(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 June 2015 (11.06.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/085147 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/00* (2011.01)

(21) International Application Number:
PCT/US2014/068746

(22) International Filing Date:
5 December 2014 (05.12.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/912,305 5 December 2013 (05.12.2013) US

(71) Applicants: **THE BROAD INSTITUTE INC.** [US/US]; 415 Main Street, Cambridge, MA 02142 (US). **DANA-FARBER CANCER INSTITUTE, INC.** [US/US]; 450 Brookline Ave., Boston, MA 02215 (US). **THE GENERAL HOSPITAL CORPORATION** [US/US]; 55 Fruit Street, Boston, MA 02114 (US).

(72) Inventors: **SHUKLA, Sachet Ashok**; 12 Cottage Ct., #9, Newton, MA 02458 (US). **WU, Catherine Ju-Ying**; 117 Mason Terrace, Brookline, MA 02446 (US). **GETZ, Gad**; 70 Hoitt Road, Belmont, MA 02478 (US).

(74) Agents: **KOWALSKI, Thomas J.** et al.; Vedder Price P.C., 1633 Broadway, New York, NY 10019 (US).

(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, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, MK, MT, NL, NO, PL, PT, RO, RS, SE, 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))*

(54) **Title**: POLYMORPHIC GENE TYPING AND SOMATIC CHANGE DETECTION USING SEQUENCING DATA

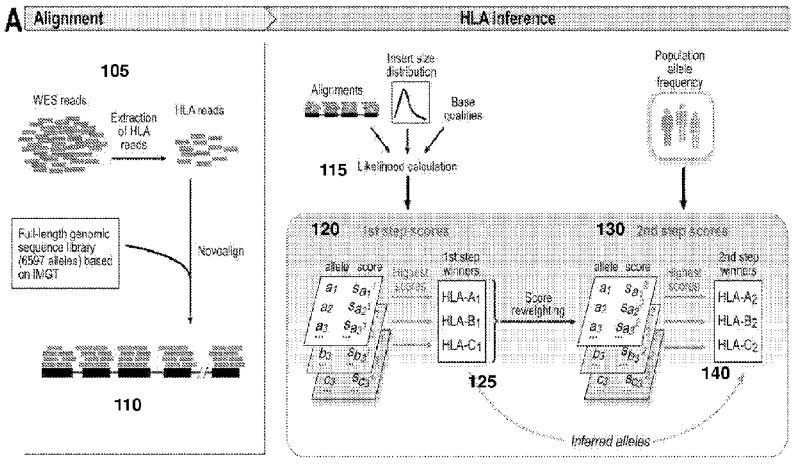

FIG. 1

(57) **Abstract**: A system and method for determining the exact pair of alleles corresponding to polymorphic genes from sequencing data and for using the polymorphic gene information in formulating an immunogenic composition. Reads from a sequencing data set mapping to the target polymorphic genes in a canonical reference genome sequence, and reads mapping within a defined threshold of the target gene sequence locations are extracted from the sequencing data set. Additionally, all reads from the set data set are matched against a probe reference set, and those reads that match with a high degree of similarity are extracted. Either one, or a union of both these sets of extracted reads are included in a final extracted set for further analysis. Ethnicity of the individual may be inferred based on the available sequencing data which may then serve as a basis for assigning prior probabilities to the allele variants. The extracted reads are aligned to a gene reference set of all known allele variants. The allele variant that maximizes a first posterior probability or posterior probability derived score is selected as the first allele variant. A second posterior probability or posterior probability derived score is calculated for reads that map to one or more other allele variants and the first allele variant using a weighting factor. The allele that maximizes the second posterior probability or posterior probability score is selected as the second allele variant. A system and method for identifying somatic changes

[Continued on next page]

WO 2015/085147 A1

in polymorphic loci using WES data. The exact pair of alleles corresponding to the polymorphic gene are determined as described using a normal or germline sample from an individual. A tumor or otherwise diseased sample is also retrieved from the individual and the corresponding WES data is generated. Reads corresponding to the polymorphic gene are extracted as described in the paragraph above. These reads are then aligned to the inferred pair of allele sequences. The alignment of the germline or normal reads to the inferred pair of alleles, along with the alignment of the tumor or diseased reads to the inferred pair of alleles are simultaneously used as inputs to somatic change detection algorithms to identify somatic changes with greater precision and sensitivity.

POLYMORPHIC GENE TYPING AND SOMATIC CHANGE DETECTION USING SEQUENCING DATA

RELATED APPLICATIONS AND INCORPORATION BY REFERENCE

[0001] This application claims benefit of U.S. provisional application Serial No. 61/912,305 filed December 5, 2013.

[0002] The foregoing applications, and all documents cited therein or during their prosecution ("appln cited documents") and all documents cited or referenced in the appln cited documents, and all documents cited or referenced herein ("herein cited documents"), and all documents cited or referenced in herein cited documents, together with any manufacturer's instructions, descriptions, product specifications, and product sheets for any products mentioned herein or in any document incorporated by reference herein, are hereby incorporated herein by reference, and may be employed in the practice of the invention. More specifically, all referenced documents are incorporated by reference to the same extent as if each individual document was specifically and individually indicated to be incorporated by reference.

FIELD OF THE INVENTION

[0003] This disclosure relates generally to gene typing and mutation detection of polymorphic genes using sequencing data, including whole exome sequencing data.

FEDERAL FUNDING LEGEND

[0004] This invention was made with government support under 1R01CA155010-04 awarded by NIH/NCI. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

[0005] The human genome comprises multiple highly polymorphic gene loci such as the Human Leukocyte Antigen (HLA) locus. Human leukocyte antigens (HLAs) are highly polymorphic proteins that present peptides to T cell receptors to initiate the adaptive immune response and to set the boundaries between self and non-self. Exact determination of an individual's gene type for these highly polymorphic genes has numerous applications including identification of compatible organ donors, understanding autoimmunity and immune biology, and design of personalized medicines. Gene typing is typically a focused effort informed by directed experimental protocols. This is commonly performed by sequencing exons 2–4 of Class

I genes (HLA-A, -B and -C) and exons 2 and/or 3 of Class II genes (HLA-DRB1 and -DQB1) (Chang et al., ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Research, 2013, 1–8; Lind et al., Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum. Immunol. 2010;71:1033-1042.). Due to the extreme diversity of HLA alleles in the population, sequence ambiguities frequently arise when the polymorphisms are outside the regions being typed and when different allelic combinations share the same sequence. Additional steps such as polymerase chain reaction (PCR) with sequence-specific primers are necessary to resolve these ambiguities (Erlich, HLA DNA typing: past, present, and future. Tissue Antigens. 2012;80:1-11). Although this workflow determines the HLA genotypes at high resolution, it is laborious and expensive.

[0006] Next-generation sequencing has been applied to sequencing short-range amplicons of informative exons (Gabriel C, et al., Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum. Immunol. 2009;70:960-964; Bentley et al., High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens. 2009;74:393-403.) It has also been applied to sequencing long-range amplicons of whole HLA genes on various platforms (Erlich, et al., Next-generation sequencing for HLA typing of class I loci. BMC Genomics 2011;12:42; Wang et al., High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc. Natl Acad. Sci. USA 2012;109:8676-8681; Shiina et al., Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens 2012;80:305-316.), suggesting a potential for parallel high-throughput HLA typing. Illumina sequencing of captured HLA genes is a cost-effective alternative that can bypass long-range PCRs. However, this is challenging because reads specific to target HLA genes are not readily available, read coverage may vary substantially among different exons and between heterozygous alleles owing to capturing bias, and the typical short read length and the level of polymorphism within the region increase the difficulty of differentiating near-identical alleles. Currently, there is no method to reliably accomplish this task given the challenges. Moreover, poor allelic HLA typing results from exome-seq data even at high coverage has been demonstrated (Warren et al., Derivation of HLA types from shotgun sequence datasets. Genome Med. 2012;4:95).

[0007] Whole exome sequencing (WES, capture sequencing), is a widely used technique for high-throughput sequencing of the coding regions of genes across the genome. Although the use of WES as a research and clinical tool is expanding, the non-specificity and relative low-fidelity of WES compared to directed experimental protocols makes it challenging to use this strategy for gene typing. Gene typing must be generated de novo for each subject. Accordingly, methods for producing high precision polymorphic gene typing from these types of sequencing data, such as WES data, are needed.

[0008] Citation or identification of any document in this application is not an admission that such document is available as prior art to the present invention.

SUMMARY OF THE INVENTION

[0009] In one aspect this disclosure is directed to methods for determining polymorphic gene types that may comprise generating an alignment of reads extracted from a sequencing data set to a gene reference set comprising allele variants of the polymorphic gene, determining a first posterior probability or a posterior probability derived score for each allele variant in the alignment, identifying the allele variant with a maximum first posterior probability or posterior probability derived score as a first allele variant, identifying one or more overlapping reads that aligned with the first allele variant and one or more other allele variants, determining a second posterior probability or posterior probability derived score for the one or more other allele variants using a weighting factor, identifying a second allele variant by selecting the allele variant with a maximum second posterior probability or posterior probability derived score, the first and second allele variant defining the gene type for the polymorphic gene, and providing an output of the first and second allele variant.

[0010] In certain example embodiments the sequencing data set is from massively parallel sequencing. This includes any high-throughput approach to DNA sequencing using the concept of massively parallel processing. That is technologies utilizing parallelized platforms for sequencing more than about 1 million to 43 billion short reads (50-400 bases each) per instrument run. In more specific embodiments the sequencing data is from massive parallel sequencing via spatially separated, clonally amplified DNA templates or single DNA molecules in a flow cell. In preferred embodiments sequencing data is whole exome sequencing data, RNA-Seq data, whole genome data, or targeted exome sequencing data.

[0011] In certain example embodiments, the reads in the sequencing data set may consist of reads that map to a reference genetic sequence of the polymorphic gene within a threshold base number value. The threshold base number value may be between approximately 0.5 Kb and approximately 5 Kb. In one example embodiment, the threshold base number value is 1 Kb.

[0012] In certain embodiment reads are extracted from the sequencing data set. In one embodiment the data is extracted by assembly of the short sequences de novo. In another embodiment the data is extracted by mapping to a known sequence from a subject of the same species.

[0013] In certain embodiments an alignment is generated using the extracted reads. The alignment may utilize a non-naturally occurring reference genetic sequence. The reference genetic sequence may be constructed from a library of known or inferred genomic and or cDNA sequences of the polymorphic gene or polymorphic genes to be typed. In one embodiment every extracted read is aligned with every sequence within the reference library. In certain example embodiments, the reads in the sequencing data set may consist of reads that match one or more sequences from a reference genetic sequence. The reads may match one or more sequences from the reference genetic sequence in the 5' to 3' direction or the 3' to 5' orientation. In certain example embodiments, the reads have between approximately 90% and approximately 100% sequence identity to one or more sequences in the reference genetic sequence. In one example embodiment the reads have approximately 100% sequence identity to one or more sequences in the reference genetic sequence.

[0014] In certain example embodiments, the reads in the sequencing data set may consist of reads that match one or more probes from a polymorphic gene probe set. The reads may match one or more probes from the polymorphic gene set in the 5' to 3' direction or the 3' to 5' orientation. In one example embodiment, the probes are derived from a library of known or inferred genomic and or cDNA sequences of the polymorphic gene. In certain example embodiments, the reads have between approximately 90% and approximately 100% sequence identity to one or more probes in the polymorphic gene probe set. In one example embodiment the reads have approximately 100% sequence identity to one or more probes in the polymorphic gene probe set. In certain example embodiments, the probes in the polymorphic gene probe set have a size between approximately 25 mer and approximately 100 mer. In one example embodiment, the probes in the polymorphic gene probe set have a size of 38 mer. In another

example embodiment, the probes in the polymorphic gene probe set have a size equal to half the read length in the sequencing experiment.

[0015] In certain embodiments the sequencing data sets, reference genetic sequence, or polymorphic gene probe set correspond to an animal. In one embodiment the sequencing data sets, reference genetic sequence, or polymorphic gene probe set correspond to a mammal. In another embodiment the sequencing data sets, reference genetic sequence, or polymorphic gene probe set correspond to a rodent. In a preferred embodiment the sequencing data sets, reference genetic sequence, or polymorphic gene probe set correspond to a human.

[0016] In certain example embodiments, the first and second posterior probability or posterior probability derived scores are determined based at least in part on base quality scores and an insert size probability value for each read in the alignment. The insert size probability value may be based at least in part on an insert size distribution of all reads in the data set. In one example embodiment, the first and second posterior probabilities or posterior probability derived scores are calculated based at least in part on population-based allele probability observed in a known population data set.

[0017] In certain example embodiments, the weighing factor for a given read mapping to the identified first allele variant and the other allele variant is equal to the contribution of the read to the overall posterior probability or posterior probability derived score of the other allele variant (s_1) divided by a sum of that contribution and a similar contribution of the read to the overall posterior probability or posterior probability derived score of the first identified allele variant (s_2). In one example embodiment the weighting factor $w = s_1/(s_1 + s_2)$, and the new contribution of the read to the overall posterior probability or posterior probability derived score of other allele variant = $w*s_1$.

[0018] In one embodiment the polymorphic gene is any gene in an animal that has more than one allele. In certain example embodiments, the polymorphic gene is Type I and II human leukocyte antigen gene (HLA) such as HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB1, MICA and MICB, TAP1 and TAP2, KIR, or the IGHV genes. In one example embodiment, the polymorphic gene is a human leukocyte antigen gene.

[0019] In another aspect the disclosed invention is directed to methods for determining mutations in polymorphic genes comprising extracting a first set of gene-specific reads from a

first sequencing data set obtained from normal tissue of a subject, extracting a second set of gene-specific reads from a second sequencing data set obtained from diseased tissues of the subject, determining a genotype sequence from the reads extracted from the first sequencing set using the genotyping methods disclosed herein, aligning the first set of gene-specific reads and the second set of gene-specific reads to the determined genotype sequence, detect mutations based at least in part on the generated sequence alignment. In certain example embodiments, the sequencing data may be whole exome sequencing data.

[0020] In certain embodiments an output that includes data is generated. In one embodiment the output provides data that indicates the first and second allele variants for the individual. In one embodiment the output provides data that indicates the mutations for the inferred alleles. In one embodiment the output is electronic. In one embodiment the output is printed.

[0021] In certain embodiments the methods are incorporated into a method to formulate a neoantigen immunogenic composition. The invention comprehends performing methods as in U.S. patent application No. 20110293637, incorporated herein by reference, e.g., a method of identifying a plurality of at least 4 subject-specific peptides and preparing a subject-specific immunogenic composition that upon administration presents the plurality of at least 4 subject-specific peptides to the subject's immune system, wherein the subject has a tumor and the subject-specific peptides are specific to the subject and the subject's tumor, said method comprising:

(i) identifying, including through

nucleic acid sequencing of a sample of the subject's tumor and

nucleic acid sequencing of a non-tumor sample of the subject,

a plurality of at least 4 tumor-specific non-silent mutations not present in the non-tumor sample; and

(ii) selecting from the identified non-silent mutations the plurality of at least 4 subject-specific peptides, each having a different tumor neo-epitope that is an epitope specific to the tumor of the subject, from the identified plurality of tumor specific mutations,

wherein each neo-epitope is an expression product of a tumor-specific non-silent mutation not present in the non-tumor sample, each neo-epitope binds to a HLA protein of the subject, and selecting includes

determining binding of the subject-specific peptides to the HLA protein,

and

(iii) formulating the subject-specific immunogenic composition for administration to the subject so that upon administration the plurality of at least 4 subject-specific peptides are presented to the subject's immune system,

wherein the selecting or formulating comprises at least one of:

including in the subject-specific immunogenic composition a subject-specific peptide that includes an expression product of an identified neo-ORF, wherein a neo-ORF is a tumor-specific non-silent mutation not present in the non-tumor sample that creates a new open reading frame, and

including in the subject-specific immunogenic composition a subject-specific peptide that includes an expression product of an identified point mutation and has a determined binding to the HLA protein of the subject with an IC50 less than 500 nM, or any other metric such as the differential of the IC50 values between the native and corresponding mutated peptide being greater than a pre-defined value,

whereby, the plurality of at least 4 subject-specific peptides are identified, and the subject-specific immunogenic composition that upon administration presents the plurality of at least 4 subject-specific peptides to the subject's immune system, wherein the subject-specific peptides are specific to the subject and the subject's tumor, is prepared; or a method of identifying a neoantigen comprising:

a. identifying a tumor specific mutation in an expressed gene of a subject having cancer;

b. wherein when said mutation identified in step (a) is a point mutation:

i. identifying a mutant peptide having the mutation identified in step (a), wherein said mutant peptide binds to a class I HLA protein with a greater affinity than a wild-type peptide; and has an IC50 less than 500 nm, or any other metric such as the differential of the IC50 values between the native and corresponding mutated peptide being greater than a pre-defined value;

c. wherein when said mutation identified in step (a) is a splice-site, frameshift, read-through or gene-fusion mutation:

i. identifying a mutant polypeptide encoded by the mutation identified in step (a), wherein said mutant polypeptide binds to a class I HLA protein; or a method of inducing a tumor specific immune response in a subject comprising administering one or more peptides or

polypeptides identified and an adjuvant; or a method of vaccinating or treating a subject for cancer comprising:

a. identifying a plurality of tumor specific mutations in an expressed gene of the subject wherein when said mutation identified is a:

i. point mutation further identifying a mutant peptide having the point mutation; and/or

ii. splice-site, frameshift, read-through or gene-fusion mutation further identifying a mutant polypeptide encoded by the mutation;

b. selecting one or more mutant peptides or polypeptides identified in step (a) that binds to a class I HLA protein;

c. selecting the one or more mutant peptides or polypeptides identified in step (b) that is capable of activating anti-tumor CD8 T-cells; and

d. administering to the subject the one or more peptides or polypeptides, autologous dendritic cells or antigen presenting cells pulsed with the one or more peptides or polypeptides selected in step (c); or preparing a pharmaceutical composition comprising one identified peptide(s), and performing method(s) as herein discussed. Thus, the neoplasia vaccine or immunogenic composition herein can be as in U.S. patent application No. 20110293637.

[0022] Accordingly, it is an object of the invention to not encompass within the invention any previously known product, process of making the product, or method of using the product such that Applicants reserve the right and hereby disclose a disclaimer of any previously known product, process, or method. It is further noted that the invention does not intend to encompass within the scope of the invention any product, process, or making of the product or method of using the product, which does not meet the written description and enablement requirements of the USPTO (35 U.S.C. §112, first paragraph) or the EPO (Article 83 of the EPC), such that Applicants reserve the right and hereby disclose a disclaimer of any previously described product, process of making the product, or method of using the product.

[0023] It is noted that in this disclosure and particularly in the claims and/or paragraphs, terms such as "comprises", "comprised", "comprising" and the like can have the meaning attributed to it in U.S. Patent law; e.g., they can mean "includes", "included", "including", and the like; and that terms such as "consisting essentially of" and "consists essentially of" have the meaning ascribed to them in U.S. Patent law, e.g., they allow for elements not explicitly recited,

but exclude elements that are found in the prior art or that affect a basic or novel characteristic of the invention. Nothing in this disclosure is to be construed as a promise.

[0024] These and other embodiments are disclosed or are obvious from and encompassed by, the following Detailed Description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0025] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0026] The following detailed description, given by way of example, but not intended to limit the invention solely to the specific embodiments described, may best be understood in conjunction with the accompanying drawings.

[0027] **Figure 1** is a process flow diagram of a method for gene typing polymorphic genes using whole exome sequence data, in accordance with certain example embodiments.

[0028] **Figure 2** is a process flow diagram of a method for mutation detection in polymorphic genes using whole exome sequencing data, in accordance with certain example embodiments.

[0029] **Figure 3** is a block diagram depicting a system for genotyping polymorphic genes using whole exome sequencing data, in accordance with certain example embodiments.

[0030] **Figure 4** is a block flow diagram depicting a method to determine a gene type of a polymorphic gene using whole exome sequencing data, in accordance with certain example embodiments.

[0031] **Figure 5** is a method for extracting reads from a WES data set that map to a gene reference set, in accordance with certain example embodiments.

[0032] **Figure 6** is a block flow diagram depicting a method to detect mutations in a polymorphic gene using whole exome sequencing data, in accordance with certain example embodiments.

[0033] **Figure 7** is a block diagram depicting a computing machine and a module, in accordance with certain example embodiments.

[0034] **Figure 8** is a graph showing the number of correctly inferred alleles using an example embodiment of the genotyping methods disclosed herein and applied to a training set of 8 CLL samples of known HLA type.

[0035] **Figure 9** is a graph showing the percent accuracy of an example embodiment of the genotyping methods disclosed herein in determining the HLA genotype of 133 HapMap samples.

[0036] **Figure 10** is a plot showing the ethnicities of 133 HapMap samples correctly inferred based on their projection in the 2-dimensional space defined by two principal components. The colored icons show the clustering of the 1,398 training samples belonging to four different ethnic groups. The black icons depict the projection of 132 HapMap samples in this space (one sample was removed as an outlier). The success rate for attributing the correct ethnicity to each sample was 100%.

[0037] **Figure 11** is graph showing the overall accuracy of an example embodiment of the genotyping methods disclosed herein in determining the HLA genotype of 133 HapMap samples as compared to other known genotyping methods.

[0038] **Figure 12** is a diagram providing a comparison of somatic HLA mutations identified across cancers using standard approaches (TCGA) and an example embodiment of the mutation detection methods disclosed herein (POLYSOLVER).

[0039] **Figure 13** is a graph showing the number of HLA mutations and the percentage of sample bearing HLA mutations per cancer type identified using standard methods (TGCA) and an example embodiment of the mutation detections methods disclosed herein (POLYSOLVER). (SKCM = melanoma, LUSC = lung squamous cell carcinoma, LUAD = lung adenocarcinoma, BLCA = bladder, HNSC = head and neck, COAD = colon adenocarcinoma, READ = rectum adenocarcinoma, UCEC = uterine, GBM = glioblastoma multiforme, OV = ovarian, BRCA = breast, CLL = chronic lymphocyte leukemia.

[0040] **Figure 14** is a graph showing validation of mutations at the transcriptome level using standard methods (TCGA) and an example embodiment of the mutation detection methods disclosed herein (POLYSOLVER).

[0041] **Figure 15** is graph showing the distribution of HLA mutations across functional domains and tumor types detected using a mutation detection method in accordance with an example embodiment. Top – Distribution of potential loss-of-function events; out of frame (blue) and nonsense mutations (red). The histogram summarizes the number of events identified at

each position. Central Panel – Pattern of mutations detected in each tumor type. Bottom – Recurrent events; recurrent positions (with disease, allele group) with frequence ≥ 3 cases / recurrent site are shown.

[0042] **Figure 16** is a graph showing a comparison of spectrum of mutations in non-HLA genes and HLA genes. The ratio of number of mutations to a particular type to the number of silent mutations is compared between the non-HLA and HLA genes for all mutation types (chi-square test, $P < 2.2 \times 10^{-16}$).

[0043] **Figure 17** is a graph showing the distribution of HLA mutations across exons.

[0044] **Figure 18** is a diagram and graph mapping detected mutations in HLA positions that are in actual physical contact with the peptide (contact residues). Left panel – The relative orientation of a 9-mer peptide with respect to the HLA and T cell molecules. Positions 2 and 9 constitute the primary anchors while position 6 forms the secondary anchor with HLA. The remaining positions interact with the T cell molecule. Right Panel – The 9 amino acids of the peptide and their corresponding HLA contact residues are indicated along the rows (orange – HLA interacting anchor positions, blue – T cell interacting positions). The histogram depicts the frequency of observed HLA mutations in contact residues corresponding to each peptide position.

DETAILED DESCRIPTION OF THE INVENTION

[0045] In one aspect, embodiments herein provide computer-implemented techniques for gene typing polymorphic genes using sequencing data. In certain example embodiments the sequencing data is whole exome sequencing data (WES), RNA-Seq data, whole genome data, targeted exome sequencing data, or any form of sequencing data that covers the polymorphic loci at either the exome, genome, or RNA levels. In a preferred embodiment the present invention provides novel computer-implemented techniques for transforming next generation sequencing or massively parallel sequencing data into allelic data for polymorphic genes. For ease of reference, the example embodiments will be described below with reference to WES data, but other sequencing data as described above may be used interchangeably. Example polymorphic genes include the Type I and II HLA loci such as HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB1 the MICA and MICB genes, the

TAP1 and TAP2 genes, KIR, and the IGHV genes. In one example embodiment, the polymorphic gene is an HLA gene.

[0046] Polymorphic gene typing can be used, for example, in identifying compatible organ donors and selecting appropriate personalized medicine treatment regimes, such as an immunogenic composition or vaccine that includes neoepitopes. Whole genome/exome sequencing may be used to identify all, or nearly all, mutated neoantigens that are uniquely present in a neoplasia/tumor of an individual patient, and that this collection of mutated neoantigens may be analyzed to identify a specific, optimized subset of neoantigens for use as a personalized cancer vaccine or immunogenic composition for treatment of the patient's neoplasia/tumor. For example, a population of neoplasia/tumor specific neoantigens may be identified by sequencing the neoplasia/tumor and normal DNA of each patient to identify tumor-specific mutations, and the patient's HLA allotype can be identified. The population of neoplasia/tumor specific neoantigens and their cognate native antigens may then be subject to bioinformatic analysis using validated algorithms to predict which tumor-specific mutations create epitopes that could bind to the patient's HLA allotype. Based on this analysis, a plurality of peptides corresponding to a subset of these mutations may be designed and synthesized for each patient, and pooled together for use as a cancer vaccine or immunogenic composition in immunizing the patient.

[0047] An overview of the process for determining polymorphic gene types is provided with reference to Figure 1. The process starts by extracting reads from a set of whole exome sequencing (WES) data that map to the polymorphic gene of interest ("target polymorphic gene") 105. In one embodiment the data is extracted by assembly of the short sequences de novo. In another embodiment the data is extracted by mapping to a known sequence from a subject of the same species. More than one polymorphic gene may be analyzed at the same time. In one embodiment, multiple genes, such as the HLA-A, HLA-B, and HLA-C can be analyzed concurrently.

[0048] The extracted reads are then aligned to a reference genetic sequence set constructed with known allele variants of the target polymorphic gene and/or genes 110. The reference genetic sequence may be constructed from a library of known or inferred genomic and or cDNA sequences of the polymorphic gene or polymorphic genes to be typed. In one embodiment every extracted read is aligned with every allele sequence corresponding to the extracted read within

the reference library. In certain example embodiments, the reads in the sequencing data set may consist of reads that match one or more sequences from a reference genetic sequence. The reads may match one or more sequences from the reference genetic sequence in the 5' to 3' direction or the 3' to 5' orientation. In certain example embodiments, the reads have between approximately 90% and approximately 100% sequence identity to one or more sequences in the reference genetic sequence. In one example embodiment the reads have approximately 100% sequence identity to one or more sequences in the reference genetic sequence.

[0049] The generated sequence alignment and other information 115, such as an insert size distribution for the aligned reads, alignment quality scores and population frequencies, are used to calculate a first posterior probability or posterior probability derived score for each allele variant 120. The allele variant that maximizes the first posterior probability or posterior probability derived score is selected as the first allele variant of the target polymorphic gene type 125. A second posterior probability or posterior probability derived score is calculated for each allele by applying a heuristic weighting strategy to the score contribution of each of its aligned reads from the first stage, taking into consideration whether a read under consideration also mapped to the first inferred allele variant. 130. The allele variant that maximizes the second posterior probability or posterior probability derived score is selected as the second allele variant 140. The first and second allele variants define the polymorphic gene type and are provided as an output. In one embodiment the allele variants are displayed on a computer screen.

[0050] In another aspect, embodiments herein provide computer-implemented techniques for detecting mutations in polymorphic genes by generating polymorphic allele data from sequencing data obtained from normal tissue and comparing the extracted reads of the normal tissue aligned to a reference set of alleles of the polymorphic gene to the extracted reads of disease tissue aligned to a reference set of allelles of the polymorphic gene. As described herein, example embodiments are described with reference to WES data, but other sequencing data types may be used interchangeably. An overview of the process for determining mutations in polymorphic genes is provided with reference to Figure 2. A WES data set is obtained from normal germline cells of the subject to be tested and a polymorphic gene type is determined according to the polymorphic gene typing method described herein 205 (POLYSOLVER). A second WES data set is obtained from diseased cells, such as cancer cells, from the subject to be tested 210. Reads from the diseased tissue WES data set mapping to the target polymorphic

gene are then extracted 215. The extracted reads are then aligned to the polymorphic gene type sequences determined at 205. The resulting alignment is then used to detect mutations in the sequences obtained from the diseased tissue sample based on the predicted polymorphic alleles.

[0051] Turning now to Figures 3-6, in which like numerals represent like (but not necessarily identical) elements throughout the figures, example embodiments are described in detail.

Example System Architectures

[0052] Figure 3 is a block diagram depicting a system for determining gene type and detecting mutations in polymorphic genes. As depicted in Figure 3, the operating environment 300 includes network devices 305 and 310 that are configured to communicate with one another via one or more networks 315. In some embodiments, a user associated with a device must install an application and/or make a feature selection to obtain the benefits of the techniques described herein.

[0053] Each network 315 includes a wired or wireless telecommunication means by which network devices (including devices 305 and 310) can exchange data. For example, each network 315 can include a local area network ("LAN"), a wide area network ("WAN"), an intranet, an Internet, a mobile telephone network, or any combination thereof. Throughout the discussion of example embodiments, it should be understood that the terms "data" and "information" are used interchangeably herein to refer to text, images, audio, video, or any other form of information that can exist in a computer-based environment.

[0054] Each network device 305 and 310 includes a device having a communication module capable of transmitting and receiving data over the network 315. For example, each network device 305 and 310 can include a server, desktop computer, laptop computer, tablet computer, a television with one or more processors embedded therein and / or coupled thereto, smart phone, handheld computer, personal digital assistant ("PDA"), or any other wired or wireless, processor-driven device.

[0055] The whole exome sequencing device 305 sequences nucleic acid material extracted from a biological sample to generate a whole exome sequencing data file containing information on the coding regions of genes across the sample genome. In one example embodiment, the whole exome sequencing device 305 may directly communicate the WES data file to the polymorphic gene typing system 310 across the network 310 and the gene typing or mutation

detection analysis is conducted in line with the sequencing analysis. In another example embodiment, the WES data file may be stored on a data storage medium and later uploaded to the polymorphic gene typing system 310 for further analysis.

[0056] The polymorphic gene typing system 310 may comprise an alignment module 311, gene typing module 312, a mutation detection module 313, an ethnicity inference module 316, and an allele variant sequence library 314. The alignment module 311 extracts and aligns reads from the whole exome sequencing data file to a gene reference set. The gene reference set comprises reference sequences for the polymorphic gene being analyzed. The sequence information for the gene reference set is stored in the allele sequence library 314. The ethnicity inference module 316 infers the ethnicity of the individual which then serves as the basis for selection of prior probabilities by the gene typing module. The gene typing module 312 performs a two-stage posterior probability based analysis on the aligned reads to identify the gene type of the sample, optionally using population derived allele frequencies as prior probabilities. The mutation detection module 313 identifies mutations based on an analysis of the gene type and WES data obtained from a diseased tissue sample.

[0057] It will be appreciated that the network connections shown are example and other means of establishing a communications link between the computers and devices can be used. Moreover, those having ordinary skill in the art having the benefit of the present disclosure will appreciate that whole exome sequencing device 305 and polymorphic gene typing system 310, can have any of several other suitable computer system configurations.

Example Processes

[0058] The example methods illustrated in Figure 4-6 are described herein with respect to the components of the example operating environment 300. The example methods of Figure 4-6 may also be performed with other systems and in other environments.

[0059] Figure 4 is a block flow diagram depicting a method 400 to determine a polymorphic gene type of one or more polymorphic gene bases on whole exome sequencing data, in accordance with certain example embodiments.

[0060] Method 400 begins at block 405, wherein the alignment module 311 generates an alignment of gene reads from WES sequencing data that map to a gene reference sequence for the target gene. Block 405 will be described in further detail with reference to Figure 5.

[0061] Figure 5 is a block diagram depicting a method 405 to align reads from the WES sequence data with a gene reference set. The method 405 begins at block 505 where the alignment module 311 maps all the reads in the WES sequence data to a reference target gene sequence. The reference data sequence may be downloaded directly from a public resource such as IMGT and may comprise genomic or cDNA sequences. It may also be a set of full length genomic sequences along with inferred full length genomic sequences. Missing exons for incompletely sequenced allele cDNAs may be deduced by multiple sequence alignment of the missing allele with all available cDNA sequences. Missing introns may be inferred by alignment of the cDNA sequence with the nearest full-length genomic sequence.

[0062] At block 510, the alignment module 311 then extracts all reads that map within a defined cut-off value of the target gene to generate a first extracted read set. In one example embodiment, reads that mapped within 200 to 2000 bases, 200 to 1750 bases, 200 to 1500 bases, 200 to 1250 bases, 200 to 1000, 200 to 750, 200 to 500, 500 to 750, 500 to 1000, 500 to 1250, 500, to 1500, 500 to 1750, 500 to 2000, or 1000 to 2000 bases are extracted. In one example embodiment, the reads mapping to within 1000 bases of the target polymorphic gene are extracted. In certain example embodiments, the alignment module 311 further determines an insert size distribution based on all aligned reads in the sequence data file. This empirical insert size distribution is then utilized by the gene typing module 312 in determining a posterior probability or posterior probability derived score as described in further detail herein.

[0063] At block 515, the alignment module 311 extracts reads from the original sequence data file or the completed set of reads in a sequencing run based on comparing each read to a probe sequence set. The probe sequence set may include short probe sequences derived from a library of known genomic and/or cDNA sequences of the target polymorphic gene. Reads that match one or more short probe sequences in a probe sequence set are included in second extracted read set. The reads may match the one or more probes in the 5' to 3' or '3' o 5' orientation. The probe sequences have a size between approximately 25 to approximately 100 mer, approximately 25 to approximately 75 mer, approximately 25 mer to 50 mer, approximately 50 mer to 100 mer, approximately 50 mer to approximately 75 mer and any combination in between. In one example embodiment the probe sequences in the probe sequence library have a size of 38 mer. In one example embodiment, the aligned reads have between 90% and 100% sequence identity with one or more the probe sequences in the probe sequence library. In

another example embodiment, the aligned reads have 100% sequence identity with one or more probes in the probe sequence library. A final extracted read set is then used for further analysis and may comprise both extracted read sets 1 and 2, or either of them exclusively.

[0064] At block 520, the reads in the final extracted read set are aligned to an allele variant sequence library for the target gene to generate a final alignment. The reads may be aligned to the allele variant sequence library using a standard alignment algorithm. The allele sequence information is stored in the allele variant sequence library 314 and may contain all available allele genomic and cDNA sequences for the target gene, or inferred full length genomic sequences as described herein. In one example embodiment, the alignment algorithm is the "Novoalign" alignment algorithm (Novocraft Selengor, Malaysia). In another example embodiment, the parameters are set to report all best-scoring alignments. In another example embodiment, all alignments that meet a score threshold will be reported. Any new optical or PCR duplicates that are unmasked as a result of the final alignment may or may not be removed. For example, duplicates can be identified and removed using Picard's MarkDuplicates module (picard.sourceforge.net/). The process then proceeds to block 410 of Figure 4.

[0065] Returning to block 410 of Figure 4, where the gene typing module 312 uses the final alignment to calculate a first posterior probability or posterior probability derived score for each allele variant of the polymorphic gene. The gene typing module 312 first determines a likelihood calculation for each allele variant of the polymorphic gene. In certain example embodiments, a log likelihood score is calculated for each allele variant as follows:

[0066] Let:

$N_A \equiv$ # alleles corresponding to the HLA gene

$M \equiv$ # alleles corresponding to the polymorphic gene

$N \equiv$ # reads aligning to at least one allele

$N_m \equiv$ # reads aligning to allele a_m

$N_T \equiv$ # reads in the sequencing run

$f_m \equiv$ population based prior probability of allele m

$r_{k1} \equiv$ first segment of read pair r_k

$r_{k2} \equiv$ second segment of read pair r_k

$d_k \equiv$ insert length of read pair r_k

$l_{k1} \equiv$ length of first segment of read pair r_k

l_{k2} ≡ length of second segment of read pair r_k

q_i ≡ quality of sequenced base i

e_i ≡ probability that the sequenced base i is an error

$$e_i = 10^{-\frac{q_i}{10}}$$

The quality scores of the alignment can be used to build a model for the sequencing process. Let Y_{Ai}, Y_{Ci}, Y_{Gi} and Y_{Ti} denote random variables corresponding to observing bases A, C, G and T respectively at position *i* in read r_k in its alignment to allele a_m. Then

$$Y_{Ai}, Y_{Ci}, Y_{Gi}, Y_{Ti} \sim Multinomial(n = 1;\ \alpha_{Ai}, \alpha_{Ci}, \alpha_{Gi}, \alpha_{Ti})$$

where

$\alpha_{Bi} = 1 - e_i$ if reference base at position *i* in a_m is B

$= e_i/3$ otherwise

[0067] D denotes a random variable for the observed insert length of a paired read in the sequencing run based on alignment to the complete genome. For a given read pair r_k, the empirical insert size distribution can be used to calculate the probability of observing the insert length d_k as

$$P(D = d_k) = \frac{\sum_{l=1}^{N_T} I(d_l = d_k)}{N_T}$$

[0068] Assuming positional independence of quality scores, and independence of generated reads and their insert sizes, the probability of observing r_k given allele a_m is then

$$P(r_k|a_m) = \begin{cases} \prod_{i=1}^{l_{k1}} \alpha_i \prod_{j=1}^{l_{k2}} \alpha_j \cdot P(D = d_k) & if\ r_k\ aligns\ to\ a_m \\ s_k & otherwise \end{cases}$$

[0069] where s_k corresponds to the lowest theoretical probability achievable for read pair r'_k with perfect base qualities and segment lengths equal to those of r_k. Since 93 is the maximum achievable base quality under Illumina 1.8+ format, s_k is computed as

$$s_k = (l_{k1} + l_{k2}) \cdot \log\frac{10^{-9.3}}{3} \approx -23 \cdot (l_{k1} + l_{k2})$$

[0070] The posterior probability of allele a_m using all reads that align to it is given by

$$P(a_m|r_1, r_2, \ldots, r_N) = \frac{\prod_{k=1}^{N} P(r_k|a_m) \cdot f_m}{\prod_{k=1}^{N} P(r_k)}$$

[0071] Log transformation of the above equation yields

$$L_m = \sum_{k=1}^{N_m} \sum_{i=1}^{l_{k1}} \log \alpha_i + \sum_{k=1}^{N_m} \sum_{j=1}^{l_{k2}} \log \alpha_j + \sum_{k=1}^{N_m} \log P(D = d_k) + (N - N_m)s_k + \log f_m - \sum_{k=1}^{N} \log P(r_k)$$

[0072] Note that the terms $N \cdot s_k$ and $\sum_{k=1}^{N} \log P(r_k)$ are constants for all alleles and can be ignored.

[0073] The likelihoods of all aligned read pairs to a given allele variant may be computed based on their respective alignments, quality scores, and insert size probabilities based on the empirical insert size distribution of all read pairs in the sequencing run. Population-based allele probabilities may be used as priors in the model. For example, allele frequencies for Caucasian, Black, and Asian ethnicities may be calculated taking a sample-size weighted average of all relevant population studies in the Allele Frequency Net Database (Gonzalez-Galarza et al., 2011).

[0074] At block 415, the gene typing module 312 selects the allele variant that maximizes the log posterior probability score or the log of the posterior probability derived score from the calculations determined in block 410 as the first allele variant for the polymorphic gene type.

$$a_w = \underset{a_m}{\operatorname{argmax}}\ L_m$$

[0075] The log posterior probability or the log of the posterior probability derived score calculated in the first stage shall henceforth be referred to as L_1 score.

[0076] At block 420, the gene typing module 312 calculates a second posterior probability score or posterior probability derived index (either of which will henceforth be referred as the L_2 score) for each allele in the database based on the overlap of reads that each allele shares with the first identified allele. A read may have mapped to a position of the first identified allele sequence and also mapped to the same position in a second allele under consideration whose L_2 score is being evaluated but with a variance in alignment quality or insert size. The read may

also map to a different position in the first and second allele variant with close sequence similarity. The gene typing module 312 computes the L_2 score for an allele a_m by applying a heuristic weighting strategy to each of its aligned reads. For a given allele a_m, the log likelihood l_m^k of a read r_k that also maps to the first identified allele variant a_w with likelihood l_w^k is weighted by a factor equal to $l_m^k/(l_m^k + l_w^k)$. If only the best-scoring alignments for each read are preserved, this ratio will always be 0.5 (Figure 1, 140). Reads mapping exclusively to a_m with respect to a_w are assigned a weight of 1. The read insert size and allele prior probability components of the L_1 score are preserved from the first stage. Alternatively, the read insert size component may also be included in the heuristic reweighting.

[0077] At block 425, the gene typing module 312 identifies the second allele variant for the gene as the allele with the maximal L_2 score.

[0078] At block 430, the gene typing module 312 displays genotyping information comprising the first allele and second allele variant with maximal L_1 and L_2 scores as the gene type for the analyzed sample. The gene typing module may additionally display information such as the associated scores, and the alleles with the second highest L_1 and L_2 scores. The difference in scores between the first and second-highest scoring alleles, or some other derivative metric such as the percentage increase in score between the second-highest and highest scoring alleles in the two stages, may also be displayed. An empirical p-value based on the chosen metric may also be part of the output. The gene typing module 312 may also generate a report comprising the genotyping information. The report may be in electronic format, hard copy format, or both.

[0079] Figure 6 is a block flow diagram depicting a method 600 for identifying mutations in polymorphic genes, in accordance with certain example embodiments.

[0080] Method 600 begins at block 605, where the gene typing module 312 determines the gene type of a target polymorphic gene based on an analysis of WES data obtained from a normal tissue sample of a subject. The process for determining the gene type from the normal tissue sample is substantially the same as that described above with reference to blocks 405 to 430 of Figure 4.

[0081] At block 610 the alignment module 311 extracts reads from WES data obtained from a diseased tissue sample of the same subject and aligns the extracted reads with the sequence of the polymorphic gene types determined in block 430. The procedure for extracting and aligning

the reads from the diseased tissue WES data is substantially the same as that described above with reference to blocks 505 through 520 of Figure 5.

[0082] At block 615 a mutation module 313 detects mutations or other somatic changes including insertions, deletions, copy number changes and translocations based on the alignment generated at block 610. The mutation module may utilize standard mutation detection or other somatic change detection algorithms known in the art. In certain example embodiments, the "MuTect" method as described in International Patent Application Publication No. WO2014036167 A1 to Cibulskis *et al.*, and hereby incorporated by reference, is used to detect mutations from the alignment data. In certain embodiments, the Strelka software (Saunders et al., 2012) may be used to detect insertions and deletions in the diseased sample.

[0083] At block 620, the mutation module 313 displays the detected mutation information. The detected mutation information may include the position, mutated or alternate bases, reference bases, sequence context, codon changes, protein changes, number of reads supporting the reference and alternate bases in the tumor and normal samples, a goodness score such as a log odds score for the change. The detected mutation information may include a mapping of the mutations positions in a two-dimensional or three-dimensional schematic of the corresponding transcribed protein. For example, the detection mutation information may include a schematic like that shown in FIG. 17 that maps the mutations to key contact points between the corresponding protein and one or more binding partners of the protein. The mutation module 313 may generate a report comprising the detected mutation information described above. The report may be generated in electronic format, hard copy format, or both.

Other Example Embodiments

[0084] Figure 7 depicts a computing machine 2000 and a module 2050 in accordance with certain example embodiments. The computing machine 2000 may correspond to any of the various computers, servers, mobile devices, embedded systems, or computing systems presented herein. The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 in performing the various methods and processing functions presented herein. The computing machine 2000 may include various internal or attached components such as a processor 2010, system bus 2020, system memory 2030, storage

media 2040, input/output interface 2060, and a network interface 2070 for communicating with a network 2080.

[0085] The computing machine 2000 may be implemented as a conventional computer system, an embedded controller, a laptop, a server, a mobile device, a smartphone, a set-top box, a kiosk, a vehicular information system, one or more processors associated with a television, a customized machine, any other hardware platform, or any combination or multiplicity thereof. The computing machine 2000 may be a distributed system configured to function using multiple computing machines interconnected via a data network or bus system.

[0086] The processor 2010 may be configured to execute code or instructions to perform the operations and functionality described herein, manage request flow and address mappings, and to perform calculations and generate commands. The processor 2010 may be configured to monitor and control the operation of the components in the computing machine 2000. The processor 2010 may be a general purpose processor, a processor core, a multiprocessor, a reconfigurable processor, a microcontroller, a digital signal processor ("DSP"), an application specific integrated circuit ("ASIC"), a graphics processing unit ("GPU"), a field programmable gate array ("FPGA"), a programmable logic device ("PLD"), a controller, a state machine, gated logic, discrete hardware components, any other processing unit, or any combination or multiplicity thereof. The processor 2010 may be a single processing unit, multiple processing units, a single processing core, multiple processing cores, special purpose processing cores, co-processors, or any combination thereof. According to certain embodiments, the processor 2010 along with other components of the computing machine 2000 may be a virtualized computing machine executing within one or more other computing machines.

[0087] The system memory 2030 may include non-volatile memories such as read-only memory ("ROM"), programmable read-only memory ("PROM"), erasable programmable read-only memory ("EPROM"), flash memory, or any other device capable of storing program instructions or data with or without applied power. The system memory 2030 may also include volatile memories such as random access memory ("RAM"), static random access memory ("SRAM"), dynamic random access memory ("DRAM"), and synchronous dynamic random access memory ("SDRAM"). Other types of RAM also may be used to implement the system memory 2030. The system memory 2030 may be implemented using a single memory module or multiple memory modules. While the system memory 2030 is depicted as being part of the

computing machine 2000, one skilled in the art will recognize that the system memory 2030 may be separate from the computing machine 2000 without departing from the scope of the subject technology. It should also be appreciated that the system memory 2030 may include, or operate in conjunction with, a non-volatile storage device such as the storage media 2040.

[0088] The storage media 2040 may include a hard disk, a floppy disk, a compact disc read only memory ("CD-ROM"), a digital versatile disc ("DVD"), a Blu-ray disc, a magnetic tape, a flash memory, other non-volatile memory device, a solid state drive ("SSD"), any magnetic storage device, any optical storage device, any electrical storage device, any semiconductor storage device, any physical-based storage device, any other data storage device, or any combination or multiplicity thereof. The storage media 2040 may store one or more operating systems, application programs and program modules such as module 2050, data, or any other information. The storage media 2040 may be part of, or connected to, the computing machine 2000. The storage media 2040 may also be part of one or more other computing machines that are in communication with the computing machine 2000 such as servers, database servers, cloud storage, network attached storage, and so forth.

[0089] The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 with performing the various methods and processing functions presented herein. The module 2050 may include one or more sequences of instructions stored as software or firmware in association with the system memory 2030, the storage media 2040, or both. The storage media 2040 may therefore represent examples of machine or computer readable media on which instructions or code may be stored for execution by the processor 2010. Machine or computer readable media may generally refer to any medium or media used to provide instructions to the processor 2010. Such machine or computer readable media associated with the module 2050 may comprise a computer software product. It should be appreciated that a computer software product comprising the module 2050 may also be associated with one or more processes or methods for delivering the module 2050 to the computing machine 2000 via the network 2080, any signal-bearing medium, or any other communication or delivery technology. The module 2050 may also comprise hardware circuits or information for configuring hardware circuits such as microcode or configuration information for an FPGA or other PLD.

[0090] The input/output ("I/O") interface 2060 may be configured to couple to one or more external devices, to receive data from the one or more external devices, and to send data to the one or more external devices. Such external devices along with the various internal devices may also be known as peripheral devices. The I/O interface 2060 may include both electrical and physical connections for operably coupling the various peripheral devices to the computing machine 2000 or the processor 2010. The I/O interface 2060 may be configured to communicate data, addresses, and control signals between the peripheral devices, the computing machine 2000, or the processor 2010. The I/O interface 2060 may be configured to implement any standard interface, such as small computer system interface ("SCSI"), serial-attached SCSI ("SAS"), fiber channel, peripheral component interconnect ("PCI"), PCI express (PCIe), serial bus, parallel bus, advanced technology attached ("ATA"), serial ATA ("SATA"), universal serial bus ("USB"), Thunderbolt, FireWire, various video buses, and the like. The I/O interface 2060 may be configured to implement only one interface or bus technology. Alternatively, the I/O interface 2060 may be configured to implement multiple interfaces or bus technologies. The I/O interface 2060 may be configured as part of, all of, or to operate in conjunction with, the system bus 2020. The I/O interface 2060 may include one or more buffers for buffering transmissions between one or more external devices, internal devices, the computing machine 2000, or the processor 2010.

[0091] The I/O interface 2060 may couple the computing machine 2000 to various input devices including mice, touch-screens, scanners, biometric readers, electronic digitizers, sensors, receivers, touchpads, trackballs, cameras, microphones, keyboards, any other pointing devices, or any combinations thereof. The I/O interface 2060 may couple the computing machine 2000 to various output devices including video displays, speakers, printers, projectors, tactile feedback devices, automation control, robotic components, actuators, motors, fans, solenoids, valves, pumps, transmitters, signal emitters, lights, and so forth.

[0092] The computing machine 2000 may operate in a networked environment using logical connections through the network interface 2070 to one or more other systems or computing machines across the network 2080. The network 2080 may include wide area networks (WAN), local area networks (LAN), intranets, the Internet, wireless access networks, wired networks, mobile networks, telephone networks, optical networks, or combinations thereof. The network 2080 may be packet switched, circuit switched, of any topology, and may use any

communication protocol. Communication links within the network 2080 may involve various digital or an analog communication media such as fiber optic cables, free-space optics, waveguides, electrical conductors, wireless links, antennas, radio-frequency communications, and so forth.

[0093] The processor 2010 may be connected to the other elements of the computing machine 2000 or the various peripherals discussed herein through the system bus 2020. It should be appreciated that the system bus 2020 may be within the processor 2010, outside the processor 2010, or both. According to some embodiments, any of the processor 2010, the other elements of the computing machine 2000, or the various peripherals discussed herein may be integrated into a single device such as a system on chip ("SOC"), system on package ("SOP"), or ASIC device.

[0094] Embodiments may comprise a computer program that embodies the functions described and illustrated herein, wherein the computer program is implemented in a computer system that comprises instructions stored in a machine-readable medium and a processor that executes the instructions. However, it should be apparent that there could be many different ways of implementing embodiments in computer programming, and the embodiments should not be construed as limited to any one set of computer program instructions. Further, a skilled programmer would be able to write such a computer program to implement an embodiment of the disclosed embodiments based on the appended flow charts and associated description in the application text. Therefore, disclosure of a particular set of program code instructions is not considered necessary for an adequate understanding of how to make and use embodiments. Further, those skilled in the art will appreciate that one or more aspects of embodiments described herein may be performed by hardware, software, or a combination thereof, as may be embodied in one or more computing systems. Moreover, any reference to an act being performed by a computer should not be construed as being performed by a single computer as more than one computer may perform the act.

[0095] The example embodiments described herein can be used with computer hardware and software that perform the methods and processing functions described herein. The systems, methods, and procedures described herein can be embodied in a programmable computer, computer-executable software, or digital circuitry. The software can be stored on computer-readable media. For example, computer-readable media can include a floppy disk, RAM, ROM,

hard disk, removable media, flash memory, memory stick, optical media, magneto-optical media, CD-ROM, etc. Digital circuitry can include integrated circuits, gate arrays, building block logic, field programmable gate arrays (FPGA), etc..

[0096] The present method advantageously provides a method to allele type any polymorphic gene using the sequencing data produced by massively parallel sequencing without the use of time consuming and expensive additional experimentation. The present invention advantageously provides the ability to transform the sequencing data into allele information when a sample has been exhausted and does not exist any longer. This advantage provides, for example, the ability to analyze neoantigens that will bind to a specific HLA allele in a sample derived from a subject without having to conduct additional experimentation on a sample. In one embodiment HLA type and neoantigen presence can be obtained from sequencing data in a single sequencing run. The present invention also provides an improved method to detect mutations within polymorphic loci.

[0097] The present invention will be further illustrated in the following Examples which are given for illustration purposes only and are not intended to limit the invention in any way.

Examples

Example 1

[0098] WES from cases of chronic lymphocytic leukemia (CLL) patients has been previously reported (Landau et al., Cell. 2013, 152(4):714-26). A subset of cases within this cohort with available experimental HLA typing information (HLA-A, HLA-B, HLA-C) were selected for further analysis. A set of 133 HapMap samples comprising 15 Caucasians, 42 Black, 41 Chinese and 35 Japanese individuals with known HLA types (Erlich et al, 2011; www.1000genomes.org) was used as a validation set. For detection of somatic mutation of the HLA loci, the gene typing method described above was applied to data from 10 tumor types curated from the TCGA project including lung squamous, lung adenocarcinoma, bladder, head and neck, colon, rectum, uterine, glioblastoma, ovarian, , and breast. Two additional data sets: a melanoma data set (Hodis et al. Cell. 2012, 150(2):251-63) and a chronic lymphocytic leukemia set (Wang et al., NEJM. 2011. 365:2497-2506), was also analyzed for mutations in the HLA-A, HLA-B and HLA-C genes. For ease of reference, the example embodiment disclosed in this section is referred to hereafter as POLYSOLVER.

[0099] A reference library of known HLA alleles (6597 unique entries) based on the IMGT database ((v3.10; www.ebi.ac.uk/ipd/imgt/hla/) was constructed. Missing exons for incompletely sequenced allele cDNAs were deduced by multiple sequence alignment of the missing allele with all available cDNA sequences. Missing introns were inferred by alignment of the cDNA sequence with the nearest full-length genomic sequence. The final library comprised full-length genomic sequences of 2129 HLA-A, 2796 HLA-B and 1672 HLA-C alleles.

[00100] To generate efficiency in alignment, all reads mapping within 1 Kb of the HLA genes from the processed bam files were extracted. A further extraction was performed using a secondary 38-mer HLA sequence library to 'fish' for any reads that perfectly matched at least one tag in either orientation. Reads selected by this process are subjected to several post-processing filters (see below) prior to inference of alleles. The choice of 38 mers for creation of HLA tag library was based on maximizing the specificity of the library for HLA genes while maintaining 100% sensitivity in the context of downstream read filtering.

[00101] To assess the specificity of tag libraries of different lengths derived from the polymorphic genes, a sequence set of ~21,000 non-polymorphic genes was created and the fraction non-polymorphic genes that matched at least one tag from the library was recorded for different tag lengths. Specificity was defined as 1 minus this fraction, see the table 1.

Table 1. Specificity of tag libraries of different lengths

length	# tags	# non poly genes	Specificity
20	697364	17505	15.72%
24	788407	17184	17.26%
28	877670	16931	18.48%
32	965304	16623	19.96%
36	1051894	16212	21.94%
38	1095054	15922	23.34%
40	1138288	15550	25.13%
44	1224676	14662	29.40%
48	1310772	13309	35.92%
50	1353714	12600	39.33%
60	1568396	8749	57.87%

70	1785652	6003	71.10%

[00102] The downstream filtering worked under the hypothesis that any alignment that had more than one event (an event being any of the following: mismatch, insertion or deletion) was likely a mis-alignment and was discarded. Under this hypothesis, 100% sensitivity for picking up all HLA reads in the sequencing run would require that the maximum tag length used for extracting reads should not exceed half the read length which in this case was 76. A choice of 38 for the tag length library assured that all reads with no more than one mismatch or deletion event would be captured, while delivering a specificity of ~23% as defined above.

[00103] HLA allele names comprise the name of the gene (i.e. A, B, C) suffixed by successive sets of digits that indicate increasing sequence-level and functional resolution. The first set specifies the serological activity of the allele (allele level resolution, ex. A*01 or A*02) while the second set of digits, in conjunction with the first set of digits, specifies the protein sequence (protein level resolution, ex. A*02:01). Alleles were resolved up to the protein level resolution. A two-step inference procedure was applied; detecting the most-likely allele first and then, given the first allele, determining the second most likely allele. Inference is based on a Bayesian calculation that takes into account: (i) the qualities of the bases comprising each aligned read; (ii) the observed insert sizes of reads aligned to the allele; (iii) the number of reads aligned to the allele; and (iv) the prior probability of each allele (Fig.1b). Previous studies have suggested that knowledge of the ethnicity of the individual under consideration can increase the probability of correct typing since the population-level allele frequencies differ based on race. (Erlich et al. BMC genomics. 2001, 12, 42); Gonzalez-Galarza et al. Nucleic Acids Research. 2011 39, D913-919). These known population-level allele frequencies were harnessed and ethnicity-dependent prior probabilities were used. The posterior probability was calculated for each allele which aggregates evidence from both the likelihood computation and the prior probability. The allele with the highest posterior probability ('the winner') was inferred to be the correct first allele.

[00104] In the second step of the inference it was taken into account that: (i) an individual may be either homozygous or heterozygous for any of the HLA genes; and (ii) alleles encoding for the same protein sequence tend to have highly similar DNA sequences, thereby artificially inflating the posterior probabilities of alleles that bear significant sequence similarity to the first inferred allele. It was observed that selecting the top two alleles when the posterior probabilities

were simply sorted in order was incorrectly biased in favor of declaring homozygous winners, with 23 out of 24 HLA loci in the training set being miscalled in this fashion. On the other hand, a complete depletion of reads mapping to the first inferred allele followed by a recalculation of the posterior probability-derived scores yielded only heterozygous calls. To balance between the two extremes, a strategy that shrank allele scores in proportion to their sequence similarity with the first inferred allele was devised as disclosed herein. The allele with the highest recalculated score was then inferred as the second true allele. By using this strategy, 47 of 48 (97.9%) alleles were correctly identified at the protein level. *See* FIG. 8.

Example 2

Validation of Genotyping Method

[00105] The method disclosed herein was applied to WES data from a set of 133 HapMap samples with known HLA genotypes. 774 of 798 (97%) alleles from this validation set were correctly resolved at the protein level while allele groups were correctly typed in 787 of 798 (98.7%) instances All 42 homozygous alleles in this set were correctly identified, and no significant differences in performance were observed based on ethnicity or HLA gene (chi-squared test *P*-values 0.043 and 0.314 respectively, 95% Bonferroni corrected *P*-value threshold = 0.025). See FIG 9. However, when this method was applied to the HapMap samples without using population-level allele frequencies, only 89% accuracy was observed. To accommodate use of the method with samples of unknown ethnic origin, the following principal components (PC)-based method for exome-based ethnicity inference, which can be used prior to analysis by the genotyping method disclosed herein.

[00106] 4-digit allele frequencies for different ethnicities were calculated by taking a sample-size weighted average of all relevant population studies in the Allele Frequency Net Database (www.allelefrequencies.net/). A rapid principal components analysis (PCA) based method was developed to infer ethnicity for samples of unknown racial origin (Kiezun et al, manuscript in preparation). Exome data for samples of known (self-described) ethnicity from the 1000 Genomes and HapMap projects (n=1,398, with 911 Caucasians, 375 Blacks, 54 Asians, and 58 South Asians) was genotyped at a predefined set of 5,845 loci chosen based on considerations related to known linkage disequilibrium between different loci, representation on population genotyping platforms and consistency between genome releases. A PCA revealed distinct

segregation of Caucasian, Black, Asian, and South Asian samples in the 2-dimensional space defined by the first two principal components. Any new sample of unknown ethnicity may be projected in this space and its Euclidean distance from the clusters centroids can be computed. Ethnicity is inferred based on the cluster of minimal distance from the sample projection.

[00107] As an alternative to applying the PC-based ethnicity inference module, it was also observed that restricting inference of alleles to those having at least a 0.05% frequency in each of Caucasian, Black and Asian populations also resulted in 96% protein-level accuracy. Consistent with this finding, re-review of the sole misidentified allele within the original training set of CLL cases revealed it to be A*01:02, whose minor allele frequency is < 0.05% in Caucasians.

Example 3

Comparison to Other HLA Typing Methods

[00108] Using the 133 HapMap samples, the HLA typing method disclosed herein was compared to four recently reported algorithms for HLA typing. Overall accuracy, i.e. the percentage of alleles that were correctly called, for comparing the different approaches was used. Ambiguous calls or failure to make a call were both assessed as mistakes. ATHLATES (Liu et al. Nucleic Acid Research. 2013 41, e142) had an overall accuracy of 17% although its performance improved with increasing number of reads (which is a proxy for the average sequence coverage). HLAminer (Warren et al. Genomic Medicine. 2012, 4, 95) and HLAforest (Kim et al. PloS One. 2013, 8, e67885) 30% and 47% overall accuracies respectively, while PHLAT (Bai et al. BMC Genomics. 2014, 15) was able to correctly identify 87% of the 798 alleles. POLYSOLVER had an overall accuracy of 97% and >96% accuracy across the range of sequencing depths. *See* FIG. 11.

Example 4

Detection of Somatic Mutations within the HLA region

[00109] The standard approach for detection of somatic mutations is to first align both tumor and normal reads to a reference genome and then scan the genome and identify mutational events observed in the tumor but not in the matched normal. An example of this approach is disclosed in Cibukskis et al. Nature Biotechnology. 2013, 31, 213-219. The genotyping method disclosed above was used to significantly improve alignment of reads (in both tumor and normal) and hence improve the sensitivity and specificity of somatic mutation calling within the HLA region.

An overview of the method is diagrammed in Fig. 2. In this setting, the two inferred alleles for each HLA gene would serve as patient-specific reference 'chromosomes' against which pre-selected HLA reads from the tumor and germline samples are aligned separately followed by standard mutation calling. An analysis pipeline to call somatic mutations in the HLA genes was built that includes the following steps: (i) ethnicity detection using the normal sample; (ii) HLA typing by applying HLA typing method disclosed herein on the normal sample; (iii) re-alignment of the HLA reads in both tumor and normal to the inferred HLA-alleles while filtering out likely erroneous alignments; (iv) applying a mutation detection tool, such as MuTect (Cibukskis et al. and Saunders et al. Bioinformatics. 2012 28, 1811-1817) to detect somatic mutations by comparing the re-aligned tumor and normal HLA reads.

[00110] Regarding step (iii), prior to detection of somatic changes using the mutation detection method by comparison of tumor and normal HLA read aligned to the inferred patient-specific HLA alleles, the following changes and filters were implemented: (i) NotPrimaryAlignment bit flag was turned off from all alignments since several reads mapped to multiple alleles; (ii) mapping quality was changed to a non-zero value (=70) for all reads; (iii) alignments where both mates did not align to the same reference allele were discarded; and (iv) alignments where at least one mate had more than one mutation, insertion or deletion event compared to the reference allele were discarded. Soft-clipping of the reads was not allowed during the alignment. Alleles with multiple detected somatic changes were removed from the analysis. In cases where both inferred alleles are identical in the region of detected somatic mutation, the mutation was assigned to the more common allele in the population. All somatic events were visualized using IGV (Mutect: 'KEEP' entries in call_stats file, Strelka: All entries in all.somatic.indels.vcf file) and the ones that passed manual review were further annotated for the gene compartment (intron, exon, splice site) and protein change. Splice sites were defined as the set of splice consensus sequence positions that had a bit score of at least 1 in either the human major/U2 or human minor/U12 introns at the exon/intron boundaries (9 positions at the 5' splice donor end of the intron including the ultimate base in the upstream exon, and 2 positions at the 3' splice acceptor end of the intron. (Irimia et al. Cold Spring Harbor perspectives in biology 6 (2014).

[00111] To test this approach, a dataset of 2,545 cases of matched tumor and germline DNA spanning 12 tumor types was assembled - 10 from the The Cancer Genome Atlas project

(TCGA), and two separate genomic studies focusing on chronic leukocytic leukemia and melanoma. 59 HLA gene (including HLA-A, B and C) somatic mutations were previously detected using standard methods and reported as part of a pan-cancer analysis effort (Omberg et al., Nature genetics. 2013 45, 1121-1126; Wang et al., The New England Journal of Medicine. 2011, 365:2497-2506; Hodis et al., Cell. 2012, 150:251-263). On re-analysis of these cases with mutation detection pipeline disclosed herein, 36 of the 59 (61%) previously reported HLA mutations were detected, as well as 37 novel HLA somatic mutations; in total, 73 mutations in 64 of the 2,545 cases. *See* Figs 12 and 13. Manual review of all HLA mutation events using IGV (Robinson et al., Nature biotechnology. 2011, 29:24-26), suggested that 9 of the 23 mutations identified exclusively by TCGA were true events, of which 6 were just below the detection limit of the initial pipeline and were identified once the read filtering criteria used prior to mutation calling were slightly relaxed.

[00112] When available, matching RNA-Sequencing data was examined for orthogonal evidence of expression of the somatically mutated HLA allele that was detected by WES (indel calls were excluded from this analysis due to low reliability of indel alignment and detection by RNA-Seq). In total, RNA-Seq data for 51 of 96 mutations was evaluated, including 11 that were exclusively reported by TCGA, 18 detected only by the methods disclosed herein and 22 that were detected by both. A high rate of RNA-Seq based validation of missense, nonsense and splice-site mutations in the 22 common set was observed (8 of 8; 8 of 11; and 2 of 3 events, respectively. Figure 14. Similar high rates of validation for events identified exclusively by mutation detection methods disclosed herein were likewise observed. (7 of 9; 5 of 6; and 3 of 3 events respectively). By contrast, only 2 of the 11 mutations uniquely identified by TCGA were validated using RNA-Seq. These results support that the mutation detection methods disclosed herein provide both sensitive and specific somatic mutation detection within the highly polymorphic HLA loci.

Example 5

Patterns of somatic HLA mutation across tumor types

[00113] The mutation detection method disclosed herein was extended to a total of 3,608 TCGA tumor/normal pairs (including the original collection of 2,545 and 1,063 additional

cases). In total, 147 somatic HLA mutations in 121 of the 3,608 (3.3%) individuals were detected.

[00114] Consistent with the expected loss-of-function consequence, the somatic HLA mutations were distributed across the entire gene. *See* Fig. 15. Interestingly, differences amongst the cancer types in frequency, localization and types of somatic HLA mutations were observed. HLA mutations in HNSC (HLA-A, HLA-B) and LUSC (HLA-A) have previously been found to be significant by MutSig (Mutation Significance) analysis. (Lawrence et al. Nature. 2014 505, 495-501). Using the mutation detection methods disclosed herein, HLA-A (FDR $q=2.3x10^{-08}$) and HLA-B (FDR $q=3.9x10^{-07}$) were identified to be significantly mutated in colon adenocarcinoma. On the other hand, CLL (n=129) and OV (n=300) entirely lacked HLA mutations, and only a single mutation was detected in GBM (n=320).

[00115] It was also observed that 70 of the 147 total HLA mutations (47.6%) fell in 23 recurrent positions (amino acids that were mutated at least twice). The recurrent sites were distributed across the HLA gene (median of 2 mutated cases/recurrent site (range 2-10). *See* bottom of Fig. 15.

Example 6

Somatic class I HLA mutations are enriched for localization at sites affecting peptide-MHC interactions.

[00116] Alterations highly likely to have a functional effect, including loss-of-function events (nonsense or frameshift mutations), were significantly enriched in HLA mutations compared to non-HLA mutations (Fig. 16, chi-squared test $P< 2.2 \times 10^{-16}$) and were distributed throughout the gene (Fig. 15).

[00117] The highest frequency of mutations occurred in exon 4 (54 mutations, 36.7%) which encodes the α3 domain of the HLA protein that binds to the CD8 co-receptor of T cells. (Fayen et al. Molecular Immunology, 1995 32, 267-275 (1995), See Fig. 17. Abrogation of this function could lead to a loss of T cell recognition and thereby a loss of immune reactivity. The second highest frequency of mutations occurred in exon 3 (31 mutations, 21%) followed by exon 2 (25 mutations, 17%), which encode the α1 and α2 peptide binding domains of the HLA molecule respectively which conventionally bind 9- and 10-mer peptides for antigen presentation. (Brusic et al. Immunol Cell Biol 80, 280-285 (2002).

[00118] Analysis of the position of the mutated residues within exons 2 and 3 in relationship to their predicted interaction with binding peptide further strongly suggest alteration of immune function by these somatic HLA mutations. The two major anchor grooves in the HLA molecule bind to positions 2 and 9 respectively of the peptide and mutation in either groove would be expected to profoundly effect on the biochemical stability of the MHC-peptide complex (Brusic et al.). A secondary anchor groove that interacts primarily with the sixth amino acid of the peptide lies between the two primary anchor grooves. (Ruppert et al. Cell 74, 929-937 (1993). Overall, 32% of mutations (18 of 56) in the peptide binding domains were in residues that come in contact with the peptide and 83% (15 of 18) of these were in positions that comprised one of the two primary anchor grooves. See Figure 18.

* * *

[00119] Having thus described in detail preferred embodiments of the present invention, it is to be understood that the invention defined by the above paragraphs is not to be limited to particular details set forth in the above description as many apparent variations thereof are possible without departing from the spirit or scope of the present invention.

WHAT IS CLAIMED IS:

1. A computer-implemented method for genotyping polymorphic genes from massively parallel sequencing data, comprising:

generating, using one or more computing devices, an alignment of reads in a sequencing data set to a gene reference sequence set, each gene reference sequence in the gene reference sequence set corresponding to an allele variant of the polymorphic gene;

determining, using the one or more computer devices, a first posterior probability or first posterior probability derived score for each allele variant in the alignment, wherein the first posterior probability score or posterior probability derived score for each allele variant is based at least in part on a quality score or base quality scores of the reads that aligned to that allele variant;

identifying, using the one or more computer devices, a first allele variant by selecting the first allele variant with a maximum first posterior probability or posterior probability derived score;

identifying, using the one or more computer devices, one or more overlapping reads that aligned with the identified first allele variant and also aligned with one or more other allele variants in the alignment;

determining, using the one or more computing devices, a second posterior probability or posterior probability derived score for each allele variant in the gene reference sequence set, wherein a weighting factor is applied to the score contribution of each aligned read based on whether or not the read was also aligned to the first identified allele variant, and wherein the weighting factor is based at least in part on the corresponding first posterior probability or posterior probability derived score for each of the one or more overlapping reads; and

identifying, using the one or more computing devices, a second allele variant by selecting the allele with a maximum second posterior probability or posterior probability derived score, wherein the identified first and second allele variants indicate a polymorphic gene type.

2. The method of claim 1, wherein the reads in the massively parallel sequencing data set consists of reads that map to a reference genetic sequence of the polymorphic gene

within a threshold base number value, wherein the threshold base number value is between approximately 0.5 Kb bases and approximately 5 Kb bases.

3. The method of claim 2, wherein the threshold base number value is 1Kb.

4. The method of claim 1, wherein the reads in the massively parallel sequencing data set consist of reads that match one or more probes from a polymorphic gene probe set, wherein the reads match one or more probes in a 5' to 3' or 3' to 5' orientation, and wherein the one or more probes are derived from a library of known or inferred genomic and or cDNA sequences for the polymorphic gene.

5. The method of claim 4, wherein the reads have between approximately 90 % sequence identity and approximately 100 % sequence identity to one or more probes in the polymorphic gene probe set.

6. The method of claim 4, wherein the reads have approximately 100% sequence identity to one or more probes in the polymorphic gene probe set.

7. The method of claim 4, wherein the one or more probes in the polymorphic gene probe set have a size between approximately 25 mer and approximately 100 mer.

8. The method of claim 4, wherein the one or more probes in the polymorphic gene probe set have a size of 38 mer.

9. The method of claim 4, wherein the one or more probes in the polymorphic gene probe set have a size equal to half the read length in the sequencing experiment.

10. The method of claim 1, wherein the first and second posterior probability or posterior probability derived scores are determined based at least in part on base quality scores and an insert size probability value for each read in the alignment, wherein the insert size probability value is based at least in part on an insert size distribution of all reads in the data set.

11. The method of claim 1, wherein the first and second posterior probabilities or posterior probability derived scores are calculated based at least in part on population-based allele probabilities observed in a known population data set.

12. The method of claim 1, wherein the weighting factor for a given read mapping to the identified first allele variant and the other allele variant is equal to the contribution of the read to the overall posterior probability or posterior probability derived score of other allele variant(s_1) divided by a sum of that contribution and a similar contribution of the read to the overall posterior probability or posterior probability derived score of the first identified allele variant(s_2), wherein the weighting factor $w = s_1/(s_1 + s_2)$, and the new contribution of the read to the overall posterior probability or posterior probability derived score of other allele variant = $w*s_1$.

13. The method of claim 1, wherein the polymorphic gene is a human leukocyte antigen gene.

14. A computer program product, comprising:

a non-transitory computer-executable storage device having computer-readable program instructions embodied thereon that when executed by a computer cause the computer to determine genotypes of polymorphic genes from massively parallel sequencing data, the computer-executable program instructions comprising:

computer-executable program instructions to generate an alignment of reads in a massively parallel sequencing data set to a gene reference sequence set, each gene reference sequence in the gene reference sequence set corresponding to an allele variant of the polymorphic gene;

computer-executable program instructions to determine a first posterior probability or posterior probability derived score for each allele variant based at least in part on the reads aligned to each allele variant;

computer-executable program instructions to identify a first allele variant by selecting the first allele variant with a maximum first posterior probability or posterior probability derived score;

computer-executable program instructions to identify one or more overlapping reads that aligned with the identified first allele variant in the alignment and also aligned with one or more other allele variants in the alignment;

computer-executable program instructions to determine a second posterior probability or posterior probability derived score for each allele variant in the gene reference sequence set, wherein a weighting factor is applied to the score contribution of each aligned read based on whether or not the read was also aligned to the first identified allele variant, and wherein the weighting factor is based at least in part on the corresponding first posterior probability or posterior probability derived score for each of the one or more overlapping reads; and

computer-executable program instructions to identify a second allele variant by selecting the allele variant with a maximum second posterior probability or posterior probability derived score, wherein the identified first and second allele variants indicate a gene type of the polymorphic gene.

15. The computer program product of claim 14, wherein the reads in the massively parallel sequencing data set consists of reads that map to a reference genetic sequence of the polymorphic gene within a threshold base number value, wherein the threshold base number value is between approximately 0.5 Kb and approximately 5 Kb.

16. The computer program product of claim 15, wherein the threshold base number value is 1 Kb.

17. The computer program product of claim 14, wherein the reads in the massively parallel sequencing data set consist of reads that match one or more probes from a polymorphic gene probe set, wherein the reads match one or more probes in a 5' to 3' or 3' to 5' orientation, and wherein the one or more probes are derived from a library of known or inferred genomic or cDNA sequences for the polymorphic gene.

18. The computer program product of claim 14, wherein the first and second posterior probabilities or posterior probability derived scores are determined based at least in part on base quality scores and an insert size probability value for each read in the alignment, wherein the insert size probability value is based at least in part on an insert size distribution of all reads in the data.

19. The computer program product of claim 14, wherein the first and second posterior probabilities or posterior probability derived scores are calculated based at least in part on population-based allele probabilities observed in a known population data set.

20. The computer program product of claim 14, wherein the weighting factor for a given read mapping to the identified first allele variant and the other allele variant is equal to the contribution of the read to the overall posterior probability or posterior probability derived score of other allele variant(s_1) divided by a sum of that contribution and a similar contribution of the read to the overall posterior probability or posterior probability derived score of the first identified allele variant(s_2). Using the notation, the weighting factor $w = s_1/(s_1 + s_2)$, and the new contribution of the read to the overall posterior probability or posterior probability derived score of other allele variant = $w*s_1$.

21. A system to determine genotypes of the polymorphic genes from massively parallel sequencing data, the system comprising:

a storage device; and

a processor communicatively coupled to the storage device, wherein the processor executes application code instructions that are stored in the storage device and that cause the system to:

generate an alignment of reads in a massively parallel sequencing data set to a gene reference sequence set, each gene reference sequence in the gene reference sequence set corresponding to an allele variant of the polymorphic gene;;

determine a first posterior probability or probability derived score for each allele variant based at least in part on the reads aligned to each allele variant;

identify a first allele variant by selecting the first allele variant with a maximum first posterior probability or probability derived score;

to determine a second posterior probability or posterior probability derived score for each allele variant in the database, wherein a weighting factor is applied to the score contribution of each aligned read based on whether or not the read was also aligned to the first identified allele variant, and wherein the weighting factor is based at least in part on the corresponding first posterior probability or posterior probability derived score for each of the one or more overlapping reads; and

identify a second allele variant by selecting the allele variant with a maximum second posterior probability or posterior probability derived score, wherein the identified first and second allele variants indicate a gene type of the polymorphic gene.

22. A computer-implemented method to detect mutations in polymorphic genes, the method comprising:

determining a polymorphic gene type of a polymorphic gene using the method of claim 1, wherein the gene type is based on an analysis of a first sequencing data set obtained from a normal tissue sample of a subject;

generating, using the one or more computing devices, an alignment of reads in a second sequencing data set to the determined polymorphic gene type, the second sequencing set obtained from a cancerous or otherwise disease sample of the subject; and

detecting, using the one or more computing devices, one or more mutations based at least in part on the generated sequence alignment.

23. The method of claim 22, wherein the first sequencing data set, the second sequencing data set, or both are massively parallel sequencing data sets.

24. The method of claim 22, wherein the reads in the first and second sequencing data set consists of reads that map to a reference genetic sequence of the polymorphic gene within a threshold base number value, wherein the threshold base number value is between approximately 0.5 Kb bases and approximately 5 Kb bases.

25. The method of claim 22, wherein the reads in the first and second sequencing data set consist of reads that match one or more probes from a polymorphic gene probe set, wherein the reads match one or more probes in a 5' to 3' or 3' to 5' orientation, and wherein the one or more probes are derived from a library of known or inferred genomic or cDNA sequences for the polymorphic gene.

26. The method claim 21, wherein the reads in the first and second massively parallel sequencing data set consist of a union of reads described in claims 22 or 23.

27. The method of claim 23, wherein the reads have approximately 100% sequence identity to one or more probes in the polymorphic gene probe set.

28. A method of identifying a plurality of at least 4 subject-specific peptides and preparing a subject-specific immunogenic composition that upon administration presents the plurality of at least 4 subject-specific peptides to the subject's immune system, wherein the subject has a tumor and the subject-specific peptides are specific to the subject and the subject's tumor, said method comprising:

(i) identifying, including through

nucleic acid sequencing of a sample of the subject's tumor and

nucleic acid sequencing of a non-tumor sample of the subject,

a plurality of at least 4 tumor-specific non-silent mutations not present in the non-tumor sample; and

(ii) selecting from the identified non-silent mutations the plurality of at least 4 subject-specific peptides, each having a different tumor neo-epitope that is an epitope specific to the tumor of the subject, from the identified plurality of tumor specific mutations,

wherein each neo-epitope is an expression product of a tumor-specific non-silent mutation not present in the non-tumor sample, each neo-epitope binds to a HLA protein of the subject, and selecting includes

generating, using one or more computing devices, an alignment of reads in the sequencing data set from the non-tumor sample to a gene reference sequence set, each gene reference sequence in the gene reference sequence set corresponding to an allele variant of the polymorphic gene;

determining, using the one or more computer devices, a first posterior probability or first posterior probability derived score for each allele variant in the alignment, wherein the first posterior probability score or posterior probability derived score for each allele variant is based at least in part on a quality score or base quality scores of the reads that aligned to that allele variant;

identifying, using the one or more computer devices, a first allele variant by selecting the first allele variant with a maximum first posterior probability or posterior probability derived score;

identifying, using the one or more computer devices, one or more overlapping reads that aligned with the identified first allele variant and also aligned with one or more other allele variants in the alignment;

determining, using the one or more computing devices, a second posterior probability or posterior probability derived score for each allele variant in the gene reference sequence set, wherein a weighting factor is applied to the score contribution of each aligned read based on whether or not the read was also aligned to the first identified allele variant, and wherein the weighting factor is based at least in part on the corresponding first posterior probability or posterior probability derived score for each of the one or more overlapping reads; and

identifying, using the one or more computing devices, a second allele variant by selecting the allele with a maximum second posterior probability or posterior probability derived score,

wherein the identified first and second allele variants indicate a polymorphic gene type,

determining binding of the subject-specific peptides to the HLA proteins,

and

(iii) formulating the subject-specific immunogenic composition for administration to the subject so that upon administration the plurality of at least 4 subject-specific peptides are presented to the subject's immune system,

wherein the selecting or formulating comprises at least one of:

including in the subject-specific immunogenic composition a subject-specific peptide that includes an expression product of an identified neo-ORF, wherein a neo-ORF is a tumor-specific non-silent mutation not present in the non-tumor sample that creates a new open reading frame, and

including in the subject-specific immunogenic composition a subject-specific peptide that includes an expression product of an identified point mutation and has a determined binding to

the HLA proteins of the subject with an IC50 less than 500 nM, or the differential of the IC50 values between the native and corresponding mutated peptide being greater than a pre-defined value,

whereby, the plurality of at least 4 subject-specific peptides are identified, and the subject-specific immunogenic composition that upon administration presents the plurality of at least 4 subject-specific peptides to the subject's immune system, wherein the subject-specific peptides are specific to the subject and the subject's tumor, is prepared.

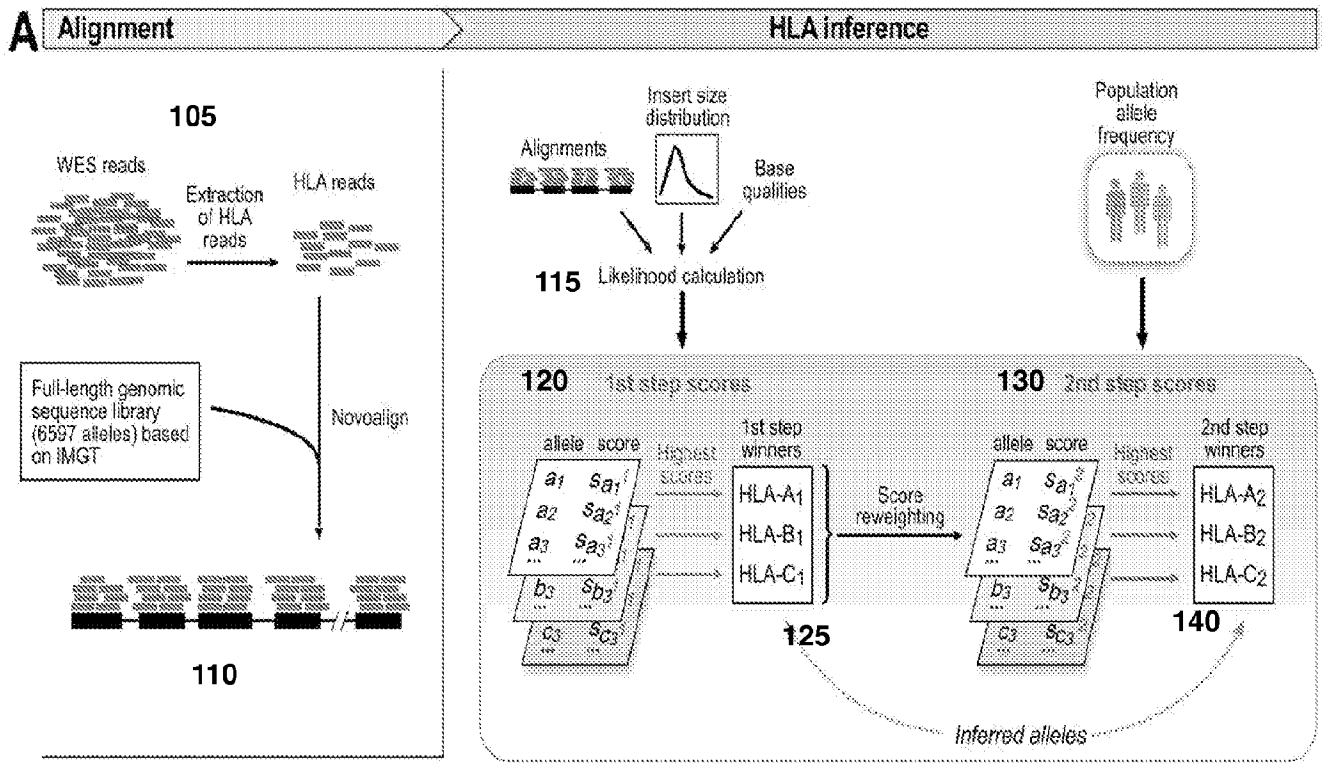
A
Alignment
HLA inference
105
WES reads
Extraction of HLA reads
HLA reads
Full-length genomic sequence library (6597 alleles) based on IMGT
Novoalign
110
Insert size distribution
Alignments
Base qualities
115
Likelihood calculation
Population allele frequency
120
1st step scores
130
2nd step scores
allele score
Highest scores
1st step winners
HLA-A1
HLA-B1
HLA-C1
Score reweighting
2nd step winners
HLA-A2
HLA-B2
HLA-C2
125
140
Inferred alleles

FIG. 1

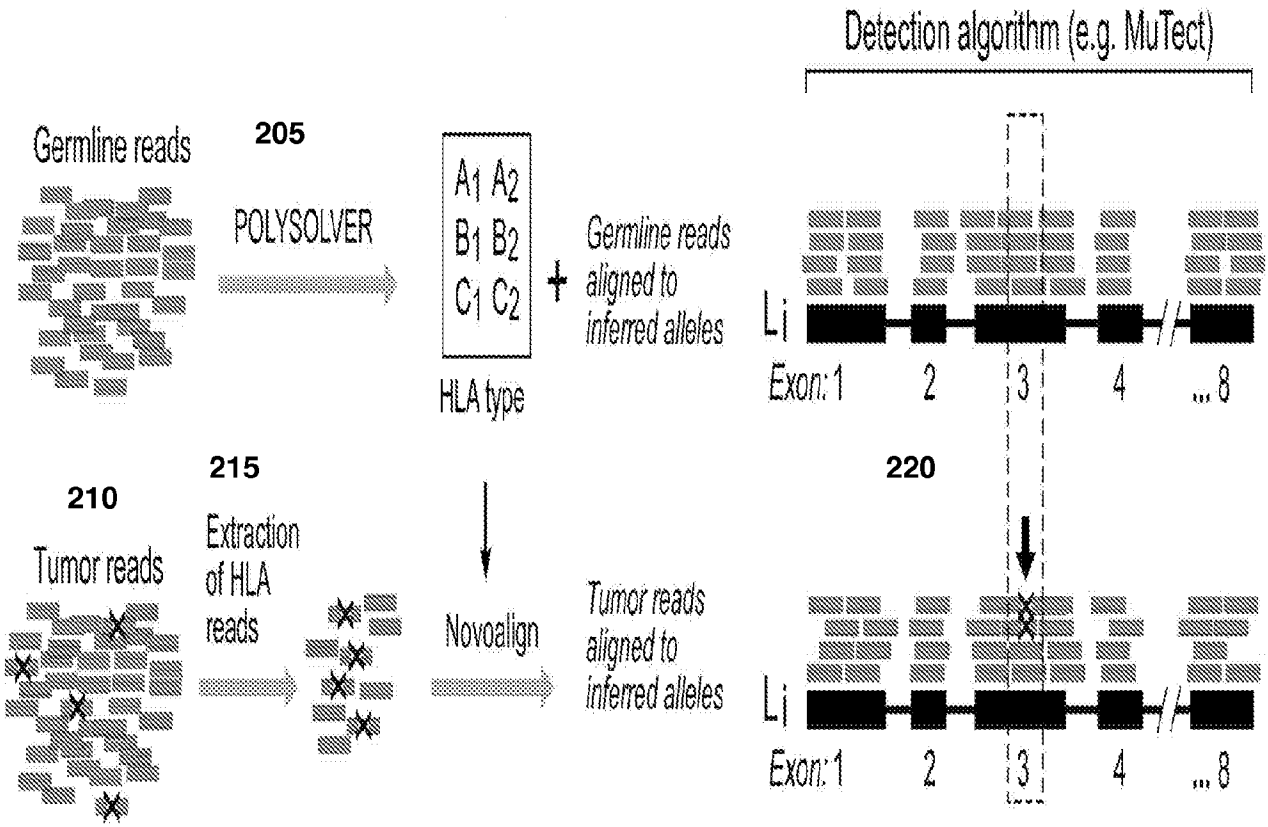
Detection algorithm (e.g. MuTect)
Germline reads
205
POLYSOLVER
A1 A2
B1 B2
C1 C2
HLA type
Germline reads aligned to inferred alleles
Li
Exon: 1 2 3 4 ... 8
210
Tumor reads
215
Extraction of HLA reads
Novoalign
220
Tumor reads aligned to inferred alleles
Li
Exon: 1 2 3 4 ... 8

FIG. 2

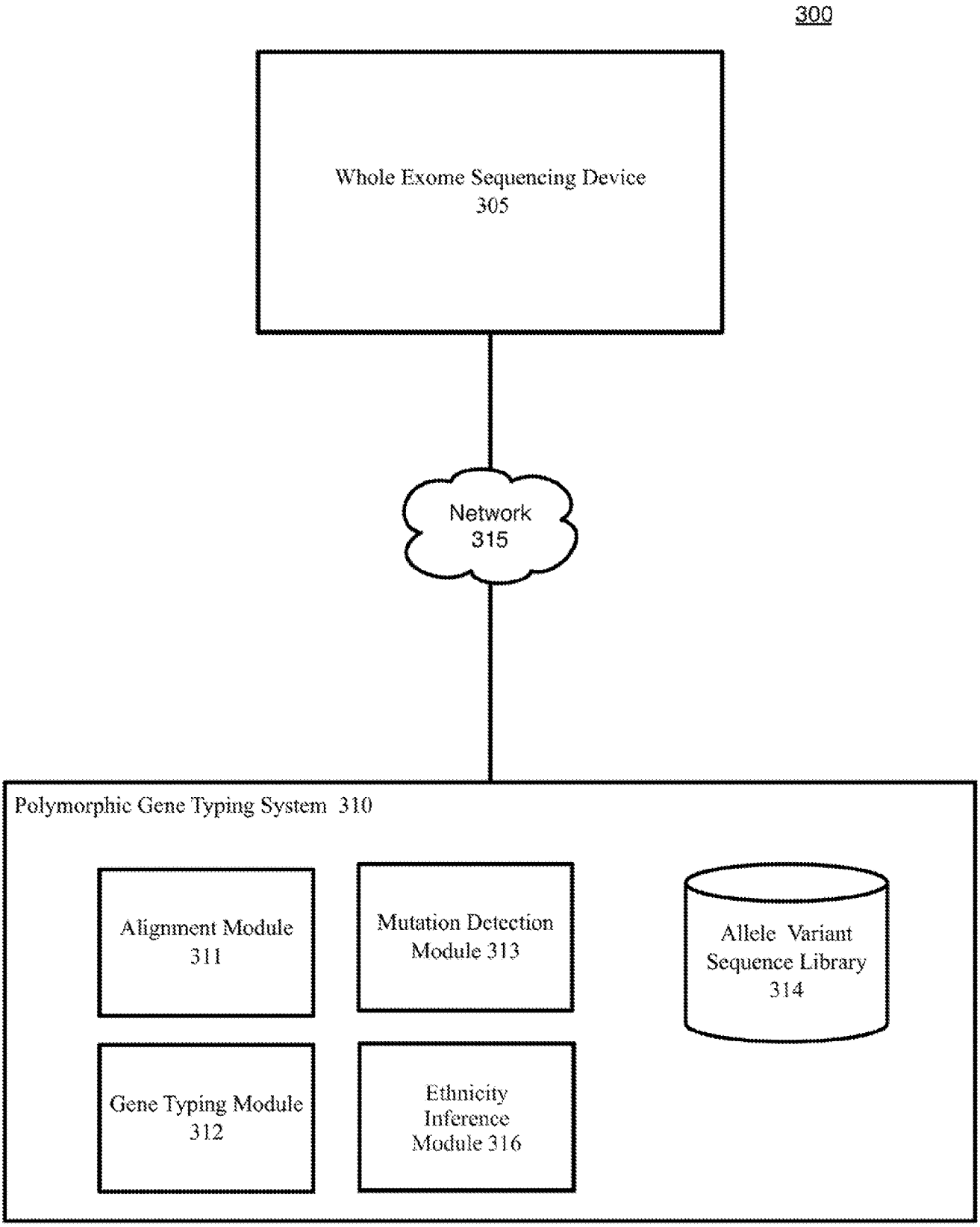
300
Whole Exome Sequencing Device
305
Network
315
Polymorphic Gene Typing System 310
Alignment Module
311
Mutation Detection
Module 313
Allele Variant
Sequence Library
314
Gene Typing Module
312
Ethnicity
Inference
Module 316

FIG. 3

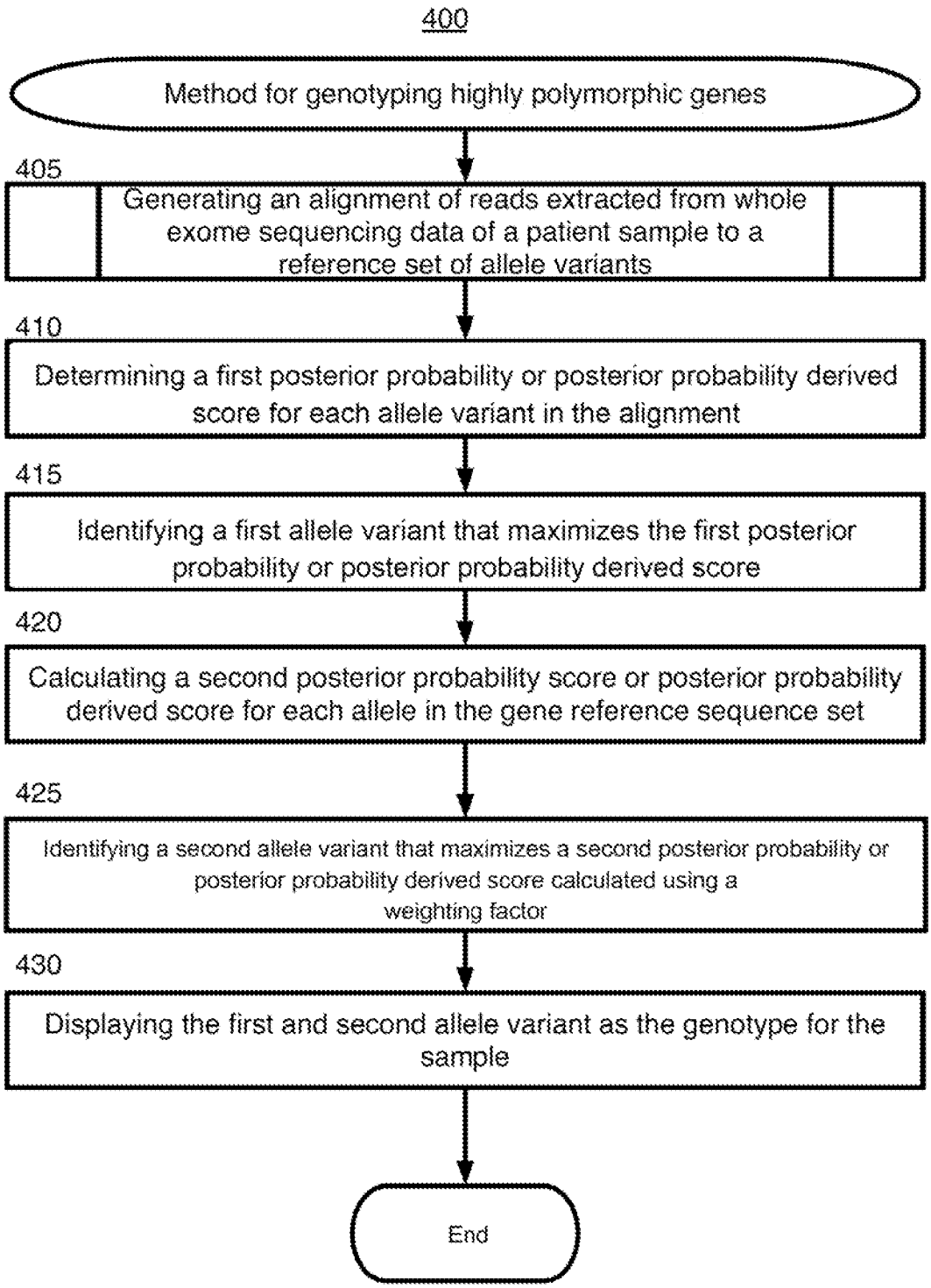
400
Method for genotyping highly polymorphic genes
405
Generating an alignment of reads extracted from whole exome sequencing data of a patient sample to a reference set of allele variants
410
Determining a first posterior probability or posterior probability derived score for each allele variant in the alignment
415
Identifying a first allele variant that maximizes the first posterior probability or posterior probability derived score
420
Calculating a second posterior probability score or posterior probability derived score for each allele in the gene reference sequence set
425
Identifying a second allele variant that maximizes a second posterior probability or posterior probability derived score calculated using a weighting factor
430
Displaying the first and second allele variant as the genotype for the sample
End

FIG. 4

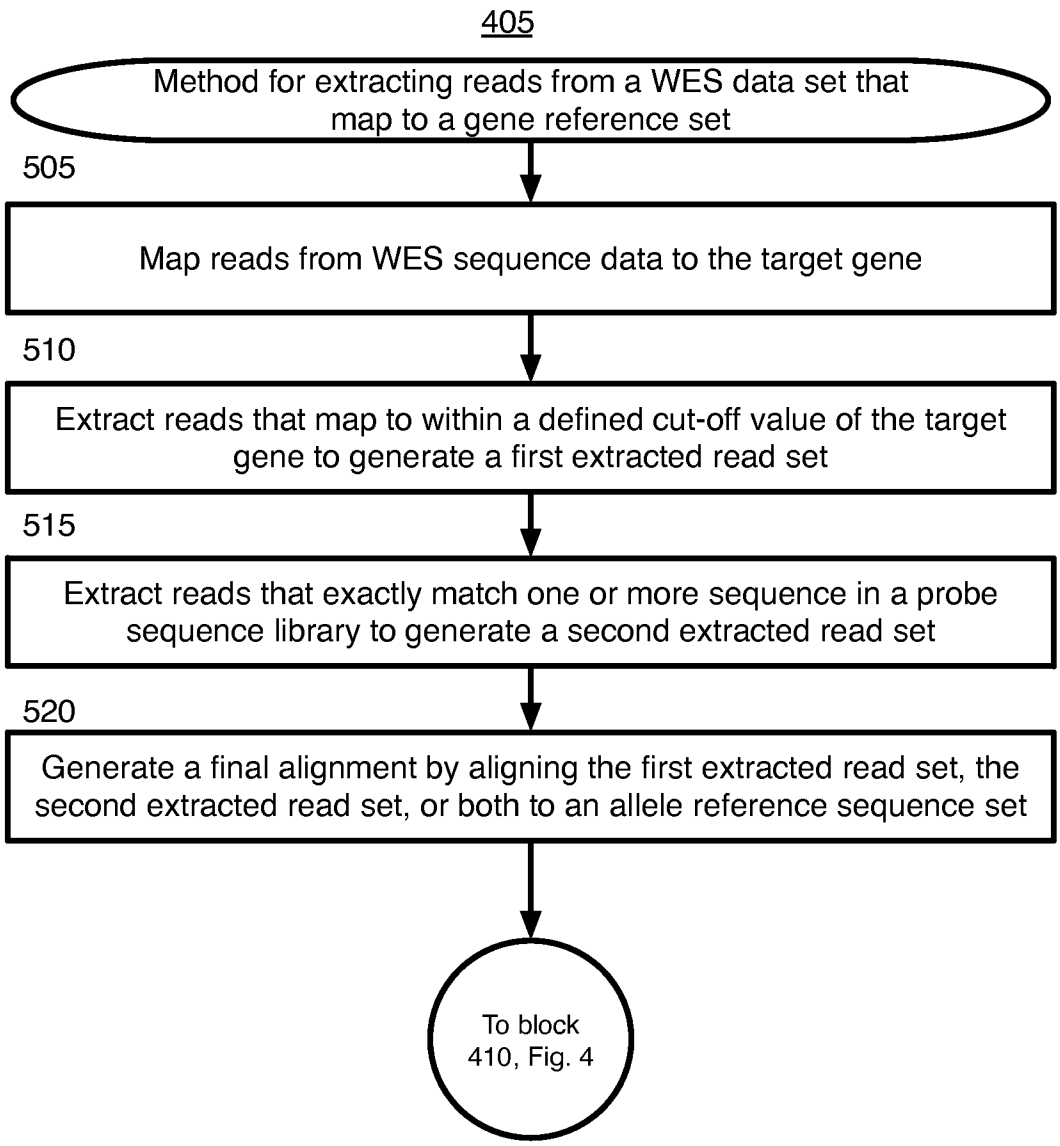
405
Method for extracting reads from a WES data set that map to a gene reference set
505
Map reads from WES sequence data to the target gene
510
Extract reads that map to within a defined cut-off value of the target gene to generate a first extracted read set
515
Extract reads that exactly match one or more sequence in a probe sequence library to generate a second extracted read set
520
Generate a final alignment by aligning the first extracted read set, the second extracted read set, or both to an allele reference sequence set
To block 410, Fig. 4

FIG. 5

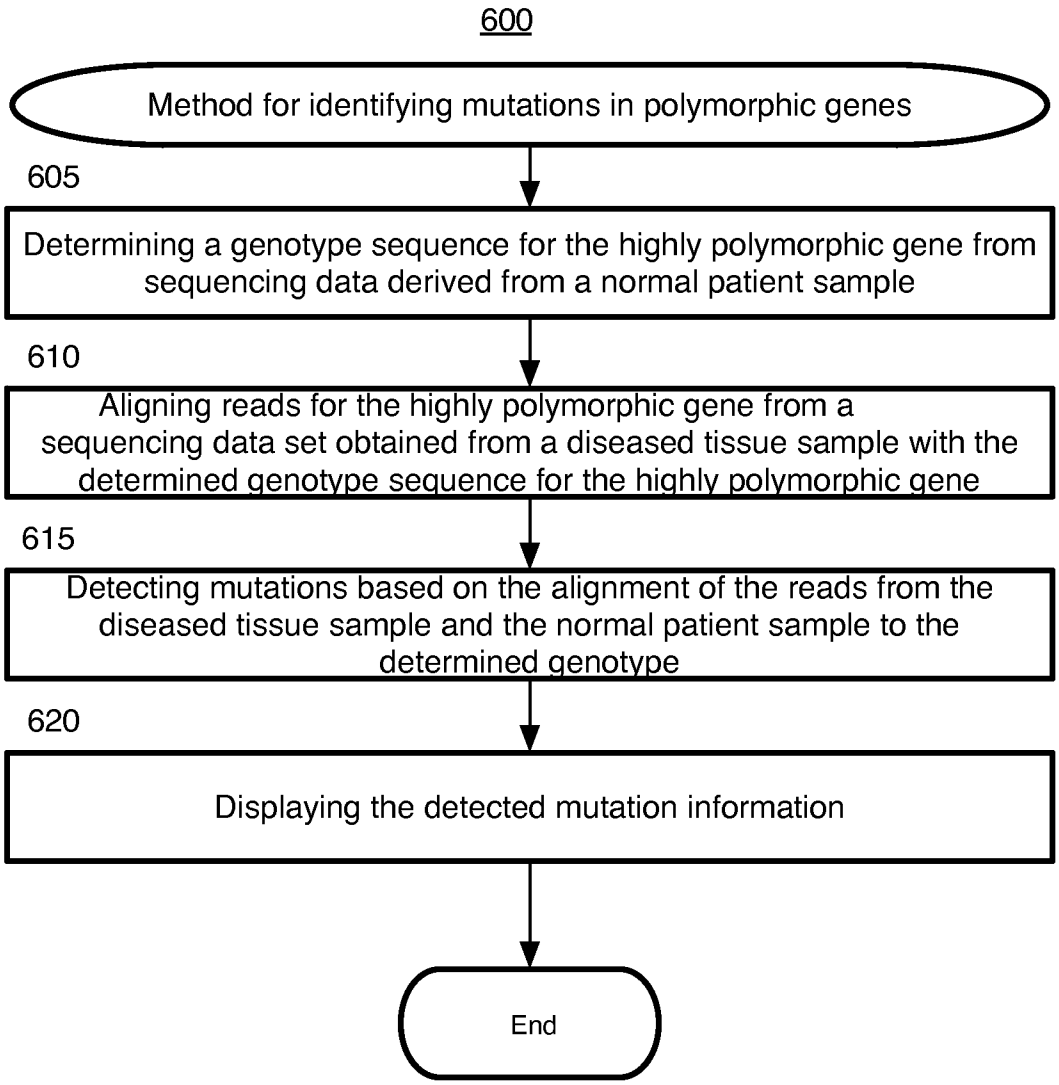
600
Method for identifying mutations in polymorphic genes
605
Determining a genotype sequence for the highly polymorphic gene from sequencing data derived from a normal patient sample
610
Aligning reads for the highly polymorphic gene from a sequencing data set obtained from a diseased tissue sample with the determined genotype sequence for the highly polymorphic gene
615
Detecting mutations based on the alignment of the reads from the diseased tissue sample and the normal patient sample to the determined genotype
620
Displaying the detected mutation information
End

FIG. 6

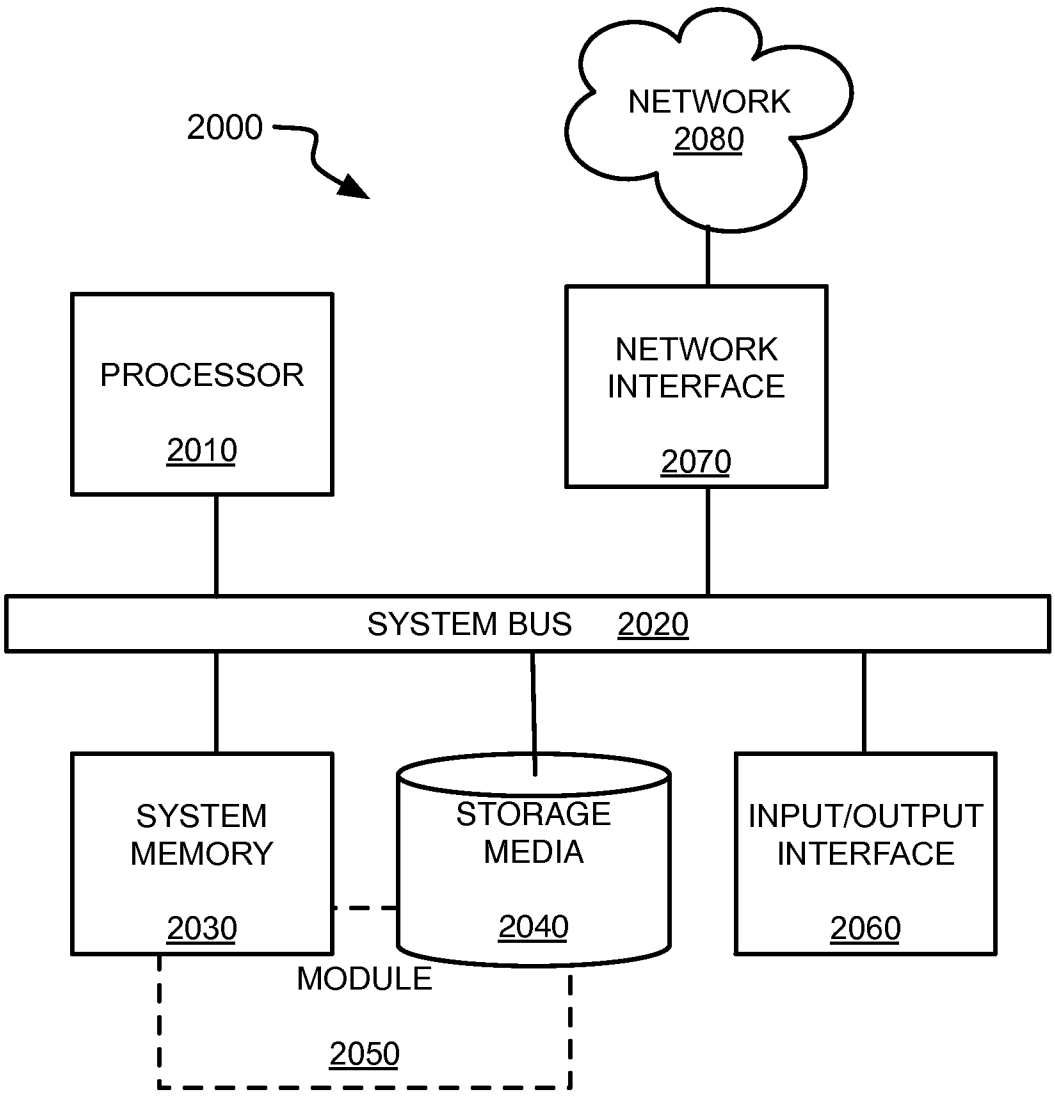
2000
NETWORK
2080
PROCESSOR
2010
NETWORK
INTERFACE
2070
SYSTEM BUS 2020
SYSTEM
MEMORY
2030
STORAGE
MEDIA
2040
INPUT/OUTPUT
INTERFACE
2060
MODULE
2050

FIG. 7

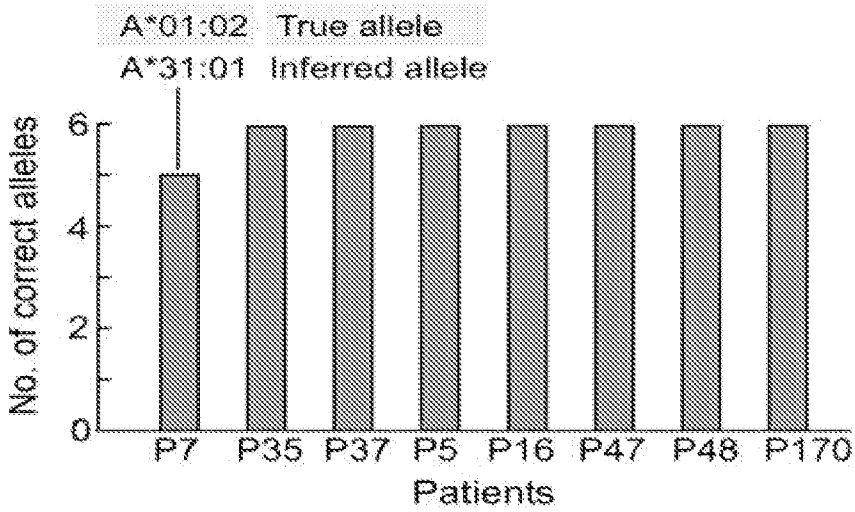
A*01:02 True allele
A*31:01 Inferred allele
No. of correct alleles
0
2
4
6
P7
P35
P37
P5
P16
P47
P48
P170
Patients

FIG. 8

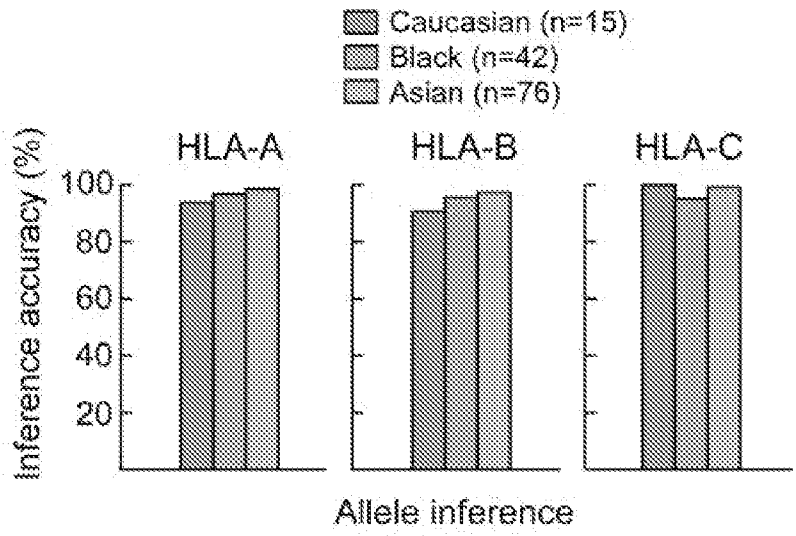
Caucasian (n=15)
Black (n=42)
Asian (n=76)
HLA-A
HLA-B
HLA-C
Inference accuracy (%)
100
80
60
40
20
Allele inference

FIG. 9

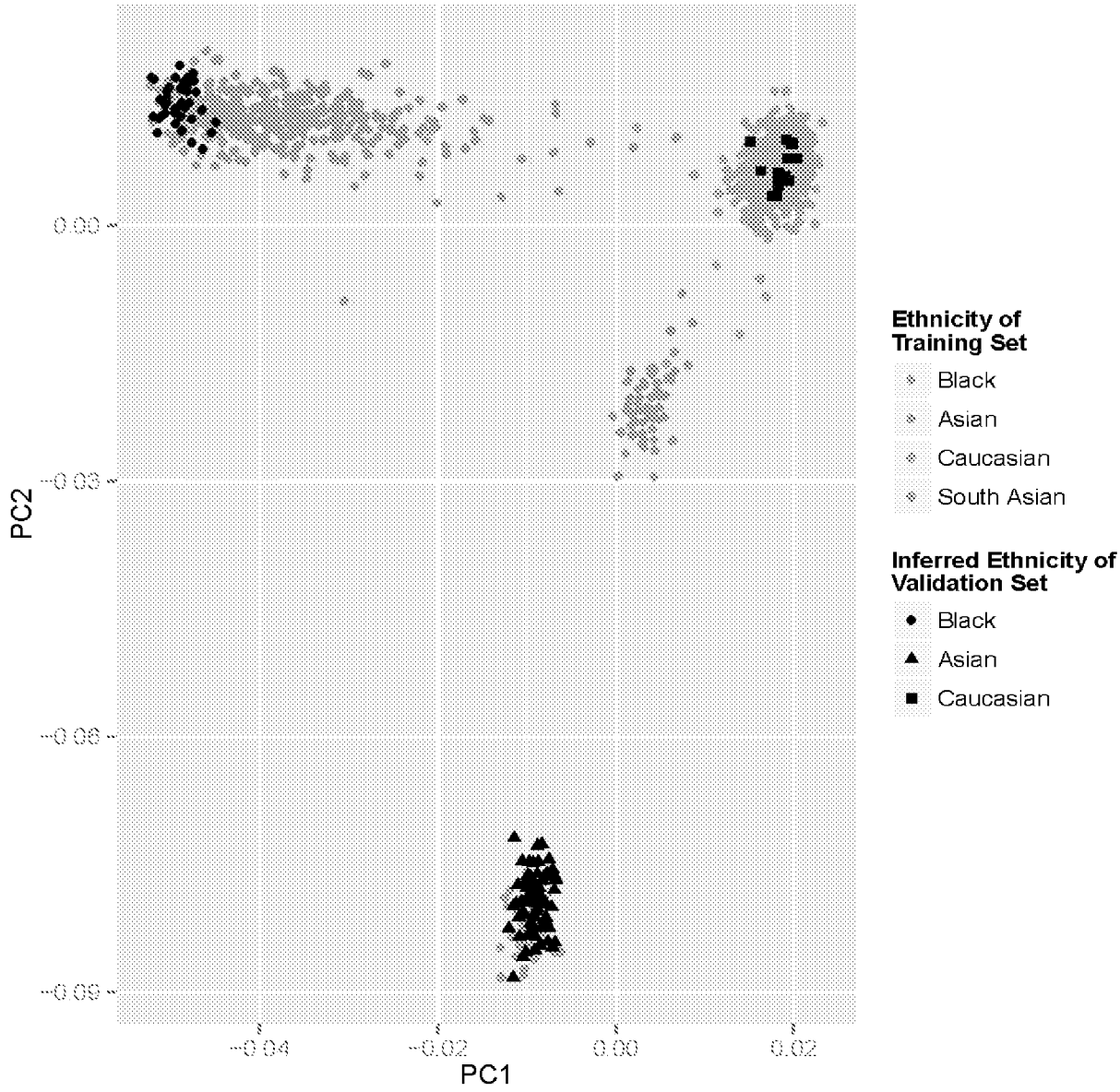
0.00
−0.03
−0.06
−0.09
PC2
−0.04
−0.02
0.00
0.02
PC1
Ethnicity of Training Set
Black
Asian
Caucasian
South Asian
Inferred Ethnicity of Validation Set
Black
Asian
Caucasian

FIG. 10

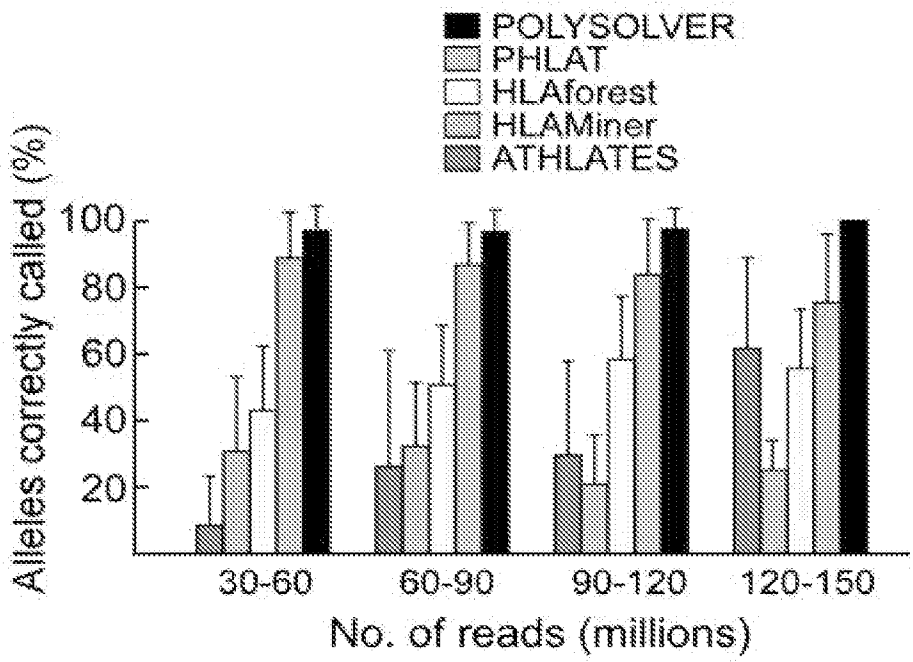
POLYSOLVER
PHLAT
HLAforest
HLAMiner
ATHLATES
Alleles correctly called (%)
100
80
60
40
20
30-60
60-90
90-120
120-150
No. of reads (millions)

FIG. 11

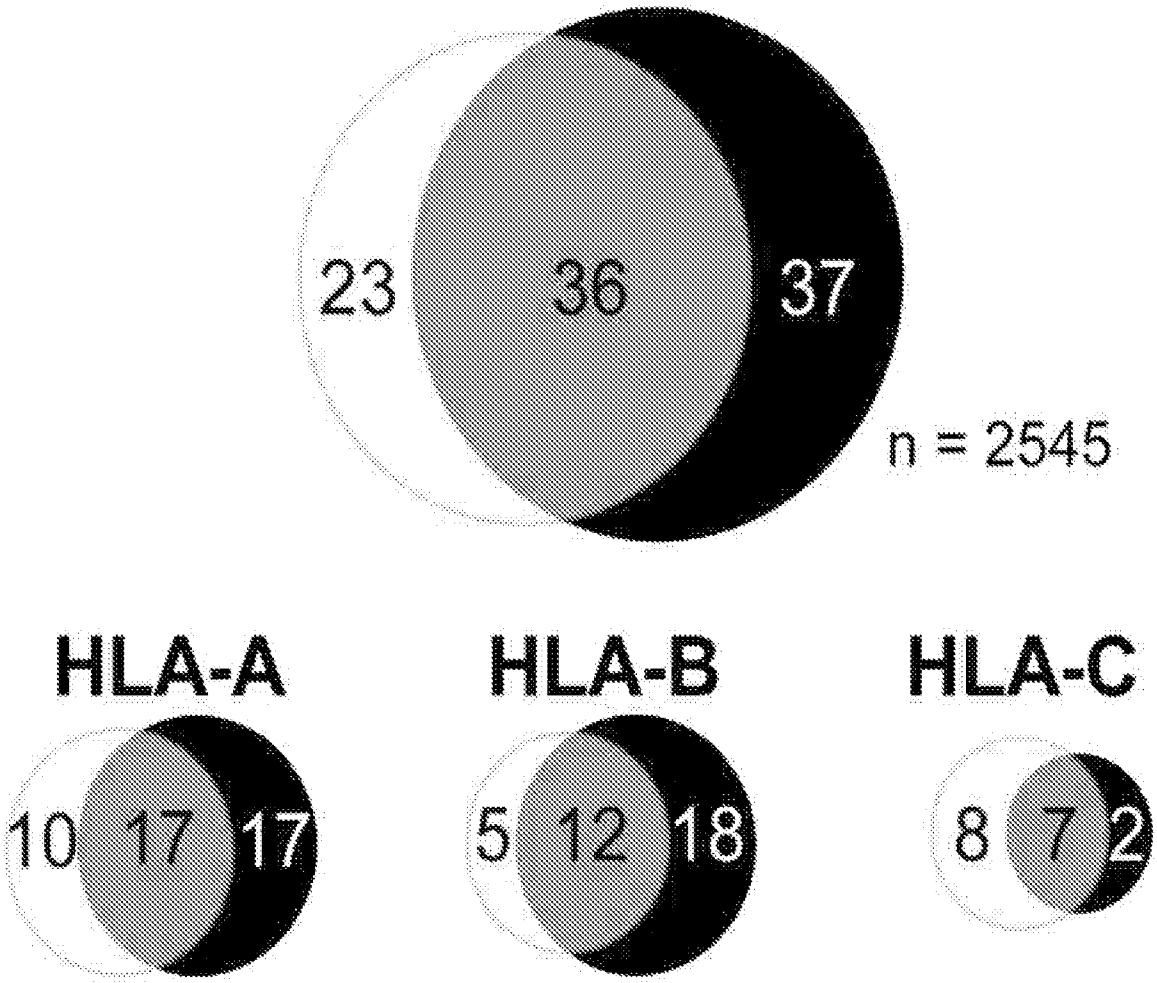

FIG. 12

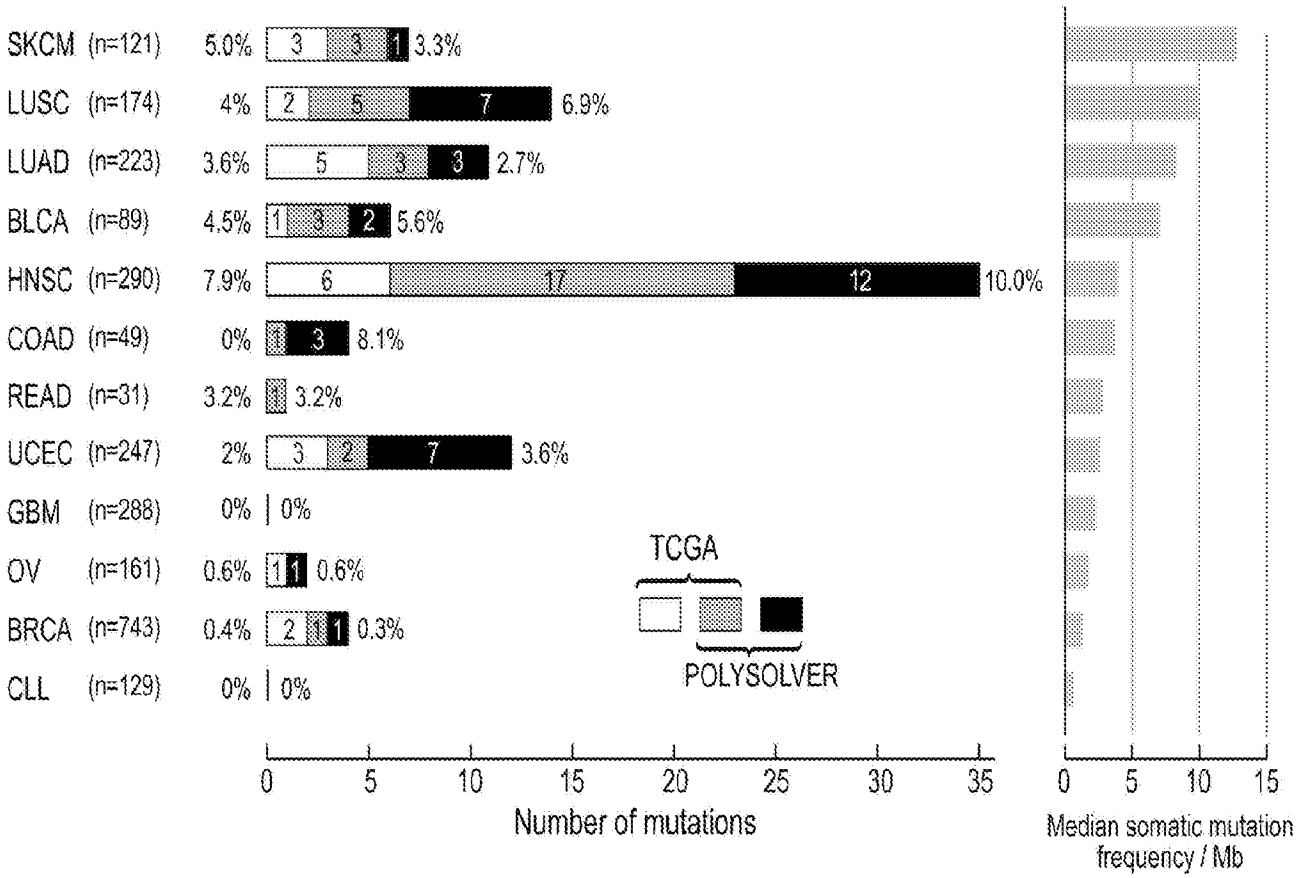
SKCM (n=121) 5.0% 3 3 1 3.3%
LUSC (n=174) 4% 2 5 7 6.9%
LUAD (n=223) 3.6% 5 3 3 2.7%
BLCA (n=89) 4.5% 1 3 2 5.6%
HNSC (n=290) 7.9% 6 17 12 10.0%
COAD (n=49) 0% 1 3 8.1%
READ (n=31) 3.2% 1 3.2%
UCEC (n=247) 2% 3 2 7 3.6%
GBM (n=288) 0% 0%
OV (n=161) 0.6% 1 1 0.6%
BRCA (n=743) 0.4% 2 1 1 0.3%
CLL (n=129) 0% 0%
TCGA
POLYSOLVER
0 5 10 15 20 25 30 35
Number of mutations
0 5 10 15
Median somatic mutation frequency / Mb

FIG. 13

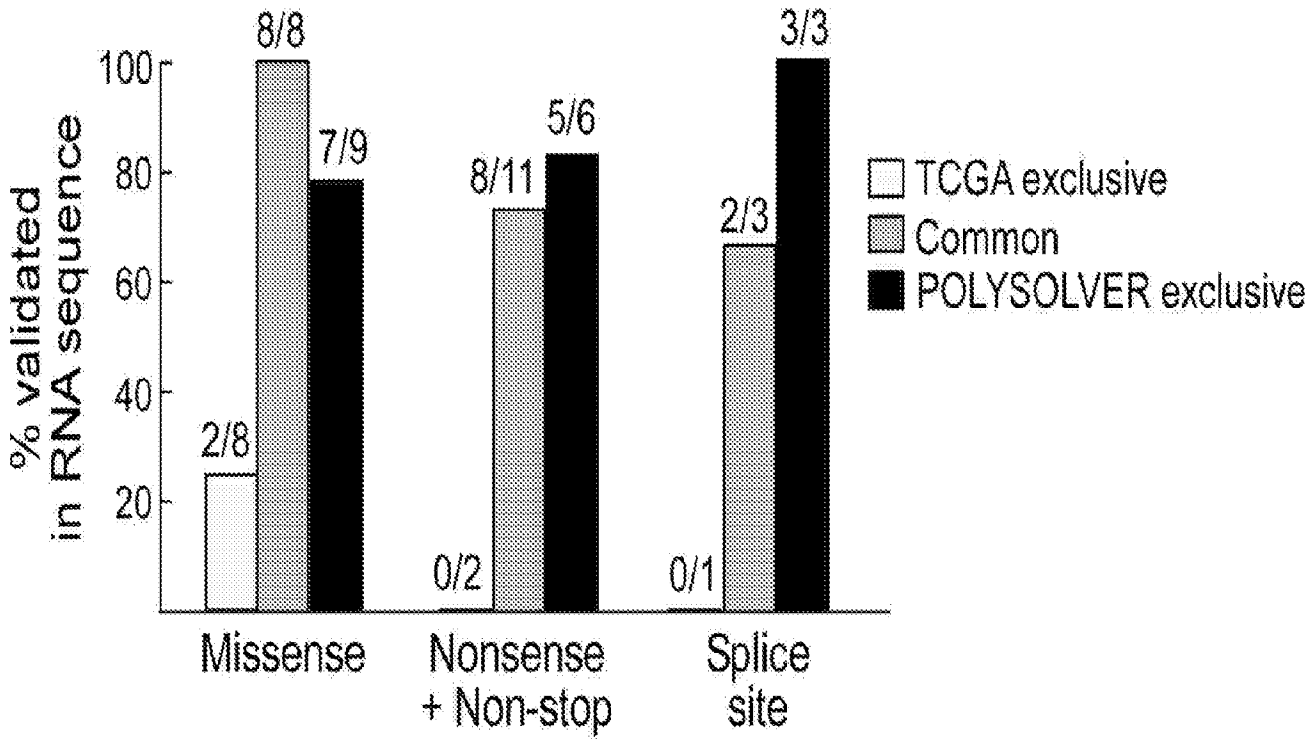
% validated in RNA sequence
100
80
60
40
20
8/8
7/9
2/8
8/11
5/6
0/2
3/3
2/3
0/1
Missense
Nonsense + Non-stop
Splice site
TCGA exclusive
Common
POLYSOLVER exclusive

FIG. 14

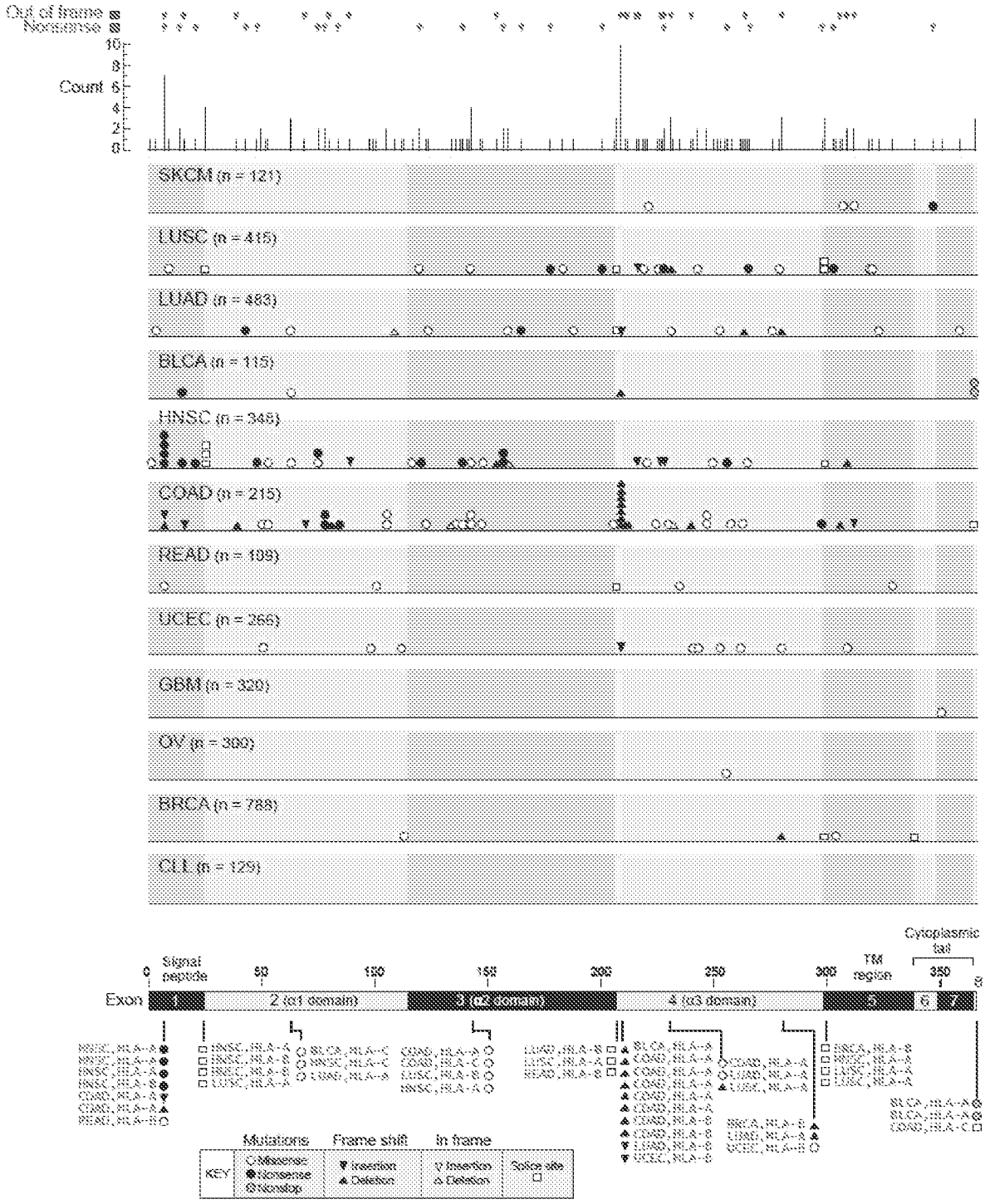
Out of frame
Nonsense
Count
SKCM (n = 121)
LUSC (n = 415)
LUAD (n = 483)
BLCA (n = 115)
HNSC (n = 348)
COAD (n = 215)
READ (n = 109)
UCEC (n = 266)
GBM (n = 320)
OV (n = 302)
BRCA (n = 788)
CLL (n = 129)
Signal peptide
TM region
Cytoplasmic tail
Exon
2 (α1 domain)
3 (α2 domain)
4 (α3 domain)
Mutations
Frame shift
In frame
KEY
Missense
Nonsense
Nonstop
Insertion
Deletion
Splice site

FIG. 15

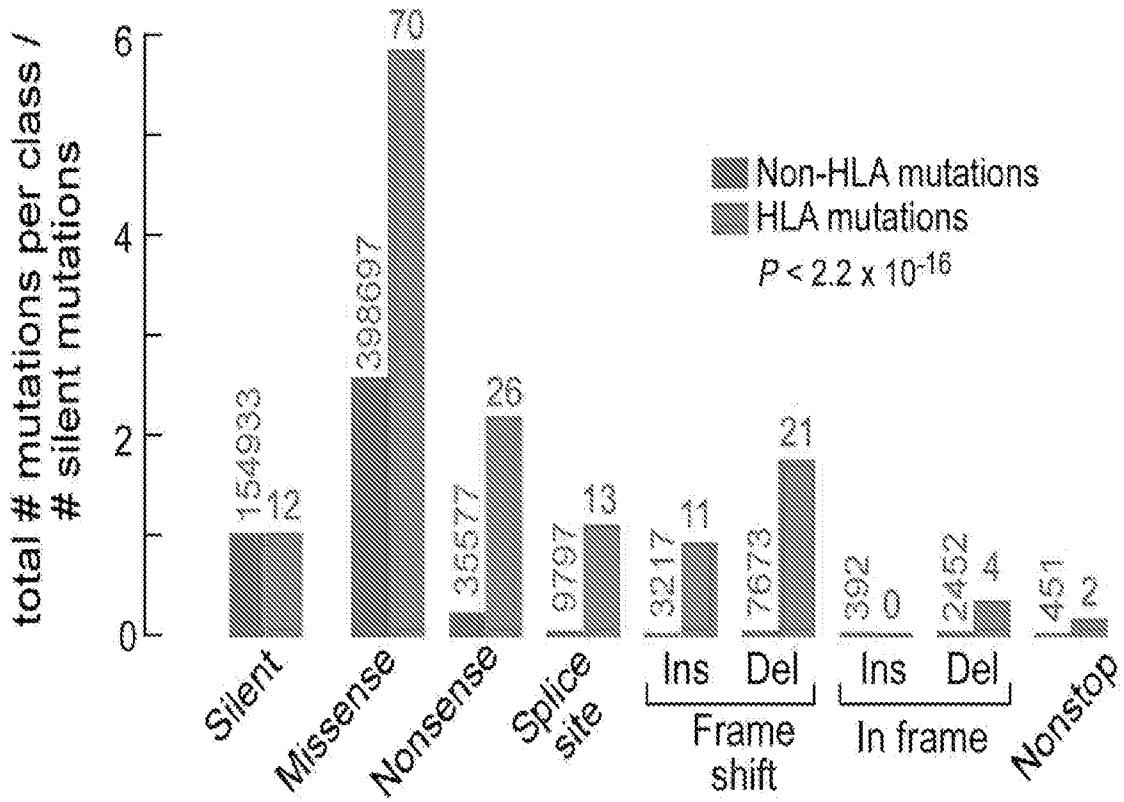
total # mutations per class / # silent mutations
0
2
4
6
1549933
12
398697
70
35577
26
9797
13
3217
11
7673
21
392
0
2452
4
451
2
Non-HLA mutations
HLA mutations
P < 2.2 x 10-16
Silent
Missense
Nonsense
Splice site
Ins
Del
Frame shift
Ins
Del
In frame
Nonstop

FIG. 16

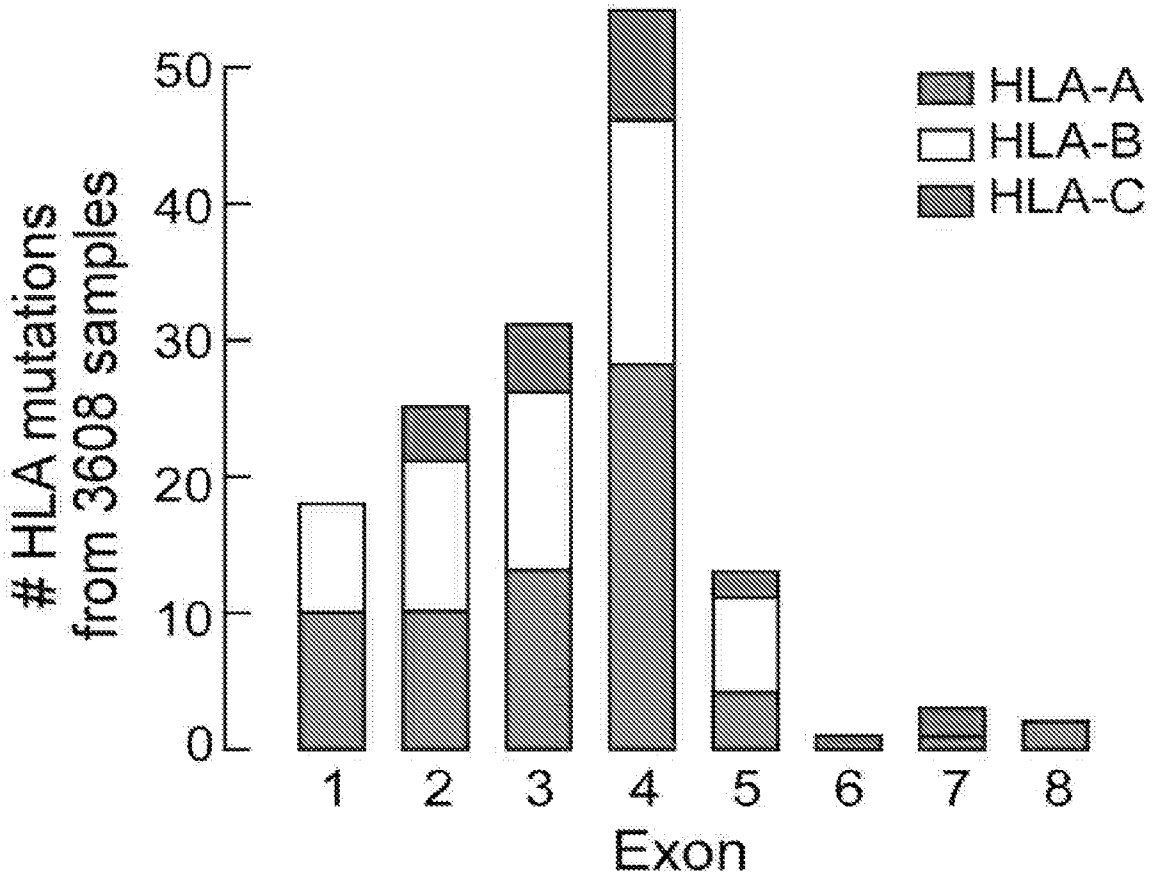
HLA mutations from 3608 samples
50
40
30
20
10
0
1
2
3
4
5
6
7
8
Exon
HLA-A
HLA-B
HLA-C

FIG. 17

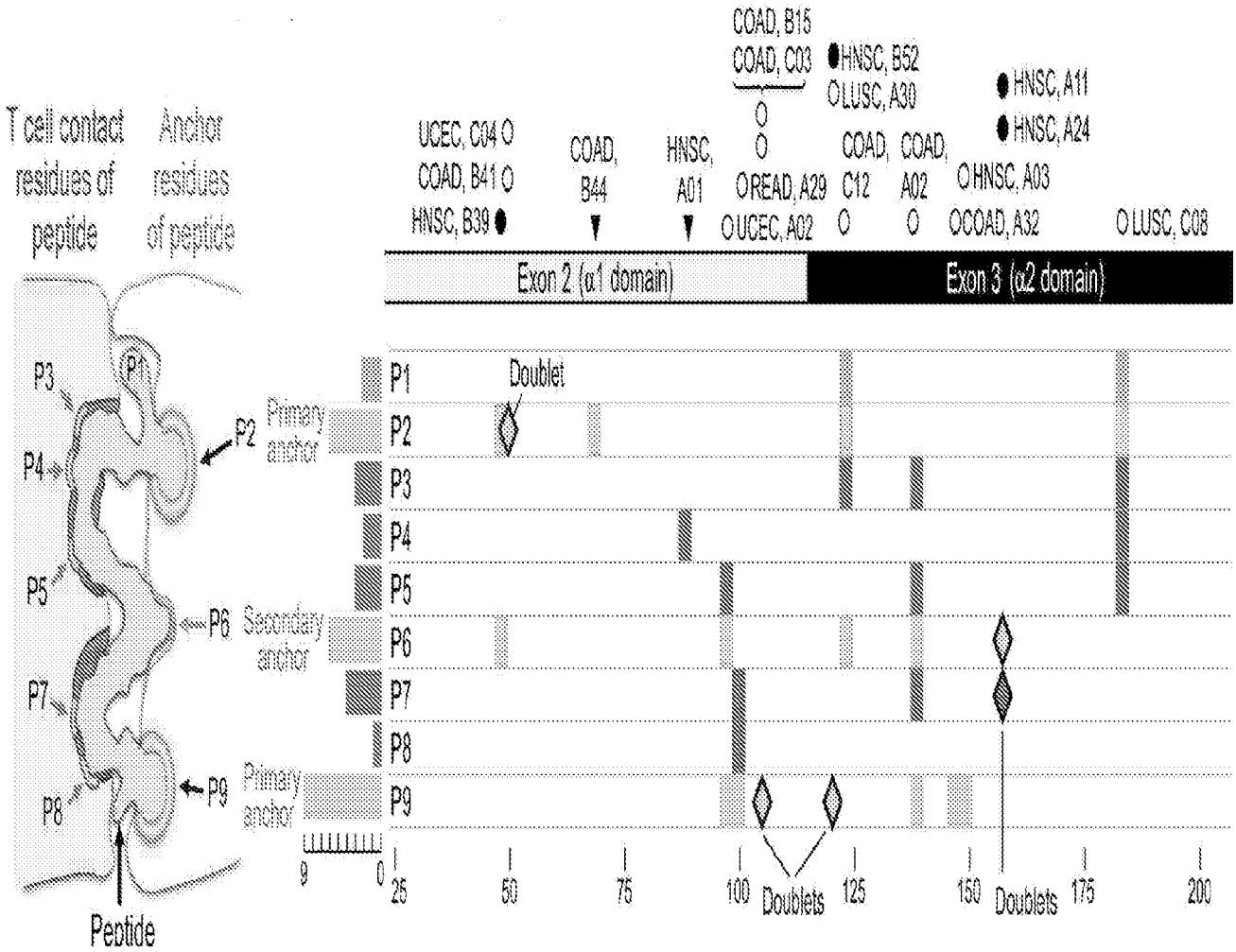
T cell contact residues of peptide
Anchor residues of peptide
P1
P2
P3
P4
P5
P6
P7
P8
P9
Primary anchor
Secondary anchor
Primary anchor
Peptide
UCEC, C04
COAD, B41
HNSC, B39
COAD, B44
HNSC, A01
COAD, B15
COAD, C03
READ, A29
UCEC, A02
HNSC, B52
LUSC, A30
COAD, C12
COAD, A02
HNSC, A11
HNSC, A24
HNSC, A03
COAD, A32
LUSC, C08
Exon 2 (α1 domain)
Exon 3 (α2 domain)
Doublet
Doublets
Doublets
25
50
75
100
125
150
175
200
9
0

FIG. 18

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2014/068746

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68 G06F19/00
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)
C12Q G06F

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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	C. LIU ET AL: "ATHLATES: accurate typing of human leukocyte antigen through exome sequencing", NUCLEIC ACIDS RESEARCH, vol. 41, no. 14, 8 June 2013 (2013-06-08), pages e142-e142, XP055171229, ISSN: 0305-1048, DOI: 10.1093/nar/gkt481 the whole document figure 1 ----- -/--	1-28

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
23 February 2015	23/03/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Sauer, Tincuta

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ERLICH RACHEL L ET AL: "Next-generation sequencing for HLA typing of class I loci", BMC GENOMICS, BIOMED CENTRAL LTD, LONDON, UK, vol. 12, no. 1, 18 January 2011 (2011-01-18), page 42, XP021086454, ISSN: 1471-2164, DOI: 10.1186/1471-2164-12-42 the whole document p. 2-3	1-28
X,P	WO 2014/168874 A2 (BROAD INST INC [US]; DANA FARBER CANCER INST INC [US]; GEN HOSPITAL CO) 16 October 2014 (2014-10-16) the whole document p. 148, ll. 1-17	1-28
X,P	Sachet Ashok Shukla: "Topics in cancer genomics", , 1 January 2014 (2014-01-01), XP055171233, ISBN: 978-1-32-102726-6 Retrieved from the Internet: URL:http://search.proquest.com/docview/1558874754 the whole document pages 7-19	1-28
T	MICHAEL S. ROONEY ET AL: "Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity", CELL, vol. 160, no. 1-2, 1 January 2015 (2015-01-01), pages 48-61, XP055170803, ISSN: 0092-8674, DOI: 10.1016/j.cell.2014.12.033 the whole document page 58	1-28

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2014/068746

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014168874 A2	16-10-2014	NONE	