(12) STANDARD PATENT
(19) AUSTRALIAN PATENT OFFICE

(11) Application No. **AU 2014284501 B2**

(54) Title
Large-scale biomolecular analysis with sequence tags

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2014284501** (22) Date of Filing: **2014.06.30**

(87) WIPO No: **WO15/002908**

(30) Priority Data

(31) Number	(32) Date	(33) Country
62/001,580	**2014.05.21**	**US**
61/841,878	**2013.07.01**	**US**

(43) Publication Date: **2015.01.08**
(44) Accepted Journal Date: **2020.05.28**

(71) Applicant(s)
Adaptive Biotechnologies Corp.

(72) Inventor(s)
Asbury, Thomas;Hervold, Kieran;Kotwaliwale, Chitra;Faham, Malek;Moorhead, Martin;Weng, Li;Wittkop, Tobias;Zheng, Jianbiao

(74) Agent / Attorney
Pizzeys Patent and Trade Mark Attorneys Pty Ltd, PO Box 291, WODEN, ACT, 2606, AU

(56) Related Art
US 20110207135 A1
WO 2013096480 A2
WO 2009152928 A2

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 January 2015 (08.01.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/002908 A1

(51) International Patent Classification: *C12Q 1/68* (2006.01)

(21) International Application Number: PCT/US2014/044971

(22) International Filing Date: 30 June 2014 (30.06.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/841,878 1 July 2013 (01.07.2013) US
62/001,580 21 May 2014 (21.05.2014) US

(71) Applicant: **SEQUENTA, INC.** [US/US]; 400 East Jamie Sourt Suite 301, South San Francisco, CA 94080 (US).

(72) Inventors: **ASBURY, Thomas**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **HERVOLD, Kieran**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **KOTWALIWALE, Chitra**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **FAHAM, Malek**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **MOORHEAD, Martin**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **WENG, Li**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **WITTKOP, Tobias**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US). **ZHENG, Jianbiao**; 400 East Jamie Court Suite 301, South San Francisco, CA 94080 (US).

(74) Agents: **MCEVOY, Michael, T.** 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, IR, 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, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

[Continued on next page]

(54) Title: LARGE-SCALE BIOMOLECULAR ANALYSIS WITH SEQUENCE TAGS

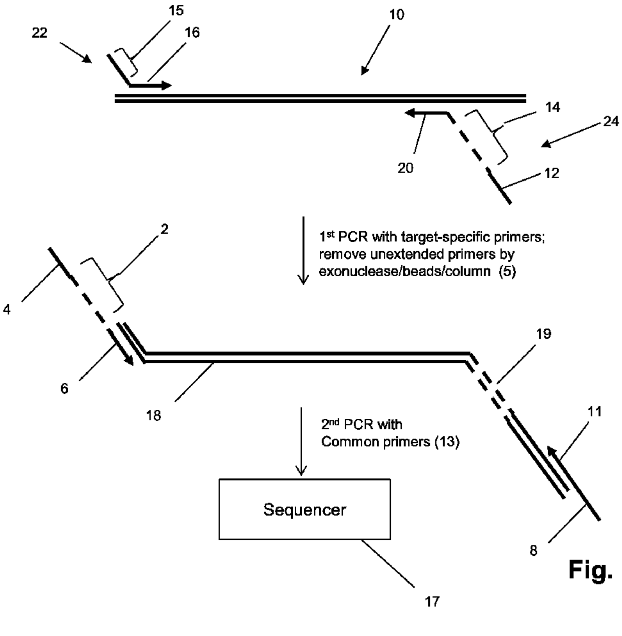

15
10
22
16
14
24
20
12
1st PCR with target-specific primers; remove unextended primers by exonuclease/beads/column (5)
2
4
19
6
18
11
2nd PCR with Common primers (13)
Sequencer
8
17

Fig. 1A

(57) **Abstract**: The invention is directed to sequence-based profiling of populations of nucleic acids by multiplex amplification and attachment of one or more sequence tags to target nucleic acids and/or copies thereof followed by high-throughput sequencing of the amplification product. In some embodiments, the invention includes successive steps of primer extension, removal of unextended primers and addition of new primers either for amplification (for example by PCR) or for additional primer extensions. Some embodiments of the invention are directed to minimal residual disease (MRD) analysis of patients being treated for cancer. Sequence tags incorporated into sequence reads provide an efficient means for determining clonotypes and at the same time provide a convenient means for detecting carry-over contamination from other samples of the same patient or from samples of a different patient which were tested in the same laboratory.

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

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

LARGE-SCALE BIOMOLECULAR ANALYSIS WITH SEQUENCE TAGS

CROSS-REFERENCE

[0001] This application claims priority from United States provisional applications Ser. No. 61/841,878 filed 01-July-2013 and Ser. No. 62/001,580 filed 21-May-2014, both of which are incorporated herein by reference in their entireties.

SEQUENCE LISTING

[0002] The instant application contains a Sequence Listing which has been submitted in ASCII format via EFS-Web and is hereby incorporated by reference in its entirety. Said ASCII copy, created on June 27, 2014, is named 848US00-SL-ST25.txt and is 2 Kilobytes in size. No new matter has been added.

BACKGROUND OF THE INVENTION

[0003] Large-scale DNA sequencing in diagnostic and prognostic applications has expanded rapidly as its speed and convenience has increased and its per-base cost has decreased, e.g. Ding et al, Nature, 481(7382): 506-510 (2012); Chiu et al, Brit. Med. J., 342: c7401 (2011); Ku et al, Annals of Neurology, 71(1): 5-14 (2012); and the like. 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. patents 8,236,503 and 8,628927; Freeman et al, Genome Research, 19: 1817-1824 (2009); Han et al, J. Immunol., 182 (1001): 42.6 (2009); Boyd et al, Sci. Transl. Med., 1(12): 12ra23 (2009); He et al, Oncotarget (March 8, 2011).

[0004] For example, patients treated for many cancers often retain a minimal residual disease (MRD) related to the cancer. That is, even though a patient may have by clinical measures a complete remission of the disease in response to treatment, a small fraction of the cancer cells may remain that have, for one reason or another, escaped destruction. The type and size of this residual population is an important prognostic factor for the patient's continued treatment, e.g. Campana, Hematol. Oncol. Clin. North Am., 23(5): 1083-1098 (2009); Buccisano et al, Blood, 119(2): 332-341 (2012). Consequently, several techniques for assessing this population have been developed, including techniques based on flow cytometry, in situ hybridization, cytogenetics, amplification of nucleic acid markers, and the like, e.g. Buccisano et al, Current Opinion in Oncology, 21: 582-588 (2009); van Dongen et al, Leukemia, 17(12): 2257-2317

(2003); and the like. The amplification of recombined nucleic acids encoding segments of immune receptors (i.e. clonotypes) from T cells and/or B cells have been particularly useful in assessing MRD in leukemias and lymphomas, because such clonotypes typically have unique sequences which may serve as molecular tags for their associated cancer cells. Such measurements are usually made by amplifying and sequencing nucleic acids encoding a single receptor chain, in part, because such amplifications are highly multiplexed and are difficult to develop. As the scale of multiplexing increases, several problems are encountered, including increased probability of spurious amplifications due to mis-hybridizaitons, primer-dimer formation, variable rates of amplification leading to biased sequence representation, and the like, e.g. Elnifro et al, Clinical Microbiology Reviews, 13(4): 559-570 (2000). Furthermore, the similarity of the target sequences and the incorporation of sequence tags into amplified sequences, either for sequence analysis, sample tracking, contamination detection, or the like, can exacerbate the above difficulties associated with large-scale amplifications. These challenges have prevented the development of large-scale one-reaction amplifications of multiple immune receptor chains, which would be highly beneficial for reducing the number of separate assays required for measuring nucleic acid sequences correlated with a minimal disease.

[0005] In view of the foregoing, it would be highly advantageous if more efficient methods were available for assessing selected nucleic acids in a single reaction, such as exons of cancer genes or clonotypes encoding sets of immune receptor chains.

SUMMARY OF THE INVENTION

[0006] The present invention is directed to methods of large-scale amplification in a single reaction, particularly by a polymerase chain reaction (PCR), of a population of target polynucleotides, such as recombined nucleic acids encoding immune receptor chains, followed by their identification using large-scale DNA sequencing. The invention includes the application of the foregoing methods for monitoring minimal residual disease of a cancer. The invention is exemplified in a number of implementations and applications, some of which are summarized below and throughout the specification.

[0007] In some embodiments, the invention is directed to methods of generating profiles of nucleic acids that encode a population of biomolecules of interest, such as immune receptor molecules. In one aspect, methods of the invention comprise attaching sequence tags to a selected population of nucleic acids in a sample to form tag-nucleic acid conjugates, amplifying the tag-nucleic acid conjugates, and sequencing amplified tag-nucleic acid conjugates to provide sequence reads each comprising both a tag sequence and a nucleic acid sequence, for which a profile of the nucleic acids is generated. In some embodiments, attaching sequence tags is enabled by one or more successive steps of primer extension and primer removal, after which the resulting product may be further amplified without bias by common forward and reverse primers.

[0008] In some embodiments, the invention is directed to methods for detecting and measuring contamination, such as carry-over contamination, in a sample from material originating from a different sample. In one embodiment, such method for detecting contamination in an individual being monitored for a minimal residual disease may comprise the following steps: (a) obtaining from an individual a tissue sample; (b) attaching sequence tags to cancer gene molecules or recombined nucleic acids to form tag-nucleic acid conjugates, wherein at least one nucleic acid or copies thereof have different sequence tags attached and wherein the cancer gene molecules are characteristic of a cancer of the individual; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing a sample of the tag-nucleic acid conjugates to provide sequence reads having error rates and comprising a tag sequence and a cancer gene sequence or recombined nucleic acid sequence; (e) comparing tag sequences to separately determined tag sequences from other tissue samples; and (f) determining the presence, absence and/or level of contamination by the identity of one or more tag sequences with any separately determined tag sequences from other tissue samples.

[0009] In another aspect, the invention is directed to a method as described above for generating clonotype profiles based on at least two chains of a B cell receptor, which method comprises amplifying in a single reaction target nucleic acids encoding two or more chains of a B cell receptor. In another aspect, such methods are employed to monitor minimal residual disease in a B cell cancer.

[0010] In another aspect, the invention is directed to a method as described above generating clonotype profiles based on at least two chains of a T cell receptor, which method comprises amplifying in a single reaction target nucleic acids encoding two or more chains of a T cell receptor. In another aspect, such methods are employed to monitor minimal residual disease in a T cell cancer.

[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 through 1C illustrate diagrammatically various embodiments of the invention. Fig. 1D illustrates a method of generating (with or without sequence tags) templates of recombined nucleic acids having a predetermined length.

[0014] Figs. 2A through 2G illustrate various methods for attaching unique sequence tags to substantially every target sequence in a sample.

[0015] Figs. 3A and 3B illustrate diagrammatically an aspect of the invention for generating clonotype profiles from nucleic acid sequences encoding IgH chains.

[0016] Fig. 4A illustrates the use of sequence tags for determining clonotype sequences from sequence reads. Fig. 4B illustrates the use of sequence tags in embodiments where multiple different sequence tags are attached to the same target polynucleotide or copies thereof.

[0017] Fig. 5A illustrates concepts of clonotypes in sequence space and distances between closely related clonotypes.

[0018] Fig. 5B is a flow chart illustrating one embodiment of a method for distinguishing genuinely different clonotypes from clonotypes that differ solely by sequencing errors (which should be coalesced).

[0019] Fig. 5C illustrates the form of a numerical function used in one embodiment for determining whether or not to coalesce related clonotypes.

[0020] Figs. 5D and 5E illustrate the use of sequence trees in a method of coalescing sequence reads.

DETAILED DESCRIPTION OF THE INVENTION

[0021] 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.

[0022] In one aspect the invention is directed to methods for producing clonotype profiles of multiple immune receptor chains by large-scale multiplex amplification of nucleic acids encoding such chains followed by high-throughput sequencing of the amplification product, or amplicon. In some embodiments, the invention overcomes common drawbacks of multiplex amplification by including successive steps of primer extension, removal of unextended, or unincorporated, primers and addition of new primers either for amplification (for example by PCR) or for additional primer extensions. Such steps also enable the use of sequence tags which otherwise would contribute to nonspecific or spurious amplifications. In another aspect, sequence tags are employed in embodiments with clinical applications, particularly minimal residual disease (MRD) analysis, for example, of samples from a patient being treated for a cancer. Sequence tags incorporated into sequence reads provide an efficient means for determining clonotypes and at the same time provide a convenient means for detecting carry-over contamination by detecting the presence or absence of sequence tags from previous assays, either from samples of the same patient or from samples of a different patient which were tested in the same laboratory. Of particular interest are methods for generating sequence-based clonotype profiles of recombined nucleic acids encoding a plurality of B cell receptor (BCR) chains by using a single amplification reaction followed by high throughput next-generation sequencing. Also of particular interest are methods for generating sequence-based clonotype profiles of recombined nucleic acids encoding a plurality of T cell receptor (TCR) chains by using a single amplification reaction followed by high throughput next generation sequencing. Methods of the invention may also be applied other large-scale amplification and sequencing of other sets of nucleic acids of interest, including, for example, sets of exons of cancer genes. In these aspects, sequence tags permit both monitoring of carry over contamination and more sensitive determination of nucleotide sequences of target polynucleotides in view of error-prone sequencing methodologies. Also in these aspects, a set of sequence tags (as discussed more fully below) is typically much larger than the number of target polynucleotides in a sample and the sequence difference among sequence tags attached to target polynucleotides is large enough so that effectively a sequence of one tag could not be transformed into another by sequencing error.

[0023] One embodiment of the invention is illustrated in Fig. 1A. In a reaction mixture, primers (22) from a first set (each primer of the first set having receptor-specific portion (16) and 5'-non-complementary portion (15) comprising a first primer binding site) anneal to one end of target polynucleotides (10) (after melting target polynucleotide (10)) and primers (24) from a second set (each primer of the second set having receptor-specific portion (20) and 5'—non-complementary portion comprising sequence tag (14) and second primer binding site (12)) anneal to another end of target polynucleotides (10). In some embodiments, as noted below, non-complementary portion (15) of primer (22) may also comprise a sequence tag. In some circumstance, two shorter sequence tags may be more advantageous than a single longer sequence tag of equivalent diversity. Thus, for example, two 8-mer random-nucleotide sequence tags may be less likely to cause spurious priming, primer-dimers, and the like, than a single 16-mer random nucleotide sequence tag. Target polynucleotides (10) are typically somatically recombined nucleic acids from T cells or B cells which encoded chains or portions of chains of T cell receptors (TCRs) or B cell receptors (e.g., portions of IgH chains or IgK chains). Thus, in some embodiments, the receptor-specific portions of primers (22) and (24) may be specific for V region sequences and J region sequences, respectively, or in other embodiments, vice versa.

[0024] In some embodiments, target polynucleotides (10) may comprise complex mixtures of nucleic acids whose sequence profiles are desired, including but not limited to, recombined nucleic acids encoding portions of immune receptor molecules, 16S rDNAs of microbial communities, metagenomic amplifications of genes encoding proteins of industrial or medical importance (such as, enzymes), human or animal genes and/or exons related to specific diseases, such as cancer, infectious disease, or the like. In embodiments relating to recombined nucleic acids encoding immune receptors, usually at least portions of a V, D or J region are present between the two binding locations of the first and second sets of primers. In some embodiments, between the two binding locations of the first and second sets of primers there is at least 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, or a VD rearrangement of TCR δ. In some embodiments, between the two binding locations of the first and second sets of primers there is at least a portion of a VDJ rearrangement of IgH, a DJ rearrangement of IgH, a VJ rearrangement of IgK, or a VJ rearrangement of IgL. In some embodiments, between the two binding locations of the first and second sets of primers there is at least a portion of 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, or a VD rearrangement of TCR δ. In still other embodiments, between the two binding locations of the first and second sets of primers there is at least a portion of a VDJ rearrangement of IgH, a DJ rearrangement of IgH, and a VJ rearrangement of IgK. And in other embodiments, between the two binding locations of the first and second sets of primers there is at least a portion of a VDJ rearrangement of TCR β, a VJ rearrangement of TCR γ, and a VDJ rearrangement of TCR δ or a VD rearrangement of TCR δ. In some embodiments, at least a portion of a VDJ rearrangement comprises the complete D or NDN portion and parts of the V and J segments sufficient for their identification. In some embodiments, at least a portion of a VDJ rearrangement comprises at least a 50 nucleotide segment comprising the complete D or NDN portion and parts of the V and J segments. In some embodiments, at least a portion of a VDJ rearrangement comprises at least a 70 nucleotide segment comprising the complete D or NDN portion and parts of the V and J segments.

[0025] In some embodiments, a first set comprises one or more primers that are each specific for a J segment or a C segment. Primers from such a first set are annealed to their target sequences and are extended, after which non-extended primers of the first set are removed. Primers from a second set that are each specific for a V segment are annealed to their target sequences and are extended. In other embodiments, a first set comprises primers that are each specific for a V segment and primers of such first set are annealed to their target sequences, and are extended, after which non-extended primers of the first set are removed, primers of a second set that are each specific for a J segment or a C segment are annealed to their target sequences and are extended. In alternatives of both of these embodiments, first and second sets may each contain a plurality of primers and each primer may be specific for a different immune receptor segment.

[0026] Returning to Fig. 1A, in some embodiments, primers of the first and second sets are extended (5) by carrying out in alternative embodiments, 1-10, or 2-10, or 3-10, or 4-10, or 5-10 cycles of melting, annealing and extension, after which nonextended primers are removed from the reaction mixture using conventional techniques. In other embodiments, primers of the first and second sets are extended (5) by carrying out in alternative embodiments, 2-5, or 3-5, or 4-5 cycles of melting, annealing and extension, after which nonextended primers are removed from the reaction mixture using conventional techniques. In still another embodiment, primers of the first and second sets are extended by carrying out two cycles of melting, annealing and extending. For example, nonextended primers may be removed by exonuclease digestion, hybridization to complementary sequences on magnetic beads, size exclusion chromatorgraphy, commercially available spin columns (e.g. Qiagen QIAquick PCR Purification Kit), or the like. In one embodiment, unextended, or unincorporated, primers are removed, for example, by digestion with an exonuclease I. Double stranded DNAs (18), which are products of extensions (5), have common first and second primer binding sites at each end, to which (in some embodiments) forward and reverse primers, with complementary sequences (6 and 11), may be added for later generation of clusters by bridge PCR. In some embodiments, double stranded DNA also has sequence tag (19) and forward or reverse primer may include sample tag (2) for identifying or tracking or associating DNA (18) with a sample or patient. In some embodiments, sequence tag (19) is substantially unique for each different recombined nucleic acid in a sample. As explained more fully below, sequence tag (19) may be used for coalescing sequence reads into clonotypes as well as used for detecting and tracking sample contamination. Forward and reverse primers may also include primer binding sites (4) and (8) for implementing (13) bridge PCR for certain sequencing protocols, e.g. on a Genome Analyzer (Illumina, San Diego)(17). In other embodiments, in which more than one extensions are carried out with sequence tag-containing primers, each different recombined nucleic acid in a sample may have copies with different sequence tags attached; thus, for example, if four separate cycles of melting, annealing and extension are carried out on target polynucleotides in accordance with the embodiment of Fig. 1A, and if the sample contains recombined nucleic acid, S_1, then at the completion of amplification (13) with common primers, the copies of S_1 will have up to four different sequence tags. Therefore, sequence reads of S_1 will have up to four different sequence tags. As explained more fully below, in such embodiments, clonotypes may be determined by a combination of aligning sequence tags and coalescing sequence reads within each subset defined by a common sequence tag.

[0027] In another embodiment, at least two extensions and two steps to remove unincorporated primers are implemented prior to PCR with common primers. As illustrated in Fig. 1B, primers (101) are annealed to one end of target polynucleotides (100), such as recombined nucleic acids encoding immune receptor chains, and extended, e.g. with a DNA polymerase. Primers (101) may each include receptor-specific portion (103) and 5' noncomplementary portion (105) which, in turn, comprises sequence tag (104) and first primer binding site (102). After extension and removal of unincorporated primers (130), as described above, to first extension product (109) in the reaction mixture is added (a) primers (125), wherein each primer comprises receptor-specific portion (106) and 5'-non-complementary portion (115) (that contains a primer binding site), and (b) primers (127), comprising portion (108) specific for first primer binding site (102) and 5' non-complementary portion (117). After primers (125) and (127) anneal to their primer binding sites, they are extended (107) to form second extension product (118), after which non-extended primers are removed. To second extension product (118) common forward primers (112) and reverse primers (110) are added and a PCR is implemented (111), after which the resulting amplicon is sequenced (120). As above with the embodiment of Fig. 1A, whenever more than one extension step is performed in the presence of sequence tag-containing primers (such as (101)), copies of the same target polynucleotide (100) may be labeled with a plurality of different sequence tags.

[0028] Fig. 1C illustrates another embodiment with V, D and J regions shown explicitly. In a reaction mixture under primer annealing conditions, to recombined nucleic acids (1200) encoding immune receptors, such as TCRs, primers (1212), a first set of primers specific to V region (1226), is added. Each primer of the first set (1212) includes a receptor-specific portion and a 5’-non-complementary portion which, in turn, comprises optionally a sequence tag and a first primer binding site (e.g., 102, 103 and 104 in Fig. 1B). Primers of first set (1212) anneal to V regions (1226) of recombined nucleic acids (1200) and primers of first set (1212) are extended (1202) through D region (1224) into at least J region (1222) and optionally to C region (1220) to form first extension products (1216) that includes optional sequence tag (1228) and first primer binding site (1230). After removing nonextended primers of first set (1212), primers of second set (1240) are added to the reaction mixture under annealing conditions so that they anneal to their respective target J regions (1222), after which they are extended (1204) to form second extension products (1232), each of which comprises sequence tag (1236) (optional) and second primer binding site (1234). Second extension products (1232) may comprise a single sequence tag located, for example, adjacent to V regions (1226), as shown by sequence tag (1228), or adjacent to J regions (1222), as shown by sequence tag (1236), or second extension products (1232) may comprise two sequence tags located in both positions. In one embodiment, second extension products (1232) comprise a single sequence tag (1228) adjacent to V regions (1226). In another embodiment, second extension products (1232) comprise a single sequence tag (1236) adjacent to J regions (1222). In some embodiments, sequence tags (1228) and/or (1236) are mosaic tags described below. After nonextended primers of second set (1240) are removed, common forward and reverse primers are added which are specific for first and second primer binding sites (1230) and (1234), respectively, and a PCR is carried out (1206). A sample of the resulting amplicon is sequenced (1208) to generate sequence reads for constructing clonotypes and clonotype profiles.

[0029] Fig. 1D illustrates a method of generating templates of a defined length and for attaching one or two sequence tags thereto. The embodiment of Fig. 1D shows messenger RNA (mRNA) as the starting material, but the method may be used with either DNA or RNA samples. To mRNA (1300) containing a VDJ region, one or more primers (1312) specific for C region (1308) ("C primers") are annealed to mRNA (1300). Usually only a single C primer is used. Alternatively, one or more primers (having a similar structure) specific for J region may be used. C primer (1312) comprises target specific segment (1313), sequence tag segment (1314) and common primer binding site (1315). Also annealed to target mRNAs (1300) are polymerase blockers (1310), which may be oligonucleotides specific for V regions (1302). In some embodiments, blockers (1310) may be a natural oligonucleotide so long as a polymerase used to extend primer (1312) does not have either strand displacement activity or 5'→3' exonuclease activity and so long as the oligonucleotide is non-extendable, e.g. it has a 3'-dideoxynucleotide. Usually, blockers (1310) are oligonucleotide analogs with enhanced binding activity and nuclease resistance, such as antisense compounds. In some embodiments, blockers (1310) may be locked nucleic acids (LNAs) or peptide nucleic acids (PNAs) or bridged nucleic acids (BNAs), which are disclosed in the following references, Wengel et al, U.S. patents 6,794,499; 7,572,582; Vester et al, Biochemistry, 43(42): 13233-13241 (2004); and the like, and Kazuyuki et al, Chem. Comm., 3765-3767 (2007); Nielson et al, Chem . Soc. Rev., 26: 73-78 (1997); and the like. Sequences of blockers (1310) are selected so that the extension of primer(s) (1312) are halted at a predetermined location on V region (1302). In some embodiments, blockers (1310) are designed so that only enough of V region (1302) is copied in the extension step so that the V region can be identified from the copied sequence. In some embodiments, obtaining blockers (1310) specific for each V region is unnecessary, as consensus sequences may be selected that permit some mismatches, so long as the progression of a polymerase is stopped. The lengths of blockers (1310) may vary widely depending on the kind of oligonucleotide or analog used. In some embodiments, blockers (1310) have lengths in the range of from 10 to 25 monomers. In some embodiments, blockers (1310) may anneal to different locations on different V region sequences.

[0030] Returning to Fig. 1D, primers (1312) are extended to blockers (1310) making a cDNA copy of a portion of VDJ region of target (1300) that has a predetermined length. In some embodiments, the predetermined length (or equivalently the binding sites of blockers (1310)) are selected so that a desired portion of the VDJ region may be covered by one or more sequence reads of the sequence technique used in the method. After extension is completed, RNA template (1300) is digested (1325) using conventional techniques, e.g. digestion with an RNAse, such as RNAse H and/or RNAse A, to give single stranded cDNA (1326). To this cDNA is added a 3' mononucleotide tail, such as a polyC tail, using terminal deoxynucleotide transferase

(TdT) in a conventional protocol. To tailed cDNA (1331), adaptor (1336) having a complementary overhang to the mononucleotide tail of cDNA (1331), after which it is extended to produce double stranded DNA (1340), which may be amplified, e.g. by PCR (1337), and the resulting amplicon sequenced (1338).

[0031] Recombined nucleic acids that undergo hypermutation, such as IgH-encoding nucleic acids, may be amplified using sets of primers that include primers that bind to different primer binding sites on the same recombined nucleic acid; that is, such sets may include primers that bind to one or more nonoverlapping primer binding sites on the same recombined nucleic acid encoding a receptor chain. Such set may comprise one or both first set of primers and second set of primers. In some embodiments, recombined nucleic acids subject to hypermutation are amplified with a first set of primers and a second set of primers wherein at least one of the two sets comprises primers specific for a plurality of nonoverlapping primer binding sites, for example, one set may contain for each different V segment a plurality of primers each specific for a different nonoverlapping primer binding on the different V segments. An embodiment applicable to amplification of recombined nucleic acids undergoing hypermutation is illustrated in Figs. 3A-3B, where nested sets of primers are employed to ensure amplification of each recombined nucleic acid in a sample under conditions, for example, of somatic hypermutation, clonal evolution, or the like. Recombined nucleic acids, e.g. encoding IgH molecules, are combined in a reaction mixture under annealing conditions with first nested set (302) of primers, comprising in this example groups (304), (306) and (308) of primers specific for different sites along V region (316) of recombined nucleic acids (300). In this embodiment, the first nested set comprises a plurality of groups of primers, each specific for a different site or location of the V region, wherein the different members of a group are specific for different variants of the V region at the site. In some embodiments, the plurality of groups is in the range of 2-4; in other embodiments, the plurality is 2 or 3. In some embodiments, each primer of first nested set (302) may have a unique sequence tag (314) and first primer binding site (312) in a 5' noncomplementary tail. Primers of first nested set (302) anneal to their target recombined nucleic acids and are extended through D region (318) and at least a portion of J region (320) to form first amplicon (323) which comprises three components (330), (332) and (334) corresponding to the three subsets of primers (304), (306) and (308), respectively. Each member of first amplicon (323) incorporates sequence tag (324) and primer binding site (326).

[0032] After nonextended primers are removed (322), second nested set of primers (340) is added to the reaction mixture under annealing conditions. As illustrated in Fig. 3A, primers of second nested set (340) comprise subsets (336) and (338) of primers which anneal at different nonoverlapping positions on J region (320) of members of first amplicon (323). In some embodiments, the second nested set of primers may contain only a single group of primers. Primers of second nested set (340) are extended to form second extension product (360) which comprises subsets (350), (352) and (354) which, in turn, each comprise two further subsets (subsubsets) corresponding to primers (336) and (338). In some embodiments, second nested set of primers (340) contain primers specific to only a single primer binding site and first nested set of primers (302) contain primers specific to at least two non-overlapping primer binding sites. After removing nonextended primers (342), common forward and reverse primers may be added to carry out PCR (356) and a sample of the resulting amplicon may be sequenced (358). In various embodiments, primers of both the first nested set and the second nested set may include sequence tags (339); primers of the first nested set but not the second nested set may include sequence tags; and primers of the second nested set but not the first nested set may include sequence tags. In some embodiments, primers of the first nested set are extended first after which non-extended primers are removed or destroyed and primers of the second nested set are annealed and extended (as illustrated in Figs. 3A-3B). In other embodiments, the order of the annealing, extending and removing steps are reversed; that is, primers of the second nested set are extended first after which non-extended primers are removed or destroyed and primers of the first nested set are annealed and extended.

[0033] In some embodiments of the above method, more than one extension step, either (322) or (342), may be implemented, for example, in order to attach sequence tags to a greater fraction of target polynucleotides in a sample. In such embodiments, more than one distinct sequence tag may be attached to a target polynucleotide and/or copies thereof. That is, a plurality of different sequence tags may be attached to a target polynucleotide and its progeny from an amplification reaction, such as PCR; thus, copies of an original target polynucleotide may be labeled with more than one sequence tag. As explained more fully below, such pluralities of sequence tags are still useful in tracking carry over contamination and in permitting more sensitive determination of target polynucleotide sequences.

[0034] Some of the embodiments described above may be carried out with the following steps. For example, a method of generating clonotype profiles from multiple, or a plurality of, T cell receptor chains may comprise the steps of: (a) combining in a reaction mixture under primer extension conditions a first set of primers with a sample of recombined nucleic acids from T-cells, wherein each primer of the first set has a receptor-specific portion with a length such that the receptor-specific portion anneals to a different recombined nucleic acid at a predetermined location or site one the target recombined nucleic acid and is extended to form a first extension product, and wherein each primer of the first set has a 5'-non-complementary end containing a first primer binding site; (b) removing from the reaction mixture non-extended primers of the first set; (c) adding to the reaction mixture under primer extension conditions a second set of primers, wherein each primer of the second set has a receptor-specific portion such that the receptor-specific portion anneals to the first extension product at a predetermined location or site and has a 5'-non-complementary end containing a second primer binding site, primers of the first set and/or primers of the second set comprising a sequence tag disposed between the receptor-specific portion and the first or second primer binding site, respectively, and wherein each primer of the second set is extended to form a second extension product, such that each second extension product comprises a first primer binding site, a second primer binding site, at least one sequence tag, and either (i) a portion of a Vβ segment and a portion of a Jβ segment of a T cell receptor chain, (ii) a portion of a Vδ segment and a portion of a J δ segment of a T cell receptor chain, or (iii) a portion of a Vγ segment and a portion of a Jγ segment of a T cell receptor chain; (d) performing a polymerase chain reaction in the reaction mixture to form an amplicon, the polymerase chain reaction using forward primers specific for the first primer binding site and reverse primers specific for the second primer binding site; and (e) sequencing the nucleic acids of the amplicon to form a clonotype profile of multiple T cell receptor chains. As used herein, "primer extension conditions" in a reaction mixture includes conditions in which substantially all primer binding sites are in a single stranded state. In some embodiments, such conditions are obtained by melting double stranded target nucleic acids so that primer binding sites are in single stranded form so that primers can anneal to them to form substrates for polymerase extension.

[0035] The predetermined locations or sites at which primers of the first and second sets bind may be determined by conventional methods known to those of ordinary skill in the art of multiplex nucleic acid amplifications, such as multiplex PCRs, as exemplified in the references cited below. For example, in the case of target polynucleotides being recombined nucleic acids encoding immune receptor molecules, Faham and Willis (cited above), Van Dongen et al, Leukemia, 17: 2257-2317 (2003), and like references provide guidance for selecting primer binding sites for multiplex amplification of such target polynucleotides. In some embodiments, selecting such predetermined locations or sites depends of several factors including (i) their effect on amplification efficiency (it is desirable that frequencies of different copies in an amplicon faithfully represent frequencies of target polynucleotides in a sample), (ii) their effect on the lengths of copies in an amplicon correspond to requirements of the DNA sequencing chemistry being employed, (iii) whether the selected primers span a portion of the recombined nucleic acids with desired diversity, e.g. a VDJ region, and the like. In relation to this aspect, in part the invention includes an appreciation and recognition that primer cross-reactivity with different target polynucleotides does not effect results of methods of the invention (as compared to, for example, in methods based solely on analog readouts of PCR amplifications, spectratyping, and the like), because a set of sequences is the readout rather than an analog signal.

[0036] In some embodiments, the step of sequencing includes the following steps: (i) providing a plurality of sequence reads each having an error rate and each comprising a nucleotide sequence and a tag sequence, and (ii) aligning groups of sequence reads having like tag sequences, after which base calls are made based on sequence reads within the groups to determine the nucleotide sequence. Such group-level nucleotide sequences may then be coalesced into the same or different clonotypes as described below. In some embodiments, in the PCR steps, the lengths of the receptor-specific portions of the primers of the first and second sets are selected so that relative levels of different recombined nucleic acids in the amplicon are substantially the same as those of recombined nucleic acids in the sample. In implementing such selection of primers the positions and lengths of the binding sites of the primers on their respective target polynucleotides may be varied. In some embodiments, sequence tags are selected from a set of sequence tags which is much larger than the number of distinct target polynucleotides in a sample, so that substantially every distinct target polynucleotide in the sample and copies thereof will have a different sequence tag (for example, in accordance with the "labeling by sampling" methodology described in Brenner, U.S. patent 7,537,897). In some embodiments, the number of sequence tags in such a set is at least 100 times the size of the population of target polynucleotides in a sample. Further, in some embodiments where substantially every original target polynucleotide and copies thereof are labeled with the same unique sequence tag, the step of sequencing includes generating sequence reads of nucleic acids of the amplicon and aligning sequence reads having the same sequence tags to determine sequence reads corresponding to the same clonotypes of the sample. Further, in some embodiments, the step of aligning further includes determining a nucleotide sequence of each clonotype by determining a majority nucleotide at each nucleotide position of the sequence reads having the same sequence tag. Further, in some embodiments, steps of removing the non-extended primers may be carried out by digesting single stranded nucleic acids in the reaction mixture using a nuclease having 3'→5' single strand exonuclease activity (which may be provided by, for example, E. coli exonuclease I, which may be conveniently inactivated by heat). In further embodiments, the above methods may be used to generate clonotype profiles for diagnosing and/or monitoring minimal residual disease of a cancer patient, such as a myeloma, lymphoma or leukemia patient. Such diagnosing and/or monitoring may be implemented with the following additional step after the above method steps: determining from the clonotype profile a presence, absence and/or level of one or more patient-specific clonotypes correlated with the cancer. Methods of this embodiment may further include steps of determining sequences of each of one or more sequence tags and comparing such sequences with sequences of sequence tags of previously determined clonotype profiles to determine a presence, absence and/or level of contaminating sequences. In some embodiments, such step of comparing includes comparing the sequences of one of more sequence tags to sequence tags of a clonotype database containing clonotypes from at least one individual other than the patient.

[0037] In still another embodiment, a method of amplifying in one reaction a plurality of recombined nucleic acids encoding β, δ and γ T cell receptor components may comprise the steps of: (a) combining in a reaction mixture under primer extension conditions a first set of primers with a sample of recombined nucleic acids from T-cells, wherein each of the recombined nucleic acids comprises at a first end at least a portion of a Jβ, Jδ or Jγ segment of a T cell receptor, and wherein each primer of the first set each has a receptor-specific portion with a length, which receptor-specific portion anneals to the first end of a different recombined nucleic acid and is extended to form a first extension product, and wherein each primer of the first set has a 5'-non-complementary end containing in a 3'→5' ordering a sequence tag and a first primer binding site, the sequence tag being different for substantially every primer of the first set; (b) removing from the reaction mixture non-extended primers of the first set; (c) adding to the reaction mixture under primer extension conditions a second set of primers, each primer of the second set having a receptor-specific portion with a length, which anneals to the first extension product and is extended to form a second extension product, wherein each second extension product comprises at least a portion of a Vβ, Vδ or Vγ segment of a T cell receptor, and wherein each primer of the second set has a 5'-non-complementary end containing a second primer binding site; and (d) performing a polymerase chain reaction in the reaction mixture to form an amplicon, the polymerase chain reaction using a forward primer specific for the first primer binding site and a reverse primer specific for the second primer binding site. The above method may further include a step of sequencing a sample of sequences of the amplicon. Typically such a sample is a "representative sample" in that it is large enough to that different clonotypes are present in the sample in approximately the same frequencies as in the original sample of biological material. In some embodiments, the step of sequencing includes providing a plurality of sequence reads each having an error rate and each comprising a nucleotide sequence and a tag sequence, and aligning sequence reads having like tag sequences to determine sequence reads corresponding to the same clonotype. Such sequence reads may be processed in further step of coalescing, as described more fully below, whenever multiple sequence tags are attached to original target polynucleotides or copies thereof.

[0038] In another embodiment, a method of generating clonotype profiles from multiple T cell receptor chains may comprise the steps of: (a) combining in a reaction mixture under primer extension conditions a first set of primers with a sample of recombined nucleic acids from T-cells, wherein each primer of the first set has a receptor-specific portion with a length such that the receptor-specific portion anneals to a different recombined nucleic acid at a predetermined location and is extended to form a first extension product, and wherein each primer of the first set has a 5'-non-complementary end containing a first primer binding site; (b) removing from the reaction mixture non-extended primers of the first set; (c) adding to the reaction mixture a second set of primers, wherein each primer of the second set has a receptor-specific portion with a length, the receptor-specific portion being specific for the first extension product at a predetermined location and having a 5'-non-complementary end containing a second primer binding site, primers of the first set and/or primers of the second set comprising a sequence tag disposed between the receptor-specific portion and the first or second primer binding site, respectively; (d) performing a first polymerase chain reaction to form a first amplicon, the first polymerase chain reaction using forward primers specific for the first primer binding site and primers of the second set, wherein each nucleotide sequence of the first amplicon comprises a first primer binding site, a second primer binding site, at least one sequence tag, and either a portion of a Vβ segment and a portion of a Jβ segment of a T cell receptor chain, a portion of a Vδ segment and a portion of a J δ segment of a T cell receptor chain, or a portion of a Vγ segment and a portion of a Jγ segment of a T cell receptor chain, and wherein the lengths of the receptor-specific portions of the primers of the first and second sets are selected so that relative levels of different recombined nucleic acids in the amplicon are substantially the same as those of different recombined nucleic acids in the sample; (e) adding reverse primers specific for the second primer binding site; (f) performing a second polymerase chain reaction in the reaction mixture to form a second amplicon, the polymerase chain reaction using forward primers specific for the first primer binding site and reverse primers specific for the second primer binding site; (g) sequencing the nucleic acids of the second amplicon to form a clonotype profile of multiple T cell receptor chains. In some embodiments, the step of sequencing includes providing a plurality of sequence reads each having an error rate and each comprising a nucleotide sequence and a tag sequence, and aligning sequence reads having like tag sequences to determine sequence reads corresponding to the same clonotype. In further embodiments where target polynucleotides and/or copies thereof are labeled with more than one sequence tag, after aligning like sequence tags, sequence reads may be processed in a further step of coalescing, as described more fully below.

[0039] In another example, a method of generating clonotype profiles from multiple B cell receptor chains may be carried out by the steps of: (a) combining in a reaction mixture under primer extension conditions a first nested set of primers with a sample of recombined nucleic acids from B-cells, the first nested set comprising one or more groups of primers, wherein each primer of each group has a receptor-specific portion with a length such that the receptor-specific portion of each primer from a different group anneals to a different recombined nucleic acid at a predetermined site that does not overlap a predetermined site of any other primer of the first nested set, and wherein each primer of each group has a 5'-non-complementary end containing a first primer binding site; (b) extending primers of the first nested set to form a first extension product; (c) removing from the reaction mixture non-extended primers of the first nested set; (d) adding to the reaction mixture under primer extension conditions a second nested set of primers, the second nested set comprising one or more groups of primers, wherein each primer of each group has a receptor-specific portion with a length such that the receptor-specific portion of each primer from a different group anneals to the first extension product at a predetermined site that does not overlap a predetermined site of any other primer of the second nested set, and wherein each primer of each group has a 5'-non-complementary end containing a second primer binding site, and wherein primers of the first nested set and/or primers of the second nested set comprise a sequence tag disposed between its receptor-specific portion and its first or second primer binding site, respectively; (e) extending primers of the second nested set to form a second extension product, such that each second extension product comprising a first primer binding site, a second primer binding site, at least one sequence tag, and either (i) a portion of a V segment and a portion of a J segment of a B cell receptor heavy chain, or (ii) a portion of a V segment and a portion of a J segment of a B cell receptor kappa light chain; (f) performing a polymerase chain reaction in the reaction mixture to form an amplicon, the polymerase chain reaction using forward primers specific for the first primer binding site and reverse primers specific for the second primer binding site; and (g) sequencing the nucleic acids of the amplicon to form a clonotype profile of multiple B cell receptor chains.

[0040] In some embodiments, more than one cycles of annealing and extending primers (after melting the extension product) may be implemented in steps (b) and/or (e), in which case copies of the original recombined nucleic acids in the sample may be labeled with one or more sequence tags. In these embodiments, sequencing step (g) may include further steps of aligning and coalescing as described below for determining clonotypes and clonotype profiles. In some embodiments, for example, where only single extensions are made in steps (b) and (e), the step of sequencing includes providing a plurality of sequence reads each having an error rate and each comprising a nucleotide sequence and a tag sequence, and aligning sequence reads having like tag sequences to determine sequence reads corresponding to the same clonotype. As above, in some embodiments, in the PCR the positions and the lengths of the receptor-specific portions of the primers of the first and second sets are selected so that relative levels of different recombined nucleic acids in the amplicon are substantially the same as those of different recombined nucleic acids in the sample.

[0041] In some of the embodiments, sequence tags are attached to a target polynucleotide or a copy thereof in a step of primer extension, wherein substantially every different target polynucleotide and copy thereof is labeled with the same sequence tag. In other embodiments, target polynucleotides of a sample or copies thereof may be labeled with more than one different sequence tags. As explained further below, in some embodiments, multiple extensions or multiple cycles of a PCR may be carried out in the presence of sequence tag-containing primers (either a first set of primers or a second set of primers), which may result in different sequence tags being attached to the same target polynucleotide and/or its copies.

Sequence Tags in Clonotype Analysis

[0042] 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 target polynucleotide, for example, recombined nucleic acid, in a sample to assist in determining whether sequence reads of such conjugates are derived from the same original target polynucleotide. 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 and/or cell-free DNA. 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 miss-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.

[0043] 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 and/or cell-free DNA; (b) attaching sequence tags to molecules of recombined nucleic acids of T-cell receptor genes or immunoglobulin genes in the sample 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 recombined nucleic acid in the sample. Samples containing B-cells or T-cells are obtained using conventional techniques. 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.

[0044] 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 TCRβ 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.

[0045] 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>>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.

[0046] 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 "$N_1N_2 \ldots N_k$" 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

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

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:

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

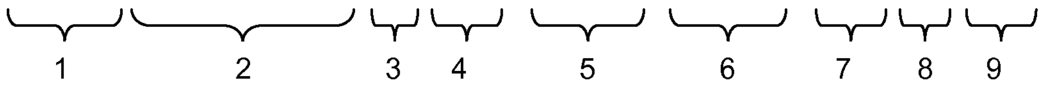
where i is selected as above and the occurrence of x' serves as a punctuation to terminate any undesired homopolymers.

Further Sequence Tags

[0047] 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: 1) 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
N  N  N  b  b  b  b  b  N  b  b  N  N  N  b  b  b  N  N  b  N
N
```



Region position

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, "***bbbbb*bb***bbb**b**", is selected to minimize the likelihood of having a perfect match in a genome of the organism making up the sample.

[0048] 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.

[0049] 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.

[0050] 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:

$$[(N_1N_2 \ldots N_{Kj})(b_1b_2 \ldots b_{Lj})]_M$$

wherein each N_i, for i= 1, 2, … 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 $N_1N_2 \ldots N_{Kj}$ are variable regions); each b_i, for i= 1, 2, … 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 Kj for every j and (ii) has the same sequences $b_1b_2 \ldots b_{Lj}$ for every j (that is, regions $b_1b_2 \ldots b_{Lj}$ 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.

Attaching Sequence Tags

[0051] A variety of different attachment reactions may be used to attach unique tags to substantially every clonotype in a sample in addition to those illustrated above. Many techniques for capturing subsets of sample nucleic acids, for example, to reduce sample complexity in microarray or genome sequencing technology, may be used with routine modification in the present invention to attached sequence tags to recombined nucleic acids. Exemplary techniques for capturing diverse sets of target nucleic acids for subsequent manipulation, including attaching sequence tags, sequencing, and the like, include the following; Willis et al, U.S. patent 7,700,323; Jones et al, U.S. patent publication 2005/0142577; Gullberg et al, U.S. patent publication 2005/0037356; Porreca et al, Nature Methods, 4(11): 931-936 (2007); Turner et al, Nature Methods, 6(5): 315-316 (2009); Church, U.S. patent 5,149,625; Macevicz, U.S. patent 8,137,936; and the like.

[0052] 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. For example, such random combining may occur in a bimolecular reaction wherein a tag-containing primer anneals to a target nucleic acid and is extended or wherein a tag-containing adaptor is ligated to the end of target nucleic acid. In some embodiments, the method of attaching tags may depend in part on the DNA sequencing approach. For example, in sequencing methods that produce relatively long accurate sequence reads, such as 454 sequencing, a cDNA library may be made from mRNA comprising recombined nucleic acids using conventional techniques, e.g. 5'-RACE, such as disclosed in Freeman et al, Genome Research, 19: 1817-1824 (2009), after which sequence tags may be attached by ligating sequence-tag-containing adaptors to one or both ends. In other embodiments, when sequencing methods, such as "Illumina" sequencing or "Ion Torrent" sequencing, are used that produce relatively short and error-prone sequence reads, further steps may be required so that amplicons for sequencing have lengths that are covered by sequence reads generated from the techniques. In such tag attachment reactions, clonotype sequences comprise linear single or double stranded polynucleotides and sequence tags are carried by reagents 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 Macevicz, 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.

[0053] Figs. 2A and 2B illustrate an attachment reaction comprising a PCR in which a population of sequence tags (T_1, T_2, T_3 … T_j, T_{j+1} … T_k, T_{k+1} … T_{n-1}, T_n) is incorporated into primers (2100). The population of sequence tags has a much greater size than that of recombined nucleic acid molecules (2102). 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 (2104), for example, in V region (2108). Since the primers (an 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 (2106) of the forward and reverse primer pair anneals to C region (2110) so that after multiple cycles of annealing, extending and melting, amplicon (2112) is formed, thereby attaching unique sequence tags to the V(D)J regions comprising the clonotypes of population (2102). That is, amplicon (2112) comprises the tag-molecule conjugates from the attachment reaction.

[0054] Figs. 2C and 2D illustrate a method for attaching a pair of sequence tags to each, or substantially each, recombined nucleic acid in a sample. As in the method of Figs. 2A and 2B, primers (2200) carrying sequence tags (T_1, T_2, T_3 … T_j, T_{j+1} … T_k, T_{k+1} … T_{n-1}, T_n) are used as downstream primers and additionally, replacing common primer (2106), primers (2206) carrying sequence tags (Tm, Tm+1, Tm+2 … Tq, Tq+1, Tq+2, … Tr, Tr+1, Tr+2, … Ts, Ts+1, Ts+2, …) are used as upstream primers. As with the downstream set of primers, the number of different sequence tags carried by upstream primers (2206) may be large compared to the number of recombined nucleic acid molecules (2202) so that substantially every recombined nucleic acid (2202) will have a unique tag after amplification. In some embodiments, each set of sequence tags in primers (2206) and (2200) need not be as large as the set of sequence tags in the embodiment of Figs. 2A and 2B. Since each recombined nucleic acid is uniquely labeled by a pair of sequence tags, sharing one sequence tag of the pair with a difference recombined nucleic acid will not detract from the substantial uniqueness to a pair of sequence tags labelling a single recombined nucleic acid. Thus, in the embodiment of Figs. 2C and 2D, sequence tags of each primer set (2200) and (2206) may be less diverse than the sequence tags of primer set (2100). For example, if random sequence tags are employed and primers (2100) contain 16-mer sequence tags, then primers (2200) and (2206) may each contain 8-mer sequence tags to provide the same total sequence tag diversity. Otherwise, the embodiment of Figs. 2C and 2D operates similarly to that of Figs. 2A and 2B. 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. As above, Fig. 2C depicts how the recombined nucleic acid molecules select, or sample, a small fraction of the total population of pairs of sequence tags by randomly annealing to the primers by way of their common primer binding regions (2204) and (2205), for example, in V region (2208) and C region (2210), respectively. Since the primers (an therefore sequence tags) combine with the recombined nucleic acid sequence molecules randomly, there is a small possibility that the same pair of sequence tags 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 pair of sequence tags attached. After multiple cycles of annealing, extending and melting, amplicon (2212) is formed, thereby attaching unique pairs of sequence tags to the V(D)J regions comprising the clonotypes of population (2202). That is, amplicon (2212) comprises the tag-molecule conjugates from the attachment reaction.

[0055] In some embodiments, circularizable probes may be used to capture and attach sequence tags to desired recombined nucleic acids, for example, with routine modification of techniques disclosed by Porreca et al (cited above); Willis et al (cited above); or like references. As illustrated in Figs. 2E and 2F, circularizable probe (2302) is provided comprising the following elements: upstream target binding segment (2304), downstream target binding segment (2306) that has 5’-phosphorylated end (2305); sequence tag (2310); second common primer binding site (2314); optional cleavage site (2308); and first common primer binding site (2312). Circularizable probe (2302) is combined in a reaction mixture under annealing conditions with a sample containing target polynucleotides (2300), which may be, for example, first or second strands of a cDNA prepared from mRNAs using conventional techniques. As shown, target polynucleotides comprise V, NDN, J and C regions of recombined nucleic acids encoding IgHs or TCRβ chains. In some embodiments, sequences of upstream and downstream target binding segments (2304) and (2306), respectively, are selected so that they span a portion of the VDJ region of the target polynucleotides. Circularizable probe (2302) and target polynucleotides (2300) form complex (2330) in the reaction mixture upon annealing of upstream and downstream target binding segments (2304 and 2306). In the presence of a DNA polymerase and dNTPs, upstream target binding segment (2304) is extended (2340) to downstream target binding segment (2306) copying (and thereby capturing) a portion of the VDJ region of the target polynucleotide. In the presence of a ligase activity, the extended upstream target binding segment is ligated to downstream target binding segment (2306), thereby forming a closed single stranded DNA circle (2342). The reaction mixture optionally may then be treated (2344) with an exonuclease to remove unreacted probe and target polynucleotides. In some embodiments, single stranded circles (2342) are linearized by cleaving cleavage site (2308), which may be, for example, a rare-cutting endonuclease recognition site, or inserting an RNA monomer in the probe and cleaving with RNase H, or the like, after which VDJ-tag inserts of the linearized probes (2348) may be amplified by primers (2350) and (2352). Primers (2350) and (2352) may include noncomplementary regions for adding elements to permit later DNA sequencing (2354). Alternatively, single stranded circle may be used to generate nanoball templates for direct sequencing, e.g. Drmanac et al, Science, 327(5961): 78-81 (2010); U.S. patent 8,445,196; and the like.

[0056] Fig. 2G illustrates another embodiment for attaching a sequence tag to a recombined nucleic acid encoding an immune receptor molecule. Guidance for implementing this embodiment may be found in Faham and Zheng, U.S. patent 7,208,295, which is incorporated herein by reference. Recombined nucleic acids (2450) are combined in a reaction mixture under annealing conditions for probes (2454) and adaptors (2456). Probes (2454) comprise receptor-specific portion (2455) and adaptor-specific portion (2457). For example, probes (2454) may comprise a mixture of probes wherein different probes have receptor-specific portions specific for different J regions, or in other embodiments, specific for different V regions. Adaptors (2456), which are 5'-phosphorylated, comprise probe-specific portion (2458) at its 5' end, sequence tag (2460) and first primer binding site (2462). The locations, sequences, and lengths of receptor-specific portion (2455) and adaptor-specific portion (2457) of probe (2454) and probe-specific portion (2458) are selected so that they hybridize with one another to form structures (2452). After structure (2452) forms, single stranded portion (2461) is cleaved from recombined nucleic acid (2450) and the free 3' end of recombined nucleic acid (2450) is ligated to the 5' phosphorylated end of adaptor (2456) to form first extension product (2459), after which probe (2454) is removed (2474). Cleavage of (2461) may be effected by a single stranded nuclease as described in Faham and Zheng. In one embodiment, probe (2454) is synthesized with thymidines replaced with uracils, e.g. in a PCR with dUTPs in place of dTTps, and it is removed by treating with uracil-DNA glycosylase (UDG), e.g. as taught by Faham et al, U.S. patent 7,208,295, which is incorporated by reference. UDG treatment cleaves probe (2454) at uracils to give fragments (2455). After free probe, adaptors and flaps are removed (2476), forward primers (2466) and reverse primers (2468) are added to extension product (2464) and PCR (2470) is carried out, after which a sample of the resulting amplicon is sequenced (2472).

[0057] In a similar embodiment to that of Fig. 2G, similar probes and adaptors may be used to attached sequence tags at predetermined sites of a target polynucleotide, wherein a flap endonuclease, such as FEN-1, is used to cleave a single stranded portion corresponding to (2461). In this embodiment besides using a different nuclease, the polarity of the probe and adaptor sequences are reversed; namely, a substrate for a flap endonuclease requires that the 3' end of the adaptor corresponding to (2454) be annealed to target sequence (2450) and that the single stranded portion corresponding to (2452) be a 5' end of the target sequence. After cleavage and removal of the probe sequence, the remaining steps are substantially the same. Guidance for using flap endonucleases in detection assays may be found in the following references: Lyamichev et al, Nature Biotechnology, 17: 292-296 (1999); Eis et al, Nature Biotechnology, 19: 673-676 (2001); and like references.

[0058] In some embodiments, recombined nucleic acids encode immune receptor molecule chains that typically form an immune repertoire which may comprise a very large set of very similar polynucleotides (e.g. >1000, but more than 10,000, and still more usually from 100,000 to 1,000,000, or more) which may have a length of less than 500 nucleotides, or in other embodiments, less than 400 nucleotides, or in still other embodiments, less than 300 nucleotides. 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.

Samples

[0059] The term “sample” refers to a quantity of biological material, which in some embodiments is obtained from a patient, and which contains cells and/or cell-free DNA; that is, the term is used interchangeably with the term “specimen,” or “tissue sample.” The term “sample” also sometime used in a statistical sense of obtaining a subset, or portion, of a larger set or quantity, respectively, of, for example, recombined nucleic acids; in particular, the statistical usage of the term “sample” may also be understood to mean “representative sample,” in that such a sample is understood to reflect, or approximate, the relative frequencies of different nucleic acids in a tissue (for example). One skilled in the art is able to distinguish the proper usage from the context of the terms.

[0060] Clonotype profiles may be obtained from samples of immune cells or fluids, such as blood, containing cell-free nucleic acids encoding immune receptor chains. For example, immune cells can include T-cells and/or B-cells. T-cells (T lymphocytes) include, for example, cells that express T cell receptors. T-cells include helper T cells (effector T cells or Th cells), cytotoxic T cells (CTLs), memory T cells, and regulatory T cells. In one aspect a sample of T cells includes at least 1,000T 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 (antibodies, B cell receptor). 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.

[0061] 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.

[0062] In some embodiments, clonotype profiles for methods of the invention are generated from a tumor or peripheral blood in the case of diagnostic samples or from peripheral blood in the case of samples for monitoring residual disease. One or more clonotypes correlated with a disease, such as a lymphoid or myeloid proliferative disorder, are determined from a diagnostic sample. Usually the one or more clonotypes correlated with a lymphoid or myeloid proliferative disorder are those present in a clonotype profile with the highest frequencies. In some cases, there may be a single correlated clonotype and in other cases there may be multiple clonotypes correlated with a lymphoid or myeloid proliferative disorder. Tumor samples may be taken from any tissue affected by such a disorder, which includes lymph nodes or other tissues outside of the lymphatic system. As mentioned above, clonotype profiles for monitoring residual disease may be generated from a sample of nucleic acids extracted from peripheral blood. The nucleic acids of the sample may from B-cells from a cell-containing fraction of the peripheral blood or from a cell free fraction of the peripheral blood, such as plasma or serum. In one embodiment, a peripheral blood sample includes at least 1,000 B cells; but more typically, such 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. In some embodiments, the number of cells in a sample sets a limit on the sensitivity of a measurement. That is, greater sensitivity of detecting a residual disease is achieved by using a larger sample of peripheral blood. For example, in a sample containing 1,000 B cells, the lowest frequency of clonotype detectable is 1/1000 or .001, regardless of how many sequencing reads are obtained when the DNA of such cells is analyzed by sequencing. The nucleic acids of the sample may from T-cells from a cell-containing fraction of the peripheral blood or from a cell free fraction of the peripheral blood, such as plasma or serum. In one embodiment, a peripheral blood sample includes at least 1,000 T cells; but more typically, such 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 T cells. In some embodiments, the number of cells in a sample sets a limit on the sensitivity of a measurement. That is, greater sensitivity of detecting a residual disease is achieved by using a larger sample of peripheral blood. For example, in a sample containing 1,000 T cells, the lowest frequency of clonotype detectable is 1/1000 or .001, regardless of how many sequencing reads are obtained when the DNA of such cells is analyzed by sequencing.

[0063] A sample for use with the invention can include 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 provided invention, the amount of RNA or DNA from a subject that can be analyzed includes, for example, as low as a single cell in some applications (e.g., a calibration test with other cell selection criteria, e.g. morphological criteria) and as many as 10 million of cells or more, which translates into a quantity of DNA in the range of from 6pg-60ug, and a quantity of RNA in the range of from 1pg-10ug. In some embodiments, a nucleic acid sample is a DNA sample of from 6 pg to 60 ug. In other embodiments, a nucleic acid sample is a DNA sample from 100μL to 10 mL of peripheral blood; in other embodiments, a nucleic acid sample is a DNA sample from a cell free fraction of from 100μL to 10 mL of peripheral blood.

[0064] In some embodiments, a sample of lymphocytes or cell free nucleic acid is sufficiently large so that substantially every B cell or T cell with a distinct clonotype is represented therein, thereby forming a “repertoire” of clonotypes. In one embodiment, to achieve substantial representation of every distinct clonotype, 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. And 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 .00001 percent or greater. In one embodiment, a sample of B cells or T cells includes at least one half million cells, and in another embodiment such sample includes at least one million cells.

[0065] Nucleic acid samples may be obtained from peripheral blood 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). Blood samples may range in volume from 100 μL to 10 mL; in one aspect, blood sample volumes are in the range of from 100 μL 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. For example, memory B cells may be isolated by way of surface markers CD19 and CD27.

[0066] Cell-free DNA may also be extracted from peripheral blood samples using conventional techniques, e.g. Lo et al, U.S. patent 6,258,540; Huang et al, Methods Mol. Biol., 444: 203-208 (2008); and the like, which are incorporated herein by reference. By way of nonlimiting example, peripheral blood may be collected in EDTA tubes, after which it may be fractionated into plasma, white blood cell, and red blood cell components by centrifugation. DNA from the cell free plasma fraction (e.g. from 0.5 to 2.0 mL) may be extracted using a QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA), or like kit, in accordance with the manufacturer's protocol.

[0067] In one aspect, a sample of lymphocytes for generating a clonotype profile is sufficiently large that substantially every T cell or B cell with a distinct clonotype is represented therein. 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 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 .00001 percent or greater. In other embodiments, a sample is taken that contains with a probability of ninety-five 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-five percent every clonotype of a population present at a frequency of .0001 percent or greater. In another embodiment, a sample is taken that contains with a probability of ninety-five percent every clonotype of a population present at a frequency of .00001 percent or greater. In still another 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.

[0068] 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.

[0069] 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). Blood samples may range in volume from 100 μL to 10 mL; in one aspect, blood sample volumes are in the range of from 100 μL 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.

[0070] Since the identifying recombinations are present in the DNA of each individual's adaptive immunity cells as well as their associated RNA transcripts, 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 chain of 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 TCRα chain is generated by VJ recombination, and the β chain receptor is generated by V(D)J recombination. For the TCRβ 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).

[0071] 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.

[0072] As mentioned above, in accordance with the invention, primers may be selected to generate amplicons containing portions of recombined nucleic acids from lymphocytes or from cell-free nucleic acids from a tissue, such as blood. Such portions may be referred to herein as "somatically rearranged regions." Somatically 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 (for example) full V(D)J regions. Exemplary incomplete somatically rearranged regions include incomplete IgH molecules (such as, molecules containing only D-J regions), incomplete TCRδ molecules (such as, molecules containing only D-J regions), and inactive IgK (for example, comprising Kde-V regions).

Amplification of Nucleic Acid Populations

[0073] In some embodiments, primer sequences of the first and second sets of primers may be selected in accordance with conventional multiplex polymerase chain reactions (PCRs).. For example, guidance for selecting primers and for carrying out multiplex PCRs of nucleic acids encoding various immune receptor chains is found in the following references, which are incorporated by reference: Faham and Willis, U.S. patents 8,236,503 and 8,628,927; Morley, U.S. patent 5,296,351; Gorski, U.S. patent 5,837,447; Dau, U.S. patent 6,087,096; Van Dongen et al, U.S. patent publication 2006/0234234; European patent publication EP 1544308B1; Van Dongen et al, Leukemia, 17: 2257-2317 (2003); and the like. Guidance for multiplex PCRs may be found in Henegariu et al, BioTechniques, 23: 504-511 (1997), and like references. In some embodiments, primers are selected so that frequencies of amplified sequences in a final product are substantially the same as frequencies of the sequences in the starting reaction mixture. Such primer selection may include selection of primer lengths, primer binding sites and primer concentrations. As noted above, depending on the methods selected to generate sequence reads and attached sequence tags, the level of multiplexing may vary widely.

[0074] In some embodiments, a step of amplifying target nucleic acids includes linear amplification of target nucleic acids, such as, for example, by repeated cycles of annealing one set of primers (for example, a first set of "upstream" or "forward" primers), extending the primers, melting the extended strand from the template, so that the quantity of extended strands is amplified as a linear function of the number of cycles. In other words, a step of amplifying include copying a target polynucleotide (that is, at least one strand of a target polynucleotide) by repeated extensions of one set of primers. In some embodiments, such single or repeated extensions in one direction may be followed by steps of removing unextended primers and a single or repeated extensions of another set of primers in the other direction (for example, a second set of "downstream" or "reverse" primers).

[0075] The number of primer in a first set of primers and a second set of primers may vary widely depending on the number and type of immune receptor-chain nucleic acids are amplified in an assay. In some embodiments, consensus primers for various chains may be used. In other embodiments, specific primers may be designed for each different target polynucleotide amplified. Usually, both first set and second set of primers each comprise a plurality of primers. In some embodiments, the plurality of primers in the first set or the second set of primers is at least 50 primers; in other embodiments, the plurality of primers in the first set or the second set of primers is at least 100 primers; in other embodiments, the plurality of primers in the first set or the second set of primers is at least 150 primers; in other embodiments, the plurality of primers in the first set or the second set of primers is at least 200 primers; in other embodiments, the plurality of primers in the first set or the second set of primers is at least 250 primers. The number of primers in the first set may be the same or different than the number of primers in the second set.

[0076] In some embodiments, primers of the first set and the second set are selected so that the length of clonotypes are at least 30 nucleotides; in other embodiments, primers of the first set and the second set are selected so that the length of clonotypes are in the range of from 30 to 500 nucleotides; in other embodiments, primers of the first set and the second set are selected so that the length of clonotypes are in the range of from 30 to 400 nucleotides; in other embodiments, primers of the first set and the second set are selected so that the length of clonotypes are in the range of from 30 to 300 nucleotides; in other embodiments, primers of the first set and the second set are selected so that the length of clonotypes are in the range of from 30 to 200 nucleotides.

[0077] Exemplary PCR 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 μL final reaction volume; 100 ng sample DNA; 10 pmol of each primer (subject to adjustments to balance amplification as described below); dNTPs at 200 μM 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.

[0078] 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 in a sample.

In one aspect, multiplex amplifications of the invention 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.

[0079] In some embodiments, amplification bias may be avoided by carrying out a two-stage amplification (for example, as described in Faham and Willis, cited above) wherein a small number of amplification cycles (for example, 2-5, or 2-10, or 2-15 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. Prior to initiation of the second stage amplification, non-extended primers of the first stage are removed from the reaction mixture, or are otherwise inactivated. In some embodiments, the primary PCR will have a small enough number of cycles (e.g. 2-10) to minimize the differential amplification by the different primers. The secondary amplification is then done with one pair of primers, which eliminates a source of differential amplification. In some embodiments, a small percent or portion, e.g. one percent of the reaction volume, of the primary PCR is taken directly to the secondary PCR reaction mixture. In some embodiments, a total of at least thirty-five cycles allocated between a first stage amplification and a second stage amplification

[0080] In some embodiments internal standards may be combined with and amplified in the same reaction as recombined nucleic acids of a sample. Internal standard are nucleic acids with known sequences and known concentrations. For example, they may be cloned copies of a natural nucleic acid encoding portions of an immune receptor chain, or they may be synthetic nucleic acids. In some embodiments, the lengths and base compositions of the internal standards are selected to representative of the particular immune receptor chains being amplified. By monitoring changes in the relative concentrations of the internal standards after amplification, amplification bias may be detected, and conditions for non-biased amplification may be determined. For example, primer lengths, positions, and concentrations may be varied to minimize bias in the amplification product. In some embodiments, a plurality of internal standards are used in a reaction; in some embodiments, 2 to 50 different internal standards are used in a reaction; in other embodiments, from 2 to 25 different internal standards are used in a reaction; and in some embodiments, from 2 to 10 different internal standards are used in a reaction. In some embodiments, amplification bias is determined by measuring the relative frequencies of the sequences of different target nucleotides (for example, all or selected clonotypes or internal standards) in an amplification product. In other embodiments. the presence, absence or level of amplification bias may be determined by real-time quantitative PCR of selected nucleic acids, such as two or more the internal standards. Internal standards may also be used to quantify the numbers of different clonotypes in the original sample. Techniques for such molecular counting are well-known, e.g. Brenner et al, U.S. patent 7,537,897, which is incorporated herein by reference.

Generating Sequence Reads

[0081] 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.

[0082] Of particular interest are approaches using sequencing by synthesis with reversible terminators, such as 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, such as described in Bentley et al (cited above) and in manufacturer's instructions (e.g. TruSeq^{TM} 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 $5x10^5$ per cm^2; or in a density of at least 10^6 clusters per cm^2. Solexa-based sequencing also provides the capability of generating two sequence reads from the same target sequence (or template) in a cluster, one sequence read each from opposite ends of a target sequence. In some embodiments, such pairs of sequence reads may be combined and treated as a single sequence read in subsequent analysis, or such pairs may be treated separately but taking into account that they originate from the same cluster. Sometimes the pair of sequence reads from the same template are referred to as "mate pairs," and the process of sequencing from both ends of a temple is referred to "bidirectional" sequencing. In some embodiments, a step of sequencing by synthesis using reversibly terminated labeled nucleotides includes the generation of a single sequence read for each cluster or clonal population of templates and the generation of a plurality of sequence reads (including but not limited to mate pairs) for each cluster or clonal population of templates. In still further embodiments, when a plurality of sequence reads are generated for each cluster or clonal population of templates, such plurality of sequence reads may be combined to form a single effective sequence read that is used in subsequent analysis, such as a coalescing step.

[0083] In one aspect, a sequence-based clonotype profile of a sample from 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 one embodiment, each of the somatically rearranged regions comprise a V region and a J region. In another embodiment, the step of sequencing includes generating a plurality of sequence reads for each clonotype determined. In still other embodiments, the step of sequencing includes combining information or data from a plurality of sequence reads to form each clonotype. In some embodiments, such step of combining may be carried out by coalescing sequence reads as described in Faham and Willis, U.S. patent 8,628,927 (which is hereby incorporated by reference for this teaching) or by using sequence tags as described in Faham et al, U.S. patent publication 2013/0236895A1 (which is hereby incorporated by reference for this teaching). 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.

[0084] Further to the latter embodiment, at least one of the forward sequence reads and at least one of the reverse sequence reads have an overlap region such that bases of such overlap region are determined by a reverse complementary relationship between such sequence reads. In still another embodiment, each of the somatically rearranged regions comprise a V region and a J region and the step of sequencing further includes determining a sequence of each of the individual nucleic acid molecules from one or more of its forward sequence reads and at least one reverse sequence read starting from a position in a J region and extending in the direction of its associated V region. In another embodiment, individual molecules comprise nucleic acids selected from the group consisting of complete IgH molecules, incomplete IgH molecules, complete IgK complete, IgK inactive molecules, TCRβ molecules, TCRγ molecules, complete TCRδ molecules, and incomplete TCRδ molecules. In another embodiment, the step of sequencing comprises generating the sequence reads having monotonically decreasing quality scores. 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.

[0085] 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. In each of the foregoing, each clonotype per run is determined from at least 10 sequence reads.

[0086] 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.

Clonotype Determination from Sequence Data

[0087] In some embodiments of the invention, sequence tags are used to determine clonotypes and in other embodiments, sequence tags in combination with a sequence read coalescing step are used to determine clonotypes. In embodiments in which a single unique sequence tag is attached to substantially every distinct target polynucleotide, clonotype determination using sequence tags is straight forward. In such embodiments, 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. 4A illustrates an exemplary alignment and method from determining the sequence (SEQ ID NO: 2) of a clonotype associated with a unique sequence tag. In this example, eleven sequence reads are aligned by way of their respective sequence tags (4302) after which nucleotides at each position of the clonotype portions (4304) of the sequence reads, indicated as 1, 2, 3, 4, … n, are compared. For example, nucleotides at position 6 (4306) are t, t, g, t, t, t, t, t, t, c, t; that is, nine base calls are t's, one is "g" (4308) and one is "c" (4310) (SEQ ID NO: 3 and SEQ ID NO: 4). 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 (4306), 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. 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.

[0088] In some embodiments, more than one extension step may be carried out using sequence-tag containing primers in order to increase the fraction of target polynucleotides in a sample that are labeled with sequence tags prior to amplification. In such embodiments, the more than one extension steps in the presence of sequence tag-containing primers results in a target polynucleotide and /or its copies being labeled with a plurality of different sequence tags. The size of the plurality depends on the number of extension steps carried out in the presence of the sequence tag-containing primers, the efficiency of the amplification reaction, whether only one or both the forward and reverse primers have sequence tags, and the like. In some such embodiments, the plurality is in the range of from 2 to 15, or in the range of from 2 to 10, or in the range of from 2 to 5. In some such embodiments, after amplification, copies of each target polynucleotide of a sample may be divided into a plurality of groups or subsets wherein members of each group or subset is labeled with the same sequence tag and members of each different group or subset is labeled with a different sequence tag; that is, members of the same group have the same sequence tag and members of different groups have different sequence tags. In other words, after amplification, in some embodiments, each copy of a target polynucleotide from a sample will be labeled with one of two different sequence tags; or in other embodiments, each copy of a target polynucleotide from a sample will be labeled with one of three different sequence tags; or in other embodiments, copies of a target polynucleotide from a sample will be labeled with one of four different sequence tags; and so on. In these embodiments, clonotypes may be determined by a combination of sequence tag alignment followed by coalescing steps for treating sequence reads within a group as originating from the same parent sequence based on a likelihood that common origin is true as a function of error rates, relative frequencies, and the like. Fig. 4B illustrates sequence reads from such an embodiment. In one approach, sequence reads are first grouped by common sequence tags (4402), which in the illustration results in three groups (4420), (4422) and (4424). In some embodiments, within each group, sequences (4404) are analyzed to determine a consensus sequence of the group; for example, as above, at each nucleotide position a base may be called as the majority base, or the highest frequency base, or the like. The group consensus sequences may then be coalesced with one another to determine clonotypes.

[0089] In some embodiments, the above aspect of the invention may be implemented in a method for profiling of virtually any population of nucleic acids in a sample. Such method may comprise the steps: (a) obtaining a sample comprising a population of nucleic acids; (b) attaching sequence tags to nucleic acids of the population to form tag-nucleic acid conjugates, wherein at least one nucleic acid of the population or copies thereof have different sequence tags attached; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing the tag-nucleic acid conjugates to generate sequence reads having error rates and comprising a nucleic acid sequence and a tag sequence; (e) aligning sequence reads having like tag sequences to form groups of sequence reads having the same sequence tags; (f) coalescing sequence reads of groups to determine sequences of the nucleic acids, wherein groups of sequence reads are coalesced into different sequences whenever said groups of sequence reads are distinct with a likelihood of at least ninety-five percent; and (g) determining the sequence profile of the population by determining the levels of the sequences. As applied to profiling a population of recombined nucleic acids, such method may be implemented by the steps: (a) obtaining a sample from an individual comprising T-cells and/or B-cells and/or cell-free DNA; (b) attaching sequence tags to recombined nucleic acid molecules of T-cell receptor genes or immunoglobulin genes from the sample to form tag-nucleic acid conjugates, wherein at least one recombined nucleic acid from the sample or copies thereof have different sequence tags attached; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing the tag-nucleic acid conjugates to provide sequence reads having error rates and comprising a tag sequence and a recombined nucleic acid sequence; (e) aligning sequence reads having like tag sequences to form groups of sequence reads having the same sequence tags; (f) coalescing sequence reads of groups to determine clonotypes, wherein groups of sequence reads are coalesced into different sequences whenever said groups of sequence reads are distinct with a likelihood of at least ninety-five percent; and (g) determining the clonotype profile of the sample by determining levels of the clonotypes.

[0090] In the above embodiments, and other embodiments disclosed herein, the step of sequencing tag-nucleic acid conjugates comprises sequencing a sample of tag-nucleic acid conjugates from an amplicon. Usually, such sample is a representative sample in that the relative frequencies of the target polynucleotides in the original sample (that is, the tissue sample, blood sample, or the like) are maintained in the sample of tag-nucleic acid conjugates from the product of an amplification reaction. In some embodiments in which population of recombined nucleic acids encoding immune receptor molecules are analyzed, a sample of tag-nucleic acid conjugates comprises at least 10^4 tag-nucleic acid conjugates; in other embodiments, such sample comprises at least 10^5 tag-nucleic acid conjugates; ; in other embodiments, such sample comprises at least 10^6 tag-nucleic acid conjugates; ; in other embodiments, such sample comprises at least 10^7 tag-nucleic acid conjugates.

Coalescing Sequence Reads

[0091] In embodiments where multiple sequence tags are attached to an original recombined nucleic acid or copies thereof, a step of coalescing sequence reads (or consensus sequence reads from groups) may be carried out for determining clonotypes. Reducing a set of sequence reads for a given sample to a set of distinct clonotypes and recording the number of reads for each clonotype would be a trivial if sequencing technology was error free. However, in the presence of sequencing errors, each genuine clonotype is surrounded by a 'cloud' of sequence reads with varying numbers of errors with respect to the its sequence. The "cloud" of sequencing errors drops off in density as the distance increases from the clonotype in sequence space. A variety of algorithms are available for converting sequence reads into clonotypes. In one aspect, coalescing of sequence reads (that is, merging candidate clonotypes determined to have one or more sequencing errors) depends on at least three factors: the number of sequences obtained for each of the clonotypes being compared; the number of bases at which they differ; and the sequencing quality score at the positions at which they are discordant. In some embodiments, a likelihood ratio may be constructed and assessed that is based on the expected error rates and a binomial distribution of errors. For example, two clonotypes, one with 150 reads and the other with 2 reads with one difference between them in an area of poor sequencing quality will likely be coalesced as they are likely to be generated by sequencing error. On the other hand two clonotypes, one with 100 reads and the other with 50 reads with two differences between them are not coalesced as they are considered to be unlikely to be generated by sequencing error. In some embodiments, the algorithm described below may be used for determining clonotypes from sequence reads. Some of these concepts are illustrated in Fig. 5A. In some embodiments of a coalescing step, sequence reads are first converted into candidate clonotypes. Such a conversion depends on the sequencing platform employed. For platforms that generate high Q score long sequence reads, the sequence read or a portion thereof may be taken directly as a candidate clonotype. For platforms that generate lower Q score shorter sequence reads, some alignment and assembly steps may be required for converting a set of related sequence reads into a candidate clonotype. For example, for Solexa-based platforms, in some embodiments, candidate clonotypes are generated from collections of paired reads from multiple clusters, e.g. 10 or more, as mentioned above.

[0092] The frequencies of candidate clonotypes may be plotted in sequence space, as illustrated in Fig. 5A, where such space is reduced to one dimension (the horizontal axis) for sake of illustration. The vertical axis gives the magnitude of each candidate clonotype's frequency, log(read count), or some like measure. In the figure, candidate clonotypes are represented by the various symbols (530). In accordance with one embodiment of the invention, whether two candidate clonotypes are coalesced depends on their respective frequencies or read counts (as noted above), the number of base differences between them (the more differences, the less likely is coalescence), and the quality scores of the bases at the locations where the respective sequences differ (higher quality scores makes coalescence less likely). Candidate clonotypes may be considered in the order of their respective frequencies. Fig. 5A shows candidate clonotype 1 (532), candidate clonotype 7 (534) and candidate clonotype 11 (536) as the three candidates with the highest three frequencies. Related to each such candidate clonotype are other candidate clonotypes that are close in sequence, but with lesser frequencies, such as (i) for candidate clonotype 1 (532) there are candidate clonotype 2 (538) and the candidate clonotypes 3, 4, 5 and 6 enclosed by cone (540); for candidate clonotype 7 (534) there are candidate clonotypes 8, 9 and 10 enclosed by cone (542); and (iii) for candidate clonotype 11, there is candidate clonotype 12 enclosed by cone (544). The cones represent likelihood boundaries within which a lesser frequency candidate clonotype would be coalesced with one of the higher frequency candidate clonotypes 1, 7 or 11. Such likelihood boundaries are functions of the frequency of the nearby candidate clonotypes (3, 4, 5 and 6 for 1; 8, 9 and 10 for 7; and 12 for 11) and their distances in sequence space from the respective higher frequency candidate clonotypes. Candidate clonotype 2 (538) is outside cone (540); thus, it would not be coalesced with candidate clonotype 1 (532). Again, the likelihood (of coalesce) boundaries are shown as cones because candidate clones with higher frequencies are more likely to be genuinely different clonotypes than those of lower frequencies and multiple differences at lower frequencies are more likely to be errors than multiple differences at higher frequencies.

[0093] The cloud of sequence reads surrounding each candidate clonotype can be modeled using the binomial distribution and a simple model for the probability of a single base error. This latter error model can be inferred from mapping V and J segments or from the clonotype finding algorithm itself, via self-consistency and convergence. A model is constructed for the probability of a given 'cloud' sequence Y with read count C2 and E errors (with respect to sequence X) being part of a true clonotype sequence X with perfect read count C1 under the null model that X is the only true clonotype in this region of sequence space. A decision is made whether or not to coalesce sequence Y into the clonotype X according the parameters C1, C2, and E. For any given C1 and E a max value C2 is pre-calculated for deciding to coalesce the sequence Y. The max values for C2 are chosen so that the probability of failing to coalesce Y under the null hypothesis that Y is part of clonotype X is less than some value P after integrating over all possible sequences Y with error E in the neighborhood of sequence X. The value P is controls the behavior of the algorithm and makes the coalescing more or less permissive.

[0094] If a sequence Y is not coalesced into clonotype X because its read count is above the threshold C2 for coalescing into clonotype X then it becomes a candidate for seeding separate clonotypes (such as with candidate clonotype 2 (538) in Fig. 5A). An algorithm implementing such principles would also make sure that any other sequences Y2, Y3, etc. which are 'nearer' to this sequence Y (that had been deemed independent of X) are not aggregated into X. This concept of 'nearness' includes both error counts with respect to Y and X and the absolute read count of X and Y, i.e. it is modeled in the same fashion as the above model for the cloud of error sequences around clonotype X. In this way 'cloud' sequences can be properly attributed to their correct clonotype if they happen to be 'near' more than one clonotype. Thus, going to Fig. 5A, if candidate clonotype 2 is deemed to be genuinely distinct from candidate clonotype 1 (532), then a special routine, or subalgorithm, would provide a rule for determining which of candidate clonotypes 1 (532) and 2 (538), candidates 4 and 5, between 1 and 2, should be coalesced to (if either).

[0095] In one embodiment, an algorithm proceeds in a top down fashion by starting with the sequence X with the highest read count. This sequence seeds the first clonotype. Neighboring sequences are either coalesced into this clonotype if their counts are below the precalculated thresholds (see above), or left alone if they are above the threshold or 'closer' to another sequence that was not coalesced. After searching all neighboring sequences within a maximum error count, the process of coalescing reads into clonotype X is finished. Its reads and all reads that have been coalesced into it are accounted for and removed from the list of reads available for making other clonotypes. The next sequence is then moved on to with the highest read count. Neighboring reads are coalesced into this clonotype as above and this process is continued until there are no more sequences with read counts above a given threshold, e.g. until all sequences with more than 1 count have been used as seeds for clonotypes.

[0096] As mentioned above, in another embodiment of the above algorithm, a further test may be added for determining whether to coalesce a candidate sequence Y into an existing clonotype X, which takes into account quality score of the relevant sequence reads. The average quality score(s) are determined for sequence(s) Y (averaged across all reads with sequence Y) were sequences Y and X differ. If the average score is above a predetermined value then it is more likely that the difference indicates a truly different clonotype that should not be coalesced and if the average score is below such predetermined value then it is more likely that sequence Y is caused by sequencing errors and therefore should be coalesced into X.

[0097] Successful implementation of the above algorithm for coalescing candidate clonotypes is dependent upon having an efficient way of finding all sequences with less than E errors (i.e. less than some sequence distance measure) from some input sequence X. This problem may be solved using a sequence tree. The implementation of such trees has some unusual features in that the nodes of the tree are not restricted to being single letters of the DNA sequences of the candidate clonotypes, as illustrated in Fig. 5D. The nodes can have arbitrarily long sequences, which allows for a more efficient use of computer memory.

[0098] All of the reads of a given sample are placed into the sequence tree. Each leaf nodes holds pointers to its associated reads. A unique sequence of a candidate clonotype is retrieved by traversing backwards in the tree from the leaf to the root node. The first sequence is placed into a simple tree with one root node and one leaf node that contains the full sequence of the read. Sequences are next added one by one. For each added sequence either a new branch is formed at the last point of common sequence between the read and the existing tree or add the read to an existing leaf node if the tree already contains the sequence. Having placed all the reads into the tree it is easy to use the tree for the following purposes: 1) Finding the highest read count: sorting leaf nodes by read count allows one to find the leaf node (i.e. sequence) with the most reads, and successively lower numbers of reads; 2) Finding neighboring leafs: for any sequence all paths through the tree which have less than X errors with respect to this sequence are searchable. A path is started at the root and branch this path into separate paths proceeding along the tree. The current error count of each path as proceeding along the tree is noted. When the error count exceeds the max allowed errors the given path is terminated. In this way large parts of the tree are pruned as early as possible. This is an efficient way of finding all paths (i.e. all leafs) within X errors from any given sequence.

[0099] Features of the above concepts are illustrated in more detail in the flow chart of Fig. 5B. A set of candidate clonotypes is obtained from sequence data obtained by sequencing recombined nucleic acids extracted from a sample of T cells or B cells. In one aspect, candidate clonotypes each include an NDN region and portions of V and J regions. These sequences are organized into a data structure (550), which may be a sequence tree. Not shown in Fig. 5B, as part of generating a set of candidate clonotypes, in one embodiment, sequence trees may also be constructed for known V regions and known J regions. Sequence reads making up a candidate clonotype may then be mapped, or aligned, to these known sequences via the sequence trees to efficiently determine the most likely known V and J sequences of the candidate clonotypes. Returning to Fig. 5B, once the candidate clonotypes are generated, a data structure, such as a sequence tree, is constructed for use in a method for distinguishing genuine clonotypes from candidate clonotypes that contain experimental or measurement errors, such as sequencing errors. The candidate clonotype that has the highest frequency of occurrence among the current candidate clonotypes ($HFCC_k$) is selected (552) from the data structure, for example a sequence tree; in other words, $HFCC_k$ is the candidate clonotype with the highest number of copies, or read counts in cycle k. Next, neighboring lesser frequency candidate clonotypes are identified (LFCCs) (554); that is, candidate clonotypes within a distance of D_k are identified. In one aspect of the invention, this identification is carried out using a sequence tree, which allows efficient sequence comparisons of relatively short (<300 bp) sequences.

[00100] In one embodiment, the comparisons, or sequence alignments, are carried out using dynamic programming, e.g. as disclosed by Gusfield (cited above). In a further embodiment, such dynamic programming is banded dynamic programming where sequences that differ from the selected HFCC by more than a predetermined distance are not considered, which speeds the computation. The candidates $HFCC_k$ and $LFCC_j$ may be compared on the basis of many different criteria or properties. In one aspect, as mentioned above, candidate clonotypes are compared on the basis of at least two properties: (i) frequency or read counts and (ii) sequence differences. In another aspect, as mentioned above, candidate clonotypes are compared on the basis of at least three properties: (i) frequency or read counts, (ii) sequence differences, and (iii) quality scores or measures of the bases where differences occur. In one embodiment, sequence differences include base substitutions; in another embodiment, sequence differences include base substitutions, deletions and insertions. The latter embodiment is especially applicable whenever sequence data is generated by sequencing-by-synthesis chemistries that do not employ terminators, such as 454 sequencers and Ion Torrent sequencers. Such sequencing approaches differentiate different sized homopolymer stretches by signal amplitude; thus, base-calling routines in such approaches are prone to insertion and deletion errors, because the difference in signal level from homopolymers differing by one nucleotide drops precipitously with increasing homopolymer size (that is, a 2-mer is readily distinguished from a 3-mer, but an 8-mer is almost indistinguishable from a 9-mer). In one aspect, comparisons of HFCCs and LFCCs may be implemented using a function (referred to herein as a "coalescence likelihood function"), such as P($HFCC_k$, $LFCC_j$, D, Q) shown in decision box (558), which depends on the quantities (i) through (iii) described above. Such a function may take many different forms, but generally the value of P changes with changes in (i), (ii) and (iii) as follows: The value of P preferably increases monotonically with the frequency of HFCC and the ratio of the frequency of HFCC to that of LFCC, such that the higher the ratio of the frequency of HFCC to that of LFCC, the higher the likelihood LFCC will be coalesced into HFCC. Likewise, the value of P preferably decreases monotonically with degree to which the sequences of HFCC and LFCC differ, so that the greater the difference between HFCC and LFCC (e.g. as measured by the minimal number of substitutions, insertions or deletions to change one to the other) the lower the likelihood LFCC will be coalesced with HFCC. Finally, the value of P preferably decreases monotonically with increasing quality scores of the locations where the sequences of HFCC and LFCC differ, so that for higher quality scores, the lower the likelihood LFCC will be coalesced with HFCC.

[00101] When the sequences of HFCC and LFCC differ at more than one location, the quality scores at the different locations may be combined in a variety of differ ways. In one embodiment, whenever there is a plurality of such differences, the plurality of quality scores is expressed as an average value, which may be either an unweighted average or a weighted average. Fig. 5C shows an exemplary function, P, computed for different quality values (curves a through e) for a given sequence difference. As illustrated in Fig. 5C, whenever HFCC is at a level of about 200 read counts (570), then if the quality scores are determined by curve (a), any LFCC with less than about 50 read counts (572) are coalesced into HFCC. The argument, D, of function P is a measure of the distance between the sequences $HFCC_k$ and $LFCC_j$ and its value may vary from cycle to cycle as an analysis progresses. (The indices "k" indicates that the values of constants with a "k" subscript may depend on the computational cycle, k.) In one embodiment, $D=D_k$, so that its value is a function of cycle number. In another embodiment, D=D(HFCC frequency), so that its value is a function of the frequency of HFCC, independent of cycle number. For example, as the frequency of HFCC decreases, then distance, D, of candidates to be compared decreases. In one embodiment, D is a Hamming distance between $HFCC_k$ and $LFCC_j$; however, other distance measures may be used. In one embodiment, D_k is a non-increasing function of k; and in another embodiment, D_k is a decreasing function of k. Decreasing the magnitude of D with increasing cycle number, or with decreasing frequency of HFCC, is advantageous in some embodiments because as a computation progresses to lower and lower frequency candidate clonotypes most such candidates are singletons, so that sequence distance (rather than frequency difference) becomes the predominant comparison. By lowering D as the computation progresses, unproductive comparisons to distant low frequency candidate clonotypes are reduced, thereby speeding up the computation. Function P may be a complicated expression depending on the number of factors being considered. Fig. 5C illustrates computed values for one embodiment of P which relates read count thresholds for coalescing an LFCC given a read count of an HFCC for different quality scores, as described above. Curves “a” through “e” represent the relationships for different quality scores (with curve “a” corresponding to the highest quality score).

[00102] Returning to Fig. 5B, if $P<P_k$, then $LFCC_j$ is not coalesced with $HFCC_k$ and another LFCC is selected (560). If $P>P_k$, then $LFCC_j$ is coalesced with $HFCC_k$ (562), in which case another LFCC is selected (566), unless there are no more LFCC left to evaluate (564). If there are no more LFCC to evaluate (564), then the current $HFCC_k$ (including all of the LFCC’s coalesced into it) is removed (568) from the data structure, such as the sequence tree. Such removal is illustrated in the simple sequence tree (590) of Figs. 5D-5E. There, path (592) (indicated by dashed line) in sequence tree (590) corresponds to HFCC (596), which is coalesced with LFCC (598). After coalescence, the segment of path (592) in shaded area (599) is removed from sequence tree (590) to give reduced sequence tree (597) shown in Fig. 5E, which is used in subsequent computations to find neighboring LFCC (554). After such removal, clonotype determination is finished if a stopping criterion (570) is met. In one embodiment, stopping criterion (570) is whether the last non-singleton candidate clonotype has been processed (552). In another embodiment, stopping criterion (570) is whether the frequency or the read counts of the selected HFCC is below that corresponding to a single lymphocyte. In one aspect of the method of the invention, an amplification step may result in each lymphocyte in a sample being represented by multiple copies of the same clonotype; thus, in one embodiment, whatever HFCC has a number of read counts below the number corresponding to a single lymphocyte, then the computation is stopped. In some embodiments, such a number of read counts (or candidate clonotype copies) is at least 10; in another embodiment, such number is at least 20; in another embodiment, such a number is at least 30; in another embodiment, such a number is at least 40. If the stopping criterion is not met, then the next HFCC is selected (572). The analytical steps summarized in the flow chart of Fig. 5B may be implemented in any suitable programming language, such as C, C++. Java, C#, Fortran, Pascal or the like.

[00103] In accordance with one aspect of the invention, the above method for determining clonotypes and/or clonotype profiles comprises steps of (a) forming a data structure of recombined immune molecules from sequence reads obtained by high throughput nucleic acid sequencing, (b) coalescing with a highest frequency candidate clonotype any lesser frequency candidate clonotypes whenever such lesser frequency is below a predetermined frequency value and a sequence difference therebetween is below a predetermined difference value to form a clonotype, (c) removing the coalesced candidate clonotype from the data structure, and (d) repeating steps (b) and (c) until a clonotype profile is formed. In one embodiment, the data structure is a sequence tree.

[00104] In accordance with another aspect of the invention, the above method of determining clonotypes may be carried out by steps comprising: (a) providing a set of sequence reads from a repertoire of recombined immune molecules each having a V region, an NDN region and a J region wherein for each such molecule at least one sequence read encompasses at least a portion of the NDN region of such molecule; (b) forming from sequence reads encompassing at least a portion of an NDN region a sequence tree having leaves representing candidate clonotypes, each leaf and its corresponding candidate clonotype having a frequency; (c) coalescing with a highest frequency candidate clonotype any lesser frequency candidate clonotypes whenever such lesser frequency is below a predetermined frequency value and a sequence difference therebetween is below a predetermined difference value to form a clonotype having a sequence of the highest frequency candidate clonotype; (d) removing leaves corresponding to the coalesced candidate clonotypes from the sequence tree; and (e) repeating steps (c) and (d) until a highest frequency of a lesser frequency candidate clonotype is below a predetermined stopping value. In one embodiment, the step of forming further includes selecting a highest frequency candidate clonotype and identifying all said lesser frequency candidate clonotypes having a sequence difference therewith less than a predetermined difference value to form a coalescence subset. Thus, in such embodiment, one may limit the total number of LFCCs that must be compared for the coalescing operation (only ones within the predetermined difference value are considered). Such value is a process input depending on the application, e.g. the size of the repertoire, how much computing time is used, and so on. As mentioned above, the function used for deciding whether to coalesce an HFCC with a LFCC can have a variety of forms. In one general aspect, for the step of coalescing, such a function may have the following properties: It depends on frequencies of HFCC, LFCC, the sequence difference therebetween (which may be expressed as a conventional string difference measure, such as a Hamming distance) and quality scores of the one or more nucleotide locations where the HFCC and LFCC differ; such that the function (i) monotonically increases with increasing ratio of frequency of HFCC and frequency of LFCC, (ii) monotonically decreases with increasing sequence difference between HFCC and LFCC, and (iii) monotonically decreases with increasing quality scores of the one or more nucleotide locations. That is, in regard to property (iii), the surer one is that HFCC and LFCC are different (e.g., because there is a high level of confidence in the base calls), then the less likely they will be coalescenced.

[00105] In some embodiments, a coalescence likelihood function is selected so that sequence reads are coalesced into different clonotypes (or target polynucleotides, such as, recombined nucleic acids) whenever such sequence reads are distinct with a likelihood of at least 95 percent; in other embodiments, a coalescence likelihood function is selected so that sequence reads are coalesced into different clonotypes whenever such sequence reads are distinct with a likelihood of at least 99 percent; in other embodiments, a coalescence likelihood function is selected so that sequence reads are coalesced into different clonotypes whenever such sequence reads are distinct with a likelihood of at least 99.9 percent. As mentioned above, in some embodiments, a coalescence likelihood function depends on an error rate of a sequencing chemistry used, the number of discrepant nucleotides in sequence reads being compared, and the relative frequencies of the sequence reads being compared; in another embodiment, a coalescence likelihood function depends on an error rate of a sequencing chemistry used, the number of discrepant nucleotides in sequence reads being compared, the relative frequencies of the sequence reads being compared, and the quality scores of the discrepant nucleotides. In the foregoing, selection of a predetermined frequency value and a predetermined difference value is a design choice that depend on particular applications. Factors affecting such choices may include details of the biology, speed of implementation, and the like.

Monitoring Applications

[00106] In one aspect, the invention is directed to methods for monitoring minimal residual disease by determining the presence, absence and/or level of nucleic acids in a sample that are characteristic or correlated with a disease. In some embodiments, such nucleic acids are somatically recombined nucleic acids, or clonotypes, which are correlated with a pre-cancerous or cancerous condition, such as a lymphoid or myeloid proliferative disorder and which can be used to monitor the status of the disorder or condition. Such nucleic acids, and in particular clonotypes, are useful for monitoring minimal residual disease of a cancer after treatment, where the result of such monitoring is a key factor in determining whether to continue, discontinue or otherwise modify treatment. In many malignant lymphoid and myeloid neoplasms, a diagnostic tissue sample, such as a peripheral blood sample or a bone marrow sample, is obtained before treatment from which a clonotype profile is generated (a "diagnostic clonotype profile"). For lymphoid or myeloid proliferative disorders, it is usually not known prior to a diagnostic sample which immune receptor chain(s) are correlated to the lymphoid or myeloid clone of the disorder or condition. Consequently, under current practice many separate amplifications and sequencings must be carried out on different recombined nucleic acids encoding different candidate immune receptor chains in order to identify clonotypes correlated with a patient's disease or condition. One or more disease-correlated clonotypes (i.e. "correlating clonotypes") are identified in clonotype profiles resulting from such amplifications and sequencing efforts. Typically, the clonotypes having the highest frequencies in the clonotype profiles are taken as the correlating clonotypes. In one aspect of the invention, the number of separate amplifications and sequencing runs necessary to identify correlating clonotypes is greatly reduced by providing larger scale multiplex amplifications in a single reaction of portions of recombined nucleic acids encoding a plurality of difference immune receptor chains. In some embodiments, such plurality is in the range of from 2 to 4 separate immune receptor chains; and in other embodiments, such plurality is in the range of from 2 to 3 separate immune receptor chains. More particularly, in some embodiments, among BCR chains the following are amplified in a single multiplex reaction: recombined nucleic acids encoding IgH including at least a portion of the VDJ region, IgH including at least a portion of the DJ region, and IgK; and in other embodiments, among TCR chains the following are amplified in a single multiplex reaction: TCRβ, TCRδ and TCRγ.

[00107] After treatment, and preferably after attainment of a complete remission of the cancer, the presence, absence or frequency of such correlating clonotypes or nucleic acids is assessed periodically to determine whether the remission is holding or whether the neoplasm is returning or relapsing, based on the presence of, or an increase in the frequency of, the correlating nucleic acids or clonotypes (or related clonotypes) in a post-treatment clonotype profile or nucleic acid profile. That is, after treatment, minimal residual disease of the cancer is assessed based on the presence, absence or frequency of the correlating clonotypes or characteristic nucleic acids. As mentioned above, when such correlating clonotypes are common or correspond to a rearranged receptor segment that lacks sufficient diversity (so that non-cancerous cells may share the clonotype), the occurrence of such clonotypes in a post-treatment clonotype profile may give rise to a false positive indication of relapse.

[00108] Methods of the invention are applicable to monitoring any proliferative disease in which a rearranged nucleic acid encoding an immune receptor or portion thereof can be used as a marker of cells involved in the disease. In one aspect, methods of the invention are applicable to lymphoid and myeloid proliferative disorders. In another aspect, methods of the invention are applicable to lymphomas and leukemias. In another aspect, methods of the invention are applicable to monitoring MRD in follicular lymphoma, chronic lymphocytic leukemia (CLL), acute lymphocytic leukemia (ALL), chronic myelogenous leukemia (CML), acute myelogenous leukemia (AML), Hodgkins's and non-Hodgkin's lymphomas, multiple myeloma (MM), monoclonal gammopathy of undetermined significance (MGUS), mantle cell lymphoma (MCL), diffuse large B cell lymphoma (DLBCL), myelodysplastic syndromes (MDS), T cell lymphoma, or the like. In a particular embodiment, a method of the invention is particularly well suited for monitoring MRD in ALL, MM or DLBCL.

[00109] In some embodiments, a patient sample, such as blood or bone marrow, is subjected to a diagnostic assay to identify which of a plurality of immune receptor chains may include the clonotype produced by a clone of a disorder (i.e. a correlating clonotype). Once the immune receptor chain of a correlating clonotype is determined, then subsequent monitoring assays may be specific for that particular immune receptor chain. For example, in some embodiments, a diagnostic assay may in the same reaction generate sequence-based clonotype profiles of a plurality of BCR chains, such as, IgH(VDJ), IgH(DJ) and IgK. If a correlating clonotype is an IgH(VDJ) chain, then subsequent monitoring assays may only generate IgH(VDJ) clonotype profiles. In some embodiments, the depth of sequencing in the diagnostic sample may be different than that of the monitoring sample. "Depth of sequencing" means the total number of sequence reads analyzed to construct clonotype profiles. For cancers, such as leukemias or lymphomas, since diagnostic assays are conducted on patient samples prior to treatment, the frequency or level of a correlating clonotype in the sample is typically high and readily identified. For example, any clonotype with a frequency over a predetermined level may be defined as a correlating clonotype. Such predetermined level may vary under with other patient indicators; however, often a predetermined level may be in the range of from 2 to 5 percent; or in some embodiments, five percent. Thus, in some embodiments, the depth of sequencing carried out is that which is necessary to reliably detect clonotypes present at a frequency of one or two percent or higher. In some embodiments, the depth of sequencing of a diagnostic sample produces at least 10,000 sequence reads; or in other embodiments, it is at least 100,000 sequence reads; in still other embodiments, the depth of sequencing of a diagnostic sample produces at least 10^6 sequence reads. In some embodiments, the depth of sequencing of

a monitoring sample is at least 100,000 sequence reads; in other embodiments, the depth of sequencing of a monitoring sample is at least 10^6 sequence reads.

[00110] In some embodiments, a lymphoid proliferative disorder, such as a leukemia or lymphoma, in a patient may be monitored by generating clonotype profiles from successively obtained samples (or tissue samples) from the patient. Such clonotype profiles may be generated as described above. In some embodiment, such monitoring may be implemented by the following steps: (a) obtaining a sample from an individual comprising T-cells and/or B-cells and/or cell-free DNA; (b) attaching sequence tags to recombined nucleic acid molecules of T-cell receptor genes or immunoglobulin genes from the sample to form tag-nucleic acid conjugates, wherein at least one recombined nucleic acid or copies thereof have different sequence tags attached; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing a sample of the tag-nucleic acid conjugates to provide sequence reads each having an error rate and each comprising a tag sequence and a recombined nucleic acid sequence; (e) aligning sequence reads having like tag sequences to form groups of sequence reads having the same sequence tags; (f) coalescing sequence reads of groups to determine clonotypes, wherein groups of sequence reads are coalesced into different recombined nucleic acid sequences whenever said groups of sequence reads are distinct with a likelihood of at least ninety-five percent; (g) determining the clonotype profile of the sample by determining levels of the clonotypes; and (h) determining the level of correlating clonotypes in the clonotype profile. In some embodiments, steps (a) through (h) may be repeated in the process of monitoring the patient to determine whether the level of correlating clonotypes is evidence of relapse of disease. In some embodiments, the steps of attaching and amplifying may comprise the following steps: (a) combining in a reaction mixture under primer extension conditions a first set of primers with a sample of recombined nucleic acids from immune cells expressing an immune receptor and/or cell-free DNA, wherein each primer of the first set has a receptor-specific portion such that the receptor-specific portion anneals to a different recombined nucleic acid at a predetermined location and is extended to form a first extension product, and wherein each primer of the first set has a 5'-non-complementary end containing a first primer binding site; (b) removing from the reaction mixture non-extended primers of the first set; and (c) adding to the reaction mixture under primer extension conditions a second set of primers, wherein each primer of the second set has a receptor-specific portion such that the receptor-specific portion anneals to the first extension product at a predetermined location and has a 5'-non-complementary end containing a second primer binding site, primers of the first set and/or primers of the second set comprising a sequence tag disposed between the receptor-specific portion and the first or second primer binding site, respectively, and wherein each primer of the second set is extended to form a second extension product, such that each second extension product comprises a first primer binding site, a second primer binding site, at least one sequence tag, and recombined nucleic acid encoding a portion of an immune cell receptor chain. In some embodiments, a step of coalescing recombined nucleic acids comprises coalescing sequence reads of different recombined nucleic acids whenever such sequence reads are distinct with a likelihood of at least ninety-nine percent; and in other embodiments, with a likelihood of at least 99.9 percent.

[00111] Methods of the invention are also applicable to monitoring minimal residual disease of a cancer in a patient, including a non-lymphoid or non-myeloid cancer, which has an identifying pattern of mutations, for example, in a selected set of cancer genes. Such a pattern of mutations, that is, the presence, absence and/or level of genes containing such mutations, can indicate a likelihood of disease recurrence. In some embodiments, target polynucleotides for such monitoring may be exons, portions of exons, selected introns and/or gene expression control regions, e.g. promoters, of a plurality of genes (referred to herein as "cancer gene molecules"). Cancer gene molecules may be isolated from a tissue sample using conventional techniques, such as exon capture techniques, e.g. TruSeq^{TM} exome enrichment kit (Illumina, San Diego, CA); Frampton et al, Nature Biotechnology, 31(11): 1023-1031 (2013); and the like. After such cancer gene molecules are obtained, sequence tags are attached to form tag-nucleic acid conjugate, tag-nucleic acid conjugate are amplified and sequenced in accordance with the invention.

[00112] Recent cancer genome sequencing studies have shown that there is significant heterogeneity in mutation patterns among different cancers, among different patients with the same cancer, among cells of the same tumor, and among cells of different metastatic sites in the same patient; however, within the same patient, the heterogeneous cancer cells typically evolve from a common ancestor, so that they share mutations and the evolutionary relationship among the cancerous cells may be discerned in a succession of measurements over time, e.g. Vogelstein et al, Science, 339: 1546-1558 (2013); Ding et al, Nature, 481(7382): 506-510 (2012); and the like; therefore, a pattern of mutations correlated with a cancer measured in a diagnostic sample provide a means to detect a recurrence of the same cancer or a clonally evolved version of it.

[00113] Cancer gene molecules may be selected from a wide variety of genes, including, but not limited to, the genes in Table I.

Table I

Exemplary Cancer Genes

ABL1	AKT1	ALK	APC	ATM
BRAF	CDH1	CSF1R	CTNNB1	EGFR
ERBB2	ERBB4	FBXW7	FGFR1	FGFR2
FGFR3	FLT3	GNA11	GNAQ	GNAS
HNF1A	HRAS	IDH1	JAK2	JAK3
KDR	KIT	KRAS	MET	MLH1
MPL	NOTCH1	NPM1	NRAS	PGGFRA
PIK3CA	PTEN	PTPN11	RB1	RET
SMAD4	SMO	SRC	STK	TP53
VHL				

[00114] In some embodiments, the above method of monitoring a minimal residual disease of a cancer may comprise the following steps: (a) obtaining from an individual a tissue sample; (b) attaching sequence tags to each of a plurality of cancer gene molecules in the sample to form tag-nucleic acid conjugates, wherein at least one nucleic acid or copies thereof have different sequence tags attached and wherein the cancer gene molecules are characteristic of a cancer of the individual; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing a sample of the tag-nucleic acid conjugates to provide sequence reads having error rates and comprising a tag sequence and a cancer gene sequence; (e) aligning sequence reads having like tag sequences to form groups of sequence reads having the same sequence tags; (f) coalescing cancer gene sequences of groups to determine sequences of cancer gene molecules, wherein groups of sequence reads are coalesced into different cancer gene molecules whenever said groups of cancer gene sequences are distinct with a likelihood of at least ninety-five percent; and (g) detecting in a profile of the cancer gene molecules the presence, absence and/or level of cancer gene molecules characteristic of the cancer of the individual. In some embodiments, a step of coalescing cancer gene sequences comprises coalescing sequence reads of different cancer gene molecules whenever such sequence reads are distinct with a likelihood of at least ninety-nine percent; and in other embodiments, with a likelihood of at least 99.9 percent.

Use of Sequence Tags to Detect Carry Over Contamination

[00115] 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.

[00116] Carry over contamination is especially challenging in a clinical setting when measuring highly complex populations of related nucleic acids, such as populations of recombined 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 king 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.

[00117] A wide variety of search methods or algorithms may be used to carry out the step of comparing measured clonotypes to database clonotypes. Many conventional sequence alignment and searching algorithms are publicly available and have been described in the following references which are incorporated by reference: Mount, Bioinformatics Sequence and Genome Analysis, Second Edition (Cold Spring Harbor Press, 2004); Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005); Altschul et al, J. Mol. Biol., 215(3): 403-410 (1990); Needleman and Wunsch, J. Mol. Biol., 48: 443-453 (1970); Smith and Waterman, Advances in Applied Mathematics, 2: 482-489 (1981); and the like.

[00118] In some embodiments, the above methods for detecting and measuring contamination, such as carry-over contamination, in a sample from material originating from a different sample may comprise the following steps: (a) obtaining from an individual a tissue sample; (b) attaching sequence tags to cancer gene molecules or recombined nucleic acids to form tag-nucleic acid conjugates, wherein at least one nucleic acid or copies thereof have different sequence tags attached and wherein the cancer gene molecules are characteristic of a cancer of the individual; (c) amplifying the tag-nucleic acid conjugates; (d) sequencing a sample of the tag-nucleic acid conjugates to provide sequence reads each having an error rate and each comprising a tag sequence and a cancer gene sequence or recombined nucleic acid sequence; (e) comparing tag sequences to separately determined tag sequences from other tissue samples; and (f) determining the presence, absence and/or level of contamination by the identity of one or more tag sequences with any separately determined tag sequences from other tissue samples. Once tag sequences are determined in an assay, they may be compared to tag sequences in a database of tag sequences recorded from assays on other patients. Such steps of comparing may be implemented at the time of an assay, or such steps may be implements retrospectively, for example, at a time after the time of the assay. In one embodiment, sequence tags are attached to recombined nucleic acids in a tissue sample, such as blood or bone marrow, from an individual suffering from a lymphoid proliferative disorder, such as a lymphoid cancer. In another embodiment, sequence tags are attached to cancer gene molecules, such as described above.

[00119] In further embodiments in which recombined nucleic acids are monitored from cross-contamination of tissue samples, the steps of attaching and amplifying may be implemented as follows: (a) combining in a reaction mixture under primer extension conditions a first set of primers with a sample of recombined nucleic acids from T-cells and/or cell-free DNA, wherein each primer of the first set has a receptor-specific portion such that the receptor-specific portion anneals to a different recombined nucleic acid at a predetermined location and is extended to form a first extension product, and wherein each primer of the first set has a 5'-non-complementary end containing a first primer binding site; (b) removing from the reaction mixture non-extended primers of the first set; (c) adding to the reaction mixture under primer extension conditions a second set of primers, wherein each primer of the second set has a receptor-specific portion such that the receptor-specific portion anneals to the first extension product at a predetermined location and has a 5'-non-complementary end containing a second primer binding site, primers of the first set and/or primers of the second set comprising a sequence tag disposed between the receptor-specific portion and the first or second primer binding site, respectively, and wherein each primer of the second set is extended to form a second extension product, such that each second extension product comprises a first primer binding site, a second primer binding site, at least one sequence tag, and recombined nucleic acid encoding a portion of a immune receptor chain; and (d) performing a polymerase chain reaction in the reaction mixture to form an amplicon, the polymerase chain reaction using forward primers specific for the first primer binding site and reverse primers specific for the second primer binding site.

Kits

[00120] The invention includes a variety of kits for carrying out methods of the invention. In some embodiments, kits comprise (a) a set of forward primers and a set of reverse primers for amplifying in a multiplex PCR recombined nucleic acids encoding a plurality of immune receptor chains wherein forward primers and/or reverse primers each have a target-specific portion, a sequence tag and a common primer binding site, and (b) a primer removal element for removing after at least a first extension unincorporated primers (i.e. non-extended primers) of the sets. In some embodiments, kits further comprise common primers specific for the common primer binding sites. In some embodiments, kits further comprise written instructions for using kit components in a method of the invention. In some embodiments, kits further comprise forward and reverse primers specific for amplifying recombined nucleic acids encoding IgH(VDJ), IgH(DJ) and IgK. In some embodiments, kits further comprise forward and reverse primers specific for amplifying recombined nucleic acids encoding TCRβ, TCRδ and TCRγ. In some embodiments, kits further comprise internal standards comprising a plurality of nucleic acids having lengths and compositions representative of the target recombined nucleic acids, wherein the internal standards are provided in known concentrations. In some embodiments, kits include a single-stranded exonuclease as a primer removal element, such as E. coli exonuclease I. In some embodiments, kits include a spin column capable of size selecting double stranded DNA as a primer removal element.

[00121] 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.

Definitions

[00122] 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).

[00123] "Aligning" means a method of comparing a test sequence, such as a sequence read, to one or more reference sequences to determine which reference sequence or which portion of a reference sequence is closest 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.

[00124] "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 polynucleotides 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. 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.

[00125] “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).

[00126] “Clonotype” means a recombined nucleic acid of a lymphocyte which encodes an immune receptor or a portion thereof. More particularly, clonotype means a recombined nucleic acid, usually extracted from a T cell or B cell, but which may also be from a cell-free source, 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.

[00127] “Clonotype profile” means a listing of distinct clonotypes and their relative abundances that are derived from a population of lymphocytes, where, for example, relative abundance may be expressed as a frequency in a given population (that is, a number between 0 and 1).. 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 x 10^6 to 1.8 x 10^6, or in the range of from 0.5 x 10^6 to 1.5 x 10^6, or in the range of from 0.8 x 10^6 to 1.2 x 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 x 10^6 to 1.8 x 10^6, or in the range of from 0.5 x 10^6 to 1.5 x 10^6, or in the range of from 0.8 x 10^6 to 1.2 x 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.

[00128] "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.

[00129] "Clonotype Database" means a collection of clonotypes formatted and arranged for ease and speed of searching, comparing and retrieving. In some embodiments, a clonotype database comprises a collection of clonotypes encoding the same region or segment of an immune receptor. In some embodiments, a clonotype database comprises clonotypes of clonotype profiles from a plurality of individuals. In some embodiments, a clonotype database comprises clonotypes of clonotype profiles of at least 10^4 clonotypes from at least 10 individuals. In some embodiments, a clonotype database comprises at least 10^6 clonotypes, or at least 10^8 clonotypes, or at least 10^9 clonotypes, or at least 10^{10} clonotypes. A clonotype database may be a public database containing clonotypes, such as IMGT database (www.imgt.org), e.g. described in Nucleic Acids Research, 31: 307-310 (2003). Clonotype databases may be in a FASTA format, and clonotype database entries may be searched or compared using a BLAST algorithm, e.g. Altschul et al, J. Mol. Biol., 215(3): 403-410 (1990), or like algorithm.

[00130] "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 and the higher frequency candidate clonotype and the lower frequency candidate clonotype are treated as the same; that is, the observed difference between them is determined to be due to error (e.g. sequencing error, amplification error, or the like). In some embodiments, the predetermined criteria is a likelihood function that depends on factors such as relative frequencies of the candidate clonotypes being compared, the number of positions at which the candidates differ, the quality scores of the positions, and the like.

[00131] "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.

[00132] “Contamination” as used herein means the presence in a tissue sample of one individual of nucleic acid from another individual. In one aspect, “contamination” means the presence of nucleic acid not originating from a patient which may affect the interpretation of a clonotype profile of the patient.

[00133] “Genetic identification” means a unique correspondence between an individual and a set of values (or states) of genetic markers from one or more genetic loci of the individual.

[00134] “Genetic marker” means a polymorphic segment of DNA at a genetic locus, which may be used to identify an individual. A genetic marker may be identified by its sequence or by adjacent or flanking sequences. Typically, a genetic marker can have a plurality of sequences, or values, in different individuals of a population. Exemplary genetic markers include, but are not limited to, short tandem repeats (STRs), single nucleotide polymorphisms (SNPs), and the like. The polymorphic segment of DNA may be genomic DNA or it may be reverse transcribed RNA. In one embodiment, the polymorphic segment is genomic DNA. In one embodiment, a genetic marker for use with the invention is identified by amplification and sequencing using conventional techniques. In another embodiment, genetic markers are amplified and sequenced together with immune molecules during the process for generating a clonotype profile.

[00135] “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.

[00136] "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.

[00137] "Minimal residual disease" means remaining cancer cells after treatment. The term is frequently used in connection with treatment of lymphomas and leukemias.

[00138] "Lymphoid or myeloid proliferative disorder" means any abnormal proliferative disorder in which one or more nucleotide sequences encoding one or more rearranged immune receptors can be used as a marker for monitoring such disorder. "Lymphoid or myeloid neoplasm" means an abnormal proliferation of lymphocytes or myeloid cells that may be malignant or non-malignant. A lymphoid cancer is a malignant lymphoid neoplasm. A myeloid cancer is a malignant myeloid neoplasm. Lymphoid and myeloid neoplasms are the result of, or are associated with, lymphoproliferative or myeloproliferative disorders, and include, but are not limited to, follicular lymphoma, chronic lymphocytic leukemia (CLL), acute lymphocytic leukemia (ALL), chronic myelogenous leukemia (CML), acute myelogenous leukemia (AML), Hodgkins's and non-Hodgkin's lymphomas, multiple myeloma (MM), monoclonal gammopathy of undetermined significance (MGUS), mantle cell lymphoma (MCL), diffuse large B cell lymphoma (DLBCL), myelodysplastic syndromes (MDS), 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). As used herein, "B cell cancer" means a lymphoid or myeloid neoplasm involving B cells or cells developed therefrom, such as plasma cells. Likewise, "T cell cancer" means a lymphoid or myeloid neoplasm involving T cells or cells developed therefrom.

[00139] "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.

[00140] “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. As used herein, the terms “forward primer” and “upstream primer” are used interchangeably, and the terms “reverse primer” and “downstream primer” are used interchangeably. Also as used herein, if a double stranded target polynucleotide is displayed with its sense strand in a 5’→3’ left-to-right orientation, a forward primer would bind to the antisense strand on the left and be extended to the right and a reverse primer would bind to the sense strand on the right and be extended to the left. 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°C, primers annealed at a temperature in the range 50-75°C, and primers extended at a temperature in the range 72-78°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. 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. "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.

[00141] "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).

[00142] "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.

[00143] "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.

[00144] "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. Each sequence tag has a nucleotide sequence which is sometimes referred to herein as a "tag sequence." 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 or circular 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. Selection of particular tag lengths and/or compositions may depend on several factors including, without limitation, the sequencing technology used to decode a tag; the number of distinguishable tags required to unambiguously identify a set of target polynucleotides, 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, and the like. In some embodiments, sequence tags can each have a length within a range of from 6 to 100 nucleotides, or from 10 to 100 nucleotides, or from 12 to 50 nucleotides, or from 12 to 25 nucleotides, respectively. In some embodiments, 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 four bases; in other embodiments, 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 five bases; in still other embodiments, 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 ten percent of their nucleotides; or in other embodiments, at least twenty-five percent of their nucleotides; or in other embodiments, at least fifty percent of their nucleotides.

Sequence Listing

```
<110>  Sequenta, Inc.
       Asbury, Thomas
       Hervold, Kieran
       Kotwaliwale, Chitra
       Faham, Malek
       Moorhead, Martin
      Weng, Li
       Wittkop, Tobias
       Zheng, Jianbiao

<120>  LARGE-SCALE BIOMOLECULAR ANALYSIS WITH SEQUENCE TAGS
<130>  848US00 (37623-739.201)
<150>  61/841878
<151>  2013-07-01
<150>  62/001580
<151>  2014-05-21
<160>  6
<170>  PatentIn version 3.5


<210>  1
<211>  24
<212>  DNA
<213>  Artificial Sequence
<220>
<223>  primer
<400>  1
agttctggct aacctgtaga gcca
24
<210>  2
<211>  24
<212>  DNA
<213>  Artificial Sequence
<220>
<223>  primer
<400>  2
```

```
agttcgggct aacctgtcga gcca
24
<210>  3
<211>  24
<212>  DNA
<213>  Artificial Sequence
<220>
<223>  primer
<400>  3
agttccggct aacctgtcga gcca
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
nnnnnnnnnn nnnnnnnnnn nn
22
<210>  5
<211>  12
<212>  DNA
<213>  Artificial Sequence
<220>
<223>  primer
<400>  5
gtattttttt 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 detecting minimal residual disease in a patient being treated for cancer, the method comprising the steps of:

(a) attaching sequence tags to each of a plurality of recombined nucleic acids in a sample containing T-cells and/or B-cells and/or cell free DNA or RNA obtained from the patient to form tag-nucleic acid conjugates, wherein at least one recombined nucleic acid or copies thereof have different sequence tags attached and are characteristic of the cancer of the patient, and wherein the attaching comprises:

(i) combining in a reaction mixture under primer extension conditions a first set of primers with the sample, wherein each primer of the first set comprises a receptor-specific portion, a 5′-non-complementary end containing a first primer binding site and a sequence tag disposed between the receptor-specific portion and the first primer binding site, wherein the receptor-specific portion anneals to a different recombined nucleic acid at a first predetermined location and is extended to form a first extension product; and

(ii) adding to the reaction mixture under primer extension conditions a second set of primers, wherein each primer of the second set has a receptor-specific portion, wherein the receptor-specific portion anneals to the first extension product at a second predetermined location, and wherein each primer of the second set is extended to form a second extension product, wherein each second extension product comprises a first primer binding site, sequence tag, and recombined nucleic acid encoding a portion of a T cell receptor chain or a B cell receptor chain;

(b) amplifying the tag-nucleic acid conjugates;

(c) sequencing a sample of the tag-nucleic acid conjugates to provide sequence reads each comprising a tag sequence and a recombined nucleic acid sequence;

(d) aligning sequence reads having like tag sequences to form groups of sequence reads having identical sequence tags;

(e) coalescing recombined nucleic acid sequences of groups to determine clonotypes, wherein groups of sequence reads are coalesced into different clonotypes whenever said groups of recombined nucleic acid sequences are distinct with a likelihood of at least 99.9 percent; and

(f) detecting in a clonotype profile the presence, absence and/or level of clonotypes correlated with the cancer of the patient, thereby detecting the minimal residual disease in the patient.

2. The method of claim 1, wherein the amplifying comprises performing a polymerase chain reaction in the reaction mixture to form an amplicon, the polymerase chain reaction using forward primers specific for the first primer binding site and reverse primers specific for the second set of primers.

3. The method of claim 1, wherein said recombined nucleic acids comprise recombined nucleic acids encoding TCRβ, TCRδ and TCRγ chains, and wherein said primers of said first set and said primer of said second set comprise primers that flank regions of said recombined nucleic acids encoding VDJ regions of TCRβ and TCRδ and primers that flank VJ regions of TCRγ.

4. The method of claim 1, wherein said recombined nucleic acids comprise recombined nucleic acids encoding IgH and IgK and wherein said primers of said first set and said primers of said second set comprise primers that flank regions of said recombined nucleic acids encoding VDJ regions of IgH, DJ regions of IgH and VJ regions of IgK.

5. The method of claim 4, wherein said primers of said first set or said second set include at least one nested set of primers specific for a plurality of different primer binding sites in V regions of said IgH chains.

6. The method of claim 1, further comprising a step of removing from the reaction mixture non-extended primers of the first set and the second set.

7. The method of claim 6, wherein the step of removing comprises adding an enzyme with exonuclease activity to the reaction mixture.

8. The method of claim 1, wherein said annealing and extension of said primers of said first set is repeated after melting said first extension product.

9. The method of claim 1, wherein said annealing and extension of said primers of said second set is repeated after melting said second extension product.

10. The method of claim 1, wherein the sequence tags are mosaic tags, wherein the mosaic tags comprise alternating constant regions and variable regions.

11. The method of claim 1, wherein the first predetermined location is a V region.

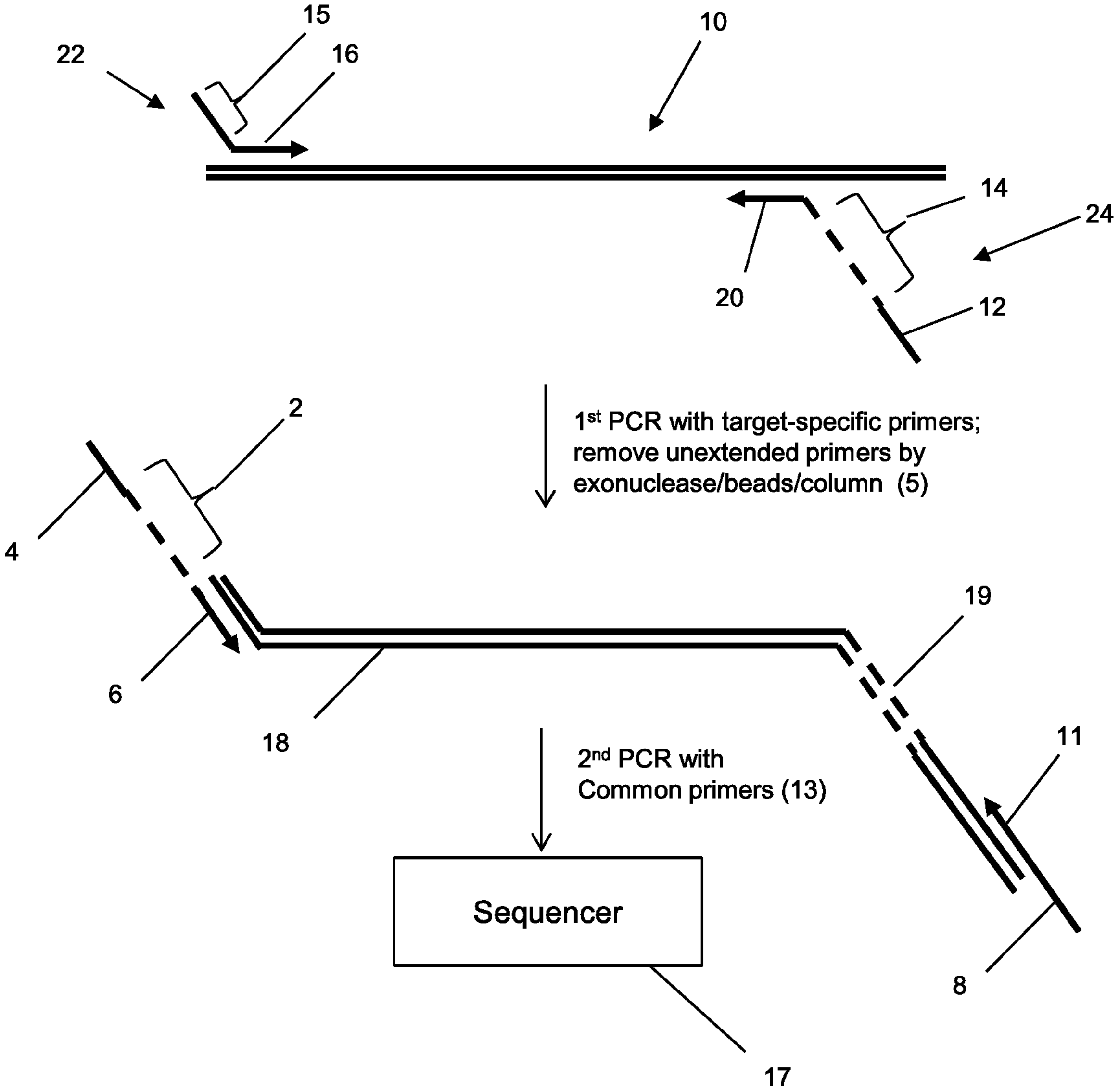
12. The method of claim 11, wherein the receptor specific portion of each of the primers of the first set anneals to different non-overlapping positions in the V region.

13. The method of claim 1, wherein the second predetermined location is a J region or a C region.

14. The method of claim 13, wherein the receptor specific portion of each of the primers of the second set anneals to different non-overlapping positions in the J region.

15. The method of claim 13, wherein the receptor specific portion of each of the primers of the second set anneals to a single primer binding site in the J region.

16. The method of claim 2, wherein the forward primers and/or the reverse primers comprise a 5′ non-complementary end comprising a sample tag.



1st PCR with target-specific primers; remove unextended primers by exonuclease/beads/column (5)

2nd PCR with Common primers (13)

Fig. 1A

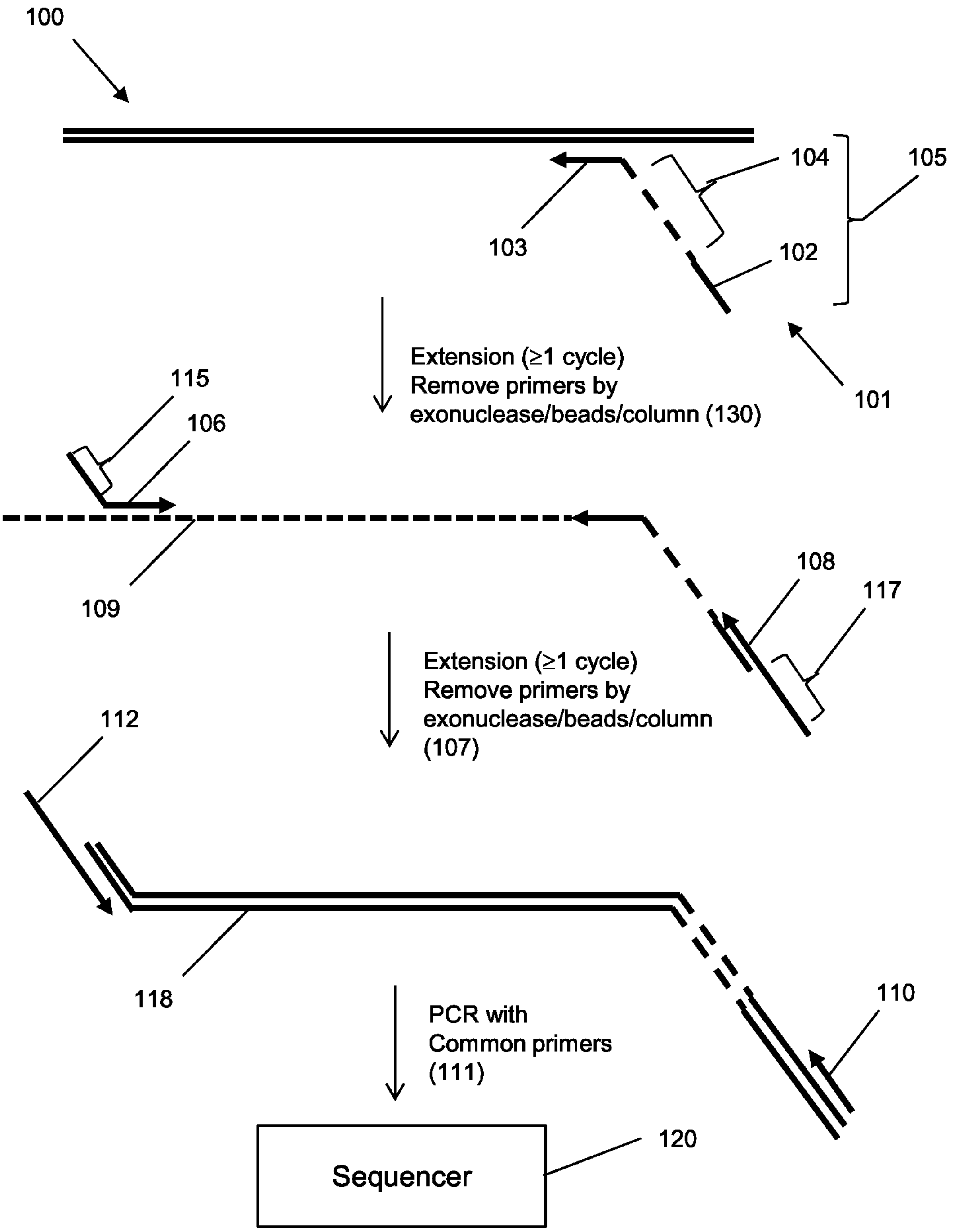

100
104
105
103
102
101
Extension (≥1 cycle)
Remove primers by
exonuclease/beads/column (130)
115
106
109
108
117
Extension (≥1 cycle)
Remove primers by
exonuclease/beads/column
(107)
112
118
110
PCR with
Common primers
(111)
Sequencer
120

Fig. 1B

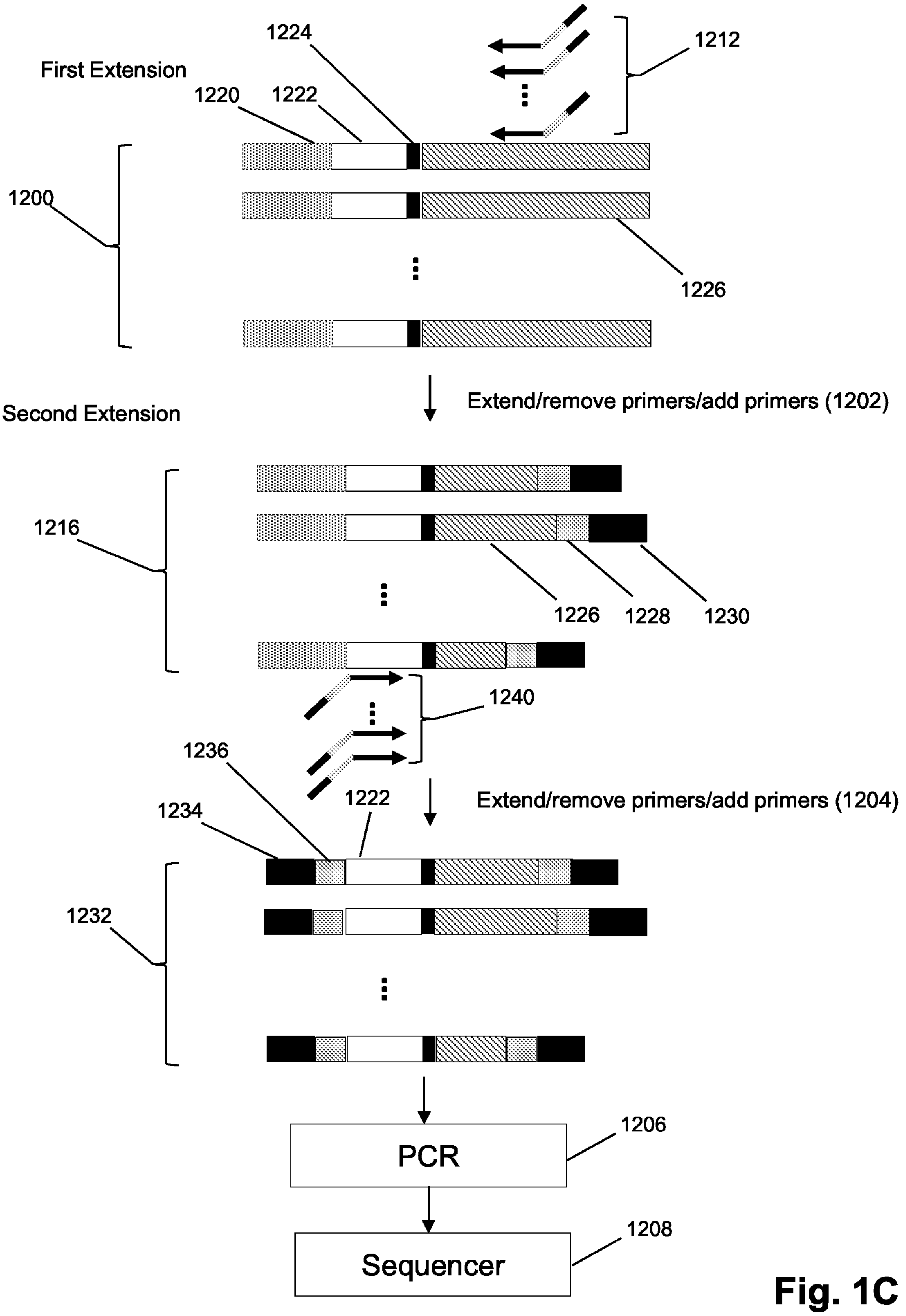

First Extension
1224
1212
1220
1222
1200
1226
Extend/remove primers/add primers (1202)
Second Extension
1216
1226
1228
1230
1240
1236
1234
1222
Extend/remove primers/add primers (1204)
1232
PCR
1206
Sequencer
1208

Fig. 1C

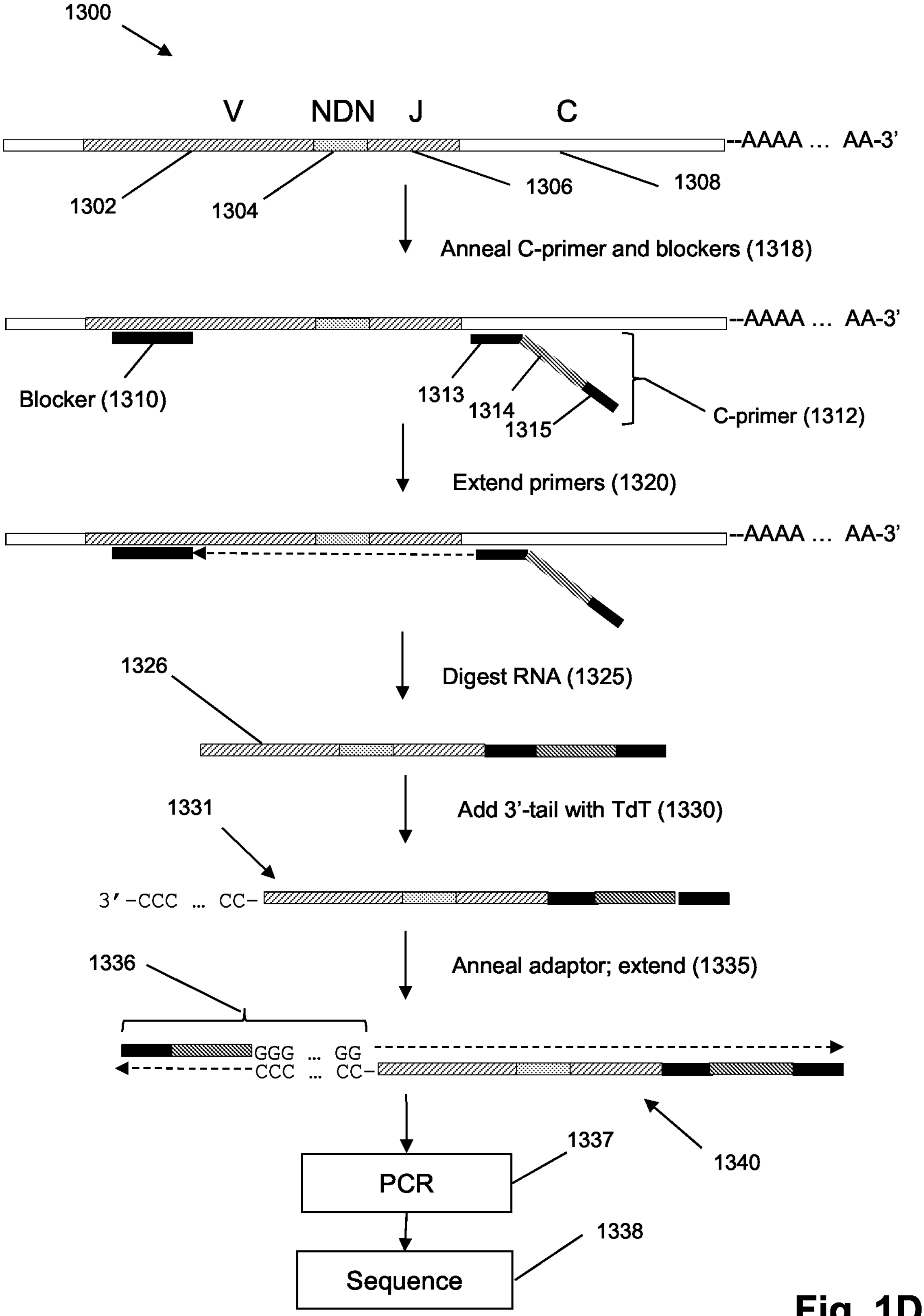

1300
V
NDN
J
C
--AAAA ... AA-3'
1302
1304
1306
1308
Anneal C-primer and blockers (1318)
--AAAA ... AA-3'
Blocker (1310)
1313
1314
1315
C-primer (1312)
Extend primers (1320)
--AAAA ... AA-3'
1326
Digest RNA (1325)
1331
Add 3'-tail with TdT (1330)
3'-CCC ... CC-
1336
Anneal adaptor; extend (1335)
GGG ... GG
CCC ... CC-
1337
1340
PCR
1338
Sequence

Fig. 1D

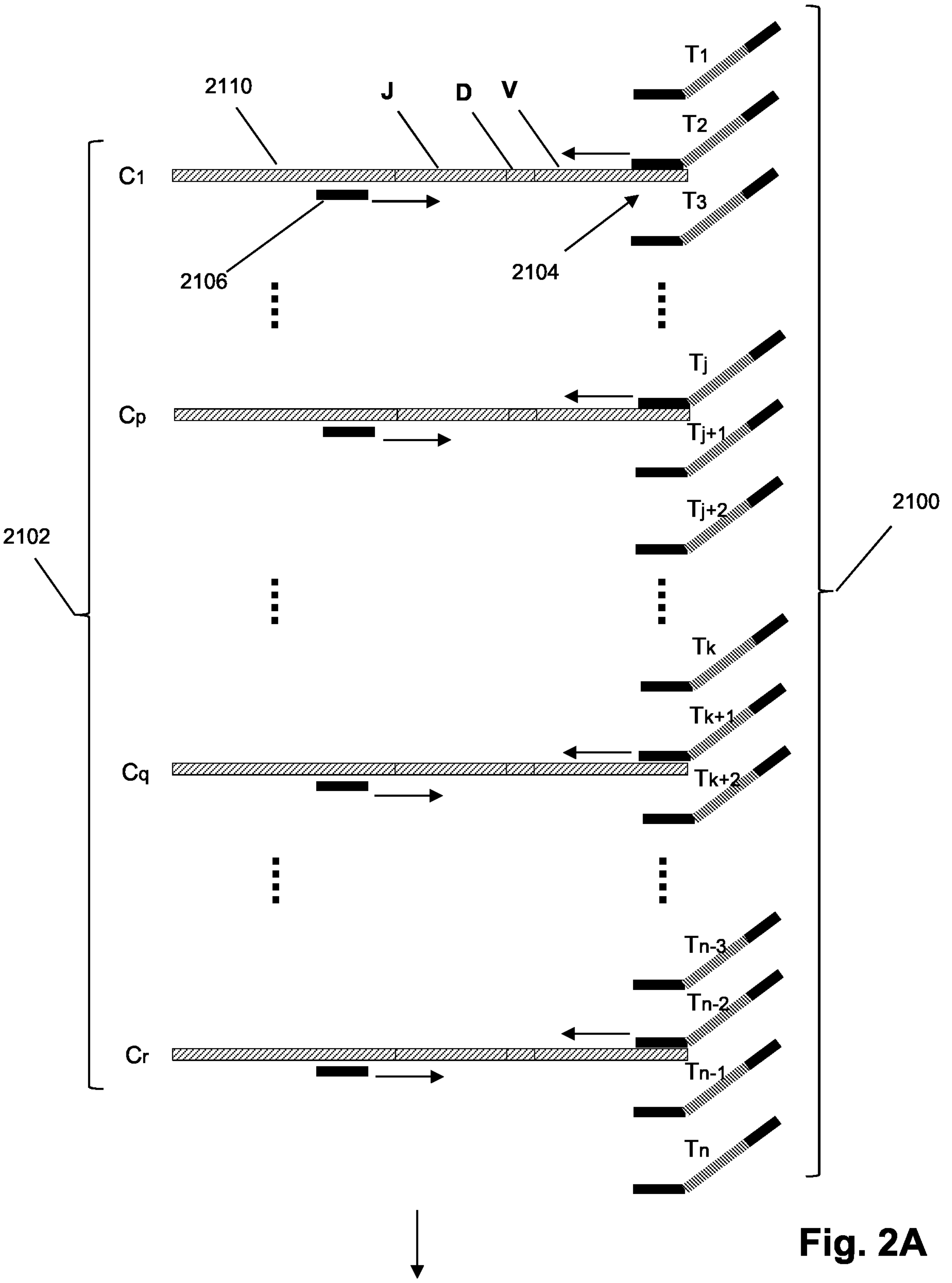

2110
J
D
V
T1
T2
T3
C1
2106
2104
Cp
Tj
Tj+1
Tj+2
2100
2102
Tk
Tk+1
Tk+2
Cq
Tn-3
Tn-2
Tn-1
Tn
Cr

Fig. 2A

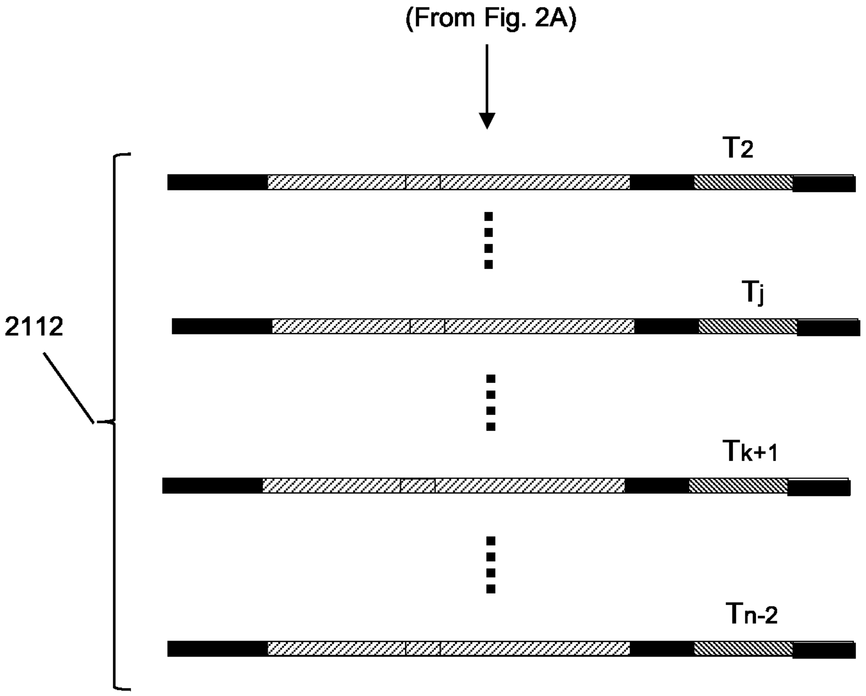


Fig. 2B

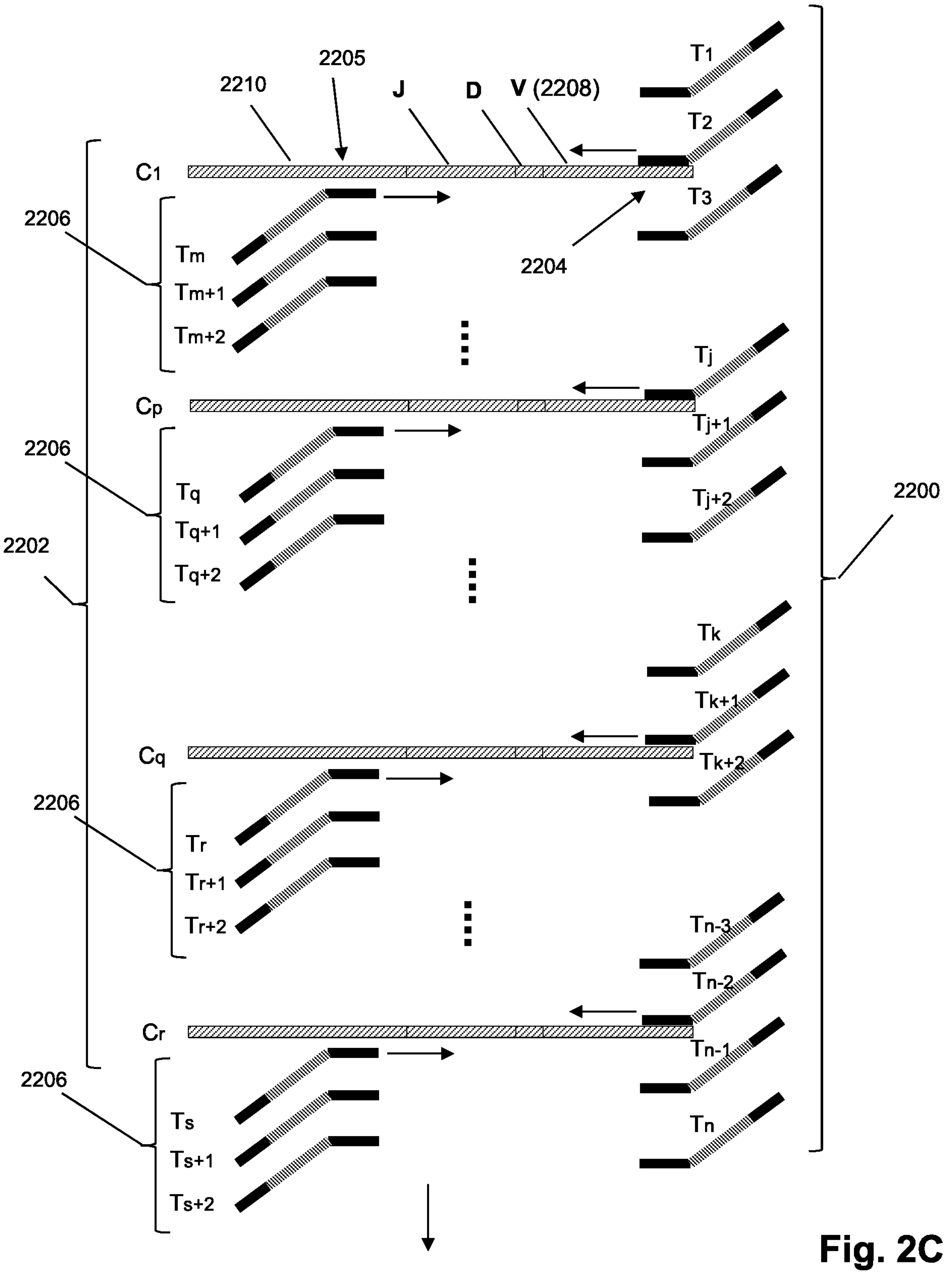

2205
2210
J
D
V (2208)
T1
T2
C1
T3
2206
Tm
Tm+1
Tm+2
2204
Tj
Cp
Tj+1
2206
Tq
Tq+1
Tq+2
Tj+2
2200
2202
Tk
Tk+1
Cq
Tk+2
2206
Tr
Tr+1
Tr+2
Tn-3
Tn-2
Cr
Tn-1
2206
Ts
Ts+1
Ts+2
Tn

Fig. 2C

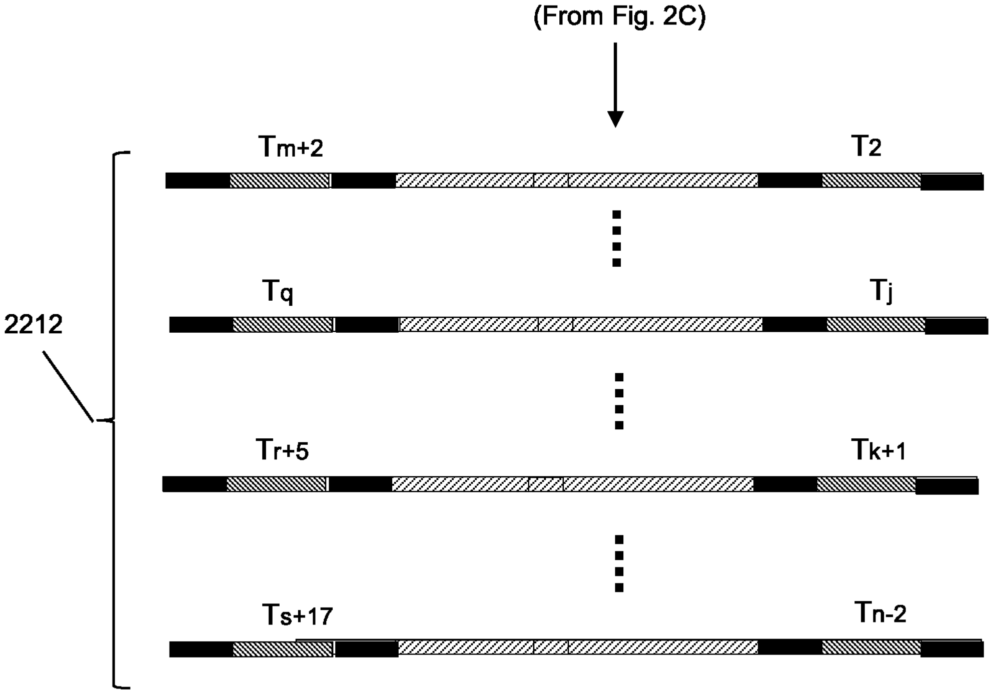


Fig. 2D

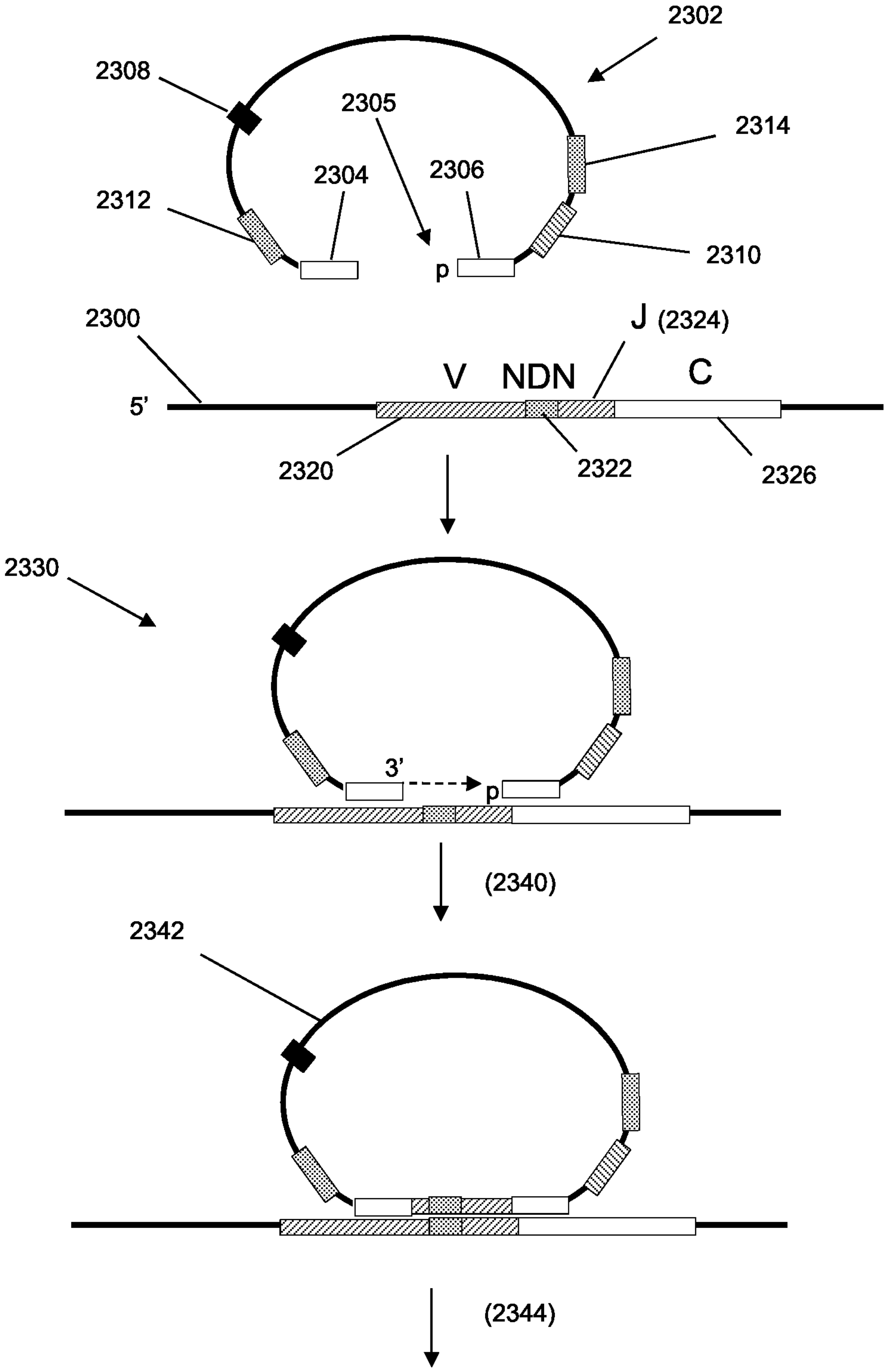

2302
2308
2305
2314
2304
2306
2312
p
2310
2300
J (2324)
V
NDN
C
5'
2320
2322
2326
2330
3'
p
(2340)
2342
(2344)

Fig. 2E

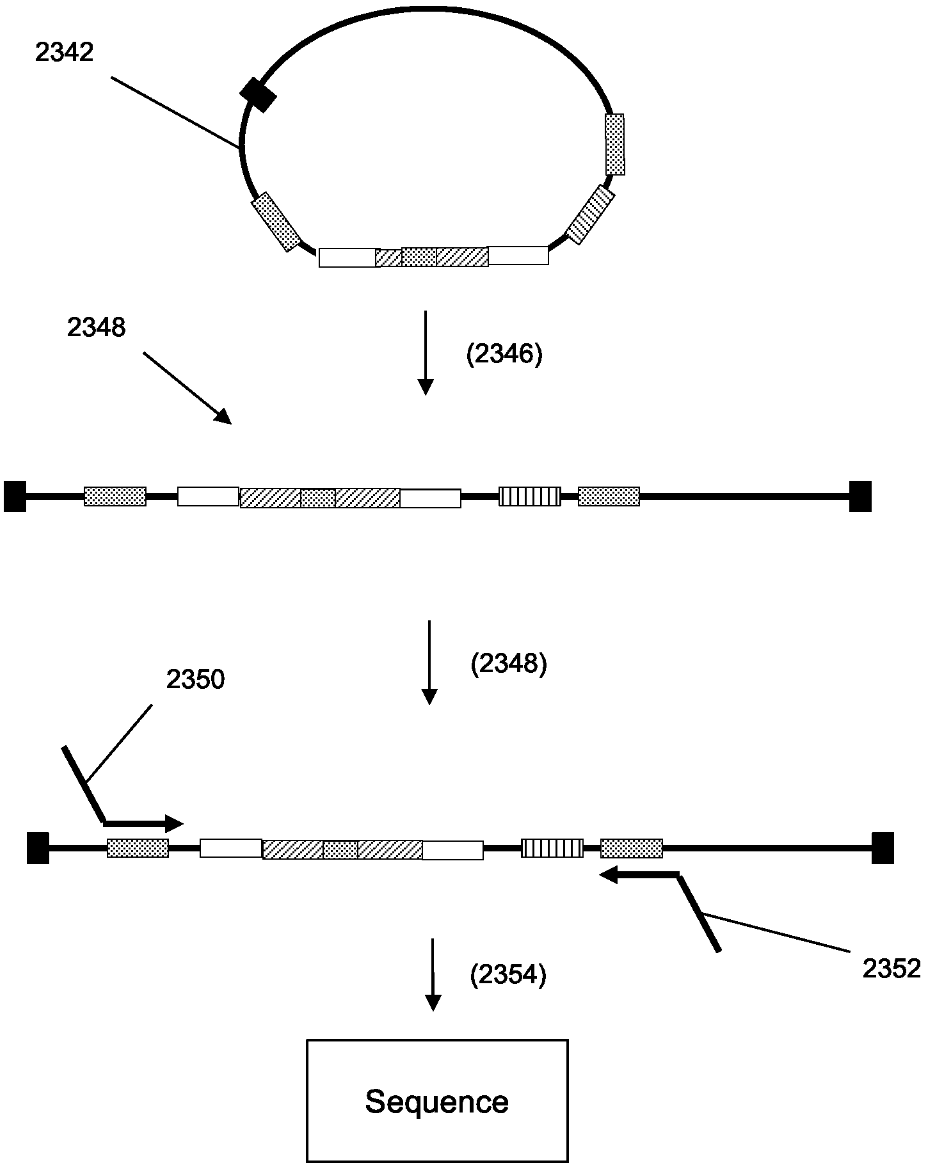


Fig. 2F

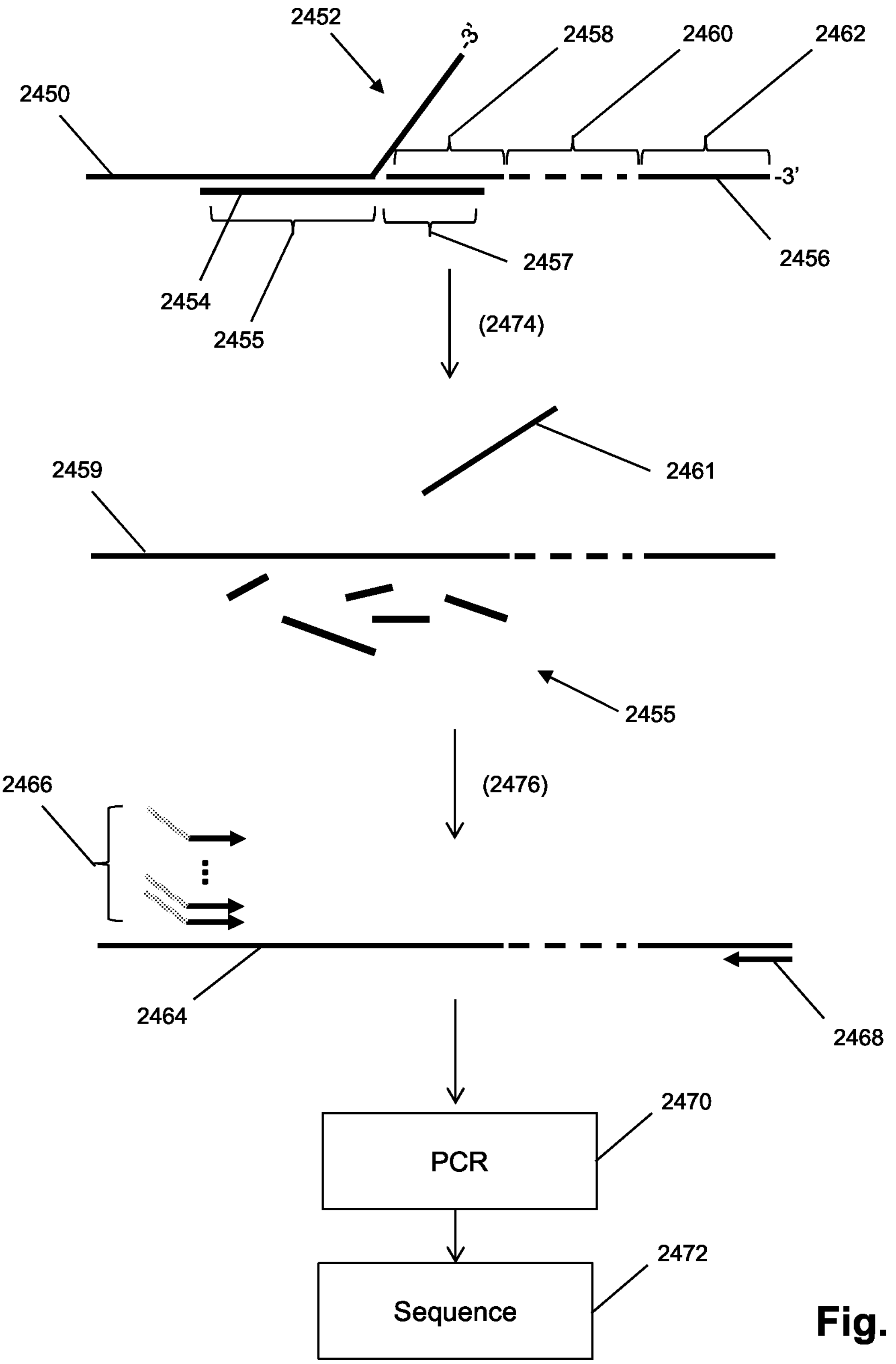

2452
3'
2458
2460
2462
2450
-3'
2457
2456
2454
2455
(2474)
2461
2459
2455
(2476)
2466
2464
2468
2470
PCR
2472
Sequence

Fig. 2G

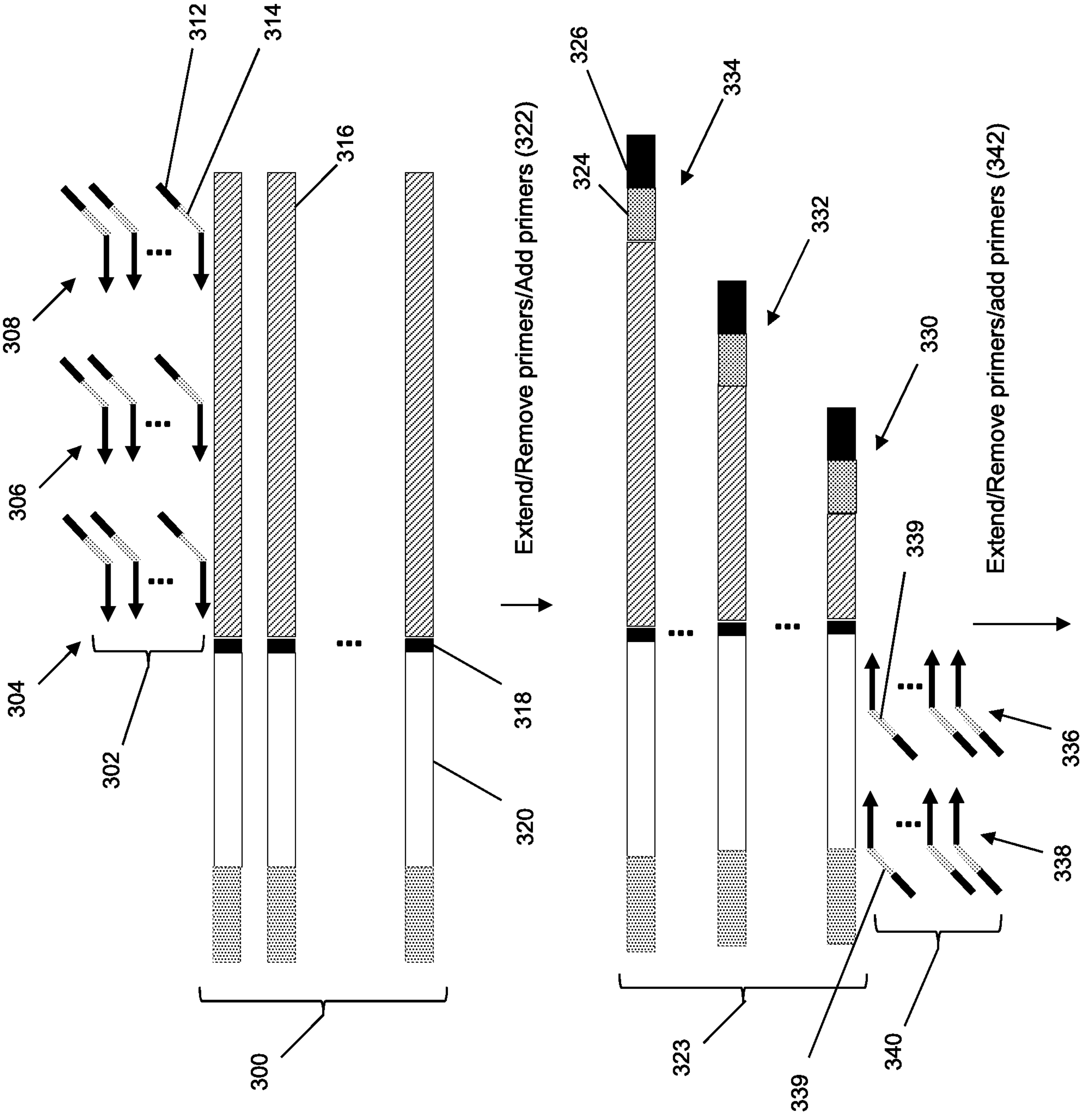
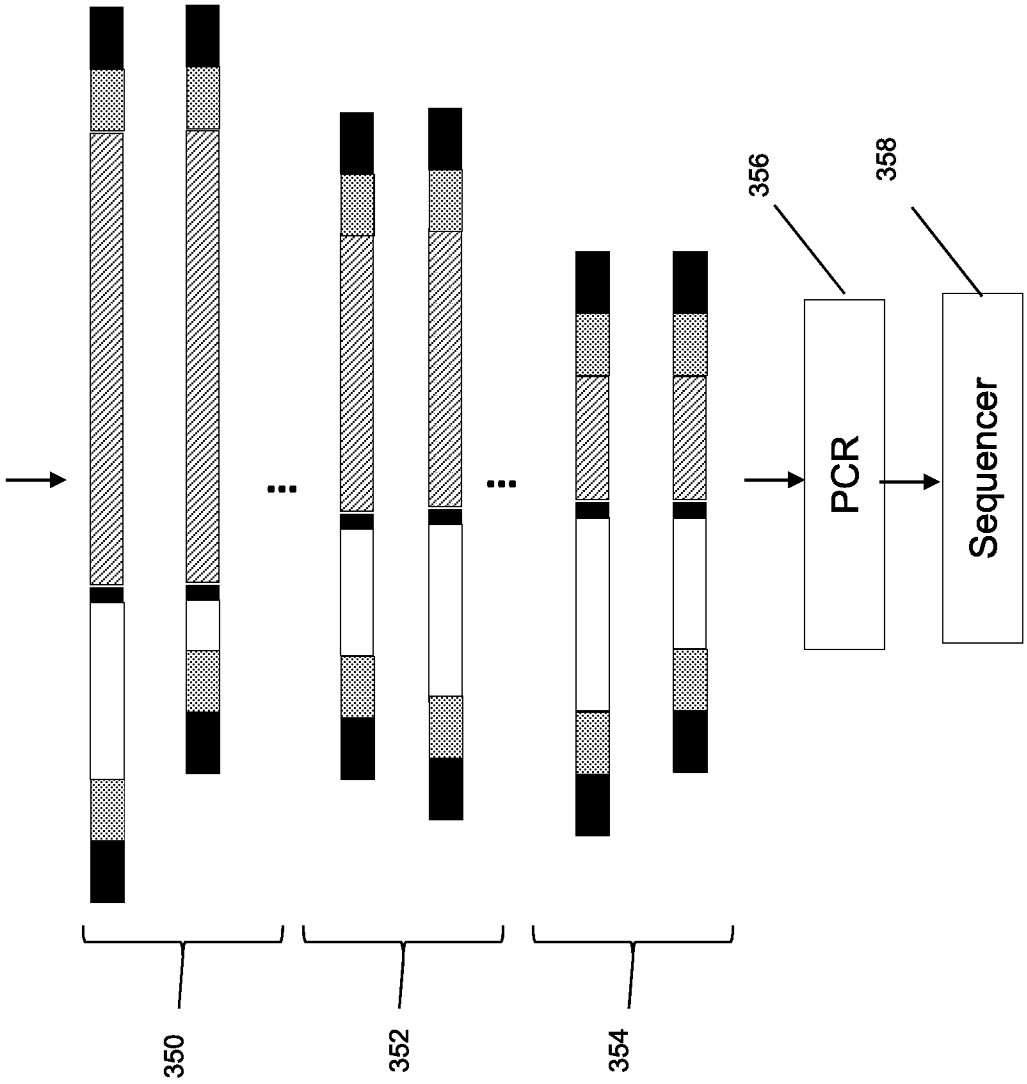

300
302
304
306
308
312
314
316
318
320
Extend/Remove primers/Add primers (322)
323
324
326
334
332
330
339
339
336
338
340
Extend/Remove primers/add primers (342)

Fig. 3A

Fig. 3B



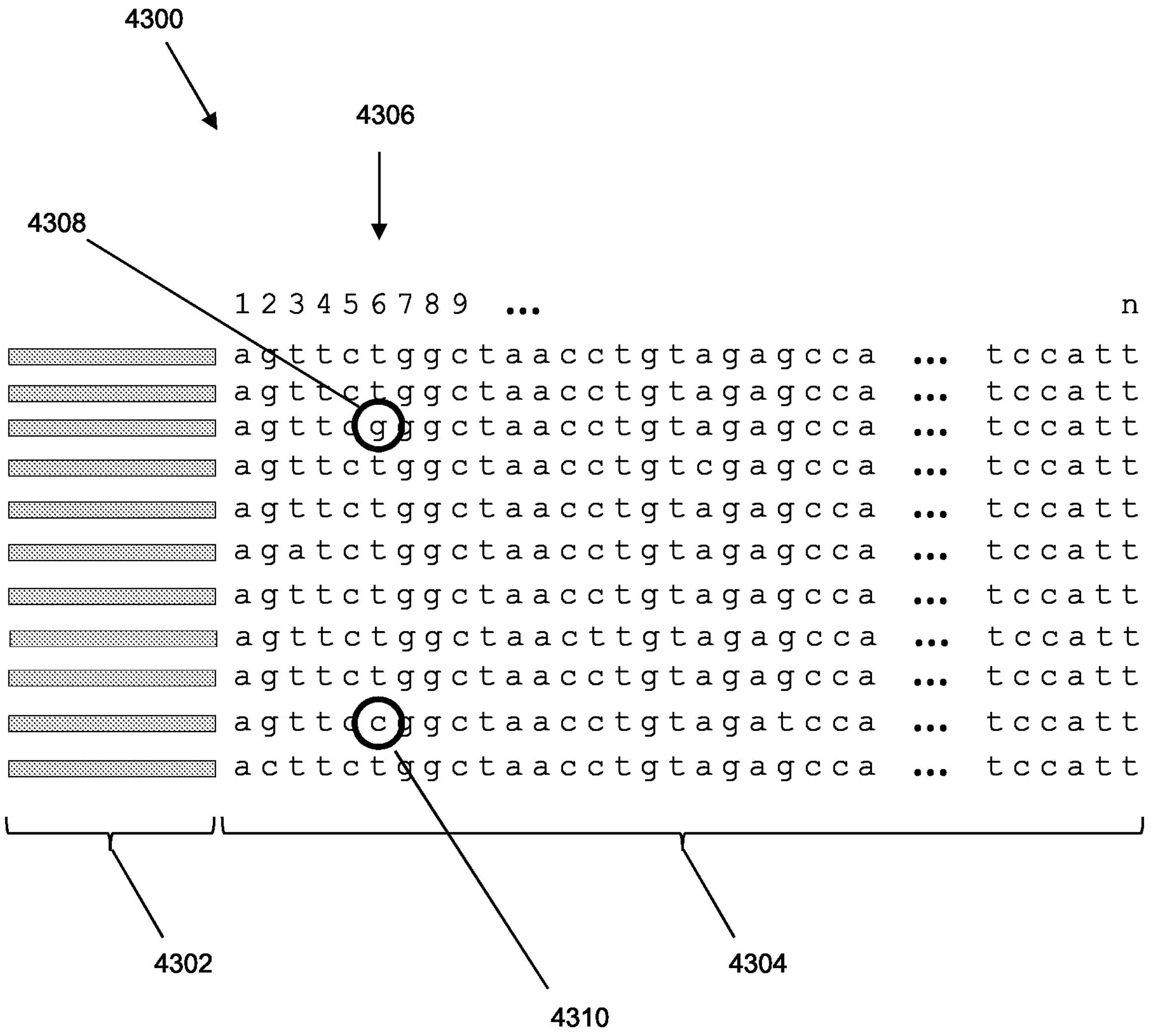

4300
4306
4308
1 2 3 4 5 6 7 8 9 ... n
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttcgggctaacctgtagagcca ... tccatt
agttctggctaacctgtcgagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agatctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacttgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttccggctaacctgtagatcca ... tccatt
acttctggctaacctgtagagcca ... tccatt
4302
4304
4310

Fig. 4A

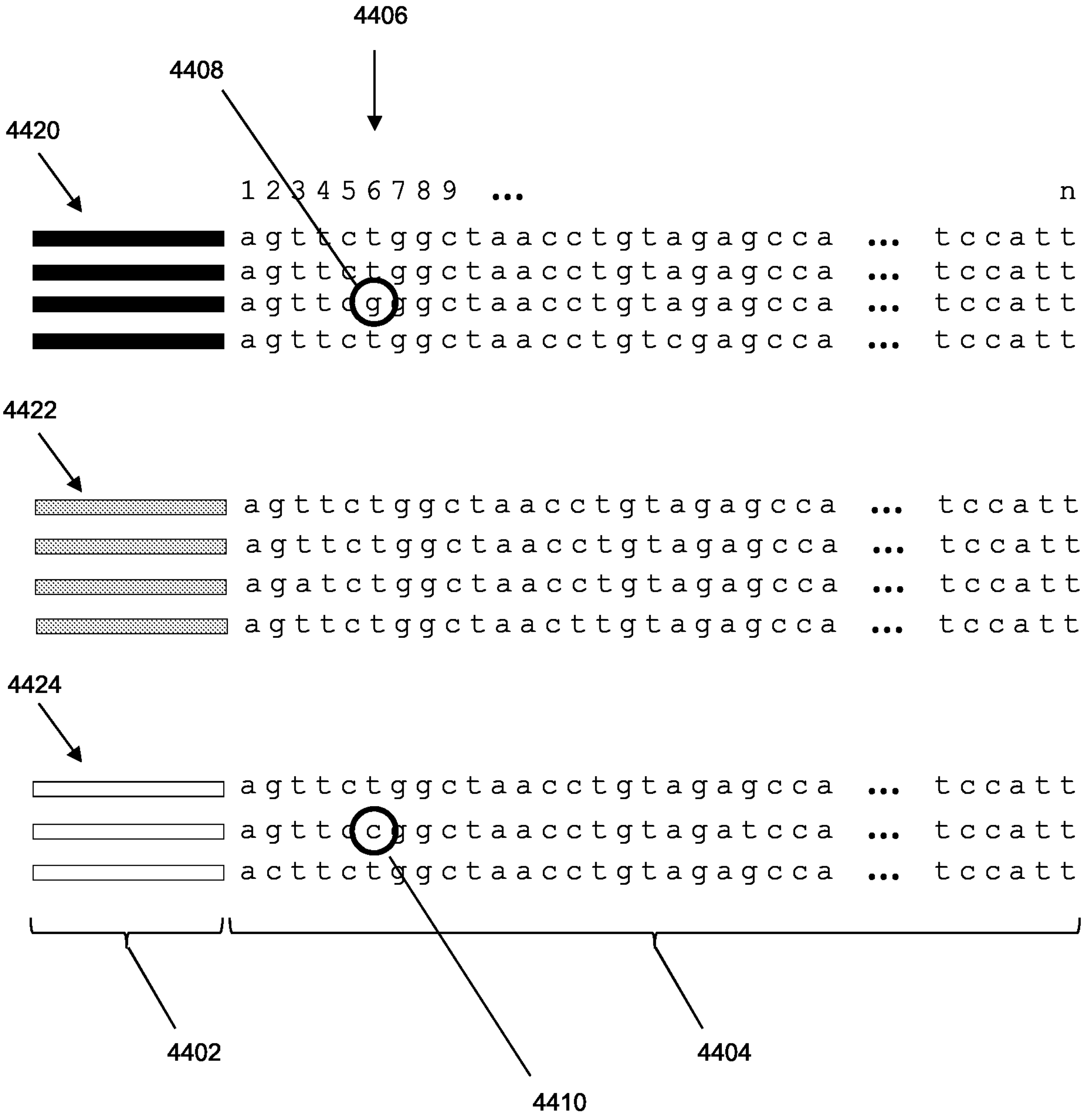

4406
4408
4420
1 2 3 4 5 6 7 8 9 ... n
a g t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c g g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c t g g c t a a c c t g t c g a g c c a ... t c c a t t
4422
a g t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g a t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c t g g c t a a c t t g t a g a g c c a ... t c c a t t
4424
a g t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
a g t t c c g g c t a a c c t g t a g a t c c a ... t c c a t t
a c t t c t g g c t a a c c t g t a g a g c c a ... t c c a t t
4402
4404
4410

Fig. 4B

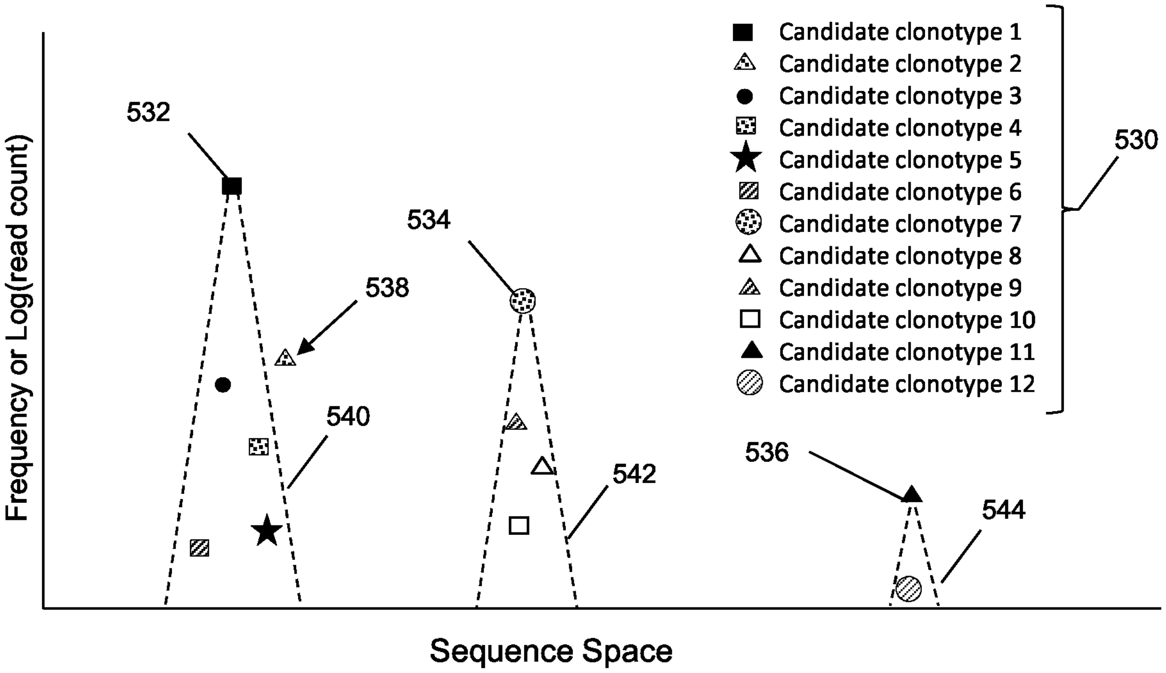

Candidate clonotype 1
Candidate clonotype 2
Candidate clonotype 3
Candidate clonotype 4
Candidate clonotype 5
Candidate clonotype 6
Candidate clonotype 7
Candidate clonotype 8
Candidate clonotype 9
Candidate clonotype 10
Candidate clonotype 11
Candidate clonotype 12
530
532
534
536
538
540
542
544
Frequency or Log(read count)
Sequence Space

Fig. 5A

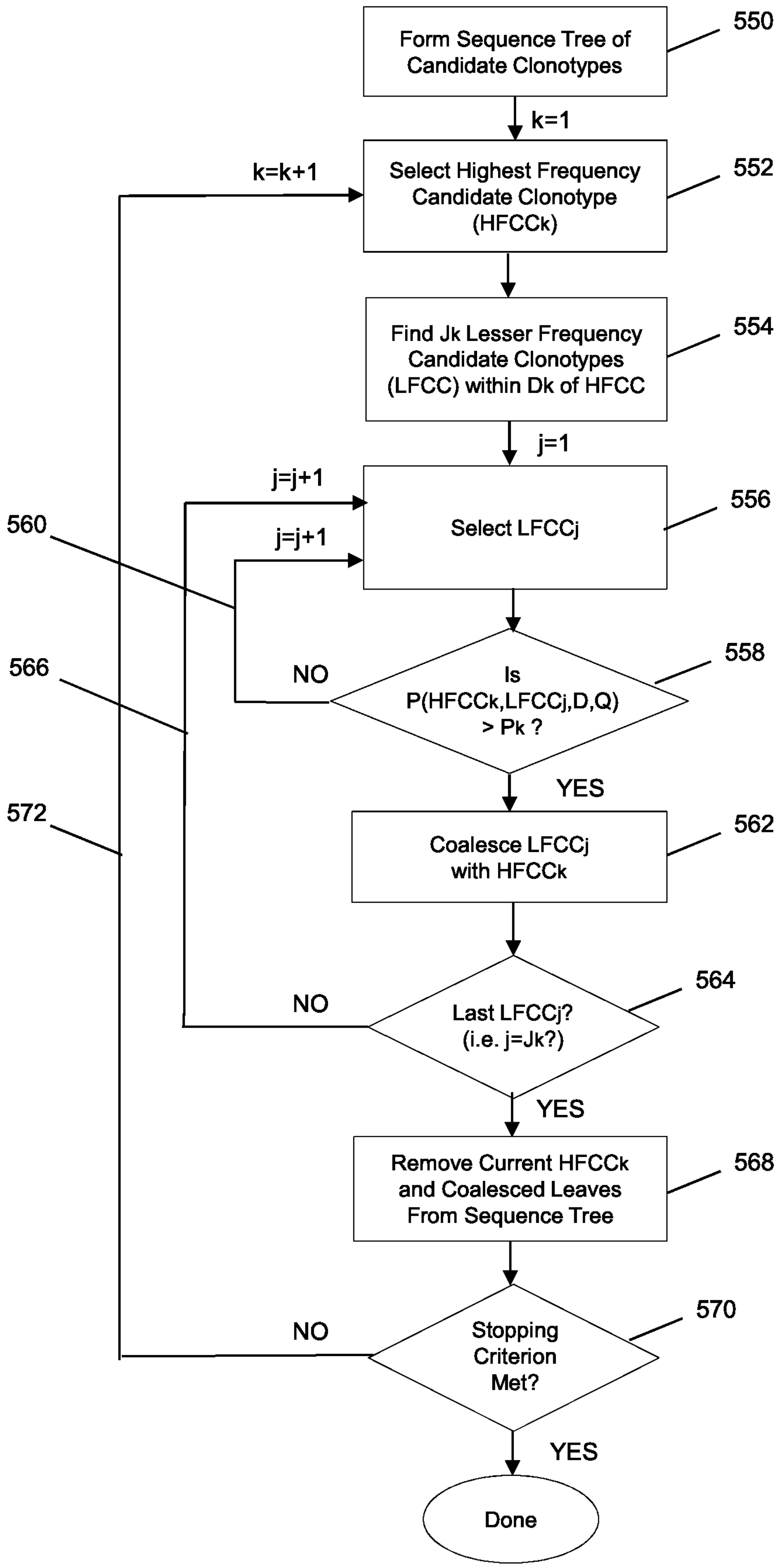

Form Sequence Tree of Candidate Clonotypes
550
k=1
Select Highest Frequency Candidate Clonotype (HFCCk)
552
k=k+1
Find Jk Lesser Frequency Candidate Clonotypes (LFCC) within Dk of HFCC
554
j=1
Select LFCCj
556
j=j+1
j=j+1
560
Is P(HFCCk,LFCCj,D,Q) > Pk ?
558
NO
YES
566
Coalesce LFCCj with HFCCk
562
572
Last LFCCj? (i.e. j=Jk?)
564
NO
YES
Remove Current HFCCk and Coalesced Leaves From Sequence Tree
568
Stopping Criterion Met?
570
NO
YES
Done

Fig. 5B

Exemplary Function For Determining Whether to Coalesce Candidate Clonotypes Depending on Read Counts, Base Differences and Q Scores

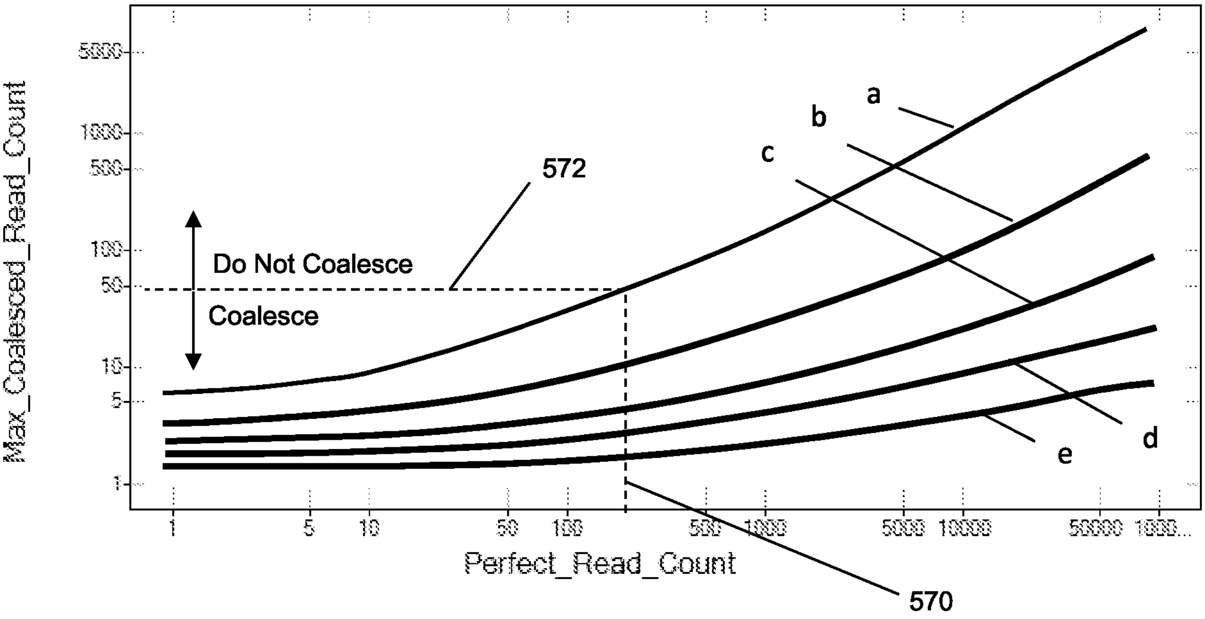


Fig. 5C

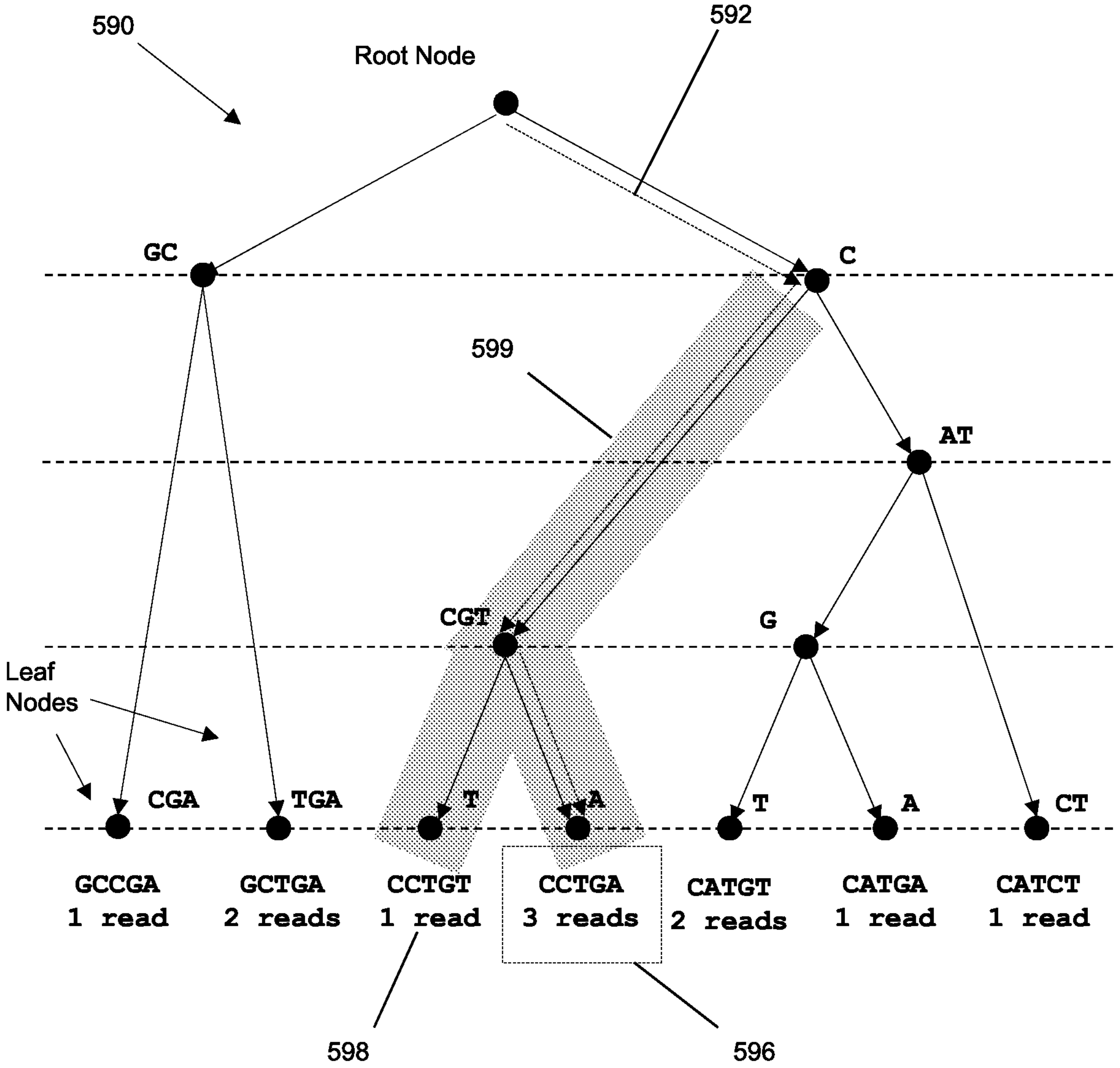

590
592
Root Node
GC
C
599
AT
CGT
G
Leaf
Nodes
CGA
TGA
T
A
T
A
CT
GCCGA
1 read
GCTGA
2 reads
CCTGT
1 read
CCTGA
3 reads
CATGT
2 reads
CATGA
1 read
CATCT
1 read
598
596

Fig. 5D

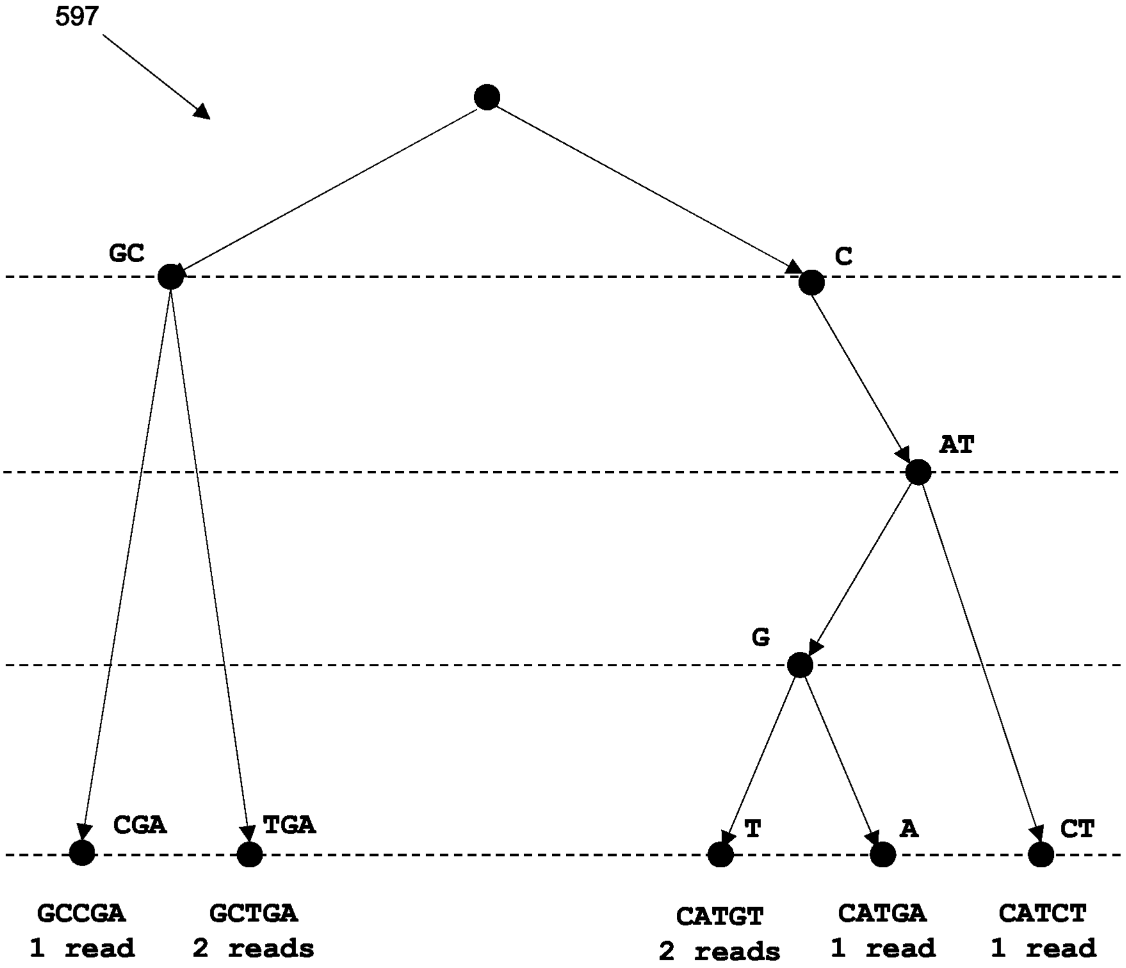

597
GC
C
AT
G
CGA
TGA
T
A
CT
GCCGA
1 read
GCTGA
2 reads
CATGT
2 reads
CATGA
1 read
CATCT
1 read

Fig. 5E

```
                         SEQUENCE LISTING

<110>  Sequenta, Inc.
       Asbury, Thomas
       Hervold, Kieran
       Kotwaliwale, Chitra
       Faham, Malek
       Moorhead, Martin
       Weng, Li
       Wittkop, Tobias
       Zheng, Jianbiao

<120>  LARGE-SCALE BIOMOLECULAR ANALYSIS WITH SEQUENCE TAGS

<130>  848US00 (37623-739.601)

<150>  61/841878
<151>  2013-07-01

<150>  62/001580
<151>  2014-05-21

<160>  6

<170>  PatentIn version 3.5

<210>  1
<211>  24
<212>  DNA
<213>  Artificial Sequence

<220>
<223>  primer

<400>  1
agttctggct aacctgtaga gcca                                              24


<210>  2
<211>  24
<212>  DNA
<213>  Artificial Sequence

<220>
<223>  primer

<400>  2
agttcgggct aacctgtcga gcca                                              24


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

<220>
<223>  primer

<400>  3
agttccggct aacctgtcga gcca                                              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
nnnnnnnnnn nnnnnnnnnn nn                                                 22


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

<220>
<223>  primer

<400>  5
gtattttttt ct                                                            12


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

<220>
<223>  primer

<400>  6
ttcagggggg gct                                                           13
```