

(51) International Patent Classification:
C12Q 1/68 (2006.01)(21) International Application Number:
PCT/US2013/045276(22) International Filing Date:
11 June 2013 (11.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/658,317	11 June 2012 (11.06.2012)	US
61/738,277	17 December 2012 (17.12.2012)	US
61/776,647	11 March 2013 (11.03.2013)	US
13/835,093	15 March 2013 (15.03.2013)	US
61/829,054	30 May 2013 (30.05.2013)	US

(71) Applicant (for all designated States except US): **SE-QUENTA, INC.** [US/US]; 400 East Jamie Court, Suite 301, South San Francisco, CA 94080 (US).

(72) Inventors; and

(71) Applicants (for US only): **FAHAM, Malek** [US/US]; 400 East Jamie Court, Suite 301, South San Francisco, CA 94080 (US). **MOORHEAD, Martin** [CH/US]; 400 East Jamie Court, Suite 301, South San Francisco, CA 94080(US). **WILLIS, Thomas** [US/US]; 400 East Jamie Court, Suite 301, South San Francisco, CA 94080 (US). **ZHENG, Jianbiao** [US/US]; 400 East Jamie Court, Suite 301, South San Francisco, CA 94080 (US).(74) Agents: **EASTERDAY, Mathew** et al.; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (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, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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,

[Continued on next page]

(54) Title: METHOD OF SEQUENCE DETERMINATION USING SEQUENCE TAGS

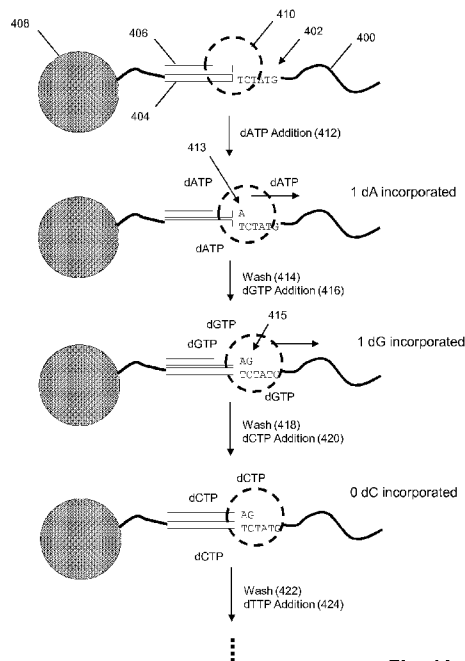


Fig. 4A

(57) Abstract: The invention is directed to the use of sequence tags to improve sequence determination of amplicons of related sequences, particularly large and complex amplicons, such as those comprising recombinant nucleic acids encoding immune receptor molecules. In one aspect, sequence reads having the same sequence tags are aligned after which final base calls are determined from a (possibly weighted) average base call from sequence read base calls at each position. Similarly, in another aspect, sequence reads comprising series of incorporation signals are aligned by common sequence tags and base calls in homopolymer regions are made as a function incorporation signal values at each "flow" position.



MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, **Published:**
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

- *without international search report and to be republished
upon receipt of that report (Rule 48.2(g))*
- *with sequence listing part of description (Rule 5.2(a))*

METHOD OF SEQUENCE DETERMINATION USING SEQUENCE TAGS

CROSS REFERENCE

[0001] This application claims priority to U.S. patent application Ser. No. 13/835,093, filed March 15, 2013, which claims priority from the following U.S. provisional patent applications: Ser. No. 61/658,317 filed 11 June 2012; Ser. No. 61/738,277 filed 17 December 2012; and Ser. No. 61/776,647 filed 11 March 2013, which applications are each hereby incorporated by reference in their entireties. This application also claims priority from U.S. provisional patent application Ser. No. 61/829,054, filed 30 May 2013, which application is hereby incorporated by reference in its entirety.

BACKGROUND OF THE INVENTION

[0002] Analysis of biological or medical samples often requires the determination of nucleic acid sequences of large and complex populations of DNA and/or RNA, e.g. Gloor et al, PLoS ONE 5(10): e15406 (2010); Petrosino et al, Clinical Chemistry, 55(5): 856-866 (2009); Arstila et al, Science, 286: 958-961 (1999). In particular, profiles of nucleic acids encoding immune molecules, such as T cell or B cell receptors, or their components, contain a wealth of information on the state of health or disease of an organism, so that the use of such profiles as diagnostic or prognostic indicators has been proposed for a wide variety of conditions, e.g. Faham and Willis, U.S. patent publication 2010/0151471; Freeman et al, Genome Research, 19: 1817-1824 (2009); Boyd et al, Sci. Transl. Med., 1(12): 12ra23 (2009); He et al, Oncotarget (March 8, 2011). Such sequence-based profiles are capable of much greater sensitivity than approaches based on size distributions of amplified target nucleic acids, sequence sampling by microarrays, hybridization kinetics curves from PCR amplicons, or other approaches, e.g. Morley et al, U.S. patent 5,418,134; van Dongen et al, Leukemia, 17: 2257-2317 (2003); Ogle et al, Nucleic Acids Research, 31: e139 (2003); Wang et al, BMC Genomics, 8: 329 (2007); Baum et al, Nature Methods, 3(11): 895-901 (2006). However, the efficient determination of clonotypes and clonotype profiles from sequence data poses challenges because of the size of populations to be analyzed, the similarity of sequences in such populations, the limited predictability of natural variability among the sequences, and noise introduced into the data by a host of sample preparation and measurement steps, e.g. Warren et al, Genome Research, 21(5): 790-797 (2011).

[0003] Sequence tags, or barcodes, have been used in a variety of ways to assist in the analysis of nucleic acid populations, including labeling, contamination monitoring, rare mutant detection, physical sorting, molecular counting, and the like, e.g. Kinde et al, Proc. Natl. Acad. Sci., 108(23): 9530-9535 (2011); Casbon et al, U.S. patent publication 2012/0071331; Brenner, U.S. patent 5,635,400; Brenner and Macevicz, U.S. patent 7,537,897; Brenner et al, Proc. Natl. Acad. Sci., 97: 1665-1670 (2000); Church et al, European patent publication 0 303 459; Shoemaker et al, Nature Genetics, 14: 450-456 (1996); Morris et al, European patent publication 0799897A1; Wallace, U.S. patent 5,981,179. Recently Kinde et al (cited above) showed how sequence tags could be used to distinguish sequencing and amplification errors from rare mutations in a reference sequence.

[0004] In view of the importance of accurate sequencing for medical and diagnostic applications, it would be highly advantageous if the use of sequence tags could be expanded for increasing the efficiency and accuracy of sequence determination in such applications.

SUMMARY OF THE INVENTION

[0005] The present invention is drawn to methods for producing sequence-based profiles of complex nucleic acid populations, particularly recombined nucleic acid populations encoding repertoires of immune molecules or portions thereof. The invention is exemplified in a number of implementations and applications, some of which are summarized below and throughout the specification.

[0006] In one aspect, the invention is directed to a method of determining clonotypes of an immune repertoire comprising the steps: (a) obtaining a sample from an individual comprising T-cells and/or B-cells; (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or immunoglobulin genes of the T-cells and/or B-cells to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; and (e) aligning sequence reads of like sequence tags to determine sequence reads corresponding to the same clonotypes of the repertoire.

[0007] In another aspect, the invention is directed to a method of detecting clonotype carry over contamination in a patient being monitored for minimal residual disease comprising the steps of: (a) monitoring a patient for minimal residual disease by periodically measuring a clonotype profile of the patient in accordance with the method of claim A1; (b) recording said sequence tags of each measurement of clonotype profiles; and (c) detecting clonotype carry

over contamination if a sequence tag of any prior clonotype profile is detected in a subsequent clonotype profile.

[0008] In another aspect, the invention is directed to a method of determining a number of lymphocytes in a sample comprising the following steps: (a) obtaining a sample from an individual comprising lymphocytes; (b) attaching sequence tags to molecules of recombinant nucleic acids of T-cell receptor genes or of immunoglobulin genes of the lymphocytes to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; (e) counting the number of distinct sequence tags to determine the number of lymphocytes in the sample.

[0009] In another aspect, the invention is directed to a method of determining nucleotide sequences of one or more polynucleotides in one or more sequencing-by-synthesis reactions, which method comprises the steps: (a) attaching a sequence tag to each of the one or more polynucleotides to form tag-polynucleotide conjugates, wherein substantially every polynucleotide of the tag-polynucleotide conjugates has a unique sequence tag; (b) amplifying the tag-polynucleotide conjugates; (c) sequencing by synthesis amplified tag-polynucleotide conjugates, wherein sequencing by synthesis comprises at least one dNTP flow; and (d) determining for each tag-polynucleotide having the same sequence tag a number of nucleotide incorporations for each dNTP flow as a function of measured incorporation signals for each such dNTP flow.

[0010] The present invention provides methods for determining clonotypes and clonotype profiles from large sets of sequence data obtained by high throughput sequencing of somatically recombined nucleic acids that encode immune molecules. In one aspect, the invention implements the above methods by labeling each somatically recombined nucleic acid molecule in a sample with a unique sequence tag, which is used to group sequence reads containing copies of the same clonotype sequence from the sample.

[0011] These above-characterized aspects, as well as other aspects, of the present invention are exemplified in a number of illustrated implementations and applications, some of which are shown in the figures and characterized in the claims section that follows. However, the above summary is not intended to describe each illustrated embodiment or every implementation of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention is obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0013] FIGS. 1A-1B illustrate an example of labeling by sampling to attach unique sequence tags to nucleic acid molecules.

[0014] FIG. 1C illustrates an IgH transcript and sources of natural variability within it.

[0015] FIGS. 2A-2C show a two-staged PCR scheme for amplifying recombined nucleic acid molecules.

[0016] FIG. 3 illustrates one implementation of the step of determining the sequence of a clonotype from sequence reads associated with the same sequence tag.

[0017] FIGS. 4A-4D illustrate an embodiment of the invention directed to determining homopolymer regions in sequencing-by-synthesis operations.

DETAILED DESCRIPTION OF THE INVENTION

[0018] The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of molecular biology (including recombinant techniques), bioinformatics, cell biology, and biochemistry, which are within the skill of the art. Such conventional techniques include, but are not limited to, sampling and analysis of blood cells, nucleic acid sequencing and analysis, and the like. Specific illustrations of suitable techniques can be had by reference to the example herein below. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV); *PCR Primer: A Laboratory Manual*; and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press); and the like.

[0019] In one aspect, the invention is directed to a method for obtaining and analyzing sequence data from a repertoire of immune molecules, such as T cell receptors (TCRs) or B cell receptors (BCRs) or defined fragments thereof, to rapidly and efficiently determine a clonotype profile. Sequence data typically comprises a large collection of sequence reads, i.e. sequences of base calls and associated quality scores, from a DNA sequencer used to analyze the immune molecules. A key challenge in constructing clonotype profiles is to

rapidly and accurately distinguish sequence reads that contain genuine differences from those that contain errors from non-biological sources, such as the extraction steps, sequencing chemistry, amplification chemistry, or the like. An aspect of the invention includes attaching a unique sequence tag to each clonotype in a sample to assist in determining whether sequence reads of such conjugates are derived from the same original clonotype. In accordance with one aspect of the invention, sequence tags are attached to the somatically recombined nucleic acid molecules to form tag-molecule conjugates wherein each recombined nucleic acid of such a conjugate has a unique sequence tag. Usually such attachment is made after nucleic acid molecules are extracted from a sample containing T cells and/or B cells. Preferably, such unique sequence tags differ as greatly as possible from one another as determined by conventional distance measures for sequences, such as, a Hamming distance, or the like. By maximizing the distance between sequence tags in tag-molecule conjugates, even with a high rate of sequencing and amplification errors, a sequence tag of a conjugate remains far closer to its ancestral tag sequence than to that of any other tag sequence of a different conjugate. For example, if 16-mer sequence tags are employed and each such tag on a set of clonotypes has a Hamming distance of at least fifty percent, or eight nucleotides, from every other sequence tag on the clonotypes, then at least eight sequencing or amplification errors would be necessary to transform one such tag into another for a mis-read of a sequence tag (and the incorrect grouping of a sequence read of a clonotype with the wrong sequence tag). In one embodiment, sequence tags are selected so that after attachment to recombined nucleic acids molecules to form tag-molecule conjugates, the Hamming distance between tags of the tag-molecule conjugates is a number at least twenty-five percent of the total length of such sequence tags (that is, each sequence tag differs in sequence from every other such tag in at least 25 percent of its nucleotides); in another embodiment, the Hamming distance between such sequence tags is a number at least 50 percent of the total length of such sequence tags.

[0020] In one aspect, the invention is implemented by the following steps: (a) obtaining a sample from an individual comprising T-cells and/or B-cells; (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or immunoglobulin genes of the T-cells and/or B-cells to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; and (e) aligning sequence reads of like sequence tags to determine sequence reads corresponding to the same clonotypes of the repertoire. Samples containing B-cells or T-cells are obtained using

conventional techniques, as described more fully below. In the step of attaching sequence tags, preferably sequence tags are not only unique but also are sufficiently different from one another that the likelihood of even a large number of sequencing or amplification errors transforming one sequence tag into another would be close to zero. After attaching sequence tags, amplification of the tag-molecule conjugate is necessary for most sequencing technologies; however, whenever single-molecule sequencing technologies are employed an amplification step is optional. Single molecule sequencing technologies include, but are not limited to, single molecule real-time (SMRT) sequencing, nanopore sequencing, or the like, e.g. U.S. patents 7,313,308; 8,153,375; 7,907,800; 7,960,116; 8,137,569; Manrao et al, Nature Biotechnology, 4(8): 2685-2693 (2012); and the like.

[0021] In another aspect, the invention includes a method for determining the number of lymphocytes in a sample by counting unique sequence tags. Even without sequence tags, clonotypes of TCR β or IgH genes, particularly those including the V(D)J regions, provide for a lymphocyte and its clones a unique marker. Whenever recombined nucleic acids are obtained from genomic DNA, then a count of lymphocytes in a sample may be estimated by the number of unique clonotypes that are counted after sequencing. This approach breaks down whenever there are significant clonal populations of identical lymphocytes associated with the same clonotype (or when recombined nucleic acids are obtained from mRNA of a sample, whose quantity of individual sequences may reflect, or depend on, expression rate as well as cell number). The use of sequence tags overcomes this short coming and is especially useful for providing counts of lymphocytes in patients suffering from many lymphoid disorders, such as lymphomas or leukemias. In accordance with one aspect of the invention, sequence tags may be used to obtain an absolute count of lymphocytes in a sample regardless of whether there is a large dominant clone present, such as with leukemia. Such a method may be implemented with the steps: (a) obtaining a sample from an individual comprising lymphocytes; (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or of immunoglobulin genes of the lymphocytes to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag; (c) amplifying the tag-molecule conjugates; (d) sequencing the tag-molecule conjugates; and (e) counting the number of distinct sequence tags to determine the number of lymphocytes in the sample. In some embodiments, the molecules of recombined nucleic acids are from genomic DNA.

[0022] In one embodiment of the invention, sequence tags are attached to recombined nucleic acid molecules of a sample by labeling by sampling, e.g. as disclosed by Brenner et al, U.S.

patent 5,846,719; Brenner et al, U.S. patent 7,537,897; Macevicz, International patent publication WO 2005/111242; and the like, which are incorporated herein by reference. In labeling by sampling, polynucleotides of a population to be labeled (or uniquely tagged) are used to sample (by attachment, linking, or the like) sequence tags of a much larger population. That is, if the population of polynucleotides has K members (including replicates of the same polynucleotide) and the population of sequence tags has N members, then $N \gg K$. In one embodiment, the size of a population of sequence tags used with the invention is at least 10 times the size of the population of clonotypes in a sample; in another embodiment, the size of a population of sequence tags used with the invention is at least 100 times the size of the population of clonotypes in a sample; and in another embodiment, the size of a population of sequence tags used with the invention is at least 1000 times the size of the population of clonotypes in a sample. In other embodiments, a size of sequence tag population is selected so that substantially every clonotype in a sample will have a unique sequence tag whenever such clonotypes are combined with such sequence tag population, e.g. in an attachment reaction, such as a ligation reaction, amplification reaction, or the like. In some embodiments, substantially every clonotype means at least 90 percent of such clonotypes will have a unique sequence tag; in other embodiments, substantially every clonotype means at least 99 percent of such clonotypes will have a unique sequence tag; in other embodiments, substantially every clonotype means at least 99.9 percent of such clonotypes will have a unique sequence tag. In many tissue samples or biopsies the number of T cells or B cells may be up to or about 1 million cells; thus, in some embodiments of the invention employing such samples, the number of unique sequence tags employed in labeling by sampling is at least 10^8 or in other embodiments at least 10^9 .

[0023] In such embodiments, in which up to 1 million clonotypes are labeled by sampling, large sets of sequence tags may be efficiently produced by combinatorial synthesis by reacting a mixture of all four nucleotide precursors at each addition step of a synthesis reaction, e.g. as disclosed in Church, U.S. patent 5,149,625, which is incorporated by reference. The result is a set of sequence tags having a structure of " $N_1N_2 \dots N_k$ " where each $N_i=A, C, G$ or T and k is the number of nucleotides in the tags. The number of sequence tags in a set of sequence tags made by such combinatorial synthesis is 4^k . Thus, a set of such sequence tags with k at least 14, or k in the range of about 14 to 18, is appropriate for attaching sequence tags to a 10^6 -member population of molecules by labeling by sampling. Sets of sequence tags with the above structure include many sequences that may introduce difficulties or errors while implementing the methods of the invention. For example, the

above combinatorially synthesized set of sequence tags includes many member tags with homopolymers segments that some sequencing approaches, such as sequencing-by-synthesis approaches, have difficulty determining with accuracy above a certain length. Therefore, the invention includes combinatorially synthesized sequence tags having structures that are efficient for particular method steps, such as sequencing. For example, several sequence tag structures efficient for sequencing-by-synthesis chemistries may be made by dividing the four natural nucleotides into disjoint subsets which are used alternatively in combinatorial synthesis, thereby preventing homopolymer segments above a given length. For example, let z be either A or C and x be either G or T, to give a sequence tag structure of

$$[(z)_1(z)_2 \dots (z)_i][(x)_1(x)_2 \dots (x)_j] \dots$$

where i and j , which may be the same or different, are selected to limit the size of any homopolymer segment. In one embodiment, i and j are in the range of from 1 to 6. In such embodiments, sequence tags may have lengths in the range of from 12 to 36 nucleotides; and in other embodiments, such sequence tags may have lengths in the range of from 12 to 24 nucleotides. In other embodiments other pairing of nucleotides may be used, for example, z is A or T and x is G or C; or z is A or G and x is T or C. Alternatively, let z' be any combination of three of the four natural nucleotides and let x' be whatever nucleotide is not a z' (for example, z' is A, C or G, and x' is T). This gives a sequence tag structure as follows:

$$[(z')_1(z')_2 \dots (z')_i]x'[(z')_1(z')_2 \dots (z')_i]x' \dots$$

where i is selected as above and the occurrence of x' serves as a punctuation to terminate any undesired homopolymers.

[0024] Further Sequence Tags

[0025] The invention uses methods of labeling nucleic acids, such as fragments of genomic DNA, with unique sequence tags, which may include “mosaic tags,” prior to amplification and sequencing. Such sequence tags are useful for identifying amplification and sequencing errors. Mosaic tags minimize sequencing and amplification artifacts due to inappropriate annealing, priming, hairpin formation, or the like, that may occur with completely random sequence tags of the prior art. In one aspect, mosaic tags are sequence tags that comprise alternating constant regions and variable regions, wherein each constant region has a position

in the mosaic tag and comprises a predetermined sequence of nucleotides and each variable region has a position in the mosaic tag and comprises a predetermined number of randomly selected nucleotides. By way of illustration, a 22-mer mosaic tag (SEQ ID NO: 4) may have the following form:

Nucleotide position:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
<u>N</u>	<u>N</u>	<u>N</u>	b	b	b	b	b	<u>N</u>	b	b	<u>N</u>	<u>N</u>	<u>N</u>	b	b	b	<u>N</u>	<u>N</u>	b	<u>N</u>	<u>N</u>
⏟			⏟					⏟	⏟	⏟			⏟			⏟	⏟	⏟			
1			2					3	4	5			6			7	8	9			

[0026] Region position

[0027] There are nine constant and variable regions, with regions 1 (nucleotides 1-3), 3 (nucleotide 9), 5 (nucleotides 12-14), 7 (nucleotides 18-19) and 9 (nucleotides 21-22) being variable (double underlined nucleotides) and regions 2 (nucleotides 4-8), 4 (nucleotides 10-11), 6 (nucleotides 15-17), and 8 (nucleotide 20) being constant. N represents a randomly selected nucleotide from the set of A, C, G or T; thus, the number of mosaic tags of this example is $4^{11} = 4,194,304$ tags. b represents a predetermined nucleotide at the indicated position. In some embodiments, the sequence of b's, "****bbbb*bb****bbb**b***", is selected to minimize the likelihood of having a perfect match in a genome of the organism making up the sample.

[0028] In one aspect, for mosaic tags of a particular embodiment of the method of the invention, all constant regions with the same position have the same length and all variable regions with the same position have the same length. This allows mosaic tags to be synthesized using partial combinatorial synthesis with conventional chemistries and instruments.

[0029] In one aspect, mosaic tags comprise from 10 to 100 nucleotides, or from 12 to 80 nucleotides, or from 15 to 60 nucleotides. In some embodiments, mosaic tags comprise at least eight nucleotide positions with randomly selected nucleotides; in other embodiments, whenever mosaic tags have a length of at least 15 nucleotides, they comprise at least 12 nucleotide positions with randomly selected nucleotides. In another aspect, no variable region within a mosaic tag may have a length that is greater than seven nucleotides.

[0030] In another aspect, mosaic tags may be used in the following steps: (i) preparing DNA templates from nucleic acids in a sample; (ii) labeling by sampling the DNA templates to form a multiplicity tag-template conjugates, wherein substantially every DNA template of a tag-template conjugate has a unique mosaic tag comprising alternating constant regions and variable regions, each constant region having a position in the mosaic tag and a length of from 1 to 10 nucleotides of a predetermined sequence and each variable region having a position in the mosaic tag and a length of from 1 to 10 randomly selected nucleotides, such that constant regions having the same positions have the same lengths and variable region having the same positions have the same lengths; (iii) amplifying the multiplicity of tag-template conjugates; (iv) generating a plurality of sequence reads for each of the amplified tag-template conjugates; and (v) determining a nucleotide sequence of each of the nucleic acids by determining a consensus nucleotide at each nucleotide position of each plurality of sequence reads having identical mosaic tags. In another aspect, mosaic tags may be used in the following steps: (a) preparing single stranded DNA templates from nucleic acids in a sample; (b) labeling by sampling the single stranded DNA templates to form tag-template conjugates, wherein substantially every single stranded DNA template of a tag-template conjugate has a unique sequence tag (that is, a mosaic tag) having a length of at least 15 nucleotides and having the following form:

$$[(N_1N_2 \dots N_{K_j})(b_1b_2 \dots b_{L_j})]_M$$

wherein each N_i , for $i = 1, 2, \dots, K_j$, is a nucleotide randomly selected from the group consisting of A, C, G and T; K_j is an integer in the range of from 1 to 10 for each j less than or equal to M (that is, regions $N_1N_2 \dots N_{K_j}$ are variable regions); each b_i , for $i = 1, 2, \dots, L_j$, is a nucleotide; L_j is an integer in the range of from 1 to 10 for each j less than or equal to M ; such that every sequence tag (i) has the same K_j for every j and (ii) has the same sequences $b_1b_2 \dots b_{L_j}$ for every j (that is, regions $b_1b_2 \dots b_{L_j}$ are constant regions); and M is an integer greater than or equal to 2; (c) amplifying the tag-template conjugates; (d) generating a plurality of sequence reads for each of the amplified tag-template conjugates; and (e) determining a nucleotide sequence of each of the nucleic acids by determining a consensus nucleotide at each nucleotide position of each plurality of sequence reads having identical sequence tags. In some embodiments, the plurality of sequence reads is at least 10^4 ; in other embodiments, the plurality of sequence reads is at least 10^5 ; in still other embodiments, the

plurality of sequence reads is at least 10^6 . In some embodiments, the total length of the above sequence tag is in the range of from 15 to 80 nucleotides.

[0031] A variety of different attachment reactions may be used to attach unique tags to substantially every clonotype in a sample. In one embodiment, such attachment is accomplished by combining a sample containing recombined nucleic acid molecules (which, in turn, comprise clonotype sequences) with a population or library of sequence tags so that members of the two populations of molecules can randomly combine and become associated or linked, e.g. covalently. In such tag attachment reactions, clonotype sequences comprise linear single or double stranded polynucleotides and sequence tags are carried by reagent such as amplification primers, such as PCR primers, ligation adaptors, circularizable probes, plasmids, or the like. Several such reagents capable of carrying sequence tag populations are disclosed in Macevitz, U.S. patent 8,137,936; Faham et al, U.S. patent 7,862,999; Landegren et al, U.S. patent 8,053,188; Unrau and Deugau, Gene, 145: 163-169 (1994); Church, U.S. patent 5,149,625; and the like, which are incorporated herein by reference.

[0032] Fig. 1A and 1B illustrate an attachment reaction comprising a PCR in which a population of sequence tags ($T_1, T_2, T_3 \dots T_j, T_{j+1} \dots T_k, T_{k+1} \dots T_{n-1}, T_n$) is incorporated into primers (100). The population of sequence tags has a much greater size than that of recombined nucleic acid molecules (102). The sequence tags are attached to the recombined nucleic acid molecules by annealing the primers to the nucleic acid molecules and extending the primers with a DNA polymerase in the first cycle of a PCR. The figure depicts how the recombined nucleic acid molecules select, or sample, a small fraction of the total population of sequence tags by randomly annealing to the primers by way of their common primer binding regions (104), for example, in V region (108). Since the primers (and therefore sequence tags) combine with the recombined nucleic acid sequence molecules randomly, there is a small possibility that the same sequence tag may be attached to different nucleic acid molecules; however, if the population of sequence tags is large as taught herein, then such possibility will be negligibly small so that substantially every recombined nucleic acid molecule will have a unique sequence tag attached. The other primer (106) of the forward and reverse primer pair anneals to C region (110) so that after multiple cycles of annealing, extending and melting, amplicon (112) is formed, thereby attaching unique sequence tags to the V(D)J regions comprising the clonotypes of population (102). That is, amplicon (112) comprises the tag-molecule conjugates from the attachment reaction.

[0033] Such immune molecules typically form an immune repertoire which comprises a very large set of very similar polynucleotides (e.g. >1000, but more usually from 100,000 to

1,000,000, or more) which are relatively short in length (e.g. usually less than 300 bp). In one aspect of the invention, the inventors recognized and appreciated that these characteristics permitted the use of highly dissimilar sequence tags to efficiently compare sequence reads of highly similar clonotypes to determine whether they are derived from the same original sequence or not.

[0034] The complexity of immune repertoires is well-known, e.g. Arstila et al, Science, 286: 958-961 (1999) and Warren et al (cited above). Fig. 1C illustrates diagrammatically a typical transcript of an IgH molecule (120) from which a clonotype profile is derived in accordance with some embodiments of the invention. Sources of natural sequence variability include modular recombination of the C, D, J and V segments from large sets carried by the genome, nucleotide additions and deletions to the ends of the D segment to produce the so-called “NDN” regions, and somatic hypermutation where substitutions are made randomly over the length of transcript (122) at a relative frequency roughly as indicated by curve (128). In one aspect of the invention, complex populations of such IgH and TCR transcripts are amplified and sequenced. In one aspect one or both operations for IgH molecules are carried out by using redundant primers annealing to different sites in the V regions (described more fully below). This is particularly advantageous where a sequencing chemistry is employed that has a relatively high error rate or where such sequence variability is difficult or impossible to know beforehand. In the latter case, primer extension for amplification or generation of sequence reads takes place even if one or more primer binding sites are inoperable, or substantially inoperable, because of mismatches caused (for example) by one or more somatic mutations. Starting from promoter P (122) relative frequency shown by curve (128) climbs through leader region (124) to a maximum over the V(D)J region (126) of the transcript after which it drops to near zero. In one aspect of the invention, a segment of recombined B cell nucleic acid is amplified by a PCR with a plurality of forward primers or a plurality of reverse primers to generate a nested set of templates, e.g. as disclosed in Faham and Willis, U.S. patent publication 2011/0207134. Templates from such a set may be further amplified on a surface to form separate amplicons (e.g. by bridge PCR using a cBot instrument, Illumina, San Diego, CA). Templates from the same nested set may be associated with one another by sequence reads generated at their common ends. Nested sets of templates allow a sequencing chemistry with relative high error rates to be used to analyze longer sequences than otherwise would be possible, while at the same time maintaining high average quality scores over the entire length of the sequence. The nested sets also ensure that

at least one sequence read is obtained from a V region even if it has been subjected to somatic hypermutation.

[0035] Somatic mutations in IgH molecules add a layer of difficulty in reconstructing clonotypes from sequence read data, because of the difficulty in determining whether a base change is due to sequencing or amplification error or is due to a natural mutation process. The use of sequence tags in accordance with the invention greatly mitigates such difficulties because every nucleic acid encoding an IgH will receive a distinct and unique sequence tag. As illustrated in Fig. 3, sequence reads (300) pursuant to the invention each comprise a copy of a sequence tag (302) and a copy of a clonotype (304) (SEQ ID NO: 1). All sequence reads having the same sequence tag are assembled so that the nucleotides of each position in the clonotype portion can be compared (e.g., SEQ ID NO: 2 and SEQ ID NO: 3). Thus, even if IgH-encoding sequences differ by no more than a single base, they will receive a distinct sequence tag, so that closely related IgH-encoding nucleic acids in a sample are not compared with one another in the clonotype determination process. As mentioned above, errors in the sequence tags are not significant because the sequences of the sequence tags associated with clonotypes are so far apart in sequence space that a huge number of base changes could be sustained without one sequence tag becoming close in sequence space to any other sequence tag.

[0036] Constructing clonotypes from sequence read data depends in part on the sequencing method used to generate such data, as the different methods have different expected read lengths and data quality. In one approach, a Solexa sequencer is employed to generate sequence read data for analysis. In one embodiment, a sample is obtained that provides at least $0.5\text{-}1.0 \times 10^6$ lymphocytes to produce at least 1 million template molecules, which after optional amplification may produce a corresponding one million or more clonal populations of template molecules (or clusters). For most high throughput sequencing approaches, including the Solexa approach, such over sampling at the cluster level is desirable so that each template sequence is determined with a large degree of redundancy to increase the accuracy of sequence determination. For Solexa-based implementations, preferably the sequence of each independent template is determined 10 times or more. For other sequencing approaches with different expected read lengths and data quality, different levels of redundancy may be used for comparable accuracy of sequence determination. Those of ordinary skill in the art recognize that the above parameters, e.g. sample size, redundancy, and the like, are design choices related to particular applications.

[0037] Figs. 2A-2C illustrate exemplary steps of attaching unique sequence tags to recombined nucleic acid molecules in a two stage PCR. Population of recombined nucleic acid molecules (250) from a sample containing T-cells or B-cells are combined in a PCR mixture with forward and reverse primers (202) and (262). Primers (262) each comprise three regions: target annealing region (263) (which in this illustration is V region (206)); sequence tag (264); and primer binding region (265) for the second stage of the two-stage PCR. In this illustration, primers (262) comprise a mixture of target annealing regions to account for the diversity of V region sequence. Thus, every different primer is prepared with a sequence tag region. Alternatively, the sequence tag element may be attached to C region primer (202) along with a primer binding region for the second PCR stage. As noted, recombined nucleic acid molecules (250) comprise constant, or C, region (203), J region (210), D region (208), and V region (206), which may represent a V(D)J segment encoding a CDR3 region of either a TCR or immunoglobulin. After a few cycles, for example, 4 to 10, first stage amplicon (266) is produced with each member polynucleotide including sequence tag (270). In the second stage PCR, polynucleotides of amplicon (266) are reamplified with new forward and reverse primers P5 (222) and P7 (220) which add further primer binding sites (224) and (223) for cluster formation using bridge PCR in a Solexa/Illumina sequencer. Primer P7 also include a secondary sequence tag (221) for optional multiplexing of samples in a single sequencing run. After the secondary PCR amplicon (280) is produced with embedded P5 and P7 sequences by which bridge PCR may be carried out.

[0038] Clonotype Determination from Sequence Data

[0039] In accordance with one aspect of the invention, clonotypes of a sample are determined by first grouping sequence reads based on their sequence tags. Such grouping may be accomplished by conventional sequence alignment methods. Guidance for selecting alignment methods is available in Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005), which is incorporated by reference. After sequence reads are assembled in groups corresponding to unique sequence tags, then the sequences of the associated clonotypes may be analyzed to determine the sequence of the clonotype from the sample. Fig. 3 illustrates an exemplary alignment and method from determining the sequence (SEQ ID NO: 1) of a clonotype associated with a unique sequence tag. In this example, eleven sequence reads are aligned by way of their respective sequence tags (302) after which nucleotides at each position of the clonotype portions of the sequence reads, indicated as 1, 2, 3, 4, ... n, are compared. For example, nucleotides at position 6 (306) are t, t, g, t, t, t, t, t, c, t; that is, nine base calls are t's, one is "g" (308) and one is "c" (310) (SEQ ID NO: 2 and SEQ ID NO:

3, respectively). In one embodiment, the correct base call of the clonotype sequence at a position is whatever the identity of the majority base is. In the example of position 6 (306), the base call is “t”, because it is the nucleotide in the majority of sequence reads at that position. In other embodiments, other factors may be taken into account to determine a correct base call for a clonotype sequence, such as quality scores of the base calls of the sequence reads, identities of adjacent bases, or the like.

[0040] Once clonotypes are determined as described above, a clonotype profile comprising the abundances or frequencies of each different clonotype of a sample may be assembled.

[0041] Homopolymer Determination in Sequence-by-Synthesis Methodologies

[0042] In accordance with an aspect of the invention, sequence tags may be used for determining the number of nucleotides in a homopolymer region from signals generated in a sequencing-by-synthesis (SBS) method. That is, whenever a target polynucleotide has been replicated and multiple replicates subjected to separate SBS reactions, each replicate may be labeled with the same sequence tag so that SBS signals from base incorporations in similar regions of the separate replicates may be analyzed by linking the different incorporation signals through the common sequence tags. SBS methodologies have a wide variety of implementations, e.g. Fuller et al, Nature Biotechnology, 27: 1013-1023 (2009); Ronaghi et al, Science, 281: 363-365 (1998); Margulies et al, Nature, 437: 376-380 (2005); Rothberg et al, Nature, 475: 348-352 (2011); Kumar et al, Scientific Reports, 2: 684 (2012); Sims et al, Nature Methods, 8: 575-580 (2011); Pourmand et al, Proc. Natl. Acad. Sci., 103: 6466-6470 (2006); Seo et al, Proc. Natl. Acad. Sci., 102: 5926-5931 (2005); and the like. Commercial DNA sequencers using SBS include 454 GS sequencers (454 Life Sciences), PGM and Proton Ion Torrent sequencers (Life Technologies), PyroMark sequencers (Qiagen), and the like. A common feature of many variants of the SBS methodologies is the generation of a sequence of signals (“incorporation signals”) each of which is proportional to (or at least monotonically related to) the number of nucleotides incorporated into an extended nucleotide chain in a template-driven synthesis reaction. A common challenge with such variants of SBS is determining the number of nucleotides incorporated in homopolymeric regions of a template when the length of the homopolymeric region exceeds 5-6 nucleotides. In one aspect of the invention, this problem is addressed by the use of sequence tags. As above, sequences of such sequence tags in this embodiment are sufficiently far apart from one another that they are readily related to their progenitor, or parent, sequences even after incurring multiple amplification and/or sequencing errors. In this manner, they are distinct from tags frequently employed in sequencing, such as tag for calibrating base-calling software (e.g. so-called

“key” sequences), designating sample or specimen origin, reducing carry-over contamination, or the like. In some embodiments, templates are labeled by sampling using mosaic tags so that the sequence tags themselves may be selected without homopolymer regions greater than the length of the longest variable region of the tag. In some embodiments, mosaic tags are used with variable regions not longer than 5 nucleotides; and in other embodiments, mosaic tags are used with variable regions not longer than 4 nucleotides. Alternatively, sequence tags suitable for SBS may be constructed as disclosed in Brenner et al, U.S. patent 7,537,897, which is hereby incorporated by reference.

[0043] Fig. 4A illustrates one variant of the SBS methodology that comprises cycles of nucleotide addition and washing, with signal generation typical going on during or after incorporation of an added nucleotide. In other words, the illustrated variant comprises cycles of sequentially adding each of the different nucleoside triphosphates and washing, such as in the following order: (a) adding dATP to a growing template-bound chain and washing; (b) adding dGTP to the growing template-bound chain and washing; (c) adding dCTP to the growing template-bound chain and washing; and (d) adding dTTP to the growing template-bound chain and washing. Many additional embodiments and variations of such cycles are possible, e.g. Schultz et al, U.S. patent publication 2012/0035062, and come within the scope of sequencing by synthesis. Typically the extensions are carried out with a nucleic acid polymerase, such as a DNA polymerase. Returning to Fig. 4A, showing one variant of SBS, single stranded template (400) having sequence region (402) adjacent to primer binding site (404) is anchored to solid phase support (408), such as a SNAPP in a microwell of an Ion Torrent semiconductor sequencing chip, e.g. Nobile et al, U.S. patent publication 2010/0300895; Hinz et al, U.S. patent publication 2011/0195253, which are incorporated herein by reference. Anchored template (400) is combined with sequencing primer (406) and polymerase (410) under conditions that permit the annealing of sequencing primer (406) to primer binding site (404) and the extension of sequencing primer (406) by polymerase (410) upon addition of nucleoside triphosphates. In the variant of Fig. 4A, the complex comprising template (400), sequencing primer (406) and polymerase (410) is exposed separately and sequentially to flows of dATP, dGTP, dCTP and dTTP. After exposure to dATP flow (412), sequencing primer (406) is extended by a single dA (413). After washing (414) to remove or destroy any residual dATP, dGTP flow (416) is introduced, so that polymerase (410) incorporates a dGTP to extend sequencing primer (406) by a single dG (415). After washing (418), dCTP flow (420) is introduced, however, because the next base in the template is not

complementary to dC, there is no nucleotide incorporation and no extension of sequencing primer (406). The cycle is completed with washing step (422) and dTTP flow (424).

[0044] During or after each of the incorporation events in this and other SBS variants, an incorporation signal is generated, which may vary widely in character and may comprise more than one physical or chemical manifestations at the same time, e.g. optical and/or electrical. Exemplary incorporation signals include, but are not limited to, changes in pH, fluorescence, chemiluminescence, resistance, and the like. Fig. 4B illustrates typical incorporation signals recorded (426, 428, 430) during or after each of the above flows of dATP, dGTP and dCTP of Fig. 4A, respectively. Such data may be analyzed to give a number for each peak that is proportional to, or monotonically related to, the number of nucleotides that are added to primer (406), or extensions thereof. The number obtained in such an analysis may vary according to particular embodiments, for example, such a number may be a function of the area of such peaks, peak heights, peak width at half maximum, or like functions known in the art. Figs. 4C and 4D illustrate the SBS process in the presence of homopolymer region (432) and the use of sequence tags for determining incorporation signals. Supports (436, 438, 440) correspond to separate SBS reactions 1, 2 and K, respectively, which have replicates of the same sequence tag-template conjugates attached (SEQ ID NO: 5). Primers 406, 406' and 406'' are extended in separate reactions by separate polymerases 410, 410' and 410'', respectively. Such separate reactions are implemented in separate reaction confinement regions which, in turn, may have a variety of forms, such as arrays of microwells of various scales, e.g. Rothberg et al, *Nature*, 475: 348-352 (2011); Leamon et al, U.S. patent 8,158,359; Rothberg et al, U.S. patent 8,313,625; which are each incorporated herein by reference. Reaction confinement regions may also be defined by clusters of template nucleic acids on beads or surfaces or within gels, e.g. as produced by bridge PCR, disclosed in Boles et al, U.S. patent 6,300,070; Balasubramanian et al, U.S. patent 6,787,308, both of which are incorporated by reference. In accordance with some embodiments of SBS methodology, different dNTP flows are introduced sequentially to each of the reactions. Typically, but not necessarily, each different reaction would be exposed to the same flow at approximately the same time, e.g. as would be the case for SBS reactions taking place in different microwells of an Ion Torrent semiconductor sequencing chip used in conjunction with an appropriate flow system, e.g. Nobile et al (cited above); Davey et al, U.S. patent publication 2012/0143531, the latter of which is incorporated herein by reference. During or after each dNTP flow, incorporation signals ($i_1, i_2, \dots i_k$) are obtained. Fig. 4C illustrates the incorporation signals for the second dATP flow (442). For homopolymers

exceeding 5-6 bases, the contribution of the last-to-be-incorporated nucleotides to the total signal becomes less and less, so that it becomes difficult to distinguish signal from an incorporated nucleotide from noise. In accordance with the invention, incorporation signals from multiple replicate templates may be analyzed together because they can be identified by their common sequence tags. As discussed above, the sequence tags are selected (such as, by a labeling by sampling process) so that even with multiple amplification or sequencing errors present, sequence tags of each tag-polynucleotide conjugate and its replicates are distinguishable from sequence tags of any other tag-polynucleotide conjugate and each of its replicates.

[0045] Fig. 4D shows multiple (1, 2, ... K) sequences of magnitudes of incorporation signals (450), wherein each sequence is associated with common sequence tag (452) and template TCGGGGGGGGACTT (SEQ ID NO: 5). The form of the illustrated data set is similar to that illustrated in Fig. 3, which illustrates how columns of base measurements could be used to determine a final base call. Fig. 4D illustrates how columns of incorporation signals can be used to make a final determination of an nucleotide incorporation number. Such determinations may be made after a separate model is used to extract nucleotide incorporate estimates from incorporation signals from individual reactions, or the determinations may be made by applying a model to all or a plurality of the incorporation signals from tag-polynucleotide conjugates with the same parent tag. As noted above, the numbers shown in Fig. 4D are the values of a function of the data sets illustrated by peaks such as (426, 428 and 430)(which show the rate of generation of an incorporation signal versus time and whose integration, or area, is one measure of nucleotide incorporation). The columns of Fig. 4D, labeled “fla, flg, flc, flt, f2a, f2g” and so on, represent cycles of dNTP flows, wherein the first cycle (fla, flg, flc, flt) comprises sequential introduction of dATP, dGTP, dCTP and dTTP to the SBS reactions, and so on.

[0046] As mentioned above, this aspect of the invention may be implement by a method having the steps: (a) attaching a sequence tag to each of the one or more polynucleotides to form tag-polynucleotide conjugates, wherein substantially every polynucleotide of the tag-polynucleotide conjugates has a unique sequence tag; (b) amplifying the tag-polynucleotide conjugates; (c) sequencing by synthesis amplified tag-polynucleotide conjugates, wherein sequencing by synthesis comprises at least one dNTP flow; and (d) determining for each tag-polynucleotide having the same sequence tag a number of nucleotide incorporations for each dNTP flow as a function of measured incorporation signals for each such dNTP flow.

Typically amplified tag-polynucleotide conjugates include flanking sequences to assist in

implementing the various steps. For example, tag-polynucleotide conjugates may include one or more flanking primer binding sites at each end to permit such operations as secondary amplification, sequencing primer annealing and extension, hybridization-based capture, and the like. As mentioned above, dNTP flows may be implemented by a wide variety of fluid delivery apparatus and control systems, such as those implemented on most commercially available DNA sequencing systems, e.g. Schultz et al, U.S. patent publication 2012/0073667, which is incorporated herein by reference. A dNTP flow is a movement of dNTP-containing reagent through a reaction site. Preferably, dNTP flows are implemented so that there is a sharp fluid boundary between successive flows. In some embodiments, washing steps may be implemented with reagents substantially identical to dNTP-containing reagent, except for the presence of a dNTP. In some embodiments, reagent for implementing washing steps may include agents to destroy or denature dNTPs of a prior flow, e.g. apyrase. In most embodiments, a dNTP flow contains all the reagents, salt levels, pH and the like, sufficient for a polymerase in a complex to extend a sequencing primer by incorporation of dNTPs of the flow in a template-driven reaction along the tag-polynucleotide conjugate. As mentioned above, a wide variety of functions or algorithms may be used to model signals (e.g. incorporation signal values) from separate reactions related by having tag-polynucleotide conjugates with the same sequence tag. In some embodiments, the function of incorporation signals from tag-polynucleotide conjugates having the same sequence tag is the whole number closest to an average of such incorporation signals measured from the separate SBS reactions. The average may be an arithmetic average or a weighted average, e.g. the latter of which may depend on characteristics of a particular sequencing system used, such as the particular sequencing chemistry used, whether labels are used, what type of signal is generated (e.g. optical, resistive pulse, pH, etc.), the location of an SBS reaction site in a dNTP flow, and the like. Guidance for alternative models that relate signals from an SBS reaction and nucleotides incorporated are disclosed in Rearick, U.S. patent publication 2012/0172241; and Hubbell, U.S. patent publication 2012/0173158, both of which are incorporated herein by reference.

[0047] As mentioned above, sequencing by synthesis may be implemented by a wide variety of process steps and reagents. In some embodiments, sequencing by synthesis may be implemented with the following steps: (a) forming for each of a plurality of tag-polynucleotide conjugates a complex comprising a sequencing primer, a nucleic acid polymerase, and a tag-polynucleotide conjugate under conditions that permit annealing of the sequencing primer to such tag-polynucleotide conjugate and extension of such sequencing

primer along the tag-polynucleotide conjugate in the presence of nucleoside triphosphates by the nucleic acid polymerase; (b) introducing a dNTP to the complexes by a dNTP flow; (c) measuring incorporation signals from each complex; (d) washing the complexes; and (e) repeating steps (b) through (d). In some embodiments, after tag-polynucleotide conjugates are amplified, a sample of the resulting amplicon is used in the sequencing by synthesis steps. In some embodiments, such sample comprises at least a number of tag-polynucleotide conjugates greater than 10^2 ; in other embodiments, such sample comprise a number of tag-polynucleotide conjugates in the range of from 10^2 to 10^{15} , or in other embodiments, in the range of from 10^2 to 10^{12} , or in other embodiments, in the range of from 10^2 to 10^9 . In other embodiments, such sample is sufficiently large so that all polynucleotides having a frequency of .001 or higher are present with a probability of ninety-nine percent; or all polynucleotides having a frequency of .0001 or higher are present with a probability of ninety-nine percent.

[0048] dNTPs employed in this aspect of the invention may be unblocked dNTPs or extension-block dNTPs. In the case of the former (unblocked dNTPs), during a single dNTP flow, if a template contains a homopolymeric region, then a number incorporations may occur that is equal to the homopolymer length. (There may be factors that cause fewer incorporations to be made (e.g. a polymerase with low processivity), but with unblocked dNTPs it is possible for such number to equal homopolymer length). For example, the Ion Torrent sequencing by synthesis chemistry employs unblocked dNTPs. In the case of extension-blocked dNTPs, only a single dNTP incorporation can take place for each flow, usually because of the presence of a chemical blocking group on an incorporated nucleotide. Thus, cycles of dNTP incorporation further include a de-blocking step in order to regenerate an extendable end on the extended sequencing primer. For example, the Illumina sequencing by synthesis chemistry employs extension-blocked dNTPs, e.g. Bentley et al (cited below). In some embodiments, a de-blocking step may be implemented as part of a washing step.

[0049] Use of Sequence Tags to Detect Carry Over Contamination

[0050] Carry over contamination is a significant problem with techniques that include amplification of nucleic acids, e.g. Borst et al, Eur. J. Clin. Microbiol. Infect. Dis., 23(4): 289-299 (2004); Aslanzadeh, Ann. Clin. Lab. Sci., 34(4): 389-396 (2004); and the like. Such contamination arises when traces of nucleic acid extraneous to a sample are unintentionally amplified in an assay of the sample and effect or impact a measured result. In a worse case, carry over contamination in a medical sample from a patient can result in a false positive interpretation of an assay result. The extraneous nucleic acid may come from a source unrelated to a particular patient; for example, it may come from the sample of another patient.

Or, the extraneous nucleic acid may come from a source related to a patient; for example, it may come from a different sample from the same patient handled in the same laboratory in the past or from an assay reaction on a different sample from the same patient which was processed in the same laboratory in the past.

[0051] Carry over contamination is especially challenging in a clinical setting when measuring highly complex populations of related nucleic acids, such as populations of recombinant nucleic acids encoding immune molecules, such as T-cell receptors or immunoglobulins. The challenge arises because it is difficult to determine whether a sequence read or clonotype is part of the genuine diversity of an intended sample or whether they originate from an extraneous source of nucleic acid, such as another patient's sample or a prior sample of the same patient, which are being processed in the same kind of assay in the same laboratory. In one aspect of the invention, such carry over contamination may be detected by using sequence tags not only to determine clonotypes from sequence reads but also to determine whether a sequence tag originated in the current sample or from another sample. This is accomplished by maintaining a record of sequence tags determined from each patient sample, then whenever a subsequent measurement is made the sequence tags of the current measurement are compared to those of prior measurements. Such records of sequence tags associated with clonotypes are conveniently maintained as electronic records on mass storage devices because of the large number of tag from each measurement and the ease of searching and comparing electronic records using conventional algorithms. If a match is found then the most likely cause is carry over contamination, provided that the populations of sequence tags employed in the measurements are sufficiently large. The same exemplary ratios of the size of sequence tag population to a clonotype population for labeling by sampling discussed above are applicable for detecting carry over contamination. In one embodiment, such ratio is 100:1 or greater.

[0052] Samples

[0053] Clonotype profiles are obtained from samples of immune cells, which are present in a wide variety of tissues. Immune cells of interest include T-cells and/or B-cells. T-cells (T lymphocytes) include, for example, cells that express T cell receptors (TCRs). B-cells (B lymphocytes) include, for example, cells that express B cell receptors (BCRs). T-cells include helper T cells (effector T cells or Th cells), cytotoxic T cells (CTLs), memory T cells, and regulatory T cells, which may be distinguished by cell surface markers. In one aspect a sample of T cells includes at least 1,000 T cells; but more typically, a sample includes at least 10,000 T cells, and more typically, at least 100,000 T cells. In another aspect, a sample

includes a number of T cells in the range of from 1000 to 1,000,000 cells. A sample of immune cells may also comprise B cells. B-cells include, for example, plasma B cells, memory B cells, B1 cells, B2 cells, marginal-zone B cells, and follicular B cells. B-cells can express immunoglobulins (also referred to as antibodies or B cell receptors). As above, in one aspect a sample of B cells includes at least 1,000 B cells; but more typically, a sample includes at least 10,000 B cells, and more typically, at least 100,000 B cells. In another aspect, a sample includes a number of B cells in the range of from 1000 to 1,000,000 B cells.

[0054] Samples (sometimes referred to as “tissue samples”) used in the methods of the invention can come from a variety of tissues, including, for example, tumor tissue, blood and blood plasma, lymph fluid, cerebrospinal fluid surrounding the brain and the spinal cord, synovial fluid surrounding bone joints, and the like. In one embodiment, the sample is a blood sample. The blood sample can be about 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, or 5.0 mL. The sample can be a tumor biopsy. The biopsy can be from, for example, from a tumor of the brain, liver, lung, heart, colon, kidney, or bone marrow. Any biopsy technique used by those skilled in the art can be used for isolating a sample from a subject. For example, a biopsy can be an open biopsy, in which general anesthesia is used. The biopsy can be a closed biopsy, in which a smaller cut is made than in an open biopsy. The biopsy can be a core or incisional biopsy, in which part of the tissue is removed. The biopsy can be an excisional biopsy, in which attempts to remove an entire lesion are made. The biopsy can be a fine needle aspiration biopsy, in which a sample of tissue or fluid is removed with a needle.

[0055] A sample or tissue sample includes nucleic acid, for example, DNA (e.g., genomic DNA) or RNA (e.g., messenger RNA). The nucleic acid can be cell-free DNA or RNA, e.g. extracted from the circulatory system, Vlassov et al, Curr. Mol. Med., 10: 142-165 (2010); Swarup et al, FEBS Lett., 581: 795-799 (2007). In the methods of the invention, the amount of RNA or DNA from a subject that can be analyzed includes varies widely. For example, DNA or RNA of a single cell may be all that is required for a calibration test (i.e. an initial measurement to determine a correlating clonotype for a disease). For generating a clonotype profile, sufficient nucleic acid must be in a sample to obtain a useful representation of an individual's immune receptor repertoire. More particularly, for generating a clonotype profile from genomic DNA at least 1 ng of total DNA from T cells or B cells (i.e. about 300 diploid genome equivalents) is extracted from a sample; in another embodiment, at least 2 ng of total DNA (i.e. about 600 diploid genome equivalents) is extracted from a sample; and in another embodiment, at least 3 ng of total DNA (i.e. about 900 diploid genome equivalents)

is extracted from a sample. One of ordinary skill would recognize that as the fraction of lymphocytes in a sample decreases, the foregoing minimal amounts of DNA must increase in order to generate a clonotype profile containing more than about 1000 independent clonotypes. For generating a clonotype profile from RNA, in one embodiment, a sufficient amount of RNA is extracted so that at least 1000 transcripts are obtained which encode distinct TCRs, BCRs, or fragments thereof. The amount of RNA that corresponds to this limit varies widely from sample to sample depending on the fraction of lymphocytes in a sample, developmental stage of the lymphocytes, and the like. In one embodiment, at least 100 ng of RNA is extracted from a tissue sample containing B cells and/or T cells for the generating of a clonotype profile; in another one embodiment, at least 500 ng of RNA is extracted from a tissue sample containing B cells and/or T cells for the generating of a clonotype profile. RNA used in methods of the invention may be either total RNA extracted from a tissue sample or polyA RNA extracted directly from a tissue sample or from total RNA extracted from a tissue sample. The above nucleic acid extractions may be carried out using commercially available kits, e.g. from Invitrogen (Carlsbad, CA), Qiagen (San Diego, CA), or like vendors. Guidance for extracting RNA is found in Liedtke et al, PCR Methods and Applications, 4: 185-187 (1994); and like references.

[0056] As discussed more fully below (Definitions), a sample containing lymphocytes is sufficiently large so that substantially every T cell or B cell with a distinct clonotype is represented therein, thereby forming a repertoire (as the term is used herein). In one embodiment, a sample is taken that contains with a probability of ninety-nine percent every clonotype of a population present at a frequency of .001 percent or greater. In another embodiment, a sample is taken that contains with a probability of ninety-nine percent every clonotype of a population present at a frequency of .0001 percent or greater. In one embodiment, a sample of B cells or T cells includes at least a half million cells, and in another embodiment such sample includes at least one million cells.

[0057] Whenever a source of material from which a sample is taken is scarce, such as, clinical study samples, or the like, DNA from the material may be amplified by a non-biasing technique, such as whole genome amplification (WGA), multiple displacement amplification (MDA); or like technique, e.g. Hawkins et al, Curr. Opin. Biotech., 13: 65-67 (2002); Dean et al, Genome Research, 11: 1095-1099 (2001); Wang et al, Nucleic Acids Research, 32: e76 (2004); Hosono et al, Genome Research, 13: 954-964 (2003); and the like.

[0058] Blood samples are of particular interest and may be obtained using conventional techniques, e.g. Innis et al, editors, PCR Protocols (Academic Press, 1990); or the like. For

example, white blood cells may be separated from blood samples using convention techniques, e.g. RosetteSep kit (Stem Cell Technologies, Vancouver, Canada). Likewise, other fractions of whole blood, such as peripheral blood mononuclear cells (PBMCs) may be isolated for use with methods of the invention using commercially available kits, e.g. Miltenyi Biotec, Auburn, CA), or the like. Blood samples may range in volume from 100 μ L to 10 mL; in one aspect, blood sample volumes are in the range of from 200 100 μ L to 2 mL. DNA and/or RNA may then be extracted from such blood sample using conventional techniques for use in methods of the invention, e.g. DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). Optionally, subsets of white blood cells, e.g. lymphocytes, may be further isolated using conventional techniques, e.g. fluorescently activated cell sorting (FACS)(Becton Dickinson, San Jose, CA), magnetically activated cell sorting (MACS)(Miltenyi Biotec, Auburn, CA), or the like.

[0059] In some embodiments, recombined nucleic acids are present in the DNA of each individual's adaptive immunity cells as well as their associated RNA transcripts, so that either RNA or DNA can be sequenced in the methods of the provided invention. A recombined sequence from a T-cell or B-cell encoding a T cell receptor or immunoglobulin molecule, or a portion thereof, is referred to as a clonotype. The DNA or RNA can correspond to sequences from T-cell receptor (TCR) genes or immunoglobulin (Ig) genes that encode antibodies. For example, the DNA and RNA can correspond to sequences encoding α , β , γ , or δ chains of a TCR. In a majority of T-cells, the TCR is a heterodimer consisting of an α -chain and β -chain. The TCR α chain is generated by VJ recombination, and the β chain receptor is generated by V(D)J recombination. For the TCR β chain, in humans there are 48 V segments, 2 D segments, and 13 J segments. Several bases may be deleted and others added (called N and P nucleotides) at each of the two junctions. In a minority of T-cells, the TCRs consist of γ and δ delta chains. The TCR γ chain is generated by VJ recombination, and the TCR δ chain is generated by V(D)J recombination (Kenneth Murphy, Paul Travers, and Mark Walport, *Janeway's Immunology* 7th edition, Garland Science, 2007, which is herein incorporated by reference in its entirety).

[0060] The DNA and RNA analyzed in the methods of the invention can correspond to sequences encoding heavy chain immunoglobulins (IgH) with constant regions (α , δ , ϵ , γ , or μ) or light chain immunoglobulins (IgK or IgL) with constant regions λ or κ . Each antibody has two identical light chains and two identical heavy chains. Each chain is composed of a constant (C) and a variable region. For the heavy chain, the variable region is composed of a

variable (V), diversity (D), and joining (J) segments. Several distinct sequences coding for each type of these segments are present in the genome. A specific VDJ recombination event occurs during the development of a B-cell, marking that cell to generate a specific heavy chain. Diversity in the light chain is generated in a similar fashion except that there is no D region so there is only VJ recombination. Somatic mutation often occurs close to the site of the recombination, causing the addition or deletion of several nucleotides, further increasing the diversity of heavy and light chains generated by B-cells. The possible diversity of the antibodies generated by a B-cell is then the product of the different heavy and light chains. The variable regions of the heavy and light chains contribute to form the antigen recognition (or binding) region or site. Added to this diversity is a process of somatic hypermutation which can occur after a specific response is mounted against some epitope.

[0061] In accordance with the invention, primers may be selected to generate amplicons of subsets of recombined nucleic acids extracted from lymphocytes. Such subsets may be referred to herein as “somatic rearranged regions.” Somatic rearranged regions may comprise nucleic acids from developing or from fully developed lymphocytes, where developing lymphocytes are cells in which rearrangement of immune genes has not been completed to form molecules having full V(D)J regions. Exemplary incomplete somatically rearranged regions include incomplete IgH molecules (such as, molecules containing only D-J regions), incomplete TCR δ molecules (such as, molecules containing only D-J regions), and inactive IgK (for example, comprising Kde-V regions).

[0062] Adequate sampling of the cells is an important aspect of interpreting the repertoire data, as described further below in the definitions of “clonotype” and “repertoire.” For example, starting with 1,000 cells creates a minimum frequency that the assay is sensitive to regardless of how many sequencing reads are obtained. Therefore one aspect of this invention is the development of methods to quantitate the number of input immune receptor molecules. This has been implemented this for TCR β and IgH sequences. In either case the same set of primers are used that are capable of amplifying all the different sequences. In order to obtain an absolute number of copies, a real time PCR with the multiplex of primers is performed along with a standard with a known number of immune receptor copies. This real time PCR measurement can be made from the amplification reaction that will subsequently be sequenced or can be done on a separate aliquot of the same sample. In the case of DNA, the absolute number of rearranged immune receptor molecules can be readily converted to number of cells (within 2 fold as some cells will have 2 rearranged copies of the specific

immune receptor assessed and others will have one). In the case of cDNA the measured total number of rearranged molecules in the real time sample can be extrapolated to define the total number of these molecules used in another amplification reaction of the same sample. In addition, this method can be combined with a method to determine the total amount of RNA to define the number of rearranged immune receptor molecules in a unit amount (say 1 μ g) of RNA assuming a specific efficiency of cDNA synthesis. If the total amount of cDNA is measured then the efficiency of cDNA synthesis need not be considered. If the number of cells is also known then the rearranged immune receptor copies per cell can be computed. If the number of cells is not known, one can estimate it from the total RNA as cells of specific type usually generate comparable amount of RNA. Therefore from the copies of rearranged immune receptor molecules per 1 μ g one can estimate the number of these molecules per cell.

[0063] One disadvantage of doing a separate real time PCR from the reaction that would be processed for sequencing is that there might be inhibitory effects that are different in the real time PCR from the other reaction as different enzymes, input DNA, and other conditions may be utilized. Processing the products of the real time PCR for sequencing would ameliorate this problem. However low copy number using real time PCR can be due to either low number of copies or to inhibitory effects, or other suboptimal conditions in the reaction.

[0064] Known amounts of one or more internal standards to cDNA or genomic DNA can be added to an assay reaction to determine absolute quantities or concentrations of cDNA or genomic DNA samples of unknown quantity. By counting the number of molecules of the internal standard and comparing it to the rest of the sequences of the same sample, one can estimate the number of rearranged immune receptor molecules in the initial cDNA sample. (Such techniques for molecular counting are well-known, e.g. Brenner et al, U.S. patent 7,537,897, which is incorporated herein by reference).

[0065] Amplification of Nucleic Acid Populations

[0066] Amplicons of target populations of nucleic acids, particularly recombined immune molecules, may be generated by a variety of amplification techniques. In one aspect of the invention, multiplex PCR is used to amplify members of a mixture of recombined immune molecules, such as T cell receptors or portions thereof or B cell receptors or portions thereof. Guidance for carrying out multiplex PCRs of such immune molecules is found in the following references, which are incorporated by reference: Faham and Willis, U.S. patent 8,236,503; Morley, U.S. patent 5,296,351; Gorski, U.S. patent 5,837,447; Dau, U.S. patent 6,087,096; Von Dongen et al, U.S. patent publication 2006/0234234; European patent

publication EP 1544308B1; and the like. In some embodiments of the invention, a step of generating a clonotype profile include steps of (a) amplifying a portion of T cell receptor genes and/or a portion of B cell receptor genes and (b) sequencing nucleic acids of the resulting amplicon. As explained elsewhere, the number of amplicon nucleic acids sequenced may vary from application to application. For example, a clonotype profile to determine whether a leukemia patient is still in remission will be large so that the limit of detection of any tumor clones will be very low. In some embodiments, the number of amplicon nucleic acids sequenced is at least 1000; in other embodiments, the number of amplicon nucleic acids sequenced is at least 10^4 ; in other embodiments, the number of amplicon nucleic acids sequenced is at least 10^5 . Such a generating step may also include further steps of coalescing sequence reads into clonotypes, enumerating or tabulating clonotypes, forming frequency distributions of clonotypes, identifying related subsets of clonotypes, displaying clonotype frequency information, and the like.

[0067] After amplification of DNA from the genome (or amplification of nucleic acid in the form of cDNA by reverse transcribing RNA), the individual nucleic acid molecules can be isolated, optionally re-amplified, and then sequenced individually. Exemplary amplification protocols may be found in van Dongen et al, Leukemia, 17: 2257-2317 (2003) or van Dongen et al, U.S. patent publication 2006/0234234, which is incorporated by reference. Briefly, an exemplary protocol is as follows: Reaction buffer: ABI Buffer II or ABI Gold Buffer (Life Technologies, San Diego, CA); 50 μ L final reaction volume; 100 ng sample DNA; 10 pmol of each primer (subject to adjustments to balance amplification as described below); dNTPs at 200 μ M final concentration; $MgCl_2$ at 1.5 mM final concentration (subject to optimization depending on target sequences and polymerase); Taq polymerase (1-2 U/tube); cycling conditions: preactivation 7 min at 95°C; annealing at 60°C; cycling times: 30s denaturation; 30s annealing; 30s extension. Polymerases that can be used for amplification in the methods of the invention are commercially available and include, for example, Taq polymerase, AccuPrime polymerase, or Pfu. The choice of polymerase to use can be based on whether fidelity or efficiency is preferred.

[0068] Real time PCR, picogreen staining, nanofluidic electrophoresis (e.g. LabChip) or UV absorption measurements can be used in an initial step to judge the functional amount of amplifiable material.

[0069] In one aspect, multiplex amplifications are carried out so that relative amounts of sequences in a starting population are substantially the same as those in the amplified population, or amplicon. That is, multiplex amplifications are carried out with minimal

amplification bias among member sequences of a sample population. In one embodiment, such relative amounts are substantially the same if each relative amount in an amplicon is within five fold of its value in the starting sample. In another embodiment, such relative amounts are substantially the same if each relative amount in an amplicon is within two fold of its value in the starting sample. As discussed more fully below, amplification bias in PCR may be detected and corrected using conventional techniques so that a set of PCR primers may be selected for a predetermined repertoire that provide unbiased amplification of any sample.

[0070] In regard to many repertoires based on TCR or BCR sequences, a multiplex amplification optionally uses all the V segments. The reaction is optimized to attempt to get amplification that maintains the relative abundance of the sequences amplified by different V segment primers. Some of the primers are related, and hence many of the primers may “cross talk,” amplifying templates that are not perfectly matched with it. The conditions are optimized so that each template can be amplified in a similar fashion irrespective of which primer amplified it. In other words if there are two templates, then after 1,000 fold amplification both templates can be amplified approximately 1,000 fold, and it does not matter that for one of the templates half of the amplified products carried a different primer because of the cross talk. In subsequent analysis of the sequencing data the primer sequence is eliminated from the analysis, and hence it does not matter what primer is used in the amplification as long as the templates are amplified equally.

[0071] In one embodiment, amplification bias may be avoided by carrying out a two-stage amplification (as described in Faham and Willis, cited above) wherein a small number of amplification cycles are implemented in a first, or primary, stage using primers having tails non-complementary with the target sequences. The tails include primer binding sites that are added to the ends of the sequences of the primary amplicon so that such sites are used in a second stage amplification using only a single forward primer and a single reverse primer, thereby eliminating a primary cause of amplification bias. Preferably, the primary PCR will have a small enough number of cycles (e.g. 5-10) to minimize the differential amplification by the different primers. The secondary amplification is done with one pair of primers and hence the issue of differential amplification is minimal. One percent of the primary PCR is taken directly to the secondary PCR. Thirty-five cycles (equivalent to ~28 cycles without the 100 fold dilution step) used between the two amplifications were sufficient to show a robust amplification irrespective of whether the breakdown of cycles were: one cycle primary and 34 secondary or 25 primary and 10 secondary. Even though ideally doing only 1 cycle in the

primary PCR may decrease the amplification bias, there are other considerations. One aspect of this is representation. This plays a role when the starting input amount is not in excess to the number of reads ultimately obtained. For example, if 1,000,000 reads are obtained and starting with 1,000,000 input molecules then taking only representation from 100,000 molecules to the secondary amplification would degrade the precision of estimating the relative abundance of the different species in the original sample. The 100 fold dilution between the 2 steps means that the representation is reduced unless the primary PCR amplification generated significantly more than 100 molecules. This indicates that a minimum 8 cycles (256 fold), but more comfortably 10 cycle (~1,000 fold), may be used. The alternative to that is to take more than 1% of the primary PCR into the secondary but because of the high concentration of primer used in the primary PCR, a big dilution factor is can be used to ensure these primers do not interfere in the amplification and worsen the amplification bias between sequences. Another alternative is to add a purification or enzymatic step to eliminate the primers from the primary PCR to allow a smaller dilution of it. In this example, the primary PCR was 10 cycles and the second 25 cycles.

[0072] Sequencing Populations of Recombined Nucleic Acids

[0073] Any high-throughput technique for sequencing nucleic acids can be used in the method of the invention. Preferably, such technique has a capability of generating in a cost-effective manner a volume of sequence data from which at least 1000 clonotypes can be determined, and preferably, from which at least 10,000 to 1,000,000 clonotypes can be determined. DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary, sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, and SOLiD sequencing. Sequencing of the separated molecules has more recently been demonstrated by sequential or single extension reactions using polymerases or ligases as well as by single or sequential differential hybridizations with libraries of probes. These reactions have been performed on many clonal sequences in parallel including demonstrations in current commercial applications of over 100 million sequences in parallel. These sequencing approaches can thus be used to study the repertoire of T-cell receptor (TCR) and/or B-cell receptor (BCR). In one aspect of the invention, high-throughput methods of sequencing are employed that comprise a step of

spatially isolating individual molecules on a solid surface where they are sequenced in parallel. Such solid surfaces may include nonporous surfaces (such as in Solexa sequencing, e.g. Bentley et al, Nature, 456: 53-59 (2008) or Complete Genomics sequencing, e.g. Drmanac et al, Science, 327: 78-81 (2010)), arrays of wells, which may include bead- or particle-bound templates (such as with 454, e.g. Margulies et al, Nature, 437: 376-380 (2005) or Ion Torrent sequencing, U.S. patent publication 2010/0137143 or 2010/0304982), micromachined membranes (such as with SMRT sequencing, e.g. Eid et al, Science, 323: 133-138 (2009)), or bead arrays (as with SOLiD sequencing or polony sequencing, e.g. Kim et al, Science, 316: 1481-1414 (2007)). In another aspect, such methods comprise amplifying the isolated molecules either before or after they are spatially isolated on a solid surface. Prior amplification may comprise emulsion-based amplification, such as emulsion PCR, or rolling circle amplification. Of particular interest is Solexa-based sequencing where individual template molecules are spatially isolated on a solid surface, after which they are amplified in parallel by bridge PCR to form separate clonal populations, or clusters, and then sequenced, as described in Bentley et al (cited above) and in manufacturer's instructions (e.g. TruSeq™ Sample Preparation Kit and Data Sheet, Illumina, Inc., San Diego, CA, 2010); and further in the following references: U.S. patents 6,090,592; 6,300,070; 7,115,400; and EP0972081B1; which are incorporated by reference. In one embodiment, individual molecules disposed and amplified on a solid surface form clusters in a density of at least 10^5 clusters per cm^2 ; or in a density of at least 5×10^5 per cm^2 ; or in a density of at least 10^6 clusters per cm^2 . In one embodiment, sequencing chemistries are employed having relatively high error rates. In such embodiments, the average quality scores produced by such chemistries are monotonically declining functions of sequence read lengths. In one embodiment, such decline corresponds to 0.5 percent of sequence reads have at least one error in positions 1-75; 1 percent of sequence reads have at least one error in positions 76-100; and 2 percent of sequence reads have at least one error in positions 101-125.

[0074] In one aspect, a sequence-based clonotype profile of an individual is obtained using the following steps: (a) obtaining a nucleic acid sample from T-cells and/or B-cells of the individual; (b) spatially isolating individual molecules derived from such nucleic acid sample, the individual molecules comprising at least one template generated from a nucleic acid in the sample, which template comprises a somatically rearranged region or a portion thereof, each individual molecule being capable of producing at least one sequence read; (c) sequencing said spatially isolated individual molecules; and (d) determining abundances of different sequences of the nucleic acid molecules from the nucleic acid sample to generate the

clonotype profile. In some embodiments, the method further includes one or more steps of amplifying the individual molecules comprising recombined nucleic acids of T-cell receptor genes or immunoglobulin genes. For example, such one or more amplification steps may include a multi-stage PCR. In another embodiment, the step of sequencing comprises bidirectionally sequencing each of the spatially isolated individual molecules to produce at least one forward sequence read and at least one reverse sequence read. In another embodiment, the above method comprises the following steps: (a) obtaining a nucleic acid sample from T-cells and/or B-cells of the individual; (b) spatially isolating individual molecules derived from such nucleic acid sample, the individual molecules comprising nested sets of templates each generated from a nucleic acid in the sample and each containing a somatically rearranged region or a portion thereof, each nested set being capable of producing a plurality of sequence reads each extending in the same direction and each starting from a different position on the nucleic acid from which the nested set was generated; (c) sequencing said spatially isolated individual molecules; and (d) determining abundances of different sequences of the nucleic acid molecules from the nucleic acid sample to generate the clonotype profile. In one embodiment, the step of sequencing includes producing a plurality of sequence reads for each of the nested sets. In another embodiment, each of the somatically rearranged regions comprise a V region and a J region, and each of the plurality of sequence reads starts from a different position in the V region and extends in the direction of its associated J region.

[0075] In one aspect, for each sample from an individual, the sequencing technique used in the methods of the invention generates sequences of least 1000 clonotypes per run; in another aspect, such technique generates sequences of at least 10,000 clonotypes per run; in another aspect, such technique generates sequences of at least 100,000 clonotypes per run; in another aspect, such technique generates sequences of at least 500,000 clonotypes per run; and in another aspect, such technique generates sequences of at least 1,000,000 clonotypes per run. In still another aspect, such technique generates sequences of between 100,000 to 1,000,000 clonotypes per run per individual sample.

[0076] The sequencing technique used in the methods of the provided invention can generate about 30 bp, about 40 bp, about 50 bp, about 60 bp, about 70 bp, about 80 bp, about 90 bp, about 100 bp, about 110, about 120 bp per read, about 150 bp, about 200 bp, about 250 bp, about 300 bp, about 350 bp, about 400 bp, about 450 bp, about 500 bp, about 550 bp, or about 600 bp per read.

[0077] While the present invention has been described with reference to several particular example embodiments, those skilled in the art will recognize that many changes may be made thereto without departing from the spirit and scope of the present invention. The present invention is applicable to a variety of sensor implementations and other subject matter, in addition to those discussed above.

[0078] Definitions

[0079] Unless otherwise specifically defined herein, terms and symbols of nucleic acid chemistry, biochemistry, genetics, and molecular biology used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, DNA Replication, Second Edition (W.H. Freeman, New York, 1992); Lehninger, Biochemistry, Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human Molecular Genetics, Second Edition (Wiley-Liss, New York, 1999); Abbas et al, Cellular and Molecular Immunology, 6th edition (Saunders, 2007).

[0080] “Aligning” in reference to two or more sequences, such as sequence reads, means a method of comparing the two or more sequences to determine how similar they are based on some sequence distance measure. An exemplary method of aligning nucleotide sequences is the Smith Waterman algorithm. Distance measures may include Hamming distance, Levenshtein distance, or the like. Distance measures may include a component related to the quality values of nucleotides of the sequences being compared.

[0081] “Amplicon” means the product of a polynucleotide amplification reaction; that is, a clonal population of polynucleotides, which may be single stranded or double stranded, which are replicated from one or more starting sequences. The one or more starting sequences may be one or more copies of the same sequence, or they may be a mixture of different sequences. Preferably, amplicons are formed by the amplification of a single starting sequence. Amplicons may be produced by a variety of amplification reactions whose products comprise replicates of the one or more starting, or target, nucleic acids. In one aspect, amplification reactions producing amplicons are “template-driven” in that base pairing of reactants, either nucleotides or oligonucleotides, have complements in a template polynucleotide that are required for the creation of reaction products. In one aspect, template-driven reactions are primer extensions with a nucleic acid polymerase or oligonucleotide ligations with a nucleic acid ligase. Such reactions include, but are not limited to, polymerase chain reactions (PCRs), linear polymerase reactions, nucleic acid sequence-based amplification (NASBAs), rolling circle amplifications, and the like, disclosed in the following references that are incorporated herein by reference: Mullis et al, U.S.

patents 4,683,195; 4,965,188; 4,683,202; 4,800,159 (PCR); Gelfand et al, U.S. patent 5,210,015 (real-time PCR with “taqman” probes); Wittwer et al, U.S. patent 6,174,670; Kacian et al, U.S. patent 5,399,491 (“NASBA”); Lizardi, U.S. patent 5,854,033; Aono et al, Japanese patent publ. JP 4-262799 (rolling circle amplification); and the like. In one aspect, amplicons of the invention are produced by PCRs. An amplification reaction may be a “real-time” amplification if a detection chemistry is available that permits a reaction product to be measured as the amplification reaction progresses, e.g. “real-time PCR” described below, or “real-time NASBA” as described in Leone et al, *Nucleic Acids Research*, 26: 2150-2155 (1998), and like references. As used herein, the term “amplifying” means performing an amplification reaction. A “reaction mixture” means a solution containing all the necessary reactants for performing a reaction, which may include, but not be limited to, buffering agents to maintain pH at a selected level during a reaction, salts, co-factors, scavengers, and the like.

[0082] “Clonality” as used herein means a measure of the degree to which the distribution of clonotype abundances among clonotypes of a repertoire is skewed to a single or a few clonotypes. Roughly, clonality is an inverse measure of clonotype diversity. Many measures or statistics are available from ecology describing species-abundance relationships that may be used for clonality measures in accordance with the invention, e.g. Chapters 17 & 18, in Pielou, *An Introduction to Mathematical Ecology*, (Wiley-Interscience, 1969). In one aspect, a clonality measure used with the invention is a function of a clonotype profile (that is, the number of distinct clonotypes detected and their abundances), so that after a clonotype profile is measured, clonality may be computed from it to give a single number. One clonality measure is Simpson’s measure, which is simply the probability that two randomly drawn clonotypes will be the same. Other clonality measures include information-based measures and McIntosh’s diversity index, disclosed in Pielou (cited above).

[0083] “Clonotype” means a recombined nucleotide sequence of a lymphocyte which encodes an immune receptor or a portion thereof. More particularly, clonotype means a recombined nucleotide sequence of a T cell or B cell which encodes a T cell receptor (TCR) or B cell receptor (BCR), or a portion thereof. In various embodiments, clonotypes may encode all or a portion of a VDJ rearrangement of IgH, a DJ rearrangement of IgH, a VJ rearrangement of IgK, a VJ rearrangement of IgL, a VDJ rearrangement of TCR β , a DJ rearrangement of TCR β , a VJ rearrangement of TCR α , a VJ rearrangement of TCR γ , a VDJ rearrangement of TCR δ , a VD rearrangement of TCR δ , a Kde-V rearrangement, or the like. Clonotypes may also encode translocation breakpoint regions involving immune receptor

genes, such as Bcl1-IgH or Bcl1-IgH. In one aspect, clonotypes have sequences that are sufficiently long to represent or reflect the diversity of the immune molecules that they are derived from; consequently, clonotypes may vary widely in length. In some embodiments, clonotypes have lengths in the range of from 25 to 400 nucleotides; in other embodiments, clonotypes have lengths in the range of from 25 to 200 nucleotides.

[0084] “Clonotype profile” means a listing of distinct clonotypes and their relative abundances that are derived from a population of lymphocytes. Typically, the population of lymphocytes are obtained from a tissue sample. The term “clonotype profile” is related to, but more general than, the immunology concept of immune “repertoire” as described in references, such as the following: Arstila et al, Science, 286: 958-961 (1999); Yassai et al, Immunogenetics, 61: 493-502 (2009); Kedzierska et al, Mol. Immunol., 45(3): 607-618 (2008); and the like. The term “clonotype profile” includes a wide variety of lists and abundances of rearranged immune receptor-encoding nucleic acids, which may be derived from selected subsets of lymphocytes (e.g. tissue-infiltrating lymphocytes, immunophenotypic subsets, or the like), or which may encode portions of immune receptors that have reduced diversity as compared to full immune receptors. In some embodiments, clonotype profiles may comprise at least 10^3 distinct clonotypes; in other embodiments, clonotype profiles may comprise at least 10^4 distinct clonotypes; in other embodiments, clonotype profiles may comprise at least 10^5 distinct clonotypes; in other embodiments, clonotype profiles may comprise at least 10^6 distinct clonotypes. In such embodiments, such clonotype profiles may further comprise abundances or relative frequencies of each of the distinct clonotypes. In one aspect, a clonotype profile is a set of distinct recombined nucleotide sequences (with their abundances) that encode T cell receptors (TCRs) or B cell receptors (BCRs), or fragments thereof, respectively, in a population of lymphocytes of an individual, wherein the nucleotide sequences of the set have a one-to-one correspondence with distinct lymphocytes or their clonal subpopulations for substantially all of the lymphocytes of the population. In one aspect, nucleic acid segments defining clonotypes are selected so that their diversity (i.e. the number of distinct nucleic acid sequences in the set) is large enough so that substantially every T cell or B cell or clone thereof in an individual carries a unique nucleic acid sequence of such repertoire. That is, preferably each different clone of a sample has different clonotype. In other aspects of the invention, the population of lymphocytes corresponding to a repertoire may be circulating B cells, or may be circulating T cells, or may be subpopulations of either of the foregoing populations, including but not limited to, CD4⁺ T cells, or CD8⁺ T cells, or other subpopulations defined by cell surface

markers, or the like. Such subpopulations may be acquired by taking samples from particular tissues, e.g. bone marrow, or lymph nodes, or the like, or by sorting or enriching cells from a sample (such as peripheral blood) based on one or more cell surface markers, size, morphology, or the like. In still other aspects, the population of lymphocytes corresponding to a repertoire may be derived from disease tissues, such as a tumor tissue, an infected tissue, or the like. In one embodiment, a clonotype profile comprising human TCR β chains or fragments thereof comprises a number of distinct nucleotide sequences in the range of from 0.1×10^6 to 1.8×10^6 , or in the range of from 0.5×10^6 to 1.5×10^6 , or in the range of from 0.8×10^6 to 1.2×10^6 . In another embodiment, a clonotype profile comprising human IgH chains or fragments thereof comprises a number of distinct nucleotide sequences in the range of from 0.1×10^6 to 1.8×10^6 , or in the range of from 0.5×10^6 to 1.5×10^6 , or in the range of from 0.8×10^6 to 1.2×10^6 . In a particular embodiment, a clonotype profile of the invention comprises a set of nucleotide sequences encoding substantially all segments of the V(D)J region of an IgH chain. In one aspect, "substantially all" as used herein means every segment having a relative abundance of .001 percent or higher; or in another aspect, "substantially all" as used herein means every segment having a relative abundance of .0001 percent or higher. In another particular embodiment, a clonotype profile of the invention comprises a set of nucleotide sequences that encodes substantially all segments of the V(D)J region of a TCR β chain. In another embodiment, a clonotype profile of the invention comprises a set of nucleotide sequences having lengths in the range of from 25-200 nucleotides and including segments of the V, D, and J regions of a TCR β chain. In another embodiment, a clonotype profile of the invention comprises a set of nucleotide sequences having lengths in the range of from 25-200 nucleotides and including segments of the V, D, and J regions of an IgH chain. In another embodiment, a clonotype profile of the invention comprises a number of distinct nucleotide sequences that is substantially equivalent to the number of lymphocytes expressing a distinct IgH chain. In another embodiment, a clonotype profile of the invention comprises a number of distinct nucleotide sequences that is substantially equivalent to the number of lymphocytes expressing a distinct TCR β chain. In still another embodiment, "substantially equivalent" means that with ninety-nine percent probability a clonotype profile will include a nucleotide sequence encoding an IgH or TCR β or portion thereof carried or expressed by every lymphocyte of a population of an individual at a frequency of .001 percent or greater. In still another embodiment, "substantially equivalent" means that with ninety-nine percent probability a repertoire of nucleotide sequences will include a nucleotide sequence encoding an IgH or TCR β or portion thereof carried or expressed by every

lymphocyte present at a frequency of .0001 percent or greater. In some embodiments, clonotype profiles are derived from samples comprising from 10^5 to 10^7 lymphocytes. Such numbers of lymphocytes may be obtained from peripheral blood samples of from 1-10 mL.

[0085] “Coalescing” means treating two candidate clonotypes with sequence differences as the same by determining that such differences are due to experimental or measurement error and not due to genuine biological differences. In one aspect, a sequence of a higher frequency candidate clonotype is compared to that of a lower frequency candidate clonotype and if predetermined criteria are satisfied then the number of lower frequency candidate clonotypes is added to that of the higher frequency candidate clonotype and the lower frequency candidate clonotype is thereafter disregarded. That is, the read counts associated with the lower frequency candidate clonotype are added to those of the higher frequency candidate clonotype.

[0086] “Complementarity determining regions” (CDRs) mean regions of an immunoglobulin (i.e., antibody) or T cell receptor where the molecule complements an antigen's conformation, thereby determining the molecule's specificity and contact with a specific antigen. T cell receptors and immunoglobulins each have three CDRs: CDR1 and CDR2 are found in the variable (V) domain, and CDR3 includes some of V, all of diverse (D) (heavy chains only) and joint (J), and some of the constant (C) domains.

[0087] “Contamination” as used herein means the presence in a sample taken from one individual at a given time of nucleic acid from another individual or of nucleic acid from the same individual from a sample taken at another time. In one aspect, “contamination” means the presence of nucleic acid not originating from a given patient tissue sample which may affect the interpretation of a clonotype profile of the patient.

[0088] “Internal standard” means a nucleic acid sequence that is processed in the same reaction as one or more target polynucleotides in order to permit absolute or relative quantification of the target polynucleotides in a sample. In one aspect the reaction is an amplification reaction, such as PCR. An internal standard may be endogenous or exogenous. That is, an internal standard may occur naturally in the sample, or it may be added to the sample prior to a reaction. In one aspect, one or more exogenous internal standard sequences may be added to a reaction mixture in predetermined concentrations to provide a calibration to which an amplified sequence may be compared to determine the quantity of its corresponding target polynucleotide in a sample. Selection of the number, sequences, lengths, and other characteristics of exogenous internal standards is a routine design choice for one of ordinary skill in the art. Endogenous internal standards, also referred to herein as

“reference sequences,” are sequences natural to a sample that correspond to minimally regulated genes that exhibit a constant and cell cycle-independent level of transcription, e.g. Selvey et al, Mol. Cell Probes, 15: 307-311 (2001). Exemplary internal standards include, but are not limited to, sequences from the following genes: GAPDH, β_2 -microglobulin, 18S ribosomal RNA, and β -actin.

[0089] “Kit” refers to any delivery system for delivering materials or reagents for carrying out a method of the invention. In the context of methods of the invention, such delivery systems include systems that allow for the storage, transport, or delivery of reaction reagents (e.g., primers, enzymes, internal standards, etc. in the appropriate containers) and/or supporting materials (e.g., buffers, written instructions for performing the assay etc.) from one location to another. For example, kits include one or more enclosures (e.g., boxes) containing the relevant reaction reagents and/or supporting materials. Such contents may be delivered to the intended recipient together or separately. For example, a first container may contain an enzyme for use in an assay, while a second container contains primers.

[0090] “Lymphoid neoplasm” means an abnormal proliferation of lymphocytes that may be malignant or non-malignant. A lymphoid cancer is a malignant lymphoid neoplasm.

Lymphoid neoplasms are the result of, or are associated with, lymphoproliferative disorders, including but not limited to, follicular lymphoma, chronic lymphocytic leukemia (CLL), acute lymphocytic leukemia (ALL), hairy cell leukemia, lymphomas, multiple myeloma, post-transplant lymphoproliferative disorder, mantle cell lymphoma (MCL), diffuse large B cell lymphoma (DLBCL), T cell lymphoma, or the like, e.g. Jaffe et al, Blood, 112: 4384-4399 (2008); Swerdlow et al, WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (e. 4th) (IARC Press, 2008).

[0091] “Minimal residual disease” means remaining cancer cells after treatment. The term is most frequently used in connection with treatment of lymphomas and leukemias.

[0092] “Percent homologous,” “percent identical,” or like terms used in reference to the comparison of a reference sequence and another sequence (“comparison sequence”) mean that in an optimal alignment between the two sequences, the comparison sequence is identical to the reference sequence in a number of subunit positions equivalent to the indicated percentage, the subunits being nucleotides for polynucleotide comparisons or amino acids for polypeptide comparisons. As used herein, an “optimal alignment” of sequences being compared is one that maximizes matches between subunits and minimizes the number of gaps employed in constructing an alignment. Percent identities may be determined with commercially available implementations of algorithms, such as that described by Needleman

and Wunsch, J. Mol. Biol., 48: 443-453 (1970) ("GAP" program of Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI), or the like. Other software packages in the art for constructing alignments and calculating percentage identity or other measures of similarity include the "BestFit" program, based on the algorithm of Smith and Waterman, Advances in Applied Mathematics, 2: 482-489 (1981) (Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI). In other words, for example, to obtain a polynucleotide having a nucleotide sequence at least 95 percent identical to a reference nucleotide sequence, up to five percent of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to five percent of the total number of nucleotides in the reference sequence may be inserted into the reference sequence.

[0093] "Polymerase chain reaction," or "PCR," means a reaction for the in vitro amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA. In other words, PCR is a reaction for making multiple copies or replicates of a target nucleic acid flanked by primer binding sites, such reaction comprising one or more repetitions of the following steps: (i) denaturing the target nucleic acid, (ii) annealing primers to the primer binding sites, and (iii) extending the primers by a nucleic acid polymerase in the presence of nucleoside triphosphates. Usually, the reaction is cycled through different temperatures optimized for each step in a thermal cycler instrument. Particular temperatures, durations at each step, and rates of change between steps depend on many factors well-known to those of ordinary skill in the art, e.g. exemplified by the references: McPherson et al, editors, PCR: A Practical Approach and PCR2: A Practical Approach (IRL Press, Oxford, 1991 and 1995, respectively). For example, in a conventional PCR using Taq DNA polymerase, a double stranded target nucleic acid may be denatured at a temperature $>90^{\circ}\text{C}$, primers annealed at a temperature in the range $50\text{-}75^{\circ}\text{C}$, and primers extended at a temperature in the range $72\text{-}78^{\circ}\text{C}$. The term "PCR" encompasses derivative forms of the reaction, including but not limited to, RT-PCR, real-time PCR, nested PCR, quantitative PCR, multiplexed PCR, and the like. The particular format of PCR being employed is discernible by one skilled in the art from the context of an application. Reaction volumes range from a few hundred nanoliters, e.g. 200 nL, to a few hundred μL , e.g. 200 μL . "Reverse transcription PCR," or "RT-PCR," means a PCR that is preceded by a reverse transcription reaction that converts a target RNA to a complementary single stranded DNA, which is then amplified, e.g. Tecott et al, U.S. patent 5,168,038, which patent is incorporated herein by reference. "Real-time PCR" means a PCR for which the amount of reaction

product, i.e. amplicon, is monitored as the reaction proceeds. There are many forms of real-time PCR that differ mainly in the detection chemistries used for monitoring the reaction product, e.g. Gelfand et al, U.S. patent 5,210,015 ("taqman"); Wittwer et al, U.S. patents 6,174,670 and 6,569,627 (intercalating dyes); Tyagi et al, U.S. patent 5,925,517 (molecular beacons); which patents are incorporated herein by reference. Detection chemistries for real-time PCR are reviewed in Mackay et al, *Nucleic Acids Research*, 30: 1292-1305 (2002), which is also incorporated herein by reference. "Nested PCR" means a two-stage PCR wherein the amplicon of a first PCR becomes the sample for a second PCR using a new set of primers, at least one of which binds to an interior location of the first amplicon. As used herein, "initial primers" in reference to a nested amplification reaction mean the primers used to generate a first amplicon, and "secondary primers" mean the one or more primers used to generate a second, or nested, amplicon. "Multiplexed PCR" means a PCR wherein multiple target sequences (or a single target sequence and one or more reference sequences) are simultaneously carried out in the same reaction mixture, e.g. Bernard et al, *Anal. Biochem.*, 273: 221-228 (1999)(two-color real-time PCR). Usually, distinct sets of primers are employed for each sequence being amplified. Typically, the number of target sequences in a multiplex PCR is in the range of from 2 to 50, or from 2 to 40, or from 2 to 30. "Quantitative PCR" means a PCR designed to measure the abundance of one or more specific target sequences in a sample or specimen. Quantitative PCR includes both absolute quantitation and relative quantitation of such target sequences. Quantitative measurements are made using one or more reference sequences or internal standards that may be assayed separately or together with a target sequence. The reference sequence may be endogenous or exogenous to a sample or specimen, and in the latter case, may comprise one or more competitor templates. Typical endogenous reference sequences include segments of transcripts of the following genes: β -actin, GAPDH, β_2 -microglobulin, ribosomal RNA, and the like. Techniques for quantitative PCR are well-known to those of ordinary skill in the art, as exemplified in the following references that are incorporated by reference: Freeman et al, *Biotechniques*, 26: 112-126 (1999); Becker-Andre et al, *Nucleic Acids Research*, 17: 9437-9447 (1989); Zimmerman et al, *Biotechniques*, 21: 268-279 (1996); Diviacco et al, *Gene*, 122: 3013-3020 (1992); Becker-Andre et al, *Nucleic Acids Research*, 17: 9437-9446 (1989); and the like.

[0094] "Polynucleotide" refers to a linear polymer of nucleotide monomers and may be DNA or RNA. Monomers making up polynucleotides are capable of specifically binding to a natural polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen

types of base pairing, or the like. Such monomers and their internucleosidic linkages may be naturally occurring or may be analogs thereof, e.g., naturally occurring or non-naturally occurring analogs. Non-naturally occurring analogs may include PNAs, phosphorothioate internucleosidic linkages, bases containing linking groups permitting the attachment of labels, such as fluorophores, or haptens, and the like. Polynucleotides may comprise the four natural nucleosides (e.g. deoxyadenosine, deoxycytidine, deoxyguanosine, deoxythymidine for DNA or their ribose counterparts for RNA) linked by phosphodiester linkages. However, they may also comprise non-natural nucleotide analogs, e.g. including modified bases, sugars, or internucleosidic linkages. It is clear to those skilled in the art that where an enzyme has specific oligonucleotide or polynucleotide substrate requirements for activity (e.g. single stranded DNA, RNA/DNA duplex, or the like), then selection of an appropriate composition for the oligonucleotide or polynucleotide substrates is well within the knowledge of one of ordinary skill, especially with guidance from treatises such as Sambrook et al, MOLECULAR CLONING, 2nd ed. (Cold Spring Harbor Laboratory, New York, 1989), and like references. As used herein, the term “oligonucleotide” refers to smaller polynucleotides, for example, having 3-60 monomeric units, or in some embodiments having from 12 to 60 monomeric units. In various embodiments, a polynucleotide or oligonucleotides may be represented by a sequence of letters (upper or lower case), such as “ATGCCTG,” and it will be understood that the nucleotides are in 5'→3' order from left to right and that “A” denotes deoxyadenosine, “C” denotes deoxycytidine, “G” denotes deoxyguanosine, and “T” denotes thymidine, “I” denotes deoxyinosine, “U” denotes uridine, unless otherwise indicated or obvious from context.

[0095] “Primer” means an oligonucleotide, either natural or synthetic that is capable, upon forming a duplex with a polynucleotide template, of acting as a point of initiation of nucleic acid synthesis and being extended from its 3' end along the template so that an extended duplex is formed. Extension of a primer is usually carried out with a nucleic acid polymerase, such as a DNA or RNA polymerase. The sequence of nucleotides added in the extension process is determined by the sequence of the template polynucleotide. Usually primers are extended by a DNA polymerase. Primers usually have a length in the range of from 14 to 40 nucleotides, or in the range of from 18 to 36 nucleotides. Primers are employed in a variety of nucleic amplification reactions, for example, linear amplification reactions using a single primer, or polymerase chain reactions, employing two or more primers. Guidance for selecting the lengths and sequences of primers for particular applications is well known to those of ordinary skill in the art, as evidenced by the following

references that are incorporated by reference: Dieffenbach, editor, PCR Primer: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Press, New York, 2003).

[0096] “Quality score” means a measure of the probability that a base assignment at a particular sequence location is correct. A variety methods are well known to those of ordinary skill for calculating quality scores for particular circumstances, such as, for bases called as a result of different sequencing chemistries, detection systems, base-calling algorithms, and so on. Generally, quality score values are monotonically related to probabilities of correct base calling. For example, a quality score, or Q, of 10 may mean that there is a 90 percent chance that a base is called correctly, a Q of 20 may mean that there is a 99 percent chance that a base is called correctly, and so on. For some sequencing platforms, particularly those using sequencing-by-synthesis chemistries, average quality scores decrease as a function of sequence read length, so that quality scores at the beginning of a sequence read are higher than those at the end of a sequence read, such declines being due to phenomena such as incomplete extensions, carry forward extensions, loss of template, loss of polymerase, capping failures, deprotection failures, and the like.

[0097] “Sequence read” means a sequence of nucleotides determined from a sequence or stream of data generated by a sequencing technique, which determination is made, for example, by means of base-calling software associated with the technique, e.g. base-calling software from a commercial provider of a DNA sequencing platform. A sequence read usually includes quality scores for each nucleotide in the sequence. Typically, sequence reads are made by extending a primer along a template nucleic acid, e.g. with a DNA polymerase or a DNA ligase. Data is generated by recording signals, such as optical, chemical (e.g. pH change), or electrical signals, associated with such extension. Such initial data is converted into a sequence read.

[0098] “Sequence tag” (or “tag”) or “barcode” means an oligonucleotide that is attached to a polynucleotide or template molecule and is used to identify and/or track the polynucleotide or template in a reaction or a series of reactions. A sequence tag may be attached to the 3’- or 5’-end of a polynucleotide or template or it may be inserted into the interior of such polynucleotide or template to form a linear conjugate, sometime referred to herein as a “tagged polynucleotide,” or “tagged template,” or “tag-polynucleotide conjugate,” “tag-molecule conjugate,” or the like. Sequence tags may vary widely in size and compositions; the following references, which are incorporated herein by reference, provide guidance for selecting sets of sequence tags appropriate for particular embodiments: Brenner, U.S. patent 5,635,400; Brenner and Macevicz, U.S. patent 7,537,897; Brenner et al, Proc. Natl. Acad.

Sci., 97: 1665-1670 (2000); Church et al, European patent publication 0 303 459; Shoemaker et al, Nature Genetics, 14: 450-456 (1996); Morris et al, European patent publication 0799897A1; Wallace, U.S. patent 5,981,179; and the like. Lengths and compositions of sequence tags can vary widely, and the selection of particular lengths and/or compositions depends on several factors including, without limitation, how tags are used to generate a readout, e.g. via a hybridization reaction or via an enzymatic reaction, such as sequencing; whether they are labeled, e.g. with a fluorescent dye or the like; the number of distinguishable oligonucleotide tags required to unambiguously identify a set of polynucleotides, and the like, and how different must tags of a set be in order to ensure reliable identification, e.g. freedom from cross hybridization or misidentification from sequencing errors. In one aspect, sequence tags can each have a length within a range of from 2 to 36 nucleotides, or from 4 to 30 nucleotides, or from 8 to 20 nucleotides, or from 6 to 10 nucleotides, respectively. In one aspect, sets of sequence tags are used wherein each sequence tag of a set has a unique nucleotide sequence that differs from that of every other tag of the same set by at least two bases; in another aspect, sets of sequence tags are used wherein the sequence of each tag of a set differs from that of every other tag of the same set by at least three bases.

[0099] "Sequence tree" means a tree data structure for representing nucleotide sequences. In one aspect, a tree data structure of the invention is a rooted directed tree comprising nodes and edges that do not include cycles, or cyclical pathways. Edges from nodes of tree data structures of the invention are usually ordered. Nodes and/or edges are structures that may contain, or be associated with, a value. Each node in a tree has zero or more child nodes, which by convention are shown below it in the tree. A node that has a child is called the child's parent node. A node has at most one parent. Nodes that do not have any children are called leaf nodes. The topmost node in a tree is called the root node. Being the topmost node, the root node will not have parents. It is the node at which operations on the tree commonly begin (although some algorithms begin with the leaf nodes and work up ending at the root). All other nodes can be reached from it by following edges or links.

SEQUENCE LISTING

<110> Sequentia, Inc.
Faham, Malek
Moorhead, Martin
Willis, Thomas

<120> METHOD OF SEQUENCE DETERMINATION USING SEQUENCE TAGS

<130> 838US00 (37623-730.201)

<140> 13/835093

<141> 2013-03-15

<150> 13/100365

<151> 2011-05-04

<150> 61/658317

<151> 2012-06-11

<150> 61/738277

<151> 2012-12-17

<150> 61/776647

<151> 2013-03-11

<160> 6

<170> PatentIn version 3.5

<210> 1

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> primer

<400> 1

agttctggct aacctgtaga gccca

24

<210> 2

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> primer

<400> 2

agttcgggct aacctgtcga gccca

24

<210> 3
<211> 24
<212> DNA
<213> Artificial Sequence

<220>
<223> primer

<400> 3
agttccggct aacctgtcga gccca

24

<210> 4
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> primer

<220>
<221> misc_feature
<222> (1)..(22)
<223> n is a, c, g, or t

<400> 4
nnnnnnnnnnn nnnnnnnnnnn nn

22

<210> 5
<211> 12
<212> DNA
<213> Artificial Sequence

<220>
<223> primer

<400> 5
gtatttttt ct

12

<210> 6
<211> 13
<212> DNA
<213> Artificial Sequence

<220>

<223> primer

<400> 6

ttcagggggg gct

13

What is claimed is:

1. A method of determining clonotypes of an immune repertoire, the method comprising the steps:
 - (a) obtaining a sample from an individual comprising T-cells and/or B-cells;
 - (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or immunoglobulin genes of the T-cells and/or B-cells to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag;
 - (c) amplifying the tag-molecule conjugates;
 - (d) sequencing the tag-molecule conjugates; and
 - (e) aligning sequence reads of like sequence tags to determine sequence reads corresponding to the same clonotypes of the repertoire.
2. The method of claim 1 wherein said step of aligning further includes determining a nucleotide sequence of each of said clonotype of each of said tag-molecule conjugate by determining a majority nucleotide at each nucleotide position of said clonotypes of said like sequence tags.
3. The method of claim 1 wherein said step of attaching includes labeling by sampling said molecules of recombined nucleic acids.
4. The method of claim 3 wherein said step of attaching is implemented in a reaction mixture such that said sequence tags are present in the reaction mixture in a concentration at least 100 times that of said molecules of recombined nucleic acid.
5. The method of claim 4 wherein said sequence tags are incorporated into primers specific for said molecules of recombined nucleic acids.
6. The method of claim 5 wherein said sequence tags are mosaic tags.
7. A method of determining a number of lymphocytes in a sample, the method comprising the steps of:
 - (a) obtaining a sample from an individual comprising lymphocytes;
 - (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or of immunoglobulin genes of the lymphocytes to form tag-molecule conjugates, wherein substantially every molecule of the tag-molecule conjugates has a unique sequence tag;
 - (c) amplifying the tag-molecule conjugates;
 - (d) sequencing the tag-molecule conjugates;

(e) counting the number of distinct sequence tags to determine the number of lymphocytes in the sample.

8. The method of claim 7 wherein said recombined nucleic acids are DNA.

9. The method of claim 7 wherein said lymphocyte is a T-cell and said recombined nucleic acids are T-cell receptor genes or fragments thereof.

10. The method of claim 7 wherein said lymphocyte is a B-cell and said recombined nucleic acids are immunoglobulin genes or fragments thereof.

11. A method of determining clonotypes of an immune repertoire, the method comprising the steps:

(a) obtaining a sample from an individual comprising T-cells and/or B-cells;

(b) labeling by sampling molecules from the T-cells and/or B-cells to form tag-molecule conjugates, wherein each tag has a sequence and each molecule comprises a recombined nucleic acid from a T-cell receptor gene or an immunoglobulin gene;

(c) sequencing the tag-molecule conjugates; and

(d) aligning sequence reads of like tags to determine like clonotypes.

12. The method of claim 11 wherein said step of aligning further includes determining a nucleotide sequence of each of said clonotype of each of said tag-molecule conjugate by determining a majority nucleotide at each nucleotide position of said clonotypes of said like sequence tags.

13. The method of claim 11 wherein said step of attaching is implemented in a reaction mixture such that said sequence tags are present in the reaction mixture in a concentration at least 100 times that of said molecules of recombined nucleic acid.

14. A method of detecting clonotype carry over contamination in a patient being monitored for minimal residual disease, the method comprising the steps of:

(a) monitoring a patient for minimal residual disease by periodically measuring a clonotype profile of the patient in accordance with the method of claim 1; and

(b) recording said sequence of each of said sequence tags of each measurement of clonotype profiles; and

(c) detecting clonotype carry over contamination if a sequence tag of any prior clonotype profile is detected in a subsequent clonotype profile.

15. A method of determining nucleotide sequences of one or more polynucleotides in one or more sequencing-by-synthesis reactions, the method comprising the steps:

(a) attaching a sequence tag to each of the one or more polynucleotides to form tag-polynucleotide conjugates, wherein substantially every polynucleotide of the tag-polynucleotide conjugates has a unique sequence tag;

(b) amplifying the tag-polynucleotide conjugates;

(c) sequencing by synthesis amplified tag-polynucleotide conjugates, wherein sequencing by synthesis comprises at least one dNTP flow; and

(d) determining for each tag-polynucleotide having the same sequence tag a number of nucleotide incorporations for each dNTP flow as a function of measured incorporation signals for each such dNTP flow.

16. The method of 15 wherein said one or more polynucleotides are a plurality of polynucleotides.

17. The method of claim 15 wherein said plurality is at least 104.

18. The method of claim 15 wherein said number of nucleotide incorporations for each of said dNTP flows is a whole number closest to an average of said measured incorporation signals.

19. The method of claim 15 wherein said step of sequencing by synthesis comprises the steps of:

(a) forming for each of said tag-polynucleotide conjugates a complex comprising a sequencing primer, a nucleic acid polymerase, and a tag-polynucleotide conjugate under conditions that permit annealing of the sequencing primer to such tag-polynucleotide conjugate and extension of such sequencing primer along the tag-polynucleotide conjugate in the presence of nucleoside triphosphates by the nucleic acid polymerase;

(b) introducing a dNTP to the complex by a dNTP flow;

(c) measuring incorporation signals;

(d) washing the complex; and

(e) repeating steps (b) through (d).

20. The method of claim 19 wherein said dNTPs are extension-blocked dNTPs and wherein said step of washing further includes a step of de-blocking incorporated extension-blocked dNTPs so that in a subsequent step of introducing a further extension-blocked dNTP may be incorporated.

21. The method of claim 15 wherein each different tag-polynucleotide conjugate is in a different reaction confinement region.
22. The method of claim 15 wherein said step of attaching includes labeling by sampling said one or more polynucleotides to form said tag-polynucleotide conjugates.
23. The method of claim 15 wherein said sequence tags are mosaic tags having variable regions with lengths in the range of from 1 to 5 nucleotides.
24. The method of claim 15 wherein said sequence tags comprise alternating regions comprising nucleotides selected from disjoint subsets of A, C, G and T, such that each alternating region has a length in the range of from 1 to 5 nucleotides.
25. The method of claim 15 wherein said step of sequencing comprises sequencing a sample of said amplified tag-polynucleotide conjugates.
26. The method of claim 15 wherein said one or more polynucleotides are molecules of recombined nucleic acids of T-cell receptor genes or immunoglobulin genes.

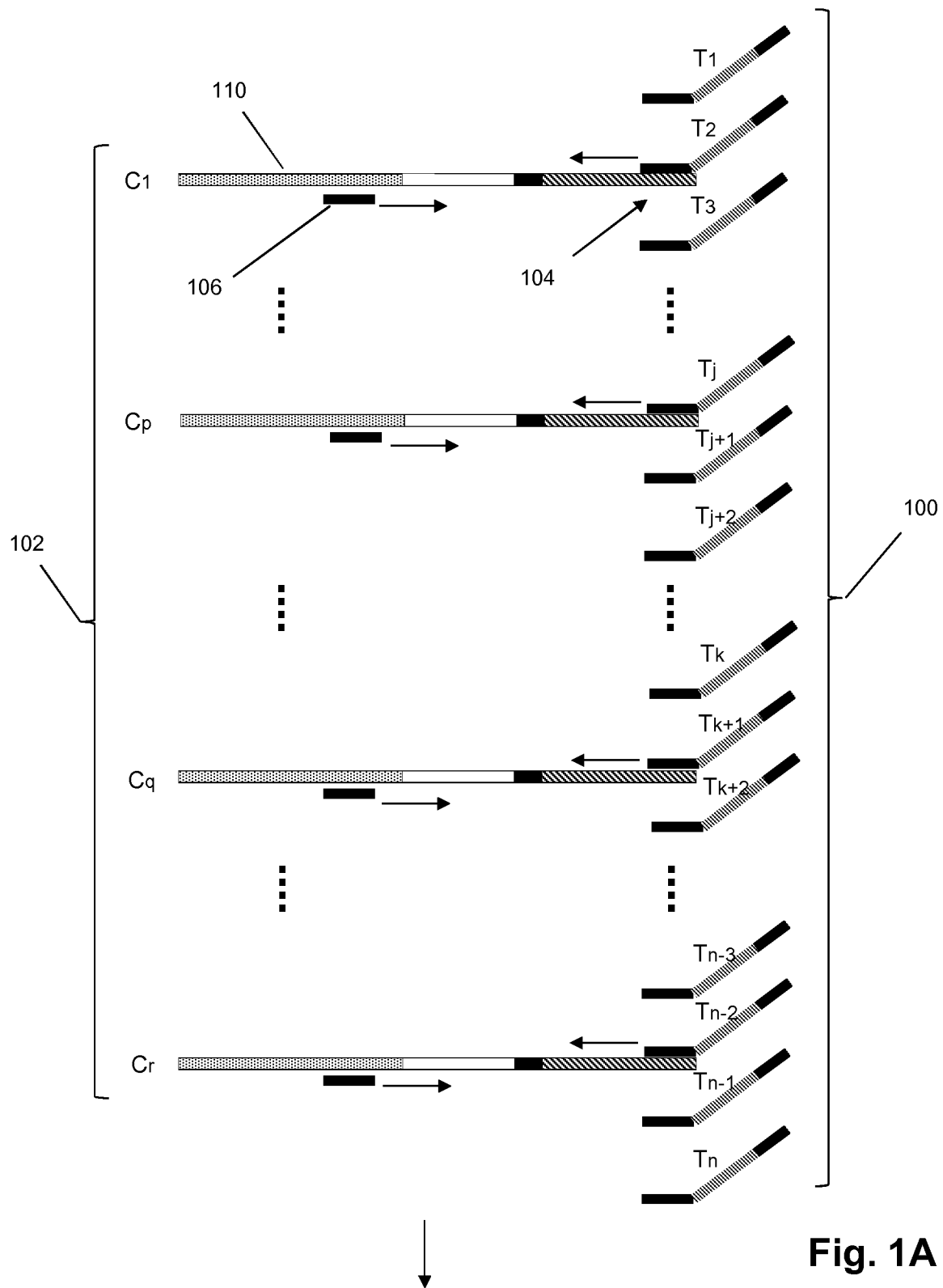


Fig. 1A

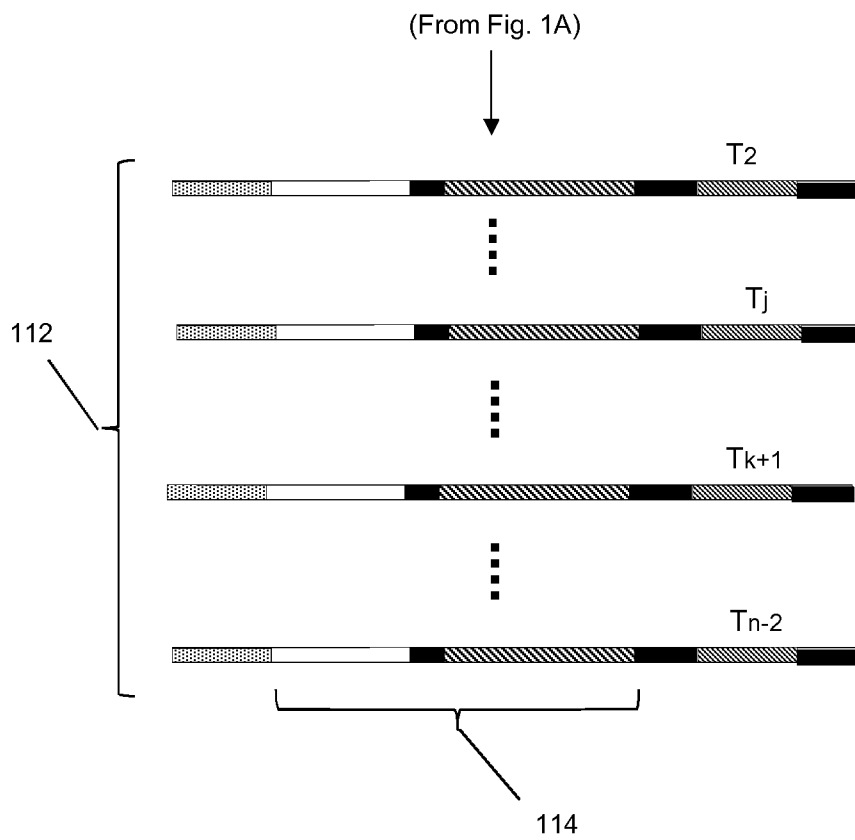


Fig. 1B

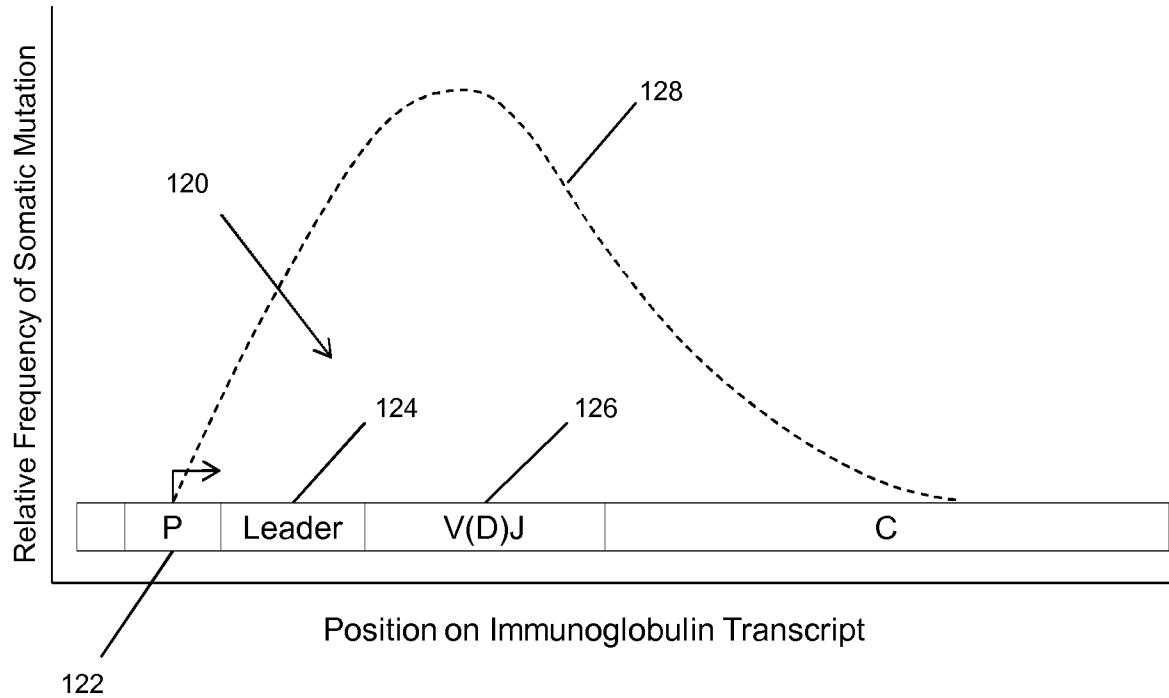


Fig. 1C

3/7

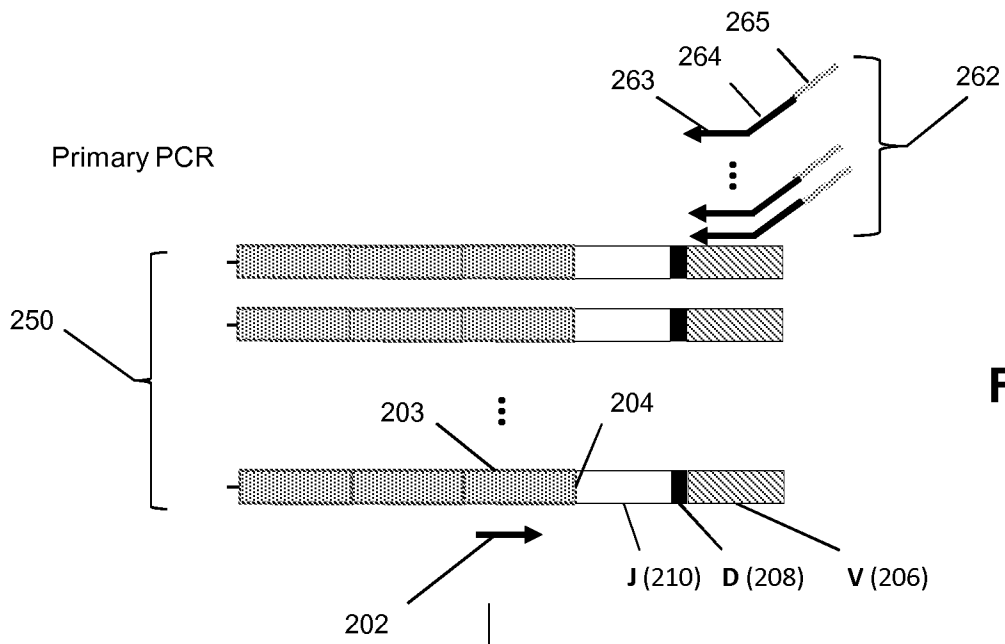


Fig. 2A

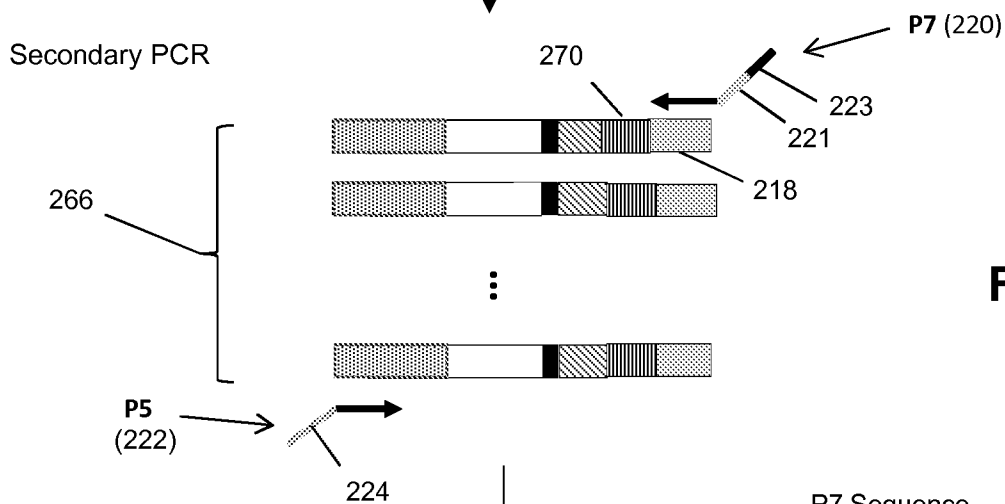


Fig. 2B

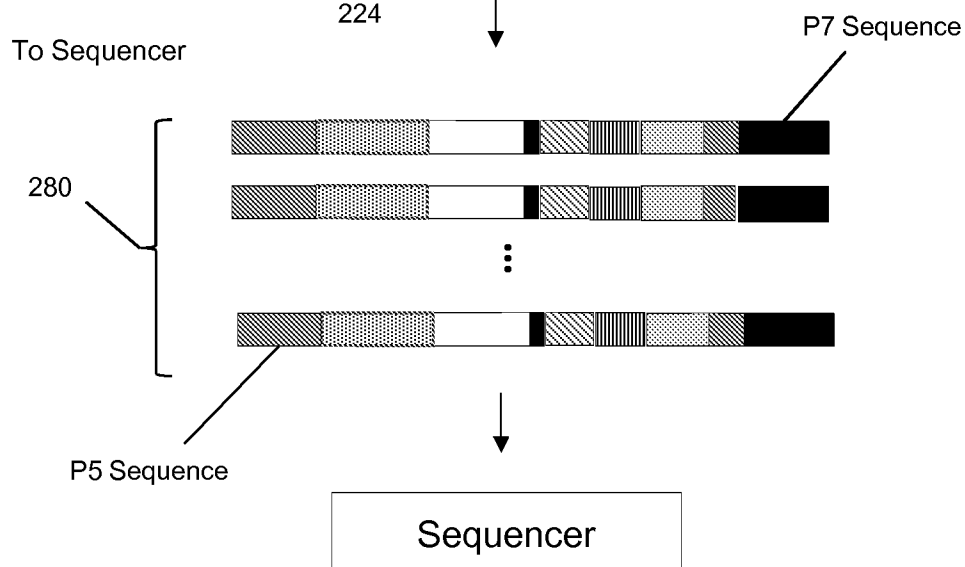
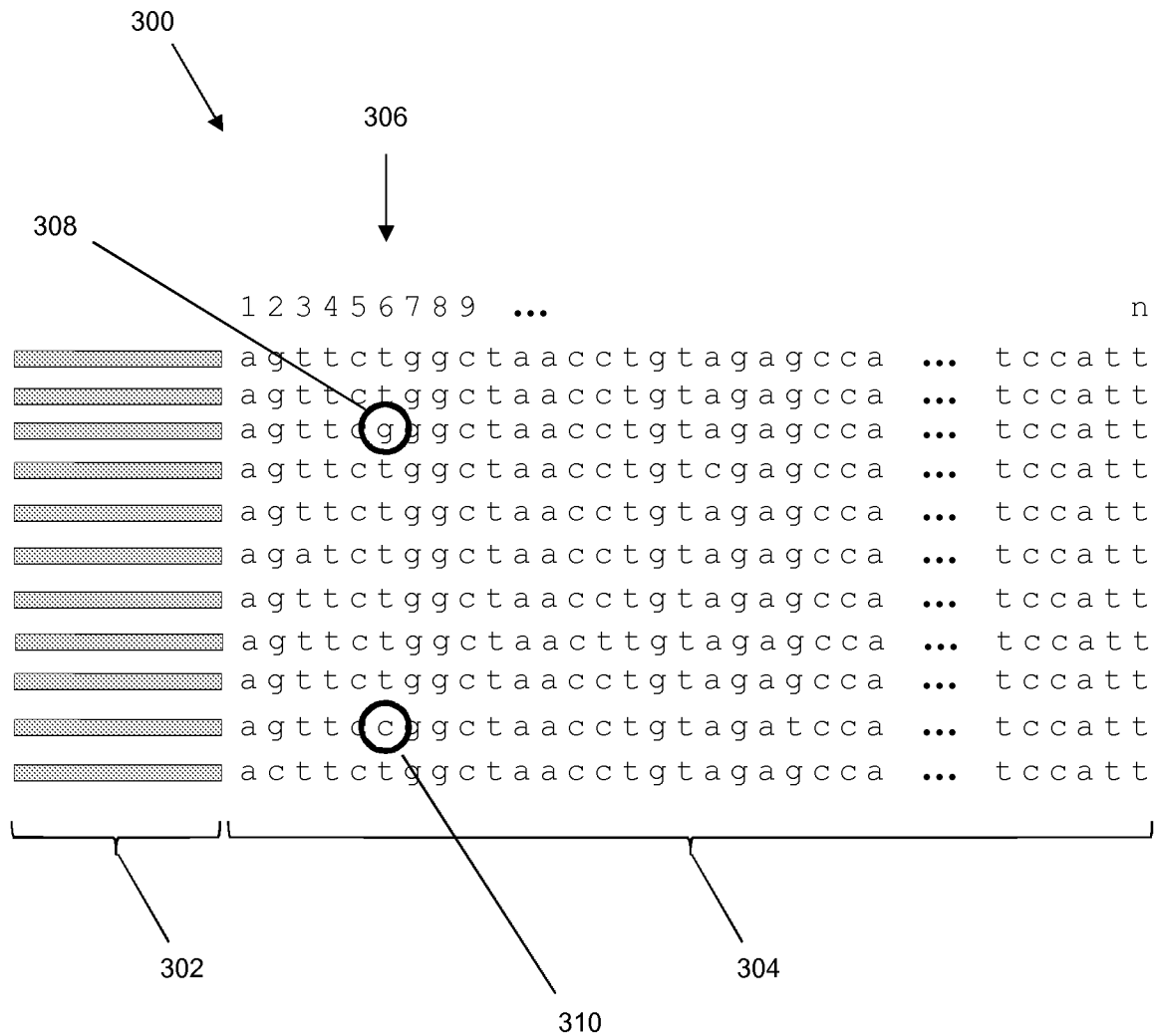


Fig. 2C

**Fig. 3**

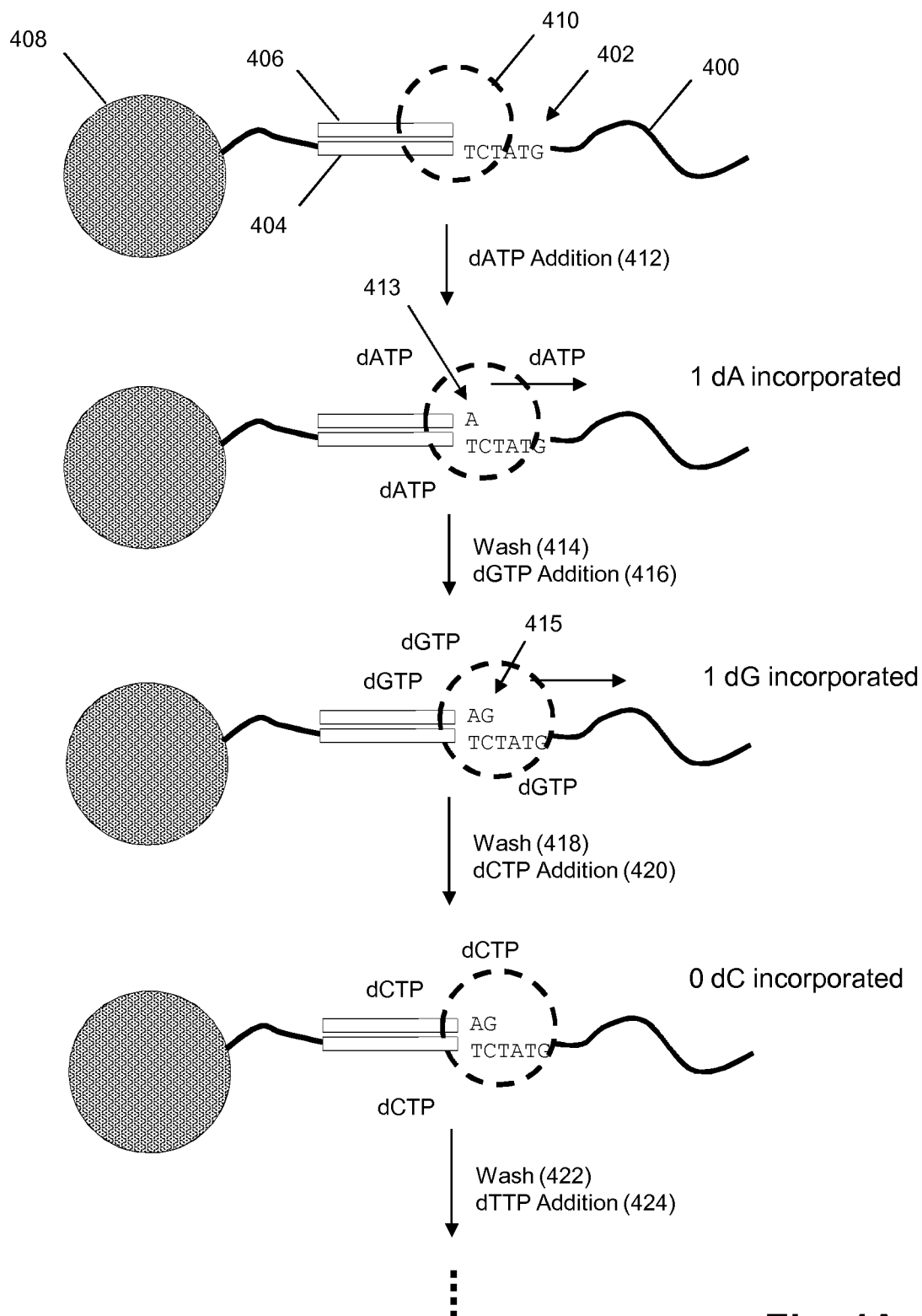
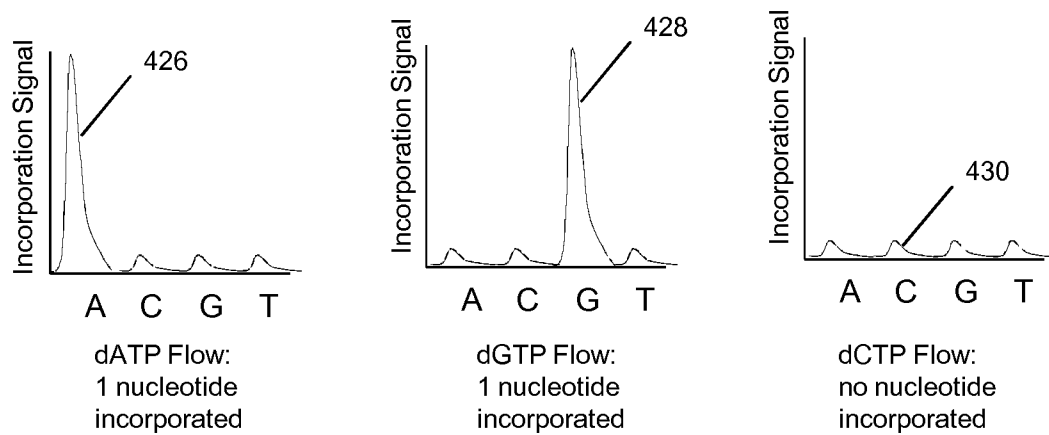
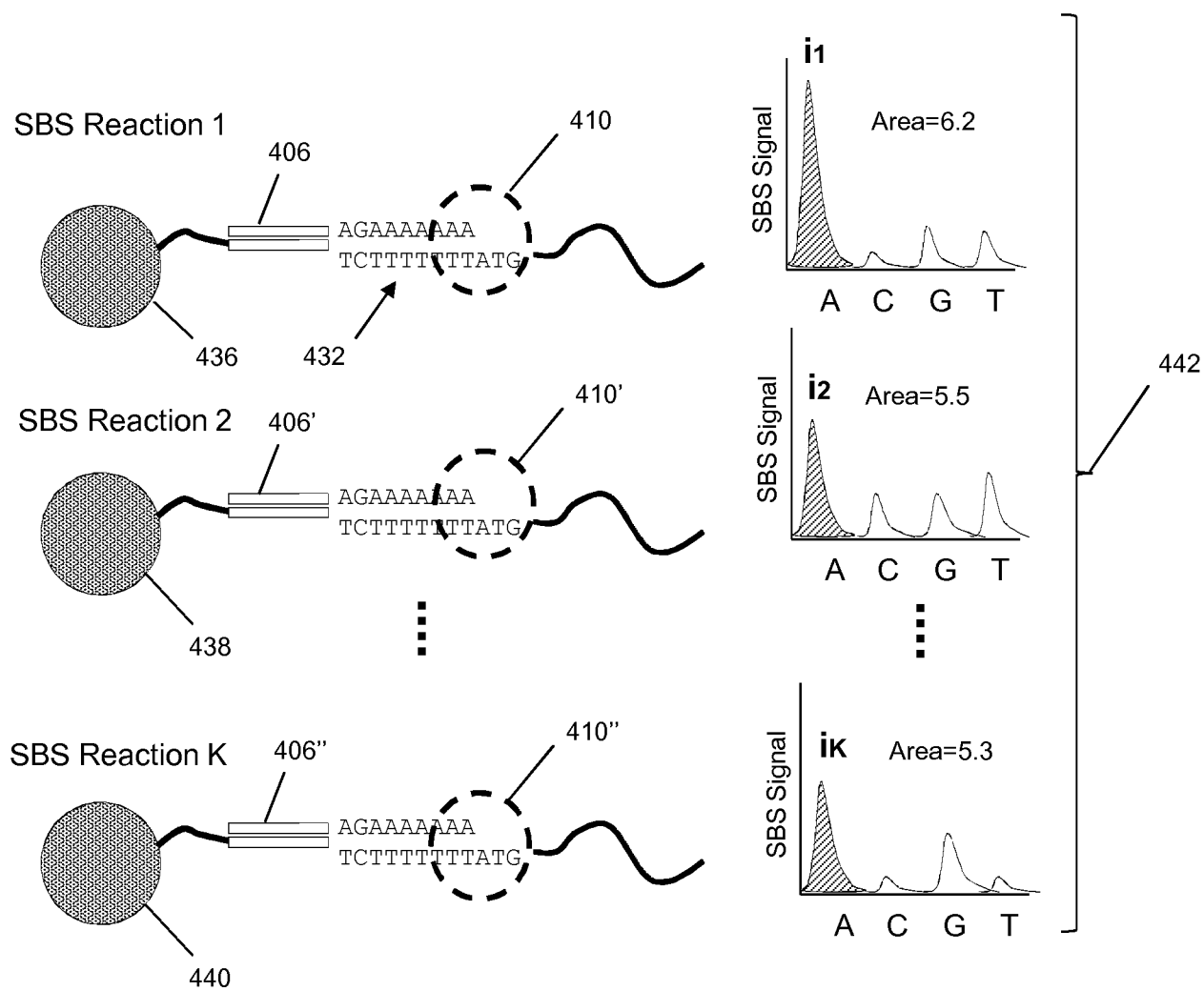


Fig. 4A

**Fig. 4B****Fig. 4C**

Template:
TCGGGGGGGGACTT

SBS Reaction	Common Sequence Tag (452)		Incorporation Signals for dNTP Flows From Separate SBS Reactions (450)					
			f 1 a	f 1 g	f 1 c	f 1 t	f 2 a	f 2 g ...
1	t c t	...	c c c t (1.1)	(1.0)	(6.5)	(0.9)	(1.1)	(2.0) ...
2	t c t	...	c c c t (0.9)	(1.1)	(7.3)	(1.9)	(1.0)	(2.3) ...
3	t c t	...	c c c t (1.0)	(0.7)	(6.8)	(0.9)	(1.0)	(2.1) ...
4	t c t	...	c c c t (1.2)	(0.6)	(6.9)	(0.9)	(1.0)	(1.3) ...
5	t c t	...	c c c t (0.8)	(1.2)	(6.8)	(0.9)	(1.2)	(2.0) ...
6	t c t	...	c c c t (1.1)	(1.0)	(7.0)	(0.9)	(0.8)	(2.0) ...
7	t c t	...	c c c t (1.0)	(0.9)	(6.2)	(1.0)	(0.9)	(1.5) ...
8	t c t	...	c c c t (1.0)	(0.7)	(7.2)	(1.1)	(1.0)	(2.0) ...
9	t c t	...	c c c t (0.9)	(1.0)	(6.9)	(1.1)	(0.5)	(2.0) ...
.								
.								
.								
K	t c t	...	c c c t (1.1)	(1.0)	(7.0)	(1.2)	(1.0)	(1.4) ...

Fig. 4D