8C, 8D illustrates the use of a method of the disclosure to assemble a peptide associated with a variant of interest that is a 5' intronic splice region mutation on the + strand. The method starts from aligned RNA sequences, illustrated here as a BAM file, and information identifying one or more variants comprising, for each variant: the variant sequence (e.g. in the illustrated embodiment the variant sequence is A) and the variant genomic coordinates in the same reference genome used to create the aligned RNA sequences. All RNA sequence reads comprising a variant of interest (illustrated as a star in the sequence reads aligned to the gene model on Figure 8C) are extracted from the BAM file and assembled into an assembled sequence. These are sequence reads that overlap with the genomic location of the variant and that include the variant sequence. The assembled sequence is then translated into all 3 possible frames (labelled as Frame 1, Frame 2, Frame 3) to obtain 3 translated sequences. A corresponding set of anchor sequences are obtained by trimming each of the 3 translated sequences to exclude all amino acids from the mutated codon (amino acid underlined in the translated sequences) onwards (in the 5'-3' direction), and any intronic amino acids between the mutated codon and the nearest upstream exon (i.e. intronic amino acids that are 5' of the mutated position). Any amino acid that is encoded by a codon that comprises at least one intronic base is considered an intronic amino acid. Translated anchor sequences that contain early stop codons may either not map to a reference protein sequence or may be mapped only to the stop codon, in which case they will typically be too short to map to a reference protein sequence and will therefore be excluded at the mapping stage. The variant information is also used to identify overlapping reference transcripts, i.e. all transcripts from a reference transcriptome that overlap with the genomic coordinates of the variant of interest. These reference transcripts are associated with corresponding reference protein sequences. The set of 3 anchors are compared with the reference protein sequences to identify the anchor that best matches to any of the reference protein sequences. A best match is identified as that with the highest sequence identity, allowing for a predetermined number of mismatches to allow for polymorphisms. The full translated sequence that corresponds to the best matching anchor sequence is identified as the neoantigen peptide associated with the variant of interest. In the illustrated embodiment it can be seen that this neoantigen includes a series of intronic amino acids that would not be present in the wild-type protein.

Figures 8E, 8F illustrates the use of a method of the disclosure to assemble a peptide associated with a variant of interest that is a 5' exonic splice region mutation on the + strand. The method starts from aligned RNA sequences, illustrated here as a BAM file, and information identifying one or more variants comprising, for each variant: the variant sequence (e.g. in the illustrated embodiment the variant sequence is T) and the variant genomic coordinates in the same reference genome used to create the aligned RNA sequences. All

RNA sequence reads comprising a variant of interest (illustrated as a star in the sequence reads aligned to the gene model on Figure 8E) are extracted from the BAM file and assembled into an assembled sequence. These are sequence reads that overlap with the genomic location of the variant and that include the variant sequence. The assembled sequence is then translated into all 3 possible frames (labelled as Frame 1, Frame 2, Frame 3) to obtain 3 translated sequences. A corresponding set of anchor sequences are obtained by trimming each of the 3 translated sequences to exclude all amino acids from the mutated codon (amino acid underlined in the translated sequences) onwards (in the 5'-3' direction). Because the variant of interest is within the exon, there are no intronic amino acids between the mutated codon and the nearest upstream exonic amino acid. The variant information is also used to identify overlapping reference transcripts, i.e. all transcripts from a reference transcriptome that overlap with the genomic coordinates of the variant of interest. These reference transcripts are associated with corresponding reference protein sequences. The set of 3 anchors are compared with the reference protein sequences to identify the anchor that best matches to any of the reference protein sequences. A best match is identified as that with the highest sequence identity, allowing for a predetermined number of mismatches to allow for polymorphisms. The full translated sequence that corresponds to the best matching anchor sequence is identified as the neoantigen peptide associated with the variant of interest. In the illustrated embodiment it can be seen that this neoantigen includes a series of intronic amino acids that would not be present in the wild-type protein.

The logic described above can also be applied to variant that affect genes on the – strand (minus strand). The same classification described on Figure 7 can be used since this classification is relative to the direction of transcription, which is always 5'-3'. However, on the minus strand the transcript is read in the reverse direction, so the reverse complement of the assembled sequence is used. Indeed, aligned reads are mapped to the sequence of the forward (+) strand and are therefore reverse complemented before inclusion in a set of aligned reads (such as a BAM or SAM file).

This is illustrated on **Figures 8G, 8H** with the example of splice donor mutations on the - strand (although the same logic applies to 5' intronic splice region mutations and 5' exonic splice region mutations). The method starts from aligned RNA sequences, illustrated here as a BAM file, and information identifying one or more variants comprising, for each variant: the variant sequence (e.g. in the illustrated embodiment the variant sequence is A) and the variant genomic coordinates in the same reference genome used to create the aligned RNA sequences. All RNA sequence reads comprising a variant of interest (illustrated as a star in the sequence reads aligned to the gene model on Figure 8G) are extracted from the BAM file and assembled into an assembled sequence. These are sequence reads that overlap with the

genomic location of the variant and that include the variant sequence. The reverse complement of the assembled sequence is then obtained.

This is then translated into all 3 possible frames (labelled as Frame 1, Frame 2, Frame 3) to obtain 3 translated sequences. A corresponding set of anchor sequences are obtained by
5 trimming each of the 3 translated sequences to exclude all amino acids from the mutated codon (amino acid underlined in the translated sequences) onwards (in the 5'-3' direction). Because the variant of interest is in the first codon of the intron, there are no intronic amino acids between the mutated codon and the nearest upstream exonic amino acid. The variant information is also used to identify overlapping reference transcripts, i.e. all transcripts from a
10 reference transcriptome that overlap with the genomic coordinates of the variant of interest. These reference transcripts are associated with corresponding reference protein sequences. The set of 3 anchors are compared with the reference protein sequences to identify the anchor that best matches to any of the reference protein sequences. A best match is identified as that with the highest sequence identity, allowing for a predetermined number of mismatches to
15 allow for polymorphisms. The full translated sequence that corresponds to the best matching anchor sequence is identified as the neoantigen peptide associated with the variant of interest. In the illustrated embodiment it can be seen that this neoantigen includes a series of intronic amino acids that would not be present in the wild-type protein.

Note that in each case the anchor sequences also have a corresponding nucleic acid
20 sequence in the assembled sequence (illustrated as a shaded area). In alternative embodiments to those illustrated above, the nucleic acid sequence corresponding to the anchor sequence is aligned with reference transcripts and the assembled sequence is translated using the translation frame of the reference transcript that best matches the nucleic acid sequence corresponding to the anchor sequence.

25 Aligned sequences typically include, for each of a plurality of reads, at least: (i) the sequence of the read, and (ii) the genomic location to which the read is best aligned, including strand information (+/- strand) and the genomic coordinate of the start of the alignment. Aligned sequences can be provided in the form of a BAM or SAM file. As mentioned above, using these file formats, reads are typically represented on the forward (+) genomic strand.
30 Therefore, sequences of reads that are mapped to the reverse (-) genomic strand are provided reverse complemented.

For splice acceptor mutations and other 3' splice variants, accurate peptide assembly depends on knowledge of the sequence or at least the length of any retained intronic sequence, as this ultimately may impact the reading frame of the 3' exon. The principles described above can
35 be used in a similar way for 3' splice variants for which the entire retained intron sequence is

included in the assembled sequence. When the retained intron sequence has a length that is not equal to a multiple of 3, the retained intron sequence and the 3' exon will therefore be read in a new reading frame (selected based on the anchor sequence for the 5' exon). A change in reading frame of the 3' exon can also be detected if the full length of the retained intron sequence is known, by using the reading frame selected based on the anchor sequence for the 5' exon for the part of the assembled sequence that maps to the 5' exon and any known 5' intronic sequence, and applying the same reading frame to the part of the assembled sequence that map to the 3' exon (i.e. starting at the first position of the 3' exon if the total retained intron length is divisible by 3, starting at the second position if the division of the total retained intron length by 3 has a residual of 2, and starting at the third position if the division of the total retained intron length by 3 has a residual of 1. Whether or not the complete sequence of a retained intron can be included in an assembled sequence depends on the type of sequencing that is performed. In typical short read exon capture sequencing this may not be possible because for all intron length because reads would typically only include the 5' end of retained introns. Long read sequencing methodologies may however comprise sufficient information to obtain assembled sequences that include the entire sequence of many retained introns.

Example 2 – Validation on synthetic data

In this example, the methods described in Example 1 were tested on a synthetic data set where reads containing exonic and intronic sequence were generated and variants randomly spiked into those reads.

Methods

The ASimulatoR R package (Manz et al., 2021) was used to create synthetic read sequences containing splicing alterations in genes found on chr22. The events that were chosen fell into one of the following Asimulator categories: intron retention (ir), exon skipping (es), alternative 3' site (a3), or alternative 5' site (a5). However for downstream analysis, only the intron retention events were considered. The ASimulatoR function "simulate_alternative_splicing" was used with the following parameters: seed = 142, ncores = 4, event_probs = 0.9, seq_depth = 5e06, max_genes = NULL, readlen = 100, error_rate = 0.001, multi_events_per_exon = FALSE, probs_as_freq = FALSE.

The output altered read FASTAs were next separated into wildtype and mutant (containing a splice event) files, based on if they had been generated *in silico* from the wildtype of mutant version of the gene.

These separated files were recombined at various frequencies to create "tumor" samples of different "purities" and aligned to chr22 using STAR (Dobin et al., 2013) with the following parameters: outFilterMismatchNmax = 10, outFilterMultimapNmax = 50, alignIntronMax = 500000, alignMatesGapMax = 100000, sjdbScore = 2, outFilterMultimapScoreRange = 3, outFilterScoreMinOverLread = 0.7, outFilterMatchNminOverLread = 0.7. For downstream analysis, only the 100% ("tumor") and 0% pure ("germline") BAM files were used.

The output STAR BAM files were next run through BamSurgeon (Ewing et al., 2015) to generate a single, proximal splice-site mutation that could have led to the splicing alterations introduced by ASimulatoR. The variants that were introduced were either: splice region variants (1-3 base pairs in the exon adjacent to the retained variant), splice donor variants (1-2 base pairs away from retained intron start), splice acceptor variant (1-2 base pairs away from retained intron end), or intron variants (3-8 base pairs within the retained intron). BamSurgeon was run with the following parameters: mutfrac = 0.9, maxdepth = 20000, aligner = STAR, seed = 123456. The output mutated, retained-intron-containing BAM files were sorted, indexed, and saved, along with the VCF file containing the splice-site mutations spiked in.

Results

Figure 9 shows results obtained using methods of the disclosure to identify mutant peptides from 5' splice region variants in a synthetic dataset. The left plot shows the number of 5' splice region variants present in the sample, divided between the 3 categories identified on Figure 7. The right plot shows the number of variants for which peptides were identified, in each of these categories. The vast majority (23 out of 24) of 5' splice region variants were correctly identified, including peptides resulting from variants in each of the 3 classes of 5' variants identified on Figure 7. The single variant for which a peptide could not be identified was a 5'intronic splice region variant falling downstream of a very short first exon that was too short to be used to anchor the assembled sequence when using the parameters used in this particular experiment. In particular, in the present examples, a matching sequence with a minimum length of 4 amino acids between an anchor sequence and a reference protein sequence was required in order to identify a reading frame for a variant. This is equivalent to requiring a mapping of at least 12 bases at the reference transcript level. However, it is possible to adjust this minimum length depending on the mappability of the region to which the anchor is to be aligned. For example, the minimum length of match required could be set to the minimum of a predetermined threshold and the length of an exon (or its corresponding protein sequence) that precedes the neojunction and to which the anchor sequence maps. In other words, the anchor sequence may be considered as matching an exon when it has a

matching sequence of at least a predetermined length or the maximum length it *could* have given the length of the exon.

Further, in the present examples, an exact match between the reference sequence and the anchor sequence was required in order to identify a reading frame. However, more permissive alternatives are possible, particularly when mapping at the reference transcript level. For example, 1 or 2 mismatches in any matching sequence could be tolerated for example to allow for the presence of SNPs. In embodiments, the number of mismatches allowed is determined depending on the minimum length of matching sequence. For example, 1 or 2 mismatches may be allowed in a 12 or more bases matching sequence, whereas 0 or 1 mismatches may be allowed in a 4 (or 5, 6) or more amino acids matching sequence may be used.

Example 3 – Validation on cell line data

In this example, the methods described in Example 1 were tested on data from a breast cancer cell line (HCC1395).

Methods

Sequence data and analysis. RNA-seq data and whole exome sequencing data for the samples were obtained for all patients. The WES data was analysed through an in-house sequence analysis pipeline comprising alignment (using BWA, Li and Durbin, 2009) variant calling (using standard tools such as e.g. Strelka2, Kim et al., 2018 and VarDict, Lai et al. 2016) by comparison to the corresponding germline sequence (HCC1395BL), and variant annotation which associates variants with transcripts. Variants were then filtered for variants that were annotated as splice region variants or splice donor/acceptor variants (splice_donor_variant, splice_acceptor_variant, splice_region_variant) by the Ensembl variant effect predictor (VEP, www.ensembl.org/info/docs/tools/vep/index.html).

Cell culture. Cells were thawed, washed, and resuspended in complete growth media on day 1 for seeding. Sub-culturing was performed as described below:

HCC1395 is an adherent cell line. It was sub-cultured at 1:2 ratio when it reached an approximately 75-80% confluence. Cells were detached using TrypLE Express Enzyme (Gibco) and growth with RPMI complete media containing the following reagents: RPMI 1640 (containing 2g/L Glucose) + 10 mM HEPES + 1 mM Sodium Pyruvate + 2.5g/L Glucose + 10% FCS + 1X Pen/Strep.

HCC1395-BL is a suspension cell line. It was maintained by addition of IMDM complete growth media, keeping cultures concentration between 5×10^5 - 2×10^6 cells/ mL. IMDM complete media: IMDM + 20% FCS + 1X Pen/Strep (stock at 100x).

HCC1395 (adherent cell lines) and HCC1395BL (suspension cell line) cells were expanded until 5.8×10^7 and 1.24×10^8 cells were obtained respectively. Once the cell lines reached the required numbers, the following dilutions of the expanded cells were prepared, to provide x38 complete sets of vials. Each vial contained a total of 1×10^6 cells: HCC1395_100%;
 5 HCC1395BL_100%; HCC1395_5%; HCC1395_10%; HCC1395_30%. Dilutions simulate tumour samples with different purities. Next, each dilution was centrifuged, resuspended in adequate volume of RLT Plus buffer (Qiagen) and aliquoted in vials containing 350 μ L each. 38 aliquots were obtained for each dilution, being technical replicates. All aliquots were stored at -80C until sequencing.

Results

Figures 10 and 11 show examples of results obtained using methods of the disclosure to identify mutant peptides from splice variants in cell line data (HCC1395). Figure 10 shows a specific example of a splice donor variant (chr12:21807476_C>A). Figure 11 shows an
 15 example of a 5' Exonic Splice Region Variant (chr4:119661818_C>T).

The data on Figure 10 shows, from the top and as a function of genomic coordinates: read coverage (the coverage in the exons is much higher than in the introns in this data which is why on the scale of the bars shown in this track it looks like there is no coverage in the intron – the read pile up shows that there is indeed coverage), a splice junction track (this displays
 20 arcs connecting alignment blocks from a single read, see software.broadinstitute.org/software/igv/splice_junctions), the aligned reads pile up, the transcript sequence (ctDNA), the sequence of the corresponding protein in all 3 possible reading frames (with the selected reading frame highlighted), the position and amino acid sequence of the genes known to be located at these coordinates, and the sequence of a
 25 neoantigen peptide (in this case a 29 amino acids peptide with the neojunction located approximately in the middle of the peptide – i.e. starting at position 15; other lengths of peptides and locations of the neojunction are possible). In the sequence of the neoantigen peptide, the anchor sequence is underlined and circled with a solid line, and the retained intronic sequence is underlined and circled with a dashed line. Note that this represents RNA
 30 data such that the coverage depends not only on the proportion of cells in the sample that have the variant but also on whether the mutated allele (in this case comprising a retained intron) is expressed at the same or different level than the wildtype allele (normal splicing).

The data on Figure 11 shows, from the top of and as a function of genomic coordinates: read coverage, a splice junction track, the aligned reads pile up, the transcript sequence (ctDNA),
 35 the sequence of the corresponding protein in all 3 possible reading frames (with the selected

reading frame highlighted), the position and amino acid sequence of the genes known to be located at these coordinates, and the sequence of a neoantigen peptide (in this case a 30 amino acids peptide with the neojunction located approximately in the middle of the peptide – i.e. starting at position 16; other lengths of peptides and locations of the neojunction are possible). The data shows the presence of a retained intronic region (circled and underlined with dashed lines) followed by a novel donor site (at the second of two consecutive GT sequences present in the intron) leading to skipping into the downstream exon (also shown in B, continuous circled and underlined sequences). The sequence circled and underlined with crosses is the anchor sequence. Figures 11A and 11B show the same type of information but Figure 11B additionally shows the downstream exon not visible on Figure 11A.

A non-evidence-based approach to identifying these peptides would have failed in this case since the presence of a GT alone does not necessarily mean that a new donor site will be present.

Example 4 – Identification of splice variant peptides in multi-region patient data

In this example, the inventors extended the approach described in Example 1 to situations where a plurality of transcriptomic datasets are available for a particular patient, and demonstrated the use of this approach to patient data.

Methods

RNA-seq data and whole exome sequencing data for human tumour samples were obtained for all patients. The WES data was analysed through an in-house sequence analysis pipeline comprising alignment (using BWA, Li and Durbin, 2009) variant calling (using standard tools such as e.g. Strelka2, Kim et al., 2018 and VarDict, Lai et al. 2016) by comparison to the corresponding germline sequence (HCC1395BL), and variant annotation which associates variants with transcripts. Variants were then filtered for variants that were annotated as splice region variants or splice donor / acceptor variants (splice_donor_variant, splice_acceptor_variant, splice_region_variant) by the Ensembl variant effect predictor (VEP, www.ensembl.org/info/docs/tools/vep/index.html).

Results

Figure 12 illustrates a method to identify neoantigen peptides from sequence data from a plurality of samples (labelled R1, R2). The method considers the union of reads from all samples over a locus when assembling a variant sequence. In other words, all RNA sequencing reads obtained from each of the samples are pooled together and aligned. For example, a single BAM file may be obtained combining all of this sequence data. In the illustrated example, the method also preserves the information about which sample each read

came from, which enables it to determine whether there are reads supporting the mutation in each sample. This can in turn be used to determine whether the mutation is ubiquitous (present in all of the plurality of samples). The sequencing reads are then grouped based on the distinct alleles that are present in the reads. In the illustrated example, a single exonic variant is present in the region considered and therefore reads are grouped in two groups: a group of reads comprising the variant allele, and a group of reads comprising the normal (wild type) allele. In other examples a plurality of somatic variants may be present, each of which may be supported by a plurality of reads. Reads that are compatible with each variant may be analysed together, i.e. reads supporting the particular variant are processed together by assembling a variant sequence. The rest of the method proceeds as explained above for each particular variant.

Figure 13A shows results obtained using methods of the disclosure to identify mutant peptides from tumour samples. On **Figure 13A**, in the top plot, each bar of the barchart represents a different cancer patient, and the height of the bar indicates the number of splice region associated variants for which peptides could be identified for each patient in each category of variant (in each bar, from top to bottom: exonic variant, splice donor variant, intronic variant). The bottom plot shows the information filtered for variants that are ubiquitously expressed. The data shows that many of the patients in the cohort have expressed splice region associated variants for which the methods of the present disclosure could identify peptides. These each represent candidate patient-specific neoantigens for immunotherapy.

Figure 13B shows, for each class of 5' splice variants analysed, the distribution of variant allele fractions (VAF, labelled as total VAF as it is obtained by merging all variants and reference reads from a plurality of samples for the same patient) of splice variants identified across the entire cohort. This data shows that splice donor variants have high VAFs irrespective of purity (as expected since we should only see reads in non-coding regions due to disrupted splicing – because reads that don't contain the variant should be spliced out), which is not the case for exonic variants (as expected since the reference allele should also be expressed – assuming that it is not silenced, resulting in a deflated VAF). This is illustrated schematically below the data. The data indicates that these variants (and splice donor variants in particular) are truly indicative of splice site disruptions and the peptides identified from them represent good candidate neoantigens. The effect is also seen to some extent in intronic variants although not to the same extent, suggesting that the intronic variants are not singly responsible for the splicing alteration observed in the RNA data.

Figure 13C shows an example of a ubiquitously expressed splice donor variant identified in a patient (chr5:140041594_T>C). This shows a retained intron (due to a mutation in a splice

donor site) followed by a novel splice donor, after which the sequence carries on to the next exon. Because the intron retained did not have a length divisible by 3, the retained intron led to a shift in reading frame, which meant that the start of the exon was translated but then hit a premature stop codon. This type of variant could not have been identified without an assembly-based approach as described herein.

References

- Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, Busby M, Duke F, Yang A, Young L, Ojo NC, Caldwell K, Abhyankar J, Boucher T, Hart MG, Makarov V, Montpreville VT, Mercier O, Chan TA, Scagliotti G, Bironzo P, Novello S, Karachaliou N, Rosell R, Anderson I, Gabrail N, Hrom J, Limvarapuss C, Choquette K, Spira A, Rousseau R, Voong C, Rizvi NA, Fadel E, Frattini M, Jooss K, Skoberne M, Francis J, Yelensky R. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol.* 2018 Dec 17.
- Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhir R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol.* 2012 May;30(5):413-21.
- Vanessa Jurtz, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters and Morten Nielsen. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* November 1, 2017, 199 (9) 3360-3368.
- Langmead, B., Trapnell, C., Pop, M. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).
- Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* 2008 Jul 1;36(Web Server issue):W509-12.
- McGranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Van Allen, E. M., Miao, D., ... Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science (New York, N.Y.)*, 351(6280), 1463–1469.
- Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell.* 2013 Feb 14;152(4):714-26. doi: 10.1016/j.cell.2013.01.019.
- Raine KM, Van Loo P, Wedge DC, Jones D, Menzies A, Butler AP, Teague JW, Tarpey P, Nik-Zainal S, Campbell PJ. asc4tNgs: Identifying Somatic Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics.* 2016 Dec 8;56:15.9.1-15.9.17. doi: 10.1002/cpbi.17.
- O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst.* 2020 Jul 22
- Lu YC, Zheng Z, Robbins PF, Tran E, Prickett TD, Gartner JJ, Li YF, Ray S, Franco Z, Bliskovsky V, Fitzgerald PC, Rosenberg SA. An Efficient Single-Cell RNA-Seq Approach to Identify Neoantigen-Specific T Cell Receptors. *Mol Ther.* 2018 Feb 7;26(2):379-389.

Pieter Moris, Joey De Pauw, Anna Postovskaya, Sofi Gielis, Nicolas De Neuter, Wout Bittremieux, BensonOgunjimi, Kris Laukens, and Pieter Meysman. Current challenges for epitope-agnostic TCR interaction prediction and a new perspective derived from image classification September 2020. Pages: 2019.12.18.880146. Section: New Results.

- 5 Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, Shah SP. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014 Apr;11(4):396-8. doi: 10.1038/nmeth.2883.
- McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, Watkins TB, Shafi S, Murugaesu N, Mitter R, Akarca AU, Linares J, Marafioti T, Henry JY, Van Allen EM, Miao D, Schilling B, Schadendorf D, Garraway LA, Makarov V, Rizvi NA, Snyder A, Hellmann MD, Merghoub T, Wolchok JD, Shukla SA, Wu CJ, Peggs KS, Chan TA, Hadrup SR, Quezada SA, Swanton C. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. 2016 Mar 25;351(6280):1463-9. doi: 10.1126/science.aaf1490.
- 10
- 15 Leko V, McDuffie LA, Zheng Z, Gartner JJ, Prickett TD, Apolo AB, Agarwal PK, Rosenberg SA, Lu YC. Identification of Neoantigen-Reactive Tumor-Infiltrating Lymphocytes in Primary Bladder Cancer. *J Immunol*. 2019 Jun 15;202(12):3458-3467.
- Hundal J, Kiwala S, Feng YY, Liu CJ, Govindan R, Chapman WC, Uppaluri R, Swamidass SJ, Griffith OL, Mardis ER, Griffith M. Accounting for proximal variants improves neoantigen prediction. *Nat Genet*. 2019 Jan;51(1):175-179. doi: 10.1038/s41588-018-0283-9.
- 20
- Quirin Manz and others, ASimulatoR: splice-aware RNA-Seq data simulation, *Bioinformatics*, Volume 37, Issue 18, 15 September 2021, Pages 3008–3010.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013 Jan 1;29(1):15-21.
- 25
- Ewing, A., Houlahan, K., Hu, Y. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623–630 (2015).
- Kim, S., Scheffler, K., Halpern, A.L. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591–594 (2018).
- 30
- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60.
- Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016 Jun 20;44(11):e108.
- 35
- All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety.

The specific embodiments described herein are offered by way of example, not by way of limitation. Various modifications and variations of the described compositions, methods, and uses of the technology will be apparent to those skilled in the art without departing from the scope and spirit of the technology as described. Any sub-titles herein are included for convenience only and are not to be construed as limiting the disclosure in any way. Unless context dictates otherwise, the descriptions and definitions of the features set out above are not limited to any particular aspect or embodiment of the invention and apply equally to all aspects and embodiments which are described.

The methods of any embodiments described herein may be provided as computer programs or as computer program products or computer readable media carrying a computer program which is arranged, when run on a computer, to perform the method(s) described above.

Throughout the specification and claims, the following terms take the meanings explicitly associated herein, unless the context clearly dictates otherwise. The phrase "in one embodiment" as used herein does not necessarily refer to the same embodiment, though it may. Furthermore, the phrase "in another embodiment" as used herein does not necessarily refer to a different embodiment, although it may. Thus, as described below, various embodiments of the invention may be readily combined, without departing from the scope or spirit of the invention.

It must be noted that, as used in the specification and the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Ranges may be expressed herein as from "about" one particular value, and/or to "about" another particular value. When such a range is expressed, another embodiment includes from the one particular value and/or to the other particular value. Similarly, when values are expressed as approximations, by the use of the antecedent "about," it will be understood that the particular value forms another embodiment. The term "about" in relation to a numerical value is optional and means for example +/- 10%.

Throughout this specification, including the claims which follow, unless the context requires otherwise, the word "comprise" and "include", and variations such as "comprises", "comprising", and "including" will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

Other aspects and embodiments of the invention provide the aspects and embodiments described above with the term "comprising" replaced by the term "consisting of" or "consisting essentially of", unless the context dictates otherwise.

“and/or” where used herein is to be taken as specific disclosure of each of the two specified features or components with or without the other. For example “A and/or B” is to be taken as specific disclosure of each of (i) A, (ii) B and (iii) A and B, just as if each is set out individually herein.

- 5 The features disclosed in the foregoing description, or in the following claims, or in the accompanying drawings, expressed in their specific forms or in terms of a means for performing the disclosed function, or a method or process for obtaining the disclosed results, as appropriate, may, separately, or in any combination of such features, be utilised for realising the invention in diverse forms thereof.

10

CLAIMS

1. A computer implemented method of identifying a neoantigen peptide associated with a tumour-specific mutation, the method comprising:

obtaining RNA sequence data from one or more samples comprising tumour genetic material;

selecting one or more RNA sequence reads from the RNA sequence data that contain the tumour-specific mutation;

assembling a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads;

identifying positions of the assembled sequence as part of intronic or exonic regions;

extracting an anchor sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions;

identifying a reading frame as the reading frame of a reference transcript: (i) that overlaps the genomic position of the tumour-specific mutation, and (ii) that the anchor sequence maps to or that has a corresponding protein that a translation of the anchor sequence maps; and

identifying the neoantigen peptide as a translation of the assembled sequence or a part thereof that uses the identified reading frame.

2. The method of claim 1, wherein the anchor sequence or translation thereof is considered to map to a reference transcript or protein corresponding to a reference transcript when the number of mismatched positions between the anchor sequence or anchor sequence translation and the reference sequence or reference sequence protein is at most a predetermined number, optionally wherein the predetermined number is 0, 1, 2 or 3.

3. The method of any preceding claim, wherein identifying a reading frame comprises:

translating at least the anchor sequence using all 3 possible reading frames; and

identifying a protein corresponding to a reference transcript that a translated anchor sequence maps to, wherein the identified reading frame is the reading frame of the translated anchor that maps to the protein corresponding to a reference transcript.

4. The method of claim 3, wherein identifying a reading frame comprises translating the entire assembled sequence using all 3 possible reading frames.

5. The method of claim 1 or claim 2, wherein identifying a reading frame comprises identifying a reference transcript that overlaps the genomic position of the tumour-specific mutation and that the anchor sequence maps to, and identifying the reading frame of the identified reference transcript.

6. The method of any preceding claim, wherein assembling a sequence comprising the tumour-specific mutation and compatible with overlapping sequences of the selected one or more RNA sequence reads comprises assembling one or more of the selected RNA sequence reads into the longest sequence comprising the tumour-specific mutation compatible with overlapping sequences of the selected one or more RNA sequence reads.

7. The method of any preceding claim, wherein the tumour-specific mutation is a splice region variant.

8. The method of any preceding claim, wherein the tumour-specific mutation is a 5' splice variant, or wherein the tumour-specific mutation is selected from: a mutation in a 5' exonic splice region, a mutation in a splice donor site, and a mutation in a 5' intronic splice region.

9. The method of any preceding claim, wherein the mutation is in a gene transcribed from the positive strand and extracting an anchor sequence from the assembled sequence as a subsequence that precedes the tumour-specific mutation and that only includes positions identified as part of exonic regions comprises extracting a sequence that is in 5' of the tumour-specific mutation.

10. The method of any of claims 1 to 8, wherein the mutation is in a gene transcribed from the negative strand and extracting an anchor sequence from the assembled sequence comprises obtaining the reverse complement of the assembled sequence, wherein the anchor sequence is in 5' of the tumour-specific mutation in the reverse complement sequence.

11. The method of any preceding claim, wherein:

obtaining RNA sequence data from one or more samples comprising tumour genetic material comprises obtaining RNA sequence data from a plurality of samples comprising tumour genetic material, and

selecting one or more RNA sequence reads from the RNA sequence data that contain the tumour-specific mutation comprises selecting reads across the RNA sequence data from all of the plurality of samples.

12. The method of claim 11, wherein the method comprises determining whether the tumour-specific mutation has support in each of the plurality of samples, and/or wherein the tumour-specific mutation is ubiquitous in the tumour from which the plurality of samples have been
5 obtained.

13. The method of any preceding claim, wherein the one or more samples are samples from the same subject and/or the same tumour.

10 14. The method of any preceding claim, wherein the one or more samples are tumour samples.

15. The method of any preceding claim, further comprising identifying the tumour-specific mutation using the RNA sequence data or DNA sequence data from one or more samples
15 comprising tumour genetic material from the same tumour.

16. The method of any preceding claim, comprising obtaining DNA sequence data from one or more samples comprising genetic material from the tumour and identifying the tumour-specific mutation by comparing the DNA sequence data from the one or more samples
20 comprising genetic material from the tumour and a reference genome sequence and/or DNA sequence data from one or more germline samples.

17. The method of any preceding claim, wherein the tumour-specific mutation is associated with genomic coordinates and selecting one or more RNA sequence reads from the RNA
25 sequence data that contain the tumour-specific mutation comprises selecting one or more RNA sequence reads that overlap with the genomic coordinates of the tumour-specific mutation and that are compatible with the presence of the tumour-specific mutation.

18. The method of any preceding claim, wherein the tumour-specific mutation is ubiquitous in
30 the one or more tumour samples, and/or wherein the tumour-specific mutation is a clonal mutation or a mutation assumed to be clonal in the tumour of the subject.

19. The method of any preceding claim, further comprising:
identifying a plurality of tumour-specific mutations,

35 for each of the plurality of tumour-specific mutations, determining one or more of:

whether the tumour-specific mutations is likely to be clonal in the tumour, and

whether the tumour-specific mutation is likely to be expressed in the tumour, and

selecting a tumour-specific mutation using the result of the determining.

20. The method of any preceding claim, wherein the tumour-specific mutation is a single nucleotide variant, multiple nucleotide variant, insertion, deletion, fusion, or combination thereof.

21. A method of identifying one or more neoantigens in a subject, the method comprising:

identifying a plurality of tumour-specific mutations in the subject;

identifying a plurality of candidate neoantigen peptides associated with one or more of the plurality of tumour-specific mutations using the method of any of claims 1 to 20; and

identifying a neoantigen associated with at least one of the one or more of the tumour-specific mutations, wherein a neoantigen is a candidate neoantigen peptide that satisfies one or more predetermined criteria selected from: the candidate neoantigen peptide being likely to bind to an MHC molecule, the candidate neoantigen peptide being likely to be presented by an MHC molecule, the candidate neoantigen peptide being likely to be immunogenic, the candidate peptide neoantigen peptide being likely to interact with a TCR, the candidate neoantigen peptide being likely to form a stable complex with an MHC molecule, and the candidate neoantigen peptide being likely to be expressed in the tumour.

22. A method of providing an immunotherapy for a subject that has been diagnosed as having cancer, the method comprising:

identifying one or more tumour-specific mutations in the subject;

identifying one or more neoantigen peptides associated with one or more tumour-specific mutations using the method of any of claims 1 to 21;

optionally selecting one or more of the neoantigen peptides identified using one or more criteria that apply to the neoantigen peptides; and

designing an immunotherapy that targets the one or more neoantigens associated with the, optionally selected, one or more neoantigen peptides.

23. A method of treating a subject that has been diagnosed as having cancer, the method comprising: identifying one or more neoantigens by:

identifying a plurality of tumour-specific mutations in the subject;

identifying one or more neoantigen peptides associated with one or more tumour-specific mutations using the method of any of claims 1 to 21;

optionally selecting one or more of the neoantigen peptides using one or more predetermined criteria that apply to the neoantigen peptides; and

treating the subject with an immunotherapy that targets a neoantigen associated with one or more of the optionally selected neoantigen peptides.

24. A system comprising: a processor; and a computer readable medium comprising instructions that, when executed by the processor, cause the processor to perform the steps of any method described herein, such as a method according to any of claims 1 to 22.
- 5 25. One or more non-transitory computer readable media comprising instructions that, when executed by one or more processors, cause the one or more processors to perform the steps of any method described herein, such as a method according to any of claims 1 to 22.

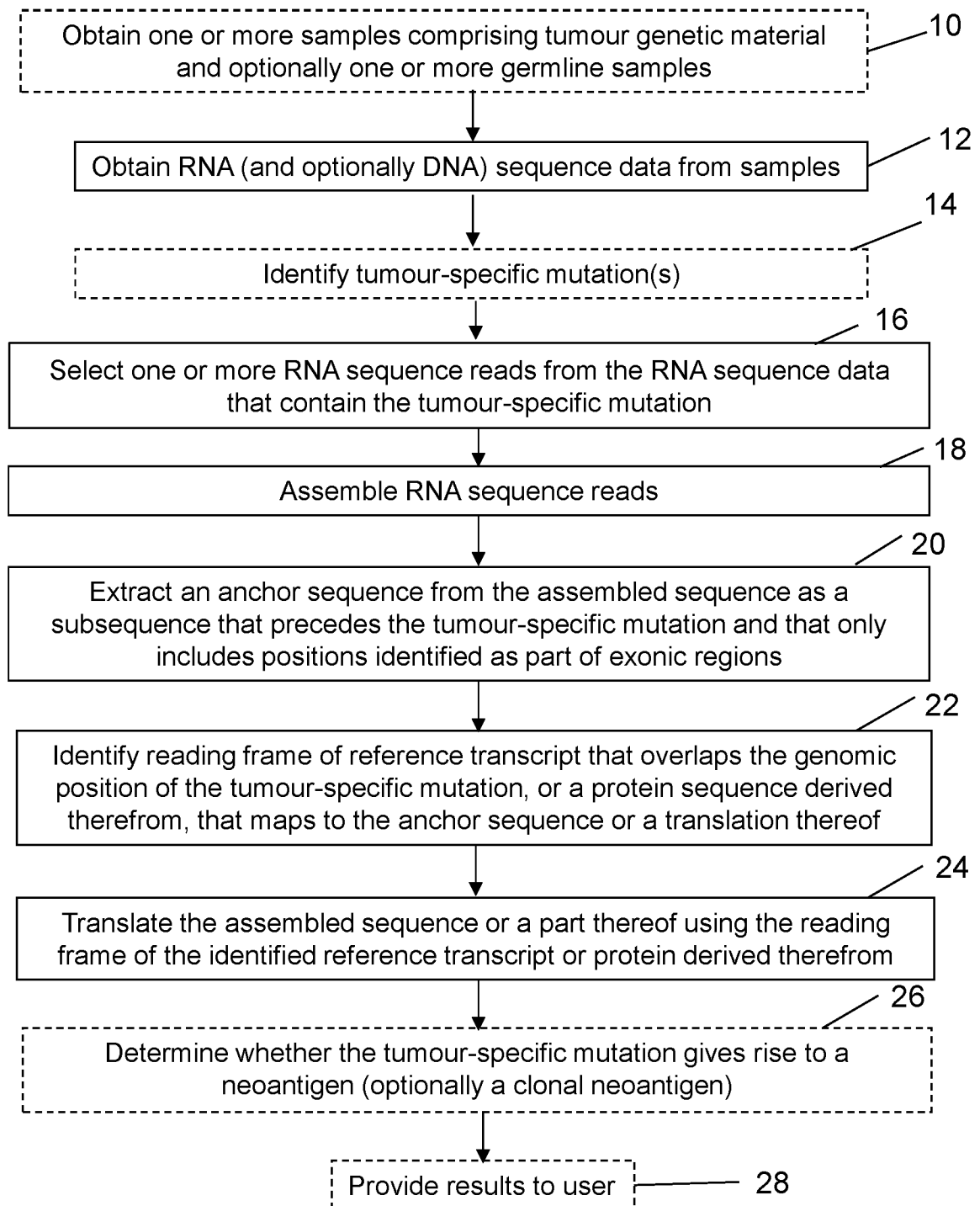


Figure 1

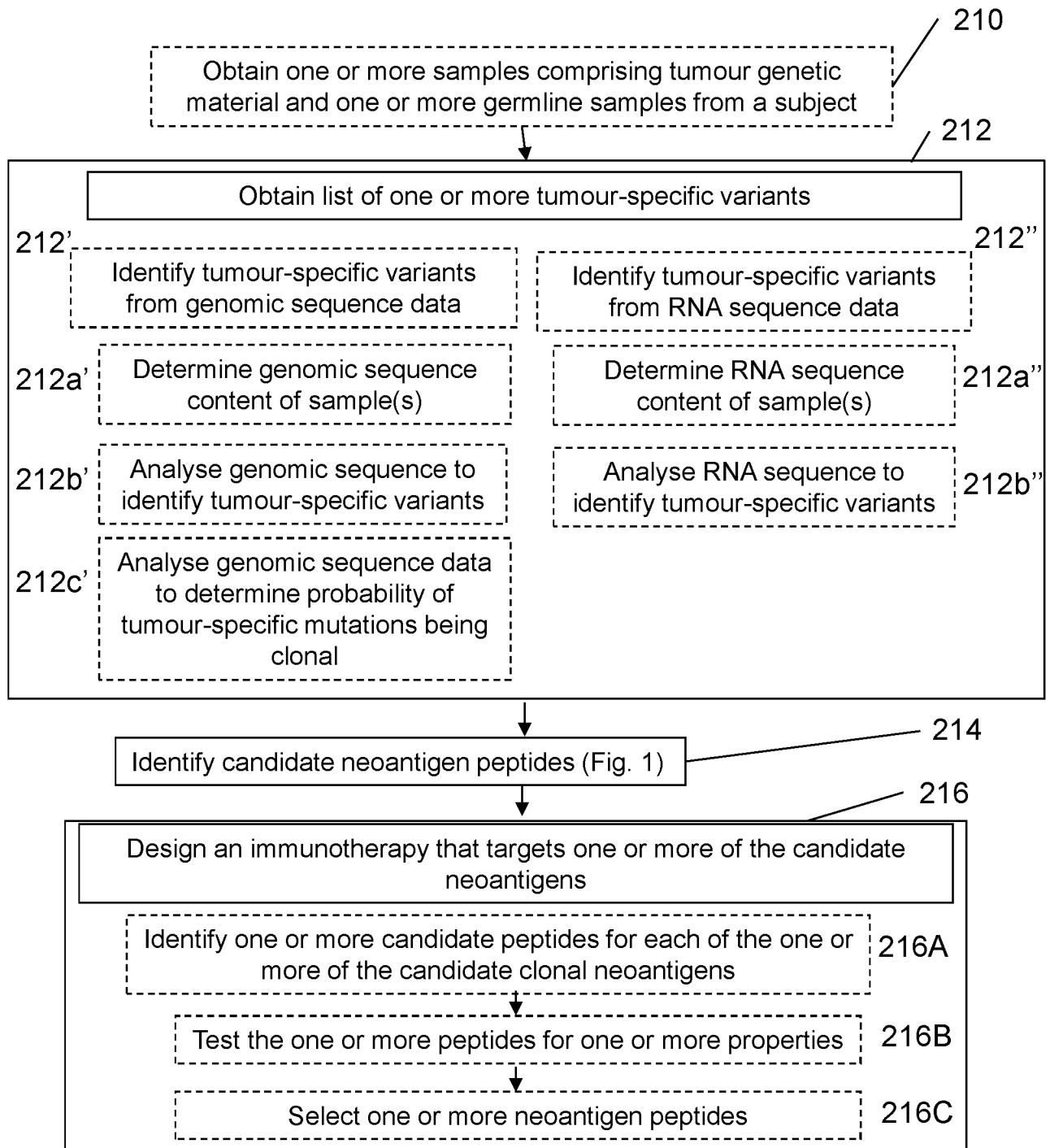


Figure 2A

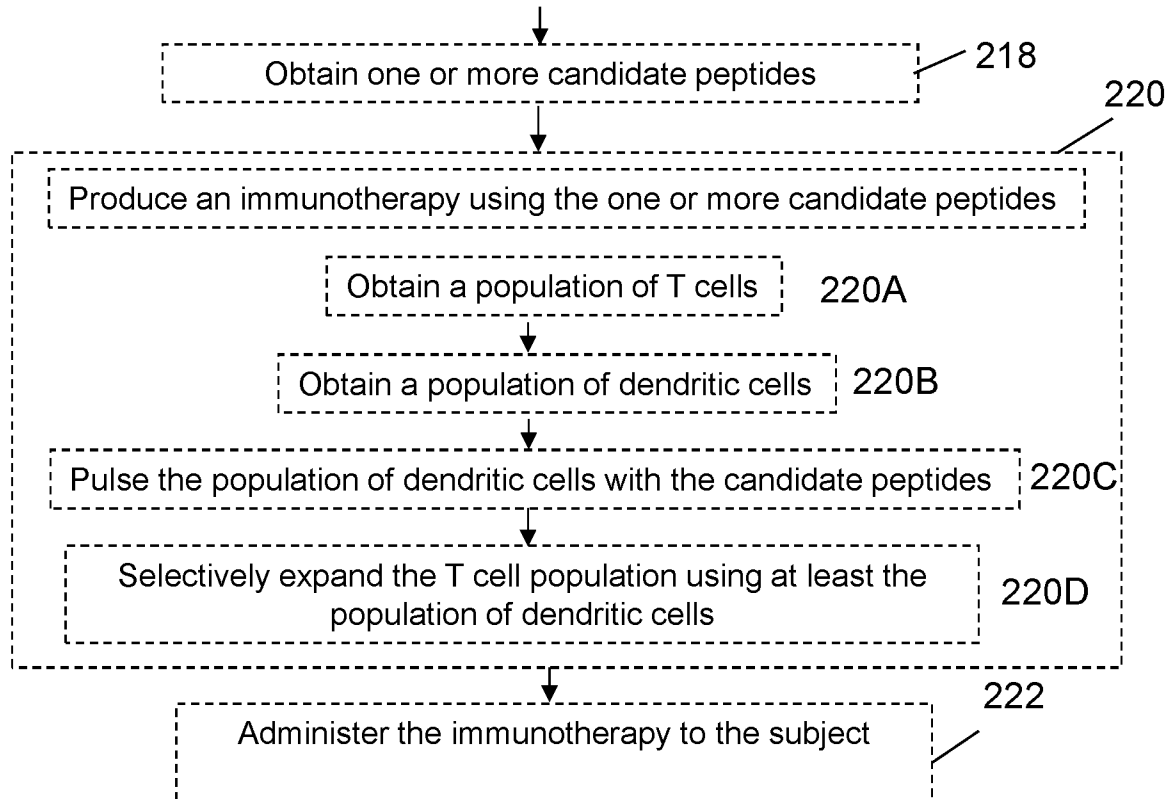


Figure 2B

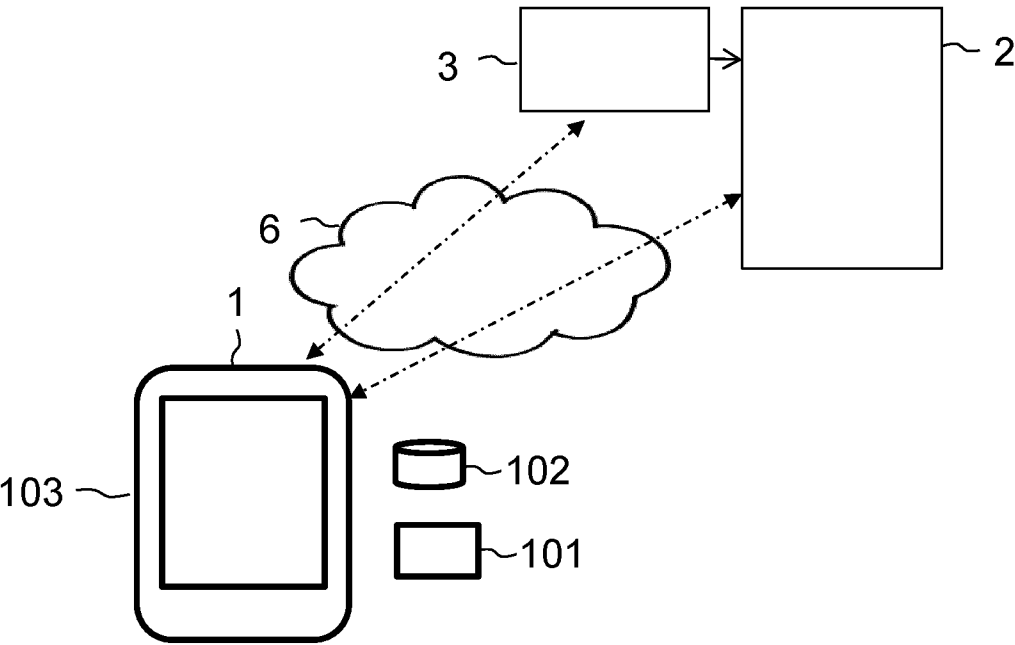


Fig. 3

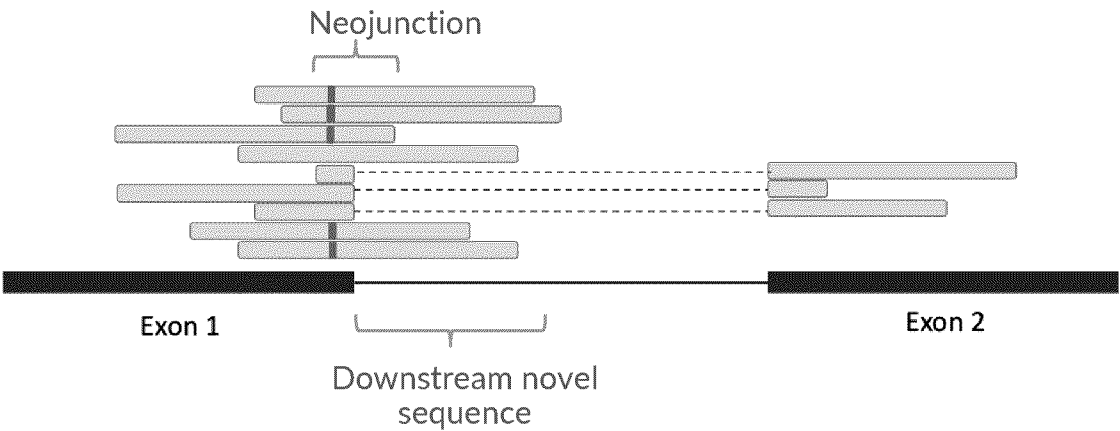


Figure 4

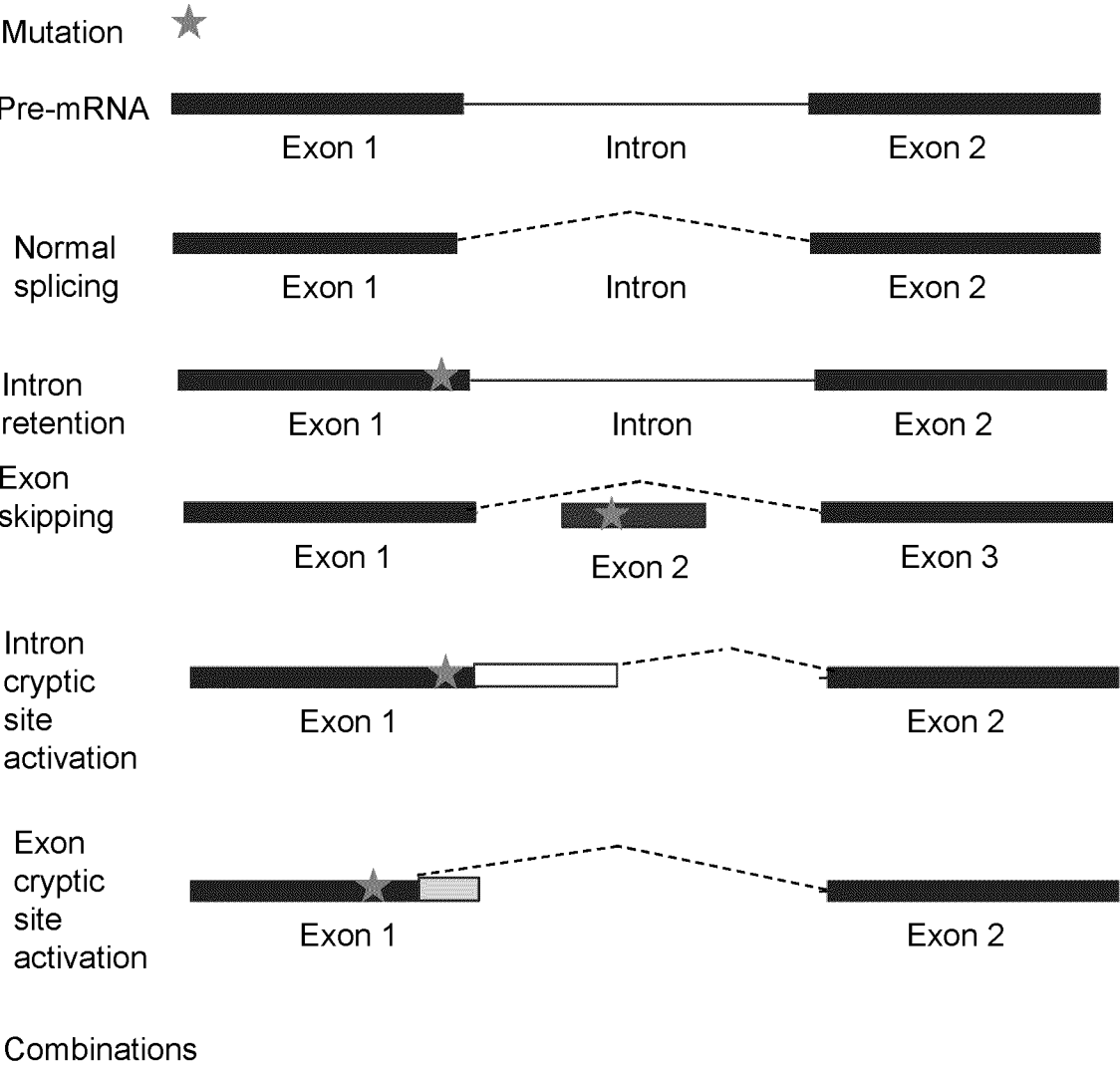


Figure 5

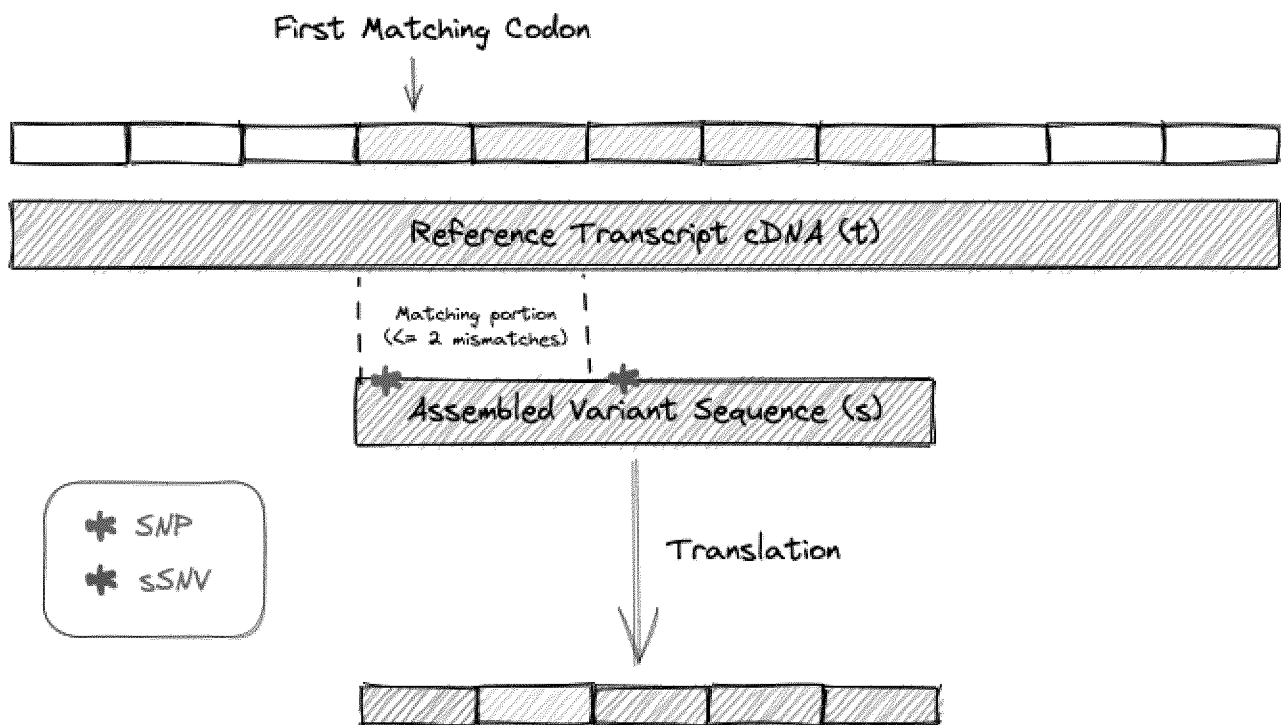


Figure 6A

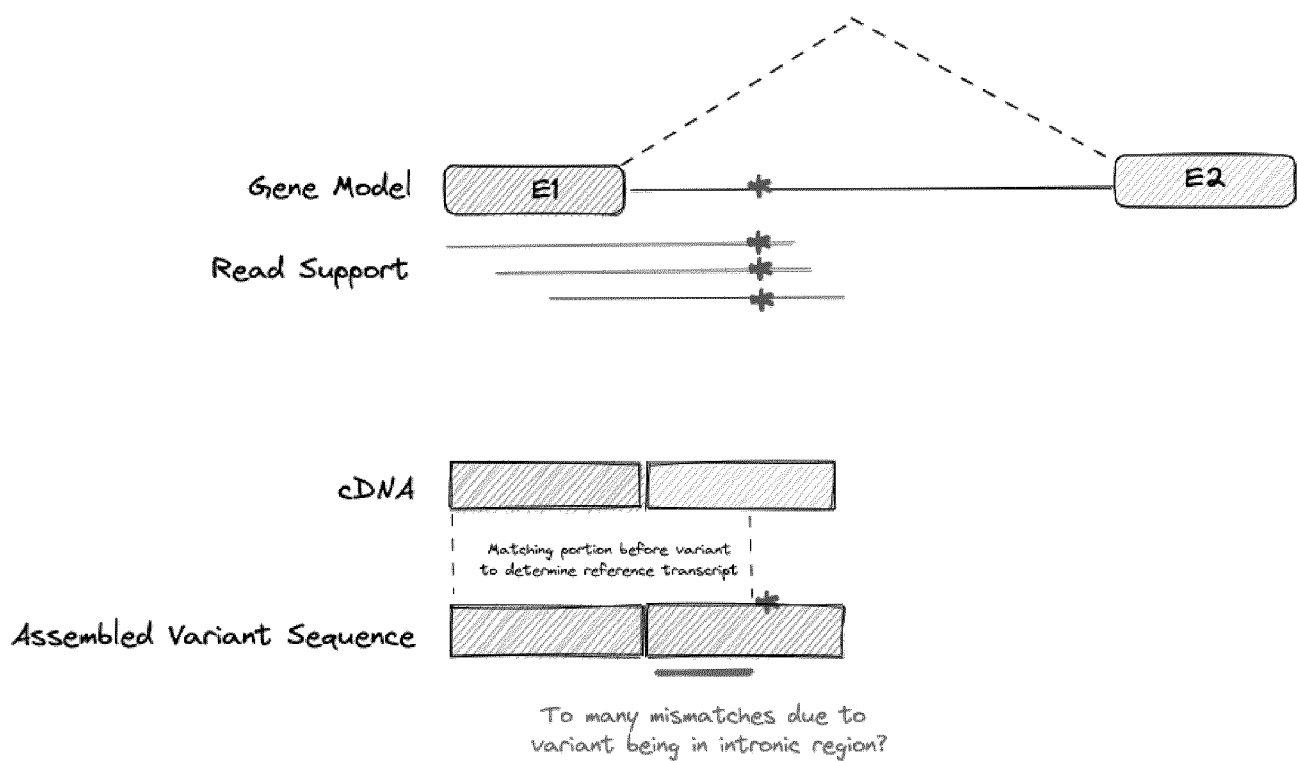


Figure 6B

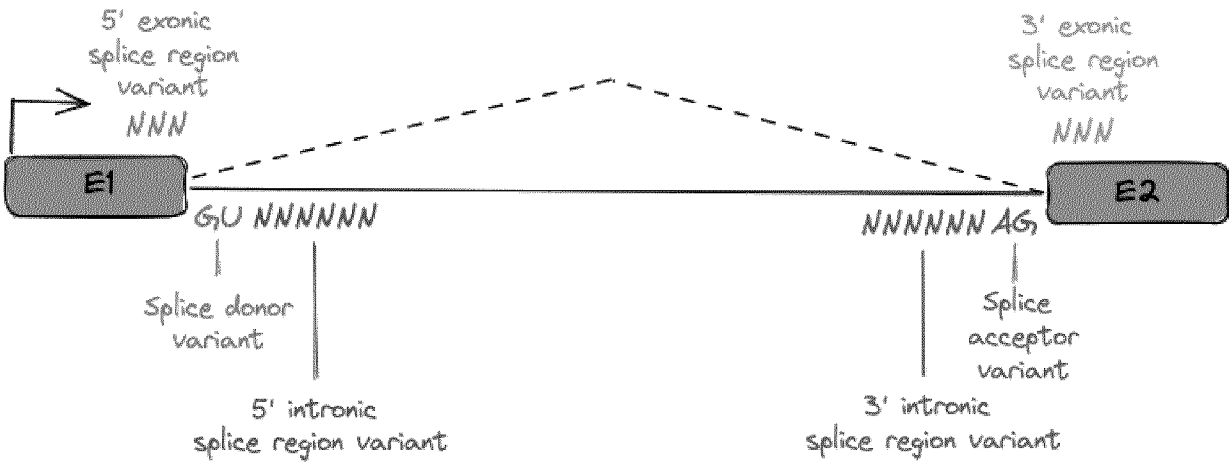


Fig. 7

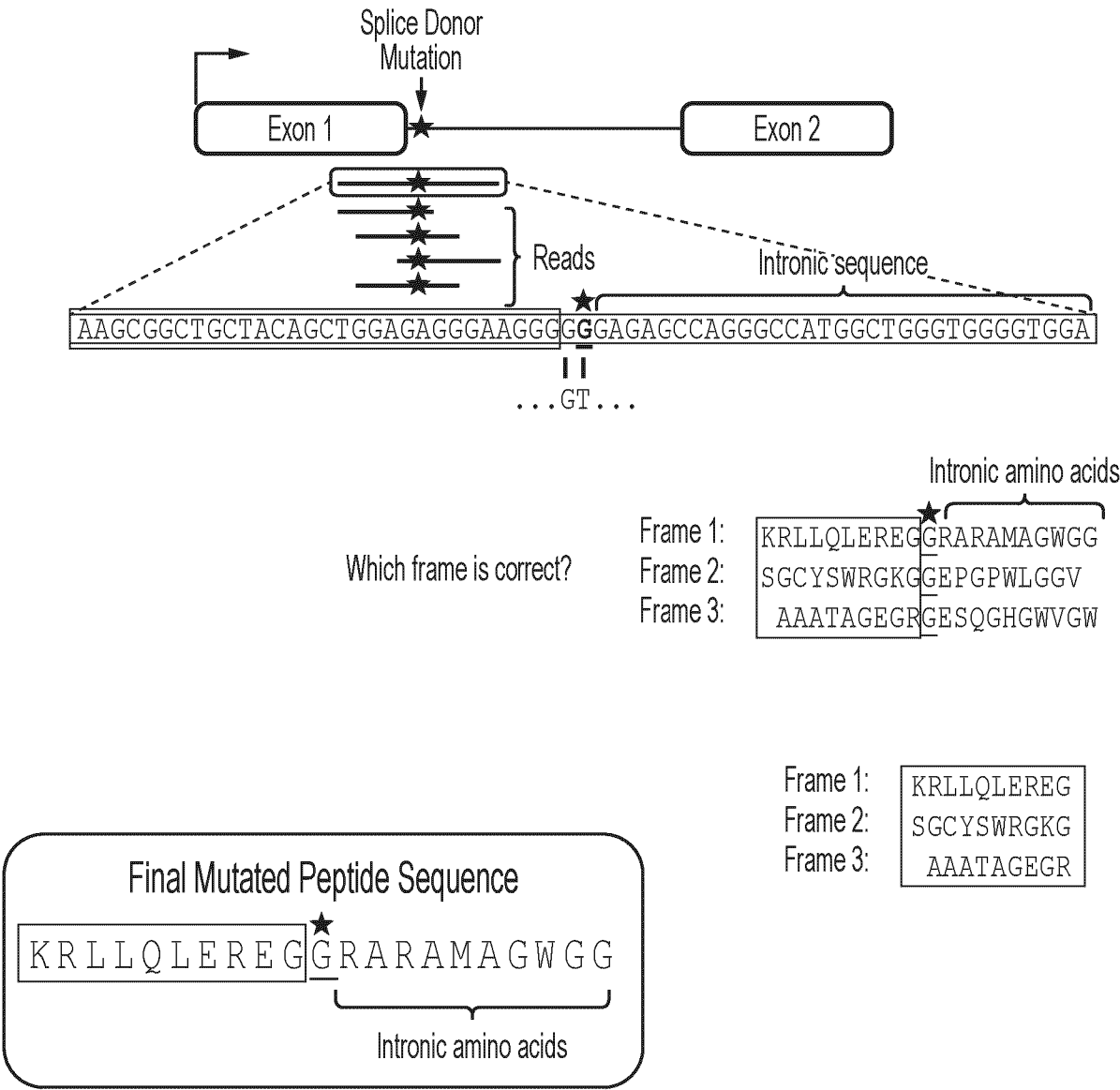


FIG. 8A

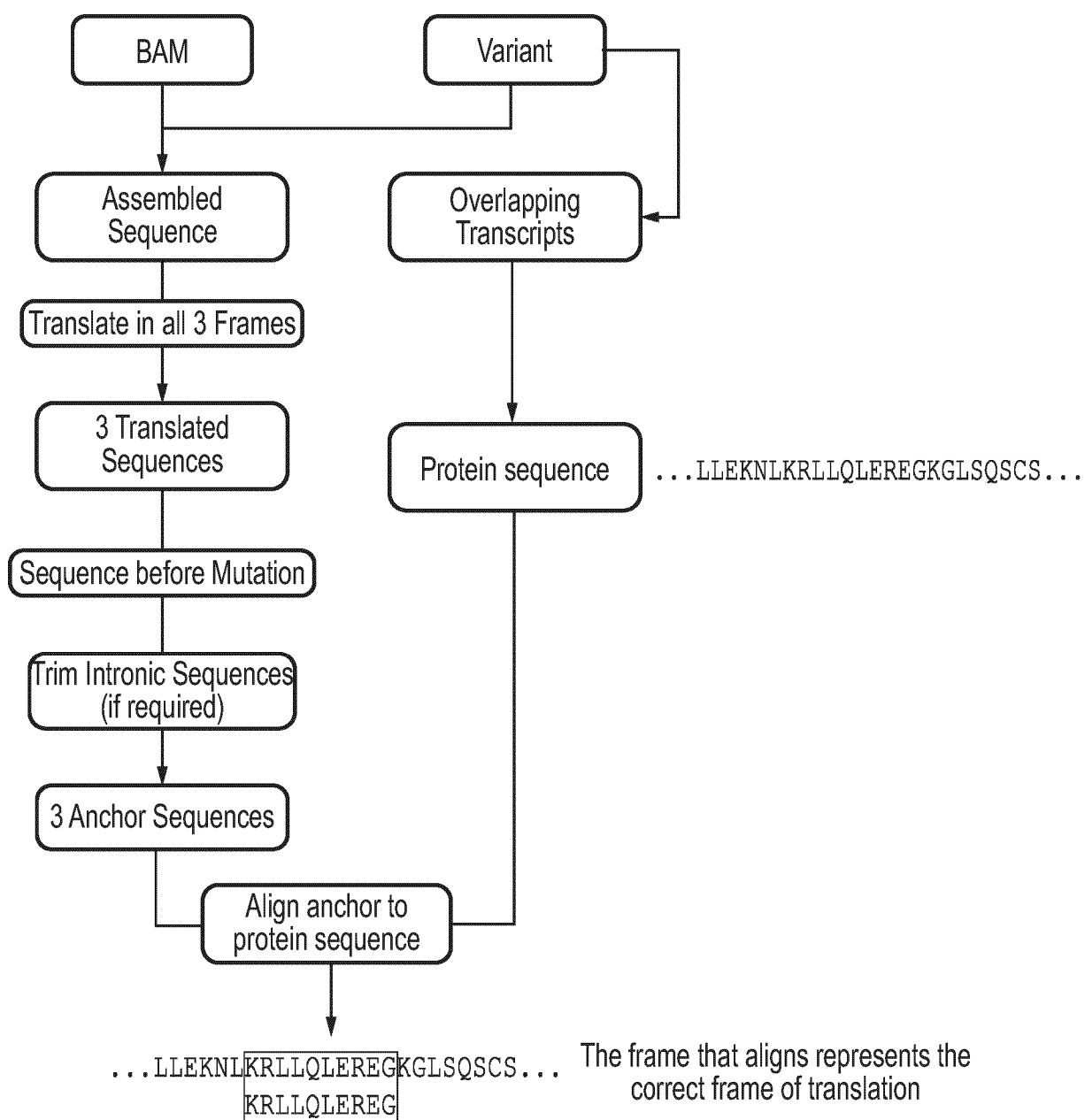


FIG. 8B

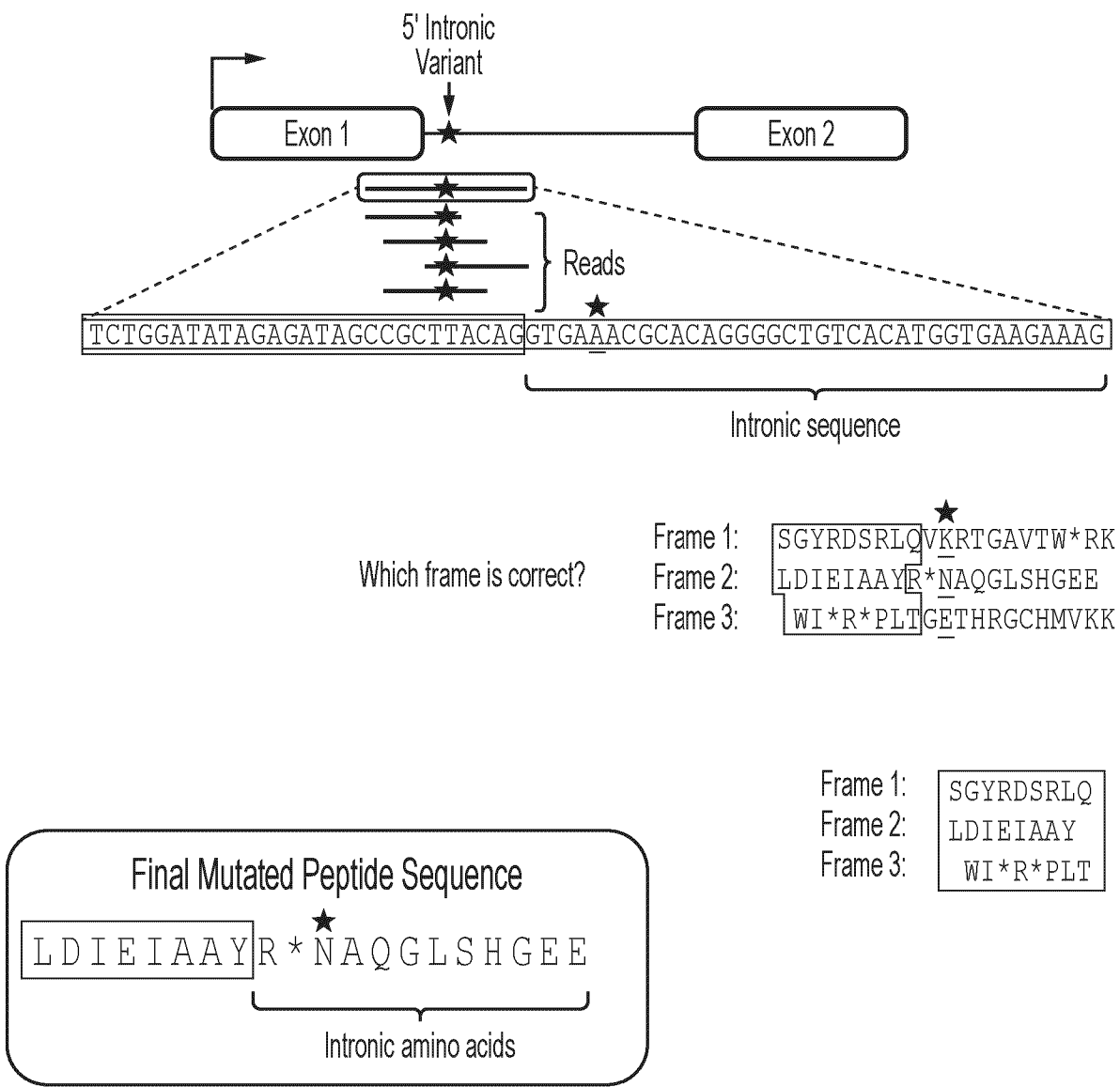


FIG. 8C

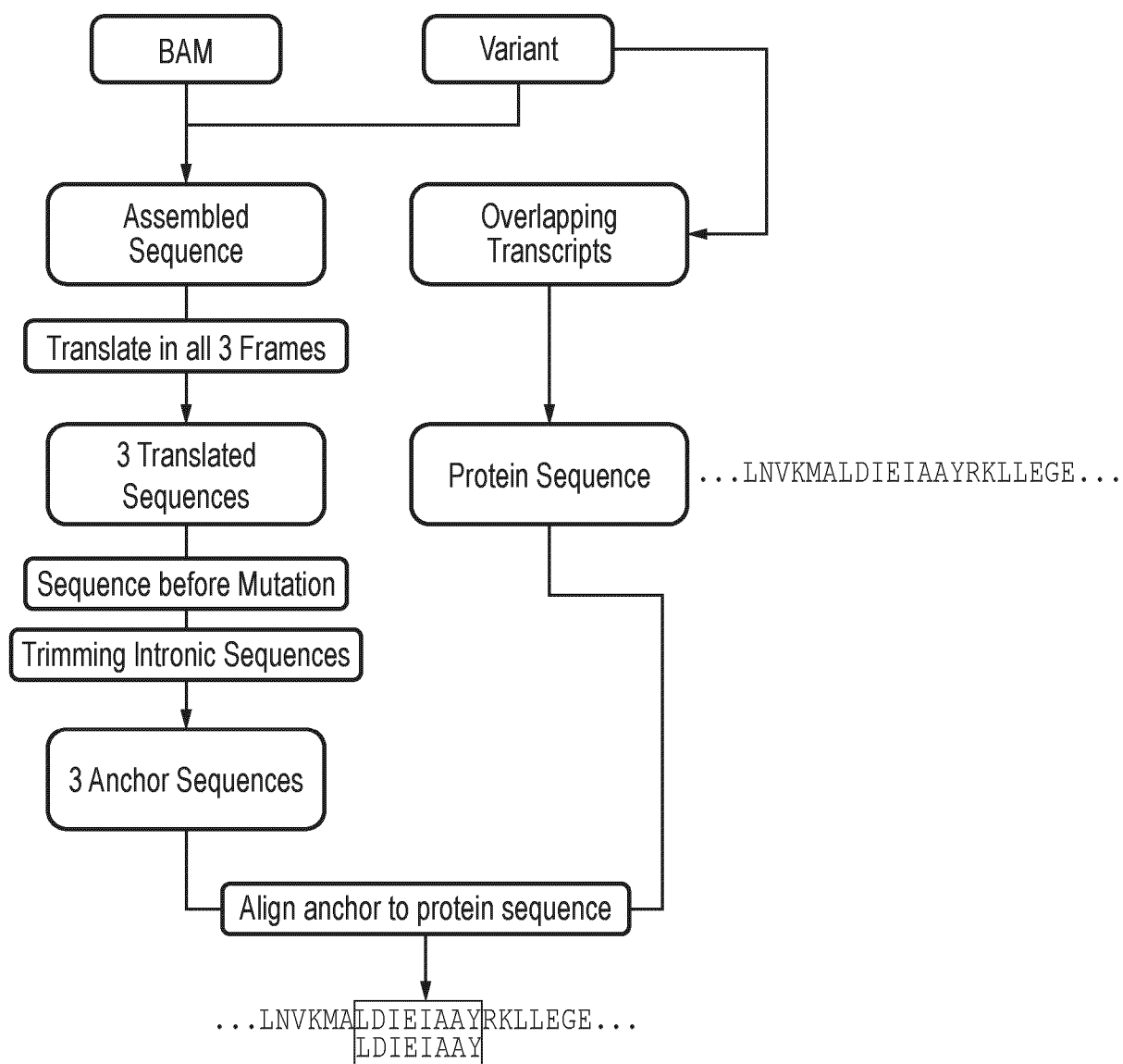


FIG. 8D

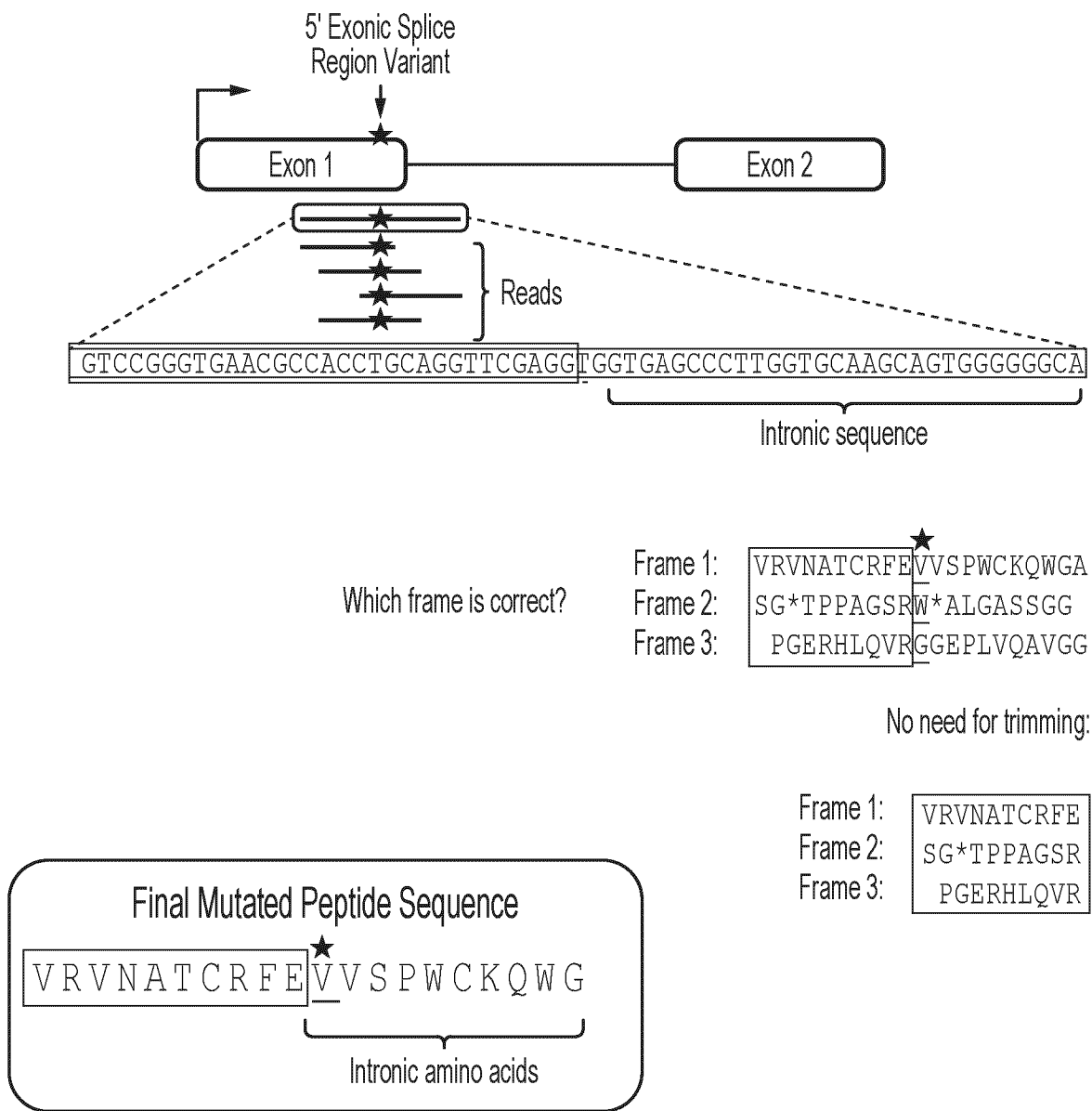


FIG. 8E

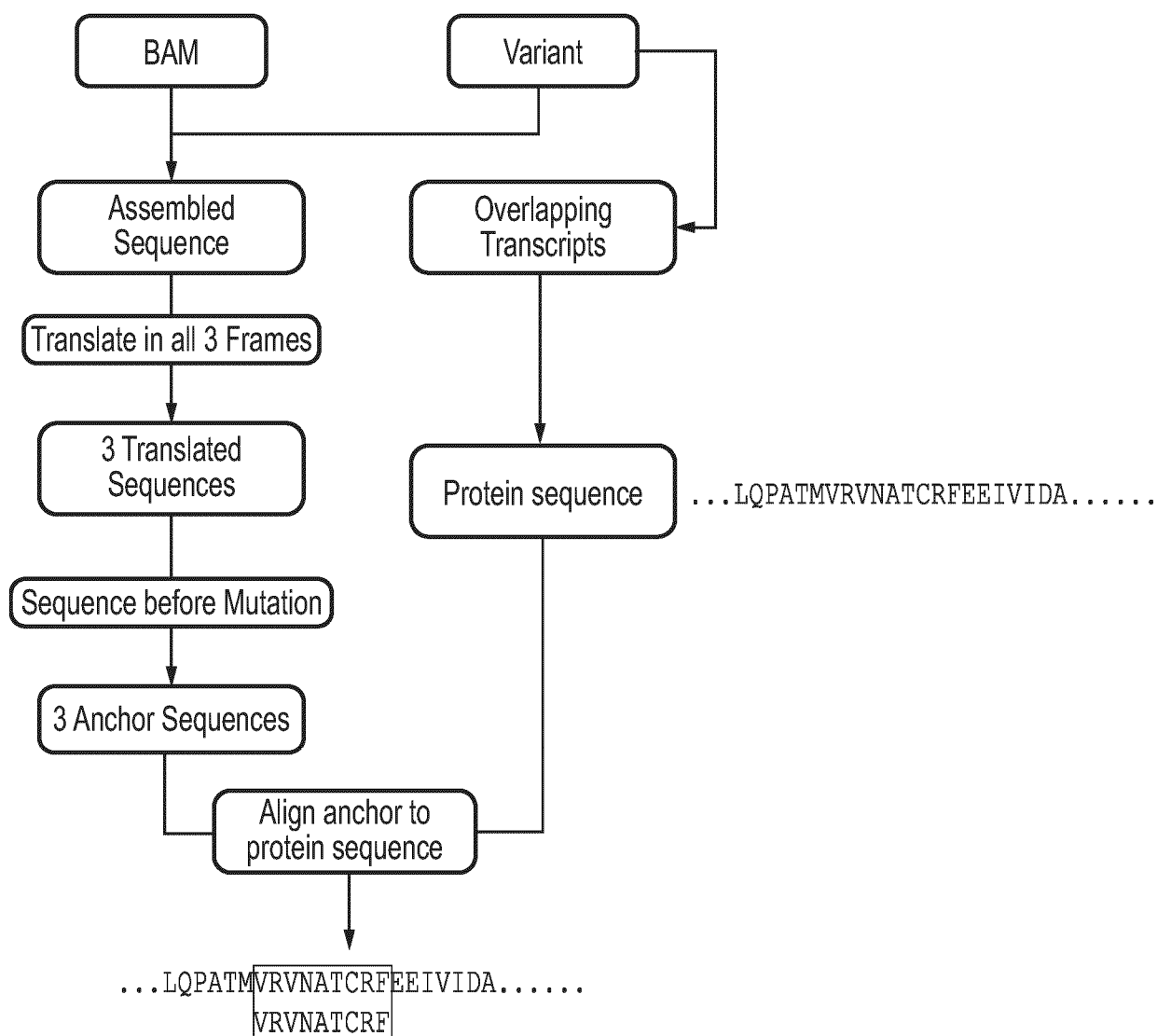


FIG. 8F

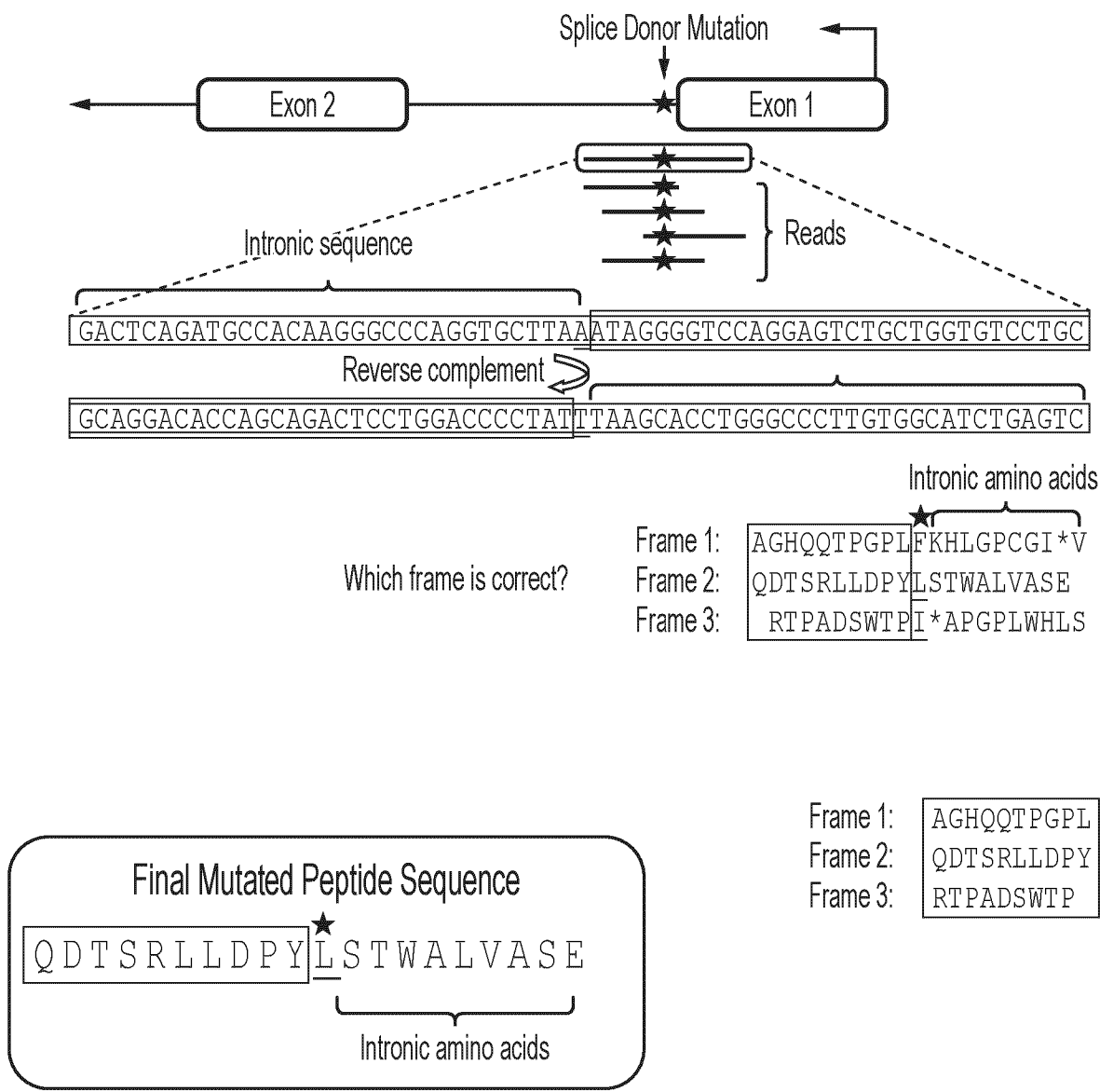


FIG. 8G

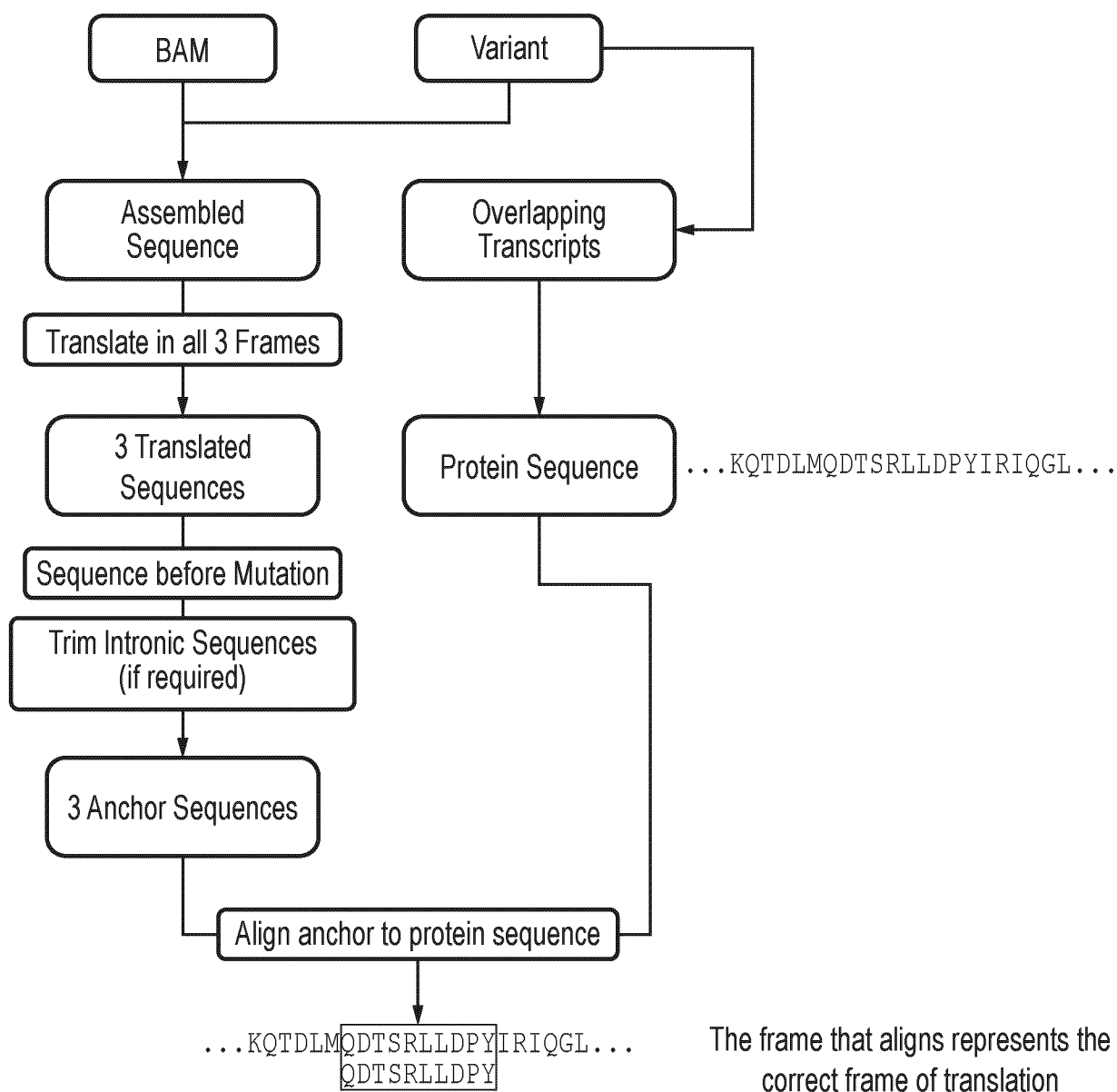


FIG. 8H

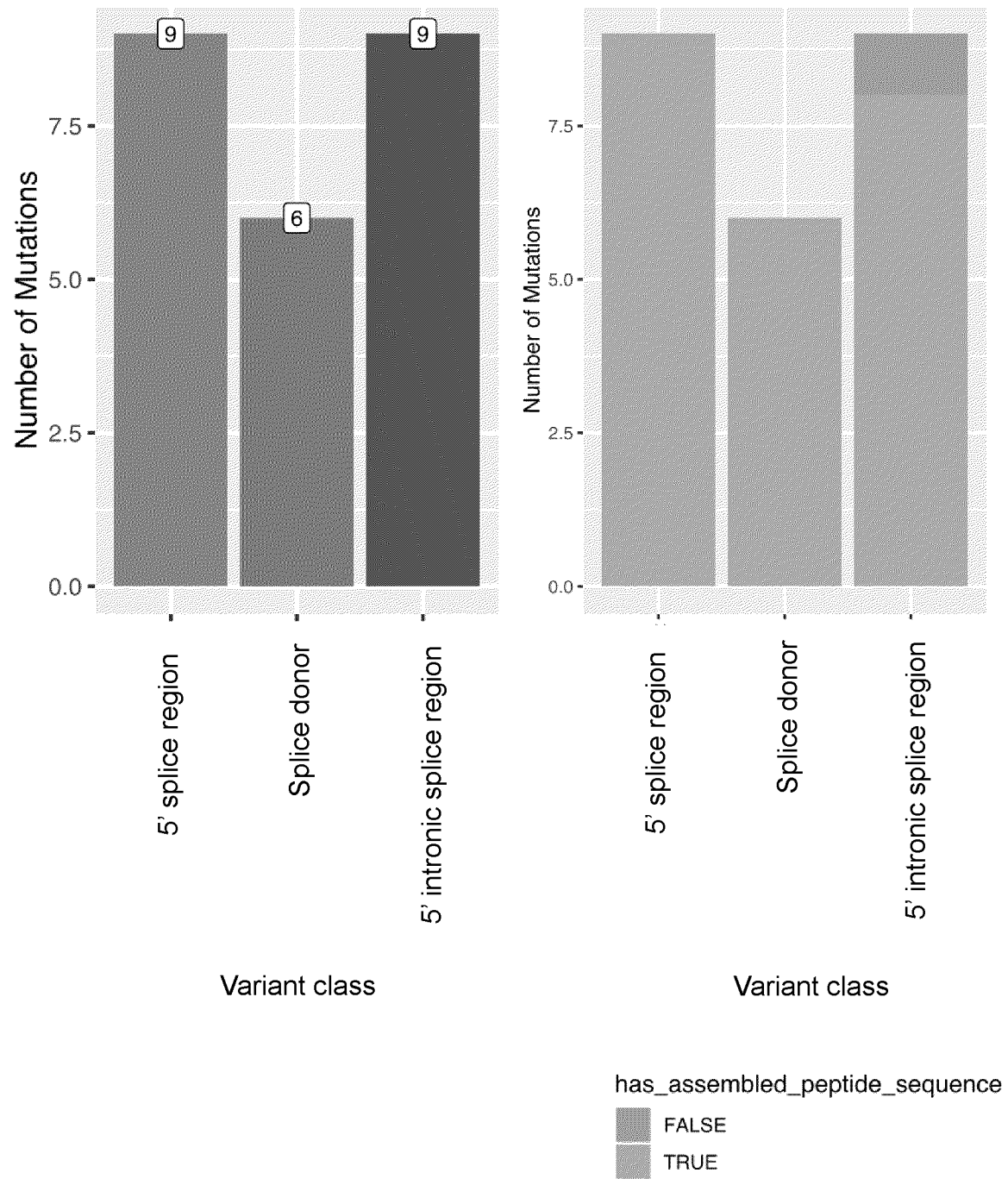
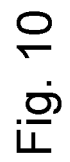


Fig. 9



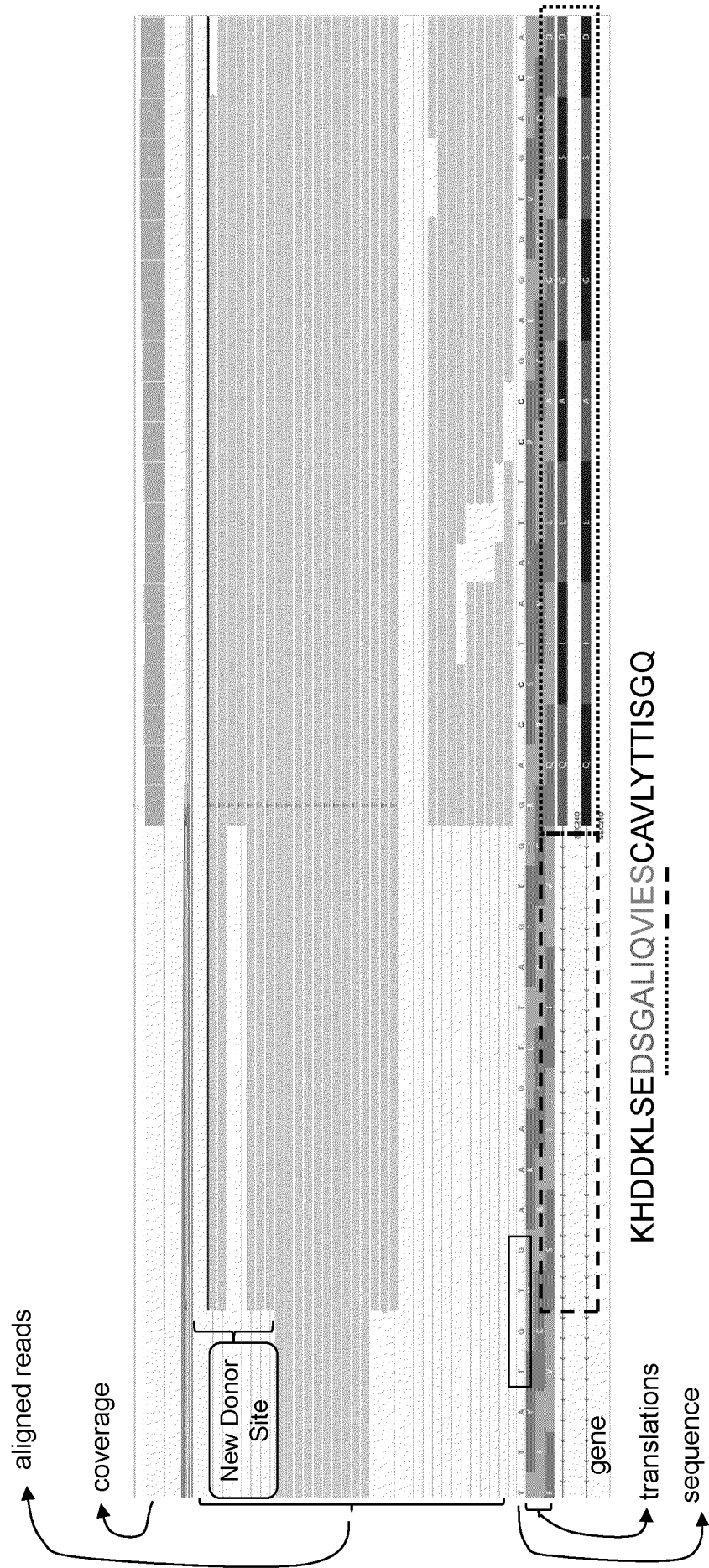


Fig. 11A

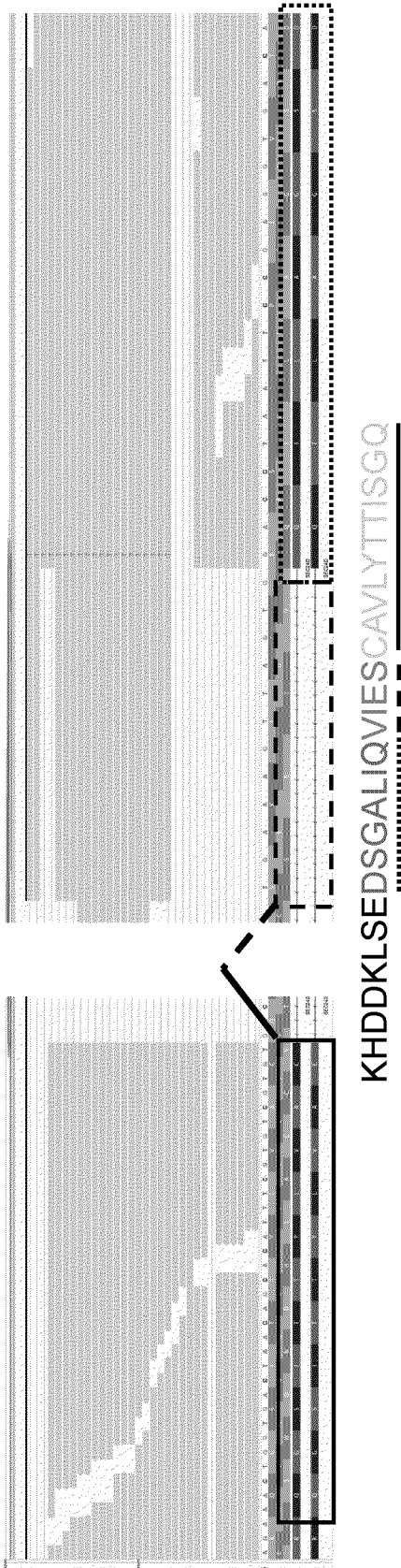


Fig. 11B

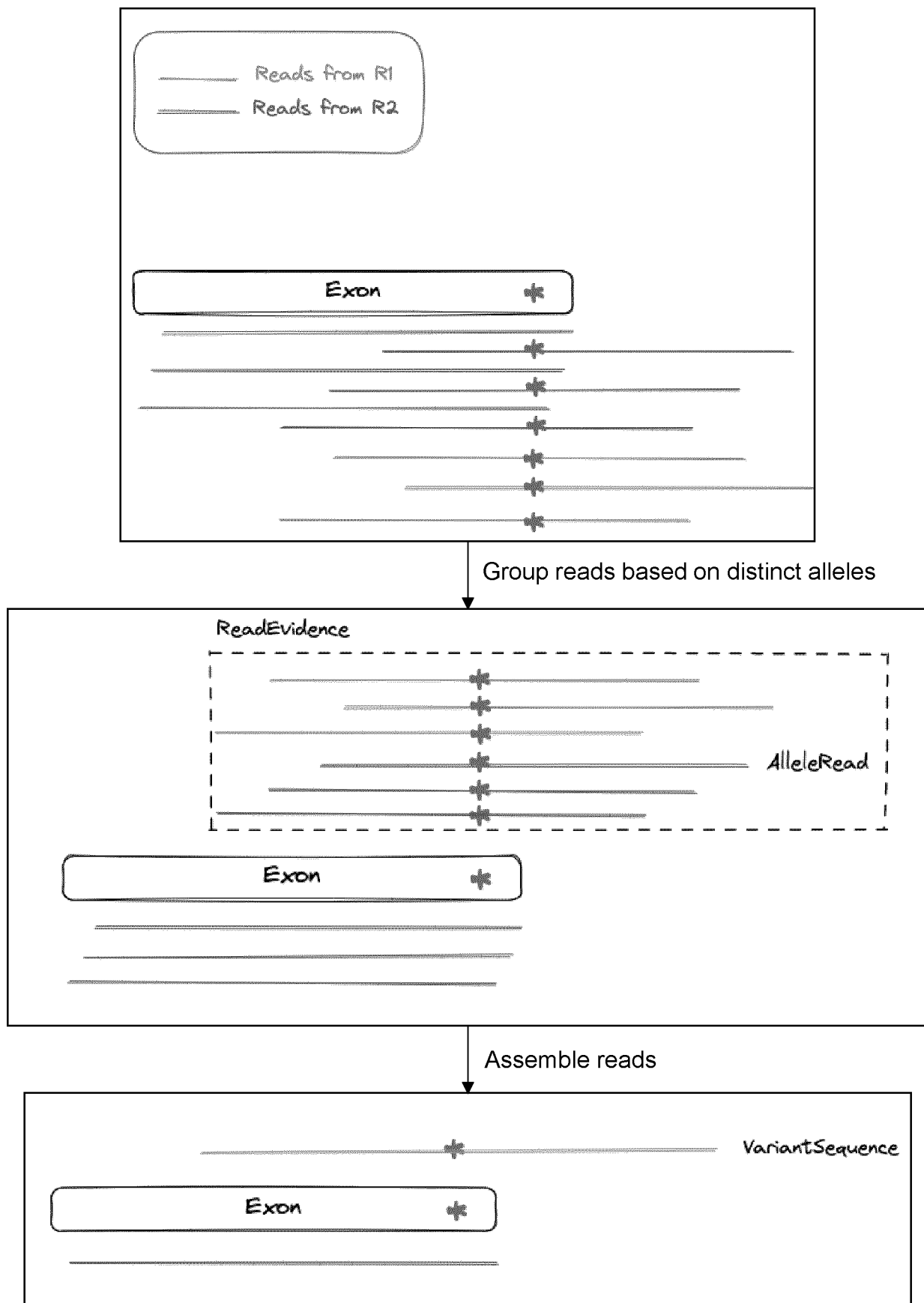


Fig. 12

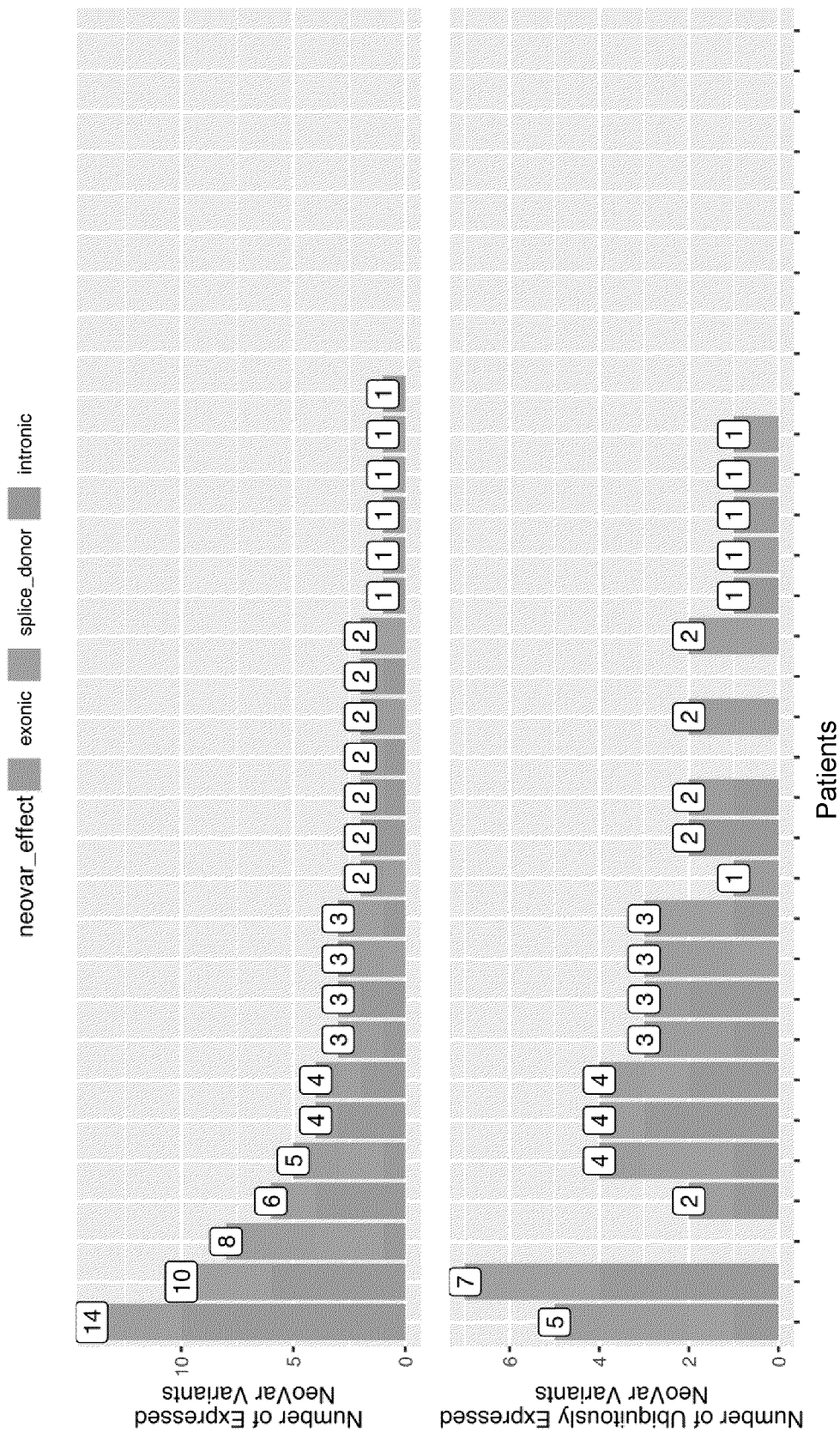


Fig. 13A

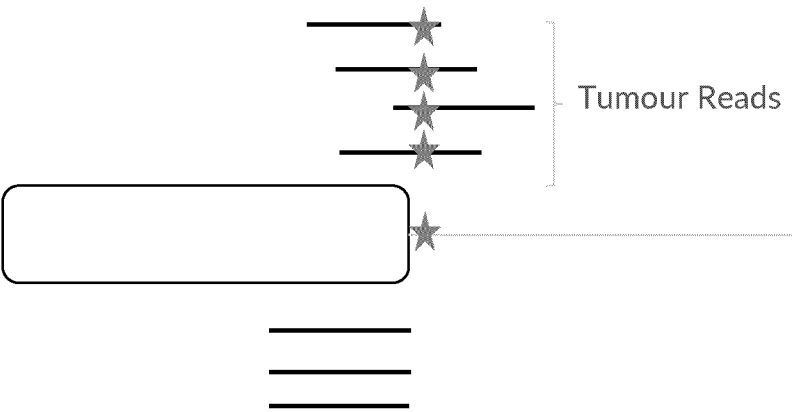
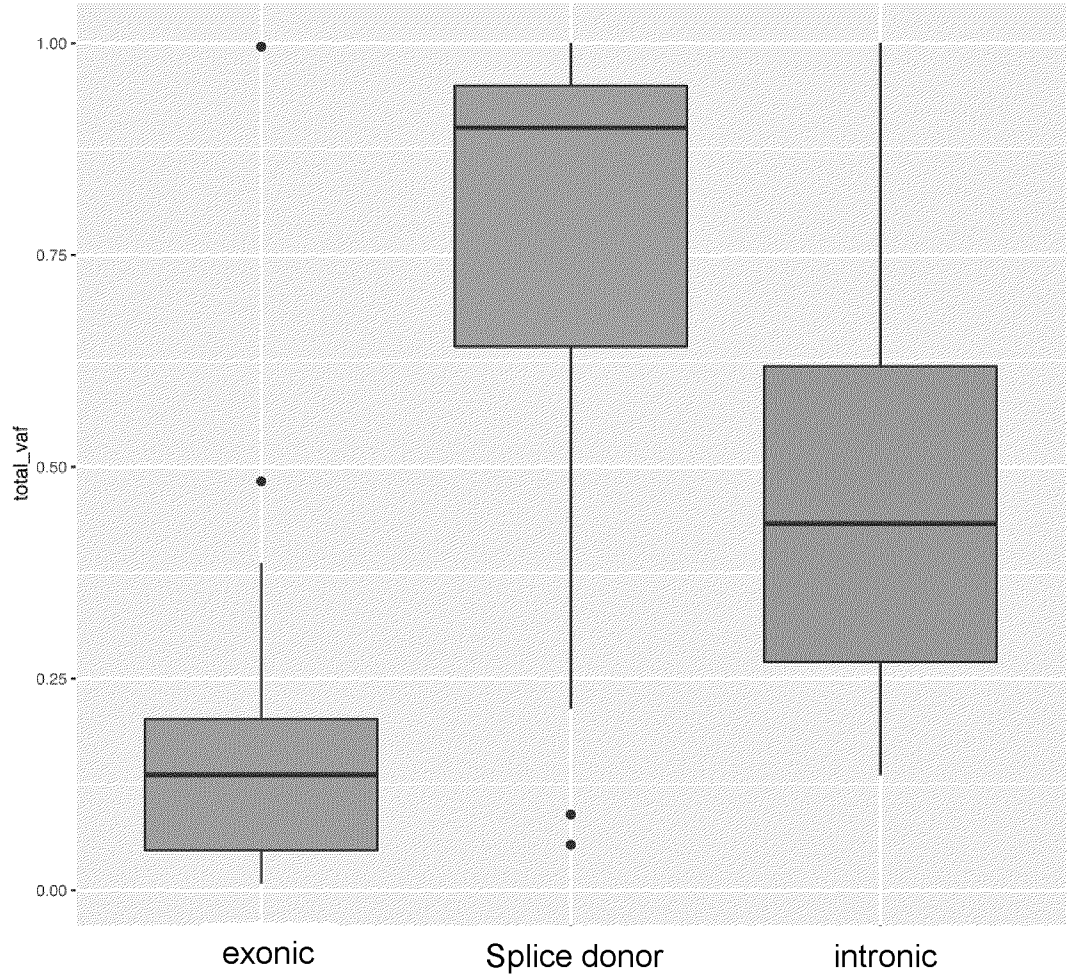


Fig. 13B

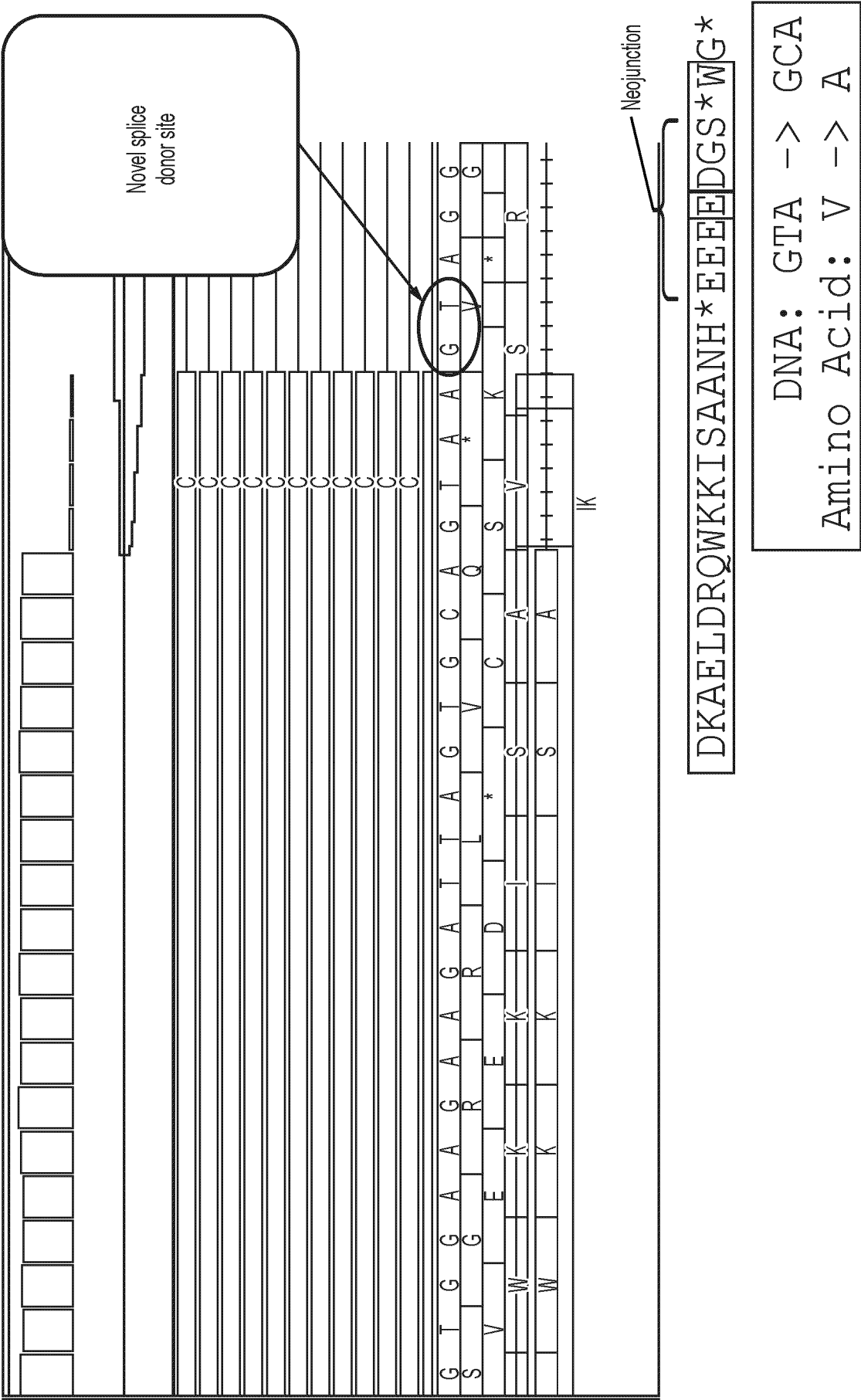


FIG. 13C

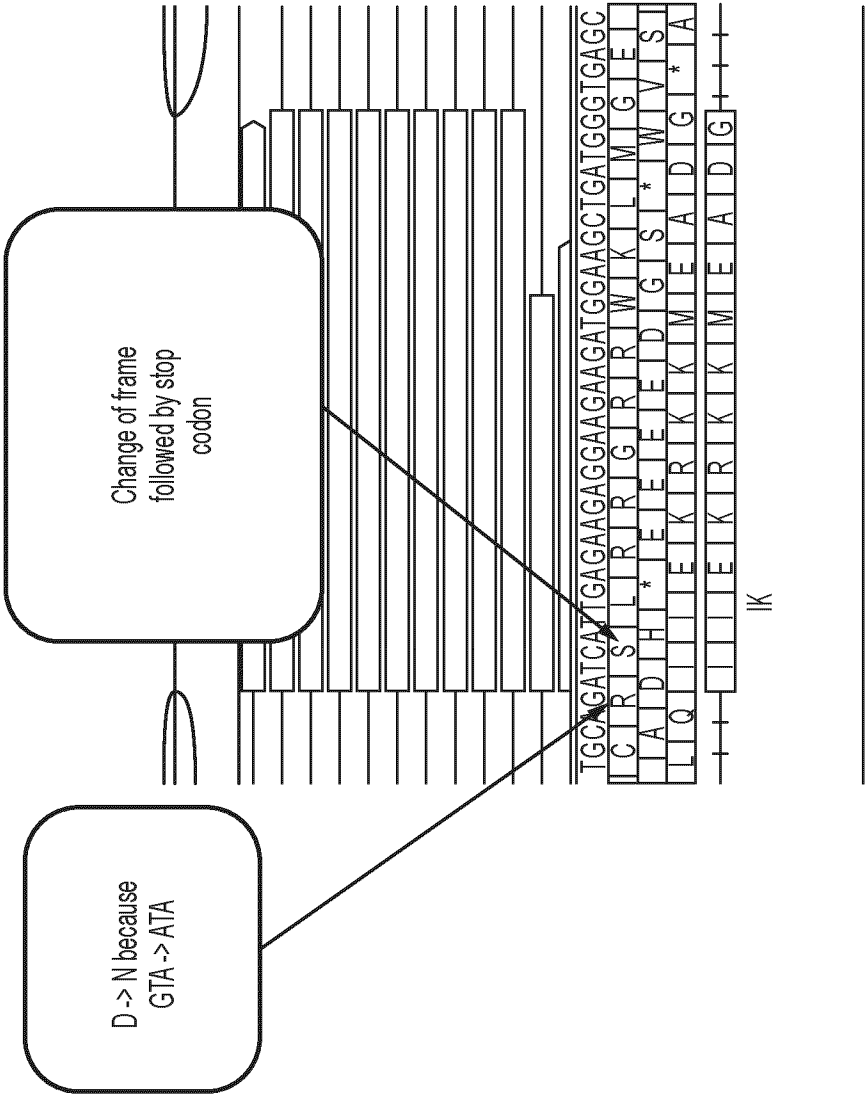


FIG. 13C (continued)

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2024/066271

A. CLASSIFICATION OF SUBJECT MATTER INV. G16B20/20 G16B20/30 G16B30/10 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G16B Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PAN YAANG ET AL: "IRIS: Discovery of cancer immunotherapy targets arising from pre-mRNA alternative splicing", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, 16 May 2023 (2023-05-16), XP093203076, ISSN: 0027-8424 abstract page 2/12, column 2, paragraph 2 page 9/12, column 2, paragraph 2 - page 10/12, column 1, paragraph 5; figure 1 page 10/12, column 2, paragraph 4 ----- -/-	1-25
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 10 September 2024		Date of mailing of the international search report 24/09/2024
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Thumb, Werner

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2024/066271

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2023/091256 A1 (KLOOSTERMAN WIGARD PIETER [NL] ET AL) 23 March 2023 (2023-03-23) abstract paragraphs [0007], [0008], [0016] - [0031], [0036] - [0044], [0180], [0296], [0306], [0307], [0318] - [0321], [0426] - [0434]; figures 3-8; example 15 -----	1 - 25
A	US 2023/071113 A1 (ROTH ANDREW [GB] ET AL) 9 March 2023 (2023-03-09) the whole document -----	1 - 25

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2024/066271

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2023091256 A1	23-03-2023	EP 4110956 A1	04-01-2023
		US 2023091256 A1	23-03-2023
		WO 2021172990 A1	02-09-2021

US 2023071113 A1	09-03-2023	AU 2022251011 A1	05-10-2023
		CA 3213222 A1	06-10-2022
		CN 117120080 A	24-11-2023
		EP 4313125 A1	07-02-2024
		IL 306136 A	01-11-2023
		JP 2024513781 A	27-03-2024
		KR 20230165259 A	05-12-2023
		US 2022323499 A1	13-10-2022
		US 2023071113 A1	09-03-2023
		WO 2022207925 A1	06-10-2022
