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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 February 2001 (22.02.2001)

PCT

(10) International Publication Number
WO 01/12855 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US00/22164

(22) International Filing Date: 11 August 2000 (11.08.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/148,870 13 August 1999 (13.08.1999) US
09/544,713 6 April 2000 (06.04.2000) US

(71) Applicant: **YALE UNIVERSITY** [US/US]; 451 College Street, New Haven, CT 06511 (US).

(72) Inventors: **KAUFMAN, Joseph, C.**; 24 Farmington Road, Hamden, CT 06518 (US). **ROTH, Matthew, E.**; 20 Medley Lane, Branford, CT 06405 (US). **LIZARDI, Paul, M.**; 7 Stoney Brook Road, Wallingford, CT 06492 (US). **FENG, Li**; 1579 Rhinelander Avenue, Apartment 2T, Bronx, NY 10461 (US). **LATIMER, Darin, R.**; 123 Florence Road, Apartment 2B, Branford, CT 06405 (US).

(74) Agents: **PABST, Patrea, L.** et al.; Arnall Golden & Gregory, LLP, 2800 One Atlantic Center, 1201 West Peachtree Street, Atlanta, GA 30309-3450 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/12855 A2

(54) Title: BINARY ENCODED SEQUENCE TAGS

(57) Abstract: Disclosed is a method for the comprehensive analysis of nucleic acid samples and a detector composition for use in the method. The method, referred to as Binary Encoded Sequence Tags (BEST), involves generation of a set of nucleic acid fragments; adding an adaptor to the ends containing recognition site for cleavage at a site offset from the recognition site; cleaving the fragment to generate fragments having a plurality of sticky ends; indexing of the fragments into sets based on the sequence of sticky ends. The fragments are indexed by adding an offset adaptor to newly generated ends. A different adaptor will be coupled to each different sticky end. The resulting fragments - which will have defined ends, be of equal lengths (in preferred embodiment), and a central sequence derived from the source nucleic acid molecule - are binary sequence tags. The binary sequence tags can be used and further analyzed in numerous ways. For example, the binary sequence tags can be captured by hybridization and coupling, preferably by ligation, to a probe. The probe is preferably immobilized in an array or on sortable beads. One form of the BEST method, referred to as modification assisted analysis of binary sequence tags (MAABST), assesses modification of sequences in nucleic acid molecules by detecting differential cleavage based on the presence or absence of modification in the molecules.

BINARY ENCODED SEQUENCE TAGS

BACKGROUND OF THE INVENTION

The disclosed invention is generally in the field of nucleic acid characterization and analysis, and specifically in the area of analysis and comparison of gene expression patterns, nucleic acid samples, and genomes.

The study of differences in gene-expression patterns is one of the most promising approaches for understanding mechanisms of differentiation and development. In addition, the identification of disease-related target molecules opens new avenues for rational pharmaceutical intervention. Currently, there are two main approaches to the analysis of molecular expression patterns: (1) the generation of mRNA-expression maps and (2) examination of the 'proteome', in which the expression profile of proteins is analyzed by techniques such as two-dimensional gel electrophoresis, mass spectrometry [matrix-assisted-desorption-ionization-time-of-flight (MALDI-TOF) or electrospray] and by the ability to sequence sub-picomole amounts of protein. Classical approaches to transcript imaging, such as northern blotting or plaque hybridization, are time-consuming and material-intensive methods to analyze mRNA-expression patterns. For these reasons, other methods for high-throughput screening in industrial and clinical research have been developed.

A breakthrough in the analysis of gene expression was the development of the northern-blot technique in 1977 (Alwine et al., *Proc. Natl. Acad. Sci. U.S.A.* 74:5350-5354 (1977)). With this technique, labeled cDNA or RNA probes are hybridized to RNA blots to study the expression patterns of mRNA transcripts. Alternatively, RNase-protection assays can detect the expression of specific RNAs. These assays allow the expression of mRNA subsets to be determined in a parallel manner. For RNase-protection assays, the sequence of the analyzed mRNA has to be known in order to synthesize a labeled cDNA that forms a hybrid with the selected mRNA; such hybrids resist RNA degradation by a single-strand-specific nuclease and can be detected by gel electrophoresis. As a third approach, differential plaque-filter hybridization allows the identification of specific differences in the expression of cloned cDNAs (Maniatis *et al. Cell* 15:687-701 (1978)). Although all of these techniques are excellent tools for studying differences in gene expression, the limiting factor of

these classical methods is that expression patterns can be analyzed only for known genes.

The analysis of gene-expression patterns made a significant advance with the development of subtractive cDNA libraries, which are generated by hybridizing an mRNA pool of one origin to an mRNA pool of a different origin. Transcripts that do not find a complementary strand in the hybridization step are then used for the construction of a cDNA library (Hedrick et al., *Nature* 308:149-153 (1984)). A variety of refinements to this method have been developed to identify specific mRNAs (Swaroop et al., *Nucleic Acids Res.* 25:1954 (1991); Diatchenko *et al, Proc. Natl. Acad. Sci. U.S.A* 93:6025-6030 (1996)). One of these is the selective amplification of differentially expressed mRNAs via biotin- and restriction-mediated enrichment (SABRE; Lavery et al., *Proc. Natl. Acad. Sci. U.S.A.* 94:6831-6836 (1997)), cDNAs derived from a tester population are hybridized against the cDNAs of a driver (control) population. After a purification step specific for tester-cDNA-containing hybrids, tester-tester homohybrids are specifically amplified using an added linker, thus allowing the isolation of previously unknown genes.

The technique of differential display of eukaryotic mRNA was the first one-tube method to analyze and compare transcribed genes systematically in a bi-directional fashion; subtractive and differential hybridization techniques have only been adapted for the unidirectional identification of differentially expressed genes (Liang and Pardee, *Science* 257:967-971 (1992)). Refinements have been proposed to strengthen reproducibility, efficiency, and performance of differential display (Bauer *et al., Nucleic Acids Res.* 11:4272-4280 (1993); Liang and Pardee, *Curr. Opin. Immunol* 7:274-280 (1995); Ito and Sakaki, *Methods Mol. Biol.* 85:37-44 (1997); Praschar and Weissman, *Proc. Natl. Acad. Sci U.S.A.* 93;659-663 (1996), Shimkets *et al., Nat Biotechnol,* 17: 798-803 (1999)). Although these approaches are more reproducible and precise than traditional PCR-based differential display, they still require the use of gel electrophoresis. This often implies the exclusion of certain DNA fragments from analysis.

Originally developed to identify differences between two complex genomes, representational difference analysis (RDA) was adapted to analyze differential gene expression by taking advantage of both subtractive

hybridization and PCR (Lisitsyn et al., *Science* 259:946-951 (1993); Hubank and Schatz, *Nucleic Acids Res.* 22:5640-5648 (1994)). In the first step, mRNA derived from two different populations, the tester and the driver (control), is reverse transcribed; the tester cDNA represents the cDNA population in which differential gene expression is expected to occur. Following digestion with a frequently cutting restriction endonuclease, linkers are ligated to both ends of the cDNA. A PCR step then generates the initial representation of the different gene pools. The linkers of the tester and driver cDNA are digested and a new linker is ligated to the ends of the tester cDNA. The tester and driver cDNAs are then mixed in a 1:100 ratio with an excess of driver cDNA in order to promote hybridization between single-stranded cDNAs common in both tester and driver cDNA pools. Following hybridization of the cDNAs, a PCR exponentially amplifies only those homoduplexes generated by the tester cDNA, via the priming sites on both ends of the double-stranded cDNA (O'Neill and Sinclair, *Nucleic Acids Res.* 25:2681-2682 (1997); Wada et al., *Kidney Int.* 51:1629-1638 (1997); Edman et al., *J.* 323:113-118 (1997)).

The gene-expression pattern of a cell or organism determines its basic biological characteristics. In order to accelerate the discovery and characterization of mRNA-encoding sequences, the idea emerged to sequence fragments of cDNA randomly, direct from a variety of tissues (Adams *et al.*, *Science* 252:1651-1656 (1991); Adams *et al.*, *Nature* 377:3-16 (1995)). These expressed sequence tags (ESTs) allow the identification of coding regions in genome-derived sequences. Publicly available EST databases allow the comparative analysis of gene expression by computer. Differentially expressed genes can be identified by comparing the databases of expressed sequence tags of a given organ or cell type with sequence information from a different origin (Lee *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 92:8303-8307 (1995); Vasmatzis et al., *Proc. Natl. Acad. Sci. U. S. A.* 95:300-304 (1998)). A drawback to sequencing of ESTs is the requirement for large-scale sequencing facilities.

Serial analysis of gene expression (SAGE) is a sequence-based approach to the identification of differentially expressed genes through comparative analyses (Velculescu et al., *Science* 270:484-487 (1995)). It allows the simultaneous analysis of sequences that derive from different cell population or

tissues. Three steps form the molecular basis for SAGE: (1) generation of a sequence tag (10-14 bp) to identify expressed transcripts; (2) ligation of sequence tags to obtain concatemers that can be cloned and sequenced; and (3) comparison of the sequence data to determine differences in expression of genes that have been identified by the tags. This procedure is performed for every mRNA population to be analyzed. A major drawback of SAGE is the fact that corresponding genes can be identified only for those tags that are deposited in gene banks, thus making the efficiency of SAGE dependent on the extent of available databases. Alternatively, a major sequencing effort is required to complete a SAGE data set capable of providing 95 % coverage of any given mRNA population, simply because most of the sequencing work yields repetitive reads on those tags that are present at high frequency in cellular mRNA. In other words, SAGE sequencing experiments yield diminishing returns for rare mRNAs, whose unique tags will begin to accumulate in the database only after many weeks of sequencing effort.

A different approach to the study of gene-expression profiles and genome composition is the use of DNA microarrays. Current DNA microarrays are systematically gridded at high density. Such microarrays are generated by using cDNAs (for example, ESTs), PCR products or cloned DNA, which are linked to the surface of nylon filters, glass slides or silicon chips (Schena et al., *Science* 270, 467-470 (1995)). DNA arrays can also be assembled from synthetic oligonucleotides, either by directly applying the synthesized oligonucleotides to the matrix or by a more sophisticated method that combines photolithography and solid-phase chemical synthesis (Fodor et al., *Nature* 364:555-556 (1993)). To determine differences in gene-expression, labeled cDNAs or oligonucleotides are hybridized to the DNA- or oligomer-carrying arrays. When using different fluorophores for labeling cDNAs or oligonucleotides, two probes can be applied simultaneously to the array and compared at different wavelengths. The expression of 10,000 genes and more can be analyzed on a single chip (Chee *et al.*, *Science* 274:610-614 (1996)). However, depending on the sensitivity of both cDNA and oligonucleotide arrays, the intensity of hybridization signals can leave the linear range when either weakly or abundantly expressed genes are analyzed. Thus, individual

optimization steps are required to ensure the accurate detection of differentially expressed genes. While such microarray methods may be used to address a number of interesting biological questions, they are not suitable for the discovery of new genes.

Techniques of tagging DNA fragments using sticky end-specific adaptors have been described by Burger and Schinzel, *Mol. Gen.Genet.* 189:269-274 (1983), Mandecki and Bolling, *Gene,* 68:101-107 (1988), Posfai and Szybalski, *Gene,* 74:179-181 (1988), Urlaub *et al., Proc. Natl. Acad. Sci.,* 82:1189-1193 (1985), Vermesch and Bennett, *Gene,* 54:229-238 (1987) , Unrau and Deugau, *Gene,* **145**(2):163-9 (1994)). These techniques all involve the use of existing restriction sites and produce tagged fragments of various lengths.

There is a need for a method that combines the power and convenience of array hybridization technology with the capability for gene discovery inherent in differential display or SAGE. Such a method would be most attractive if it could enable comprehensive gene expression analysis without the use of gel electrophoresis, and without the need for a redundant DNA sequencing effort.

Therefore, it is an object of the present invention to provide a method for the comprehensive analysis of nucleic acid sequence tags.

It is another object of the present invention to provide a detector composition that allows indexing of nucleic acid sequence tags.

It is another object of the present invention to provide catalogs of sequence tags from nucleic acid samples.

BRIEF SUMMARY OF THE INVENTION

Disclosed is a method for the comprehensive analysis of nucleic acid samples and a detector composition for use in the method. The method, referred to as Binary Encoded Sequence Tags (BEST), involves generation of a set of nucleic acid fragments; adding an adaptor to the ends containing a recognition site for cleavage at a site offset from the recognition site; cleaving the fragment to generate fragments having a plurality of sticky ends; indexing of the fragments into sets based on the sequence of sticky ends. Multiple sticky end sequences are generated by virtue of offset cleavage using the recognition site added as part of the adaptor. Preferably this is accomplished by subjecting the nucleic acid sample to digestion by a restriction endonuclease that cleaves at a

site different from the site of the recognition sequence. The fragments are indexed by adding an offset adaptor to newly generated ends. A different adaptor will be coupled to each different sticky end. The resulting fragments—which will have defined ends, are of equal lengths (in a preferred embodiment), and a central sequence derived from the source nucleic acid molecule—are binary sequence tags. The binary sequence tags can be used and further analyzed in numerous ways. For example, the binary sequence tags can be captured by hybridization and coupling, preferably by ligation, to a probe. The probe is preferably immobilized in an array or on sortable beads. The disclosed method differs from prior methods at least since the present method introduces an offset cleavage site into target nucleic fragment. This has the advantage that sets of sequence tags are generated that have defined lengths.

The method allows detection of the binary sequence tags where detection provides some sequence information for the tags including the sequence of the generated sticky end of each fragment, the recognition sequence of the nucleic acid cleaving reagent—preferably a restriction endonuclease—used to initially cleave nucleic acid molecules, and the central sequence of the tag. The set of binary sequence tags produced from a nucleic acid sample using particular nucleic acid cleaving reagents and adaptors will produce characteristic sets of binary sequence tags. The method allows a complex sample of nucleic acid to be cataloged quickly and easily in a reproducible and sequence-specific manner. The disclosed method also should produce two binary sequence tags for each cleavage site in the nucleic acid sample. This can allow comparisons and validation of a set of binary sequence tags.

One form of the BEST method, referred to as modification assisted analysis of binary sequence tags (MAABST), assesses modification of sequences in nucleic acid molecules by detecting differential cleavage based on the presence or absence of modification in the molecules. For example, a site that is methylated in a nucleic acid molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is insensitive to methylation will cleave at that site, thus producing a different pattern of binary sequence tags.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-1C are a listing of examples of ligator-detectors (numbered sequences) designed for use with one of two example adaptor-indexers (top). The sticky end sequences (or their complements) are shown in bold.

DETAILED DESCRIPTION OF THE INVENTION

The disclosed method, referred to as Binary Encoded Sequence Tags (BEST), allows a complex sample of nucleic acid to be quickly and easily cataloged in a reproducible and sequence-specific manner. Such a catalog can be compared with other, similarly prepared catalogs of other nucleic acid samples to allow convenient detection of differences between the samples. The catalogs, which incorporate a significant amount of information about the nucleic acid samples, can serve as fingerprints of the nucleic acid samples which can be used both for detection of related nucleic acid samples and comparison of nucleic acid samples. For example, the presence or identity of specific organisms can be detected by producing a catalog of nucleic acid of the test organism and comparing the resulting catalog with reference catalogs prepared from known organisms. Changes and differences in gene expression patterns can also be detected by preparing catalogs of mRNA from different cell samples and comparing the catalogs. The catalog of sequences can also be used to produce a set of probes or primers that is specific for the source of a nucleic acid sample.

Comparison of nucleic acid catalogs produced with the disclosed method is facilitated by the ordered nature of the sequence information produced and cataloged in the method. Use of immobilization, sorting, and/or array detection in the method allows automation of the method, the cataloging of the information, and comparisons to other catalogs. The method can result in the equivalent of a large number of sequence-specific bins that can be filled, empty, or filled to different levels, with the pattern of filled and empty bins, and/or of the amount of signal in a bin, providing information about the nucleic acid sample that has been cataloged. There is no need to assemble overlapping sequence strings into larger sequences (although this can be done). Rather, the individual sequences detected are themselves data points in the catalog.

RECTIFIED SHEET (RULE 91)
ISA/EP

The BEST method involves the following basic steps. A nucleic acid sample is incubated with one or more nucleic acid cleaving reagents, preferably restriction endonucleases, that results in a set of DNA fragments cleaved at particular sites. The sample is then mixed with one or more offset adaptors, each of which has a recognition sequence for a nucleic acid cleaving reagent that cleaves at a site offset from the recognition sequence. The offset adaptors are then covalently coupled, preferably by ligation, onto the DNA fragments. The offset adapters should have ends compatible with the ends of the nucleic acid fragments. Coupling, both here and in other steps, can be accomplished using any suitable technique, including ligation and chemical reactions. Ligation is preferred. When coupling is by ligation, there should be a 5'-phosphate capable of participating in ligation on the appropriate strand.

The nucleic acid sample is incubated with one or more nucleic acid cleaving reagents, preferably restriction endonucleases, that cleave the nucleic acid fragments using the recognition sequence of the offset adaptors and that generate fragments having sticky ends with a variety of sequences. The fragments are preferably cleaved at the same distance from the recognition sequence. In one form of the method, the sample can be divided into aliquots (referred to as index samples); preferably as many aliquots as there are sticky end sequences. Where multiple restriction endonucleases are used, the nucleic acid sample is preferably divided into index samples before cleavage. Where a single restriction endonuclease is used, the nucleic acid sample is preferably divided into index samples following cleavage. The nucleic acid samples can also be divided into index samples following addition of the offset adaptor or adaptor-indexer. Index samples themselves can be further divided into secondary index samples.

Each sample (each index sample if the nucleic acid sample was divided) is then mixed with one or more adaptor-indexers, each of which has a sticky end compatible with one of the possible sticky ends on the DNA fragments in that index sample. It is preferred that a different adaptor-indexer be mixed with each index sample. The adaptor-indexers are then covalently coupled, preferably by ligation, onto compatible DNA fragments. The resulting nucleic acid fragments are binary sequence tags.

The binary sequence tags can be analyzed in a variety of ways. For example, the binary sequence tags can be amplified, detected, identified, sequenced, cataloged, or a combination. Preferably the binary sequence tags are detected, preferably by determining, directly or indirectly, the presence, amount, presence and amount, or absence of one or more binary sequence tags. Numerous techniques and methods are known for the analysis of nucleic acid fragments which are suitable for analysis of binary sequence tags.

A preferred form of binary sequence tag analysis is indexed probe hybridization. This can be accomplished by hybridizing the binary sequence tags in each sample (or index sample) to ligator-detectors. One end of each ligator-detector has sequence matching or complementary to all or part of one of the possible sticky end sequences generated by the second nucleic acid cleaving reagent. The ligator-detector can, and preferably does, have sequence matching or complementary to all or part of the sequence adjacent to the sticky end sequence in the fragment coupled to the adaptor-indexer. The ligator-detector used in each index sample preferably matches or is complementary to all or part of the sequence, including sticky end sequence, in the adaptor-indexer sequence used in that index sample. Each sample (or index sample) is reacted with and coupled, preferably by ligation, to one or more detector probes. Preferably, the set of detector probes used include every possible sequence of a given length (for example, every possible six base sequence). The ends of the probes and the ligator-detectors are coupled only if the probe hybridizes adjacent to the end of the ligator-adaptor. The probes are preferably immobilized oligonucleotides.

Each binary sequence tag processed through indexed probe hybridization will result in a signal based on coupling of the ligator-detector to a probe. A complex nucleic acid sample will produce a unique pattern of signals. It is this pattern that allows unique cataloging of nucleic acid samples and sensitive and powerful comparisons of the patterns of signals produced from different nucleic acid samples.

Coupling of ligator-detectors to probes can be detected directly or indirectly. For example, any of the probe, the ligator-detector, or the associated adaptor-indexer or offset adaptor can be detected. Association of a ligator-detector, adaptor-indexer, or offset adaptor with a given probe is indicative of

coupling of the probe and ligator-detector. Detection of such associations can be facilitated through immobilization of the probes, detector-ligators, adaptor-indexers, or offset adaptors, and through the use of capture tags, sorting tags and detectable labels in association with the probes, detector-ligators, adaptor-indexers and/or offset adaptors. Any combination of immobilization and association with capture tags, sorting tags, and labels can be used. Preferably, the probes are immobilized in arrays and the ligator-detectors are associated with a detectable label. Thus, detection of a signal at a particular location in a particular array of immobilized probes can provide information about nucleic acid fragments indexed from the nucleic acid sample.

Where the probes are immobilized in arrays, the array, and location in the array, where a DNA fragment generates a signal identify the sequence of the sticky end of the DNA fragment and of the sequence adjacent to the sticky end. This is a ten base sequence when a four base sticky end and six base immobilized probes are used. The fixed relationship between the recognition sequence and the cleavage site of the Type IIS restriction enzyme (when used) and the identity of the recognition sequence, provide additional sequence information about the DNA fragment. The same effect can be accomplished by otherwise capturing, sorting, or detecting particular probes (via capture tags, sorting tags, and labels). That is, so long as the probe and the ligator-detector coupled to it can be identified, a pattern can be determined.

Binary sequence tags produced using the disclosed method generally occur in correlated pairs, except in a few cases where the initial cleavage site occurs near the terminus of a DNA substrate. Analysis of the tag catalogs using the disclosed method can reveal, within certain confidence limits, the identity of a subset of the correlated pairs. The identification of the subset of correlated pairs provides additional sequence information about the catalog of tags.

One form of the BEST method, referred to as modification assisted analysis of binary sequence tags (MAABST), assesses modification of sequences in nucleic acid molecules by detecting differential cleavage based on the presence or absence of modification in the molecules. For example, a site that is methylated in a nucleic acid molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is

insensitive to methylation will cleave at that site, thus producing a different pattern of sequence tags.

Materials

Nucleic Acid Samples

Any nucleic acid sample can be used with the disclosed method. Examples of suitable nucleic acid samples include genomic samples, mRNA samples, cDNA samples, nucleic acid libraries (including cDNA and genomic libraries), whole cell samples, environmental samples, culture samples, tissue samples, bodily fluids, and biopsy samples. Numerous other sources of nucleic acid samples are known or can be developed and any can be used with the disclosed method. Preferred nucleic acid samples for use with the disclosed method are nucleic acid samples of significant complexity such as genomic samples and mRNA samples.

Nucleic acid fragments are segments of larger nucleic molecules. Nucleic acid fragments, as used in the disclosed method, generally refer to nucleic acid molecules that have been cleaved. A nucleic acid sample that has been incubated with a nucleic acid cleaving reagent is referred to as a digested sample. A nucleic acid sample that has been digested using a restriction enzyme is referred to as a digested sample.

An index sample is a nucleic acid sample that has been divided into different aliquots for further processing. In the context of the disclosed method, index samples are preferably aliquots of a digested nucleic acid sample to which different adaptor-indexers are added for coupling, preferably by ligation, to nucleic acid fragments present in the digested sample. In the disclosed method, different nucleic acid fragments are processed in the different index samples based on the sticky end sequence of the fragments. Thus, it is preferred that digested nucleic acid samples be divided into as many index samples as the number of possible sticky end sequences generated by the nucleic acid cleaving reagent used to digest the sample. Where multiple different nucleic acid cleaving reagents are used to cleave a nucleic acid sample, it is preferred that the nucleic acid sample be divided into as many aliquots as nucleic acid cleaving reagents used and that the nucleic acid sample be divided prior to cleavage. Where multiple restriction endonucleases are used, the nucleic acid sample is

preferably divided into index samples before cleavage. Where a single restriction endonuclease is used, the nucleic acid sample is preferably divided into index samples following cleavage. The nucleic acid samples can also be divided into index samples following addition of the offset adaptor or adaptor-indexer. Index samples themselves can be further divided into secondary index samples.

A control nucleic acid sample is a nucleic acid sample to which another nucleic acid sample (which can be referred to as a tester nucleic acid sample) is to be compared. A control index sample is an index sample to which another index sample (which can be referred to as a tester index sample) is to be compared.

Nucleic Acid Cleaving Reagents

The disclosed method makes use of nucleic acid cleaving reagents. Nucleic acid cleaving reagents are compounds, complexes, and enzymes that cause, mediate, or catalyze cleavage on nucleic acid molecules. Preferred nucleic acid cleaving reagents are those that cleave nucleic acid molecules in a sequence-specific manner. Restriction enzymes (also referred to as restriction endonucleases) are the preferred form of nucleic acid cleaving reagents. Other nucleic acid cleaving reagents include the universal restriction endonucleases of Szybalski (Szybalski, *Gene* **40**(2-3):169-73 (1985); Podhajska and Szybalski, *Gene* **40**(2-3):175-82 (1985)[published erratum appears in *Gene* 43(3):325 (1985)]), the advanced DNA cleavage systems developed by Breaker et al. (Carmi et al., *Proc Natl Acad Sci USA* **95**(5):2233-2237 (1998)), and the use of zinc fingers to direct site recognition of restriction enzymes such as the hybrid restriction enzymes described by Kim et al., *Proc. Natl. Acad. Sci. USA* **93**(3):1156-1160 (1996), and Smith et al., *Nucleic Acids Res.* **27**(2):674-681 (1999).

Many nucleic acid cleaving reagents are known and can be used with the disclosed method. Relevant to the disclosed method, nucleic acid cleaving reagents generally have a recognition sequence and a cleavage site. Many nucleic acid cleaving reagents, especially restriction enzymes, also generate sticky ends at the cleavage site. A recognition sequence is the nucleotide sequence which, if present in a nucleic acid molecule, will direct cleavage of the

nucleic acid molecule by a cognate nucleic acid cleaving reagent. The cleavage site of a nucleic acid cleaving reagent is the site, usually in relation to the recognition sequence, where the nucleic acid cleaving reagent cleaves a nucleic acid molecule. Sticky ends (also referred to as cohesive ends, protruding ends, and 5' or 3' overhangs) are single-stranded nucleic acid segments at the end of a double-stranded nucleic acid segment.

For specific embodiments of the method, the nucleic acid cleaving reagents used will have certain properties and/or certain relationships to other restriction enzymes used in the method. For example, in some preferred embodiments of the disclosed method, nucleic acid cleaving reagents that generate sticky ends having a plurality of different sequences are preferred, with nucleic acid cleaving reagents having a cleavage site offset from the recognition sequence being most preferred. Other embodiments of the disclosed method require the use of different nucleic acid cleaving reagents that have different recognition sequences and/or generate different sticky ends than other nucleic acid cleaving reagents used on the same index sample at other stages in the method. For example, where multiple digests (that is, cleavage reactions) are used in the method, it is preferred that the nucleic acid cleaving reagents used in each of the digests have a recognition sequence different from that of the nucleic acid cleaving reagents used in the other digests. In such cases, the known properties of nucleic acid cleaving reagents can be used to select or design appropriate nucleic acid cleaving reagents.

Where a nucleic acid cleaving reagent cleaves DNA at a site different or offset from the recognition sequence, a variety of sticky ends having different sequences can be generated. This is because recognition sequences in nucleic acids can occur next to any sequence and therefore the site of cleavage can have any sequence. For example, FokI cleaves 9 (upper strand) and 13 (lower strand) nucleotides downstream from the recognition site of GGATG. The four base sticky end will have whatever sequence happens to be 10 to 13 nucleotides away from the recognition site. Given enough cleavage sites, a total of 256 different sticky end sequences (that is every possible four base sequence) can result from a FokI digestion. As a result, restriction enzymes such as Type IIS restriction

enzymes can be said to generate sticky ends having a plurality of different sequences.

As used herein, unless otherwise indicated, the terms digest, digestion, digested, and digesting refer generally to a cleavage reaction or the act of cleaving and is not intended to be limited to cleavage by a protein enzyme or by any particular mechanism. Similarly, the term restricted is intended to refer to any nucleic acid cleavage, not just cleavage by a restriction enzyme. In the context of nucleic acid cleaving reagents, sequence-specific requires only some sequence specificity, not absolute sequence specificity. That is, nucleic acid cleaving reagents having a completely or partially defined recognition sequence are preferred. Thus, nucleic acid cleaving reagents having some degeneracy in their recognition sequence are still considered sequence-specific.

A first nucleic acid cleaving reagent is a nucleic acid cleaving reagent used first to digest a nucleic acid sample. A second nucleic acid cleaving reagent is a nucleic acid cleaving reagent used to digest a fragment to which an offset adaptor has been coupled. First nucleic acid cleaving reagents are preferably Type II restriction endonucleases that cleave within the recognition sequence. Second nucleic acid cleaving reagents are preferably Type IIS restriction enzymes.

In addition to the use of restriction enzymes in a standard mode, the Type IIS enzymes can be used as universal restriction endonuclease as described by Szybalski (Szybalski, Gene **40**(2-3):169-73 (1985); Podhajska and Szybalski, Gene **40**(2-3):175-82 (1985)[published erratum appears in Gene 43(3):325 (1985)]). In the Szybalski technique single-stranded or double-stranded DNA can be cleaved at any arbitrary (but specific) site utilizing the structure described in combination with a Type IIS enzyme. More advanced DNA cleavage systems developed by Breaker et al. (Carmi et al., Proc Natl Acad Sci U S A **95**(5):2233-2237 (1998)). In these systems Breaker has shown that DNA recognize a particular sequence in a target DNA and can cleave the target DNA, single-stranded or double-stranded targets. With Breaker's system for evolution of DNA for a particular action, it is clear that given reasonable time and effort a suitable DNA for a recognition and particular cleavage result is practical.

Offset Adaptors

Offset adaptors are double-stranded nucleic acids that contain a recognition site for a nucleic acid cleaving reagent that cleaves at a site offset from the recognition site. Offset adaptors preferably contain a single-stranded portion and a double-stranded portion. The single-stranded portion is at one end of the offset adaptor and constitutes a sticky end. It is preferable that the protruding single strand (sticky end) have two, three, four, or five nucleotides. The double-stranded portion of offset adaptor may have any convenient sequence or length. In general, the sequence and length of the double-stranded portion is selected to be adapted to subsequent steps in the method. For example, sequences in the offset adaptor may be used for primer or probe hybridization. If the fragments in the samples to which adaptors have been ligated are to be amplified, the offset adaptors can provide sequence for primer hybridization. Thus, preferred sequence composition and length for the double-stranded portion of offset adaptors will generally be those that are useful for primer hybridization.

It is preferred that offset adaptors not have any sequences that are self-complementary. It is considered that this condition is met if there are no complementary regions greater than six nucleotides long without a mismatch or gap. A set of offset adaptors for use in the disclosed method can include different offset adaptors where the single-stranded portions each have a different nucleotide sequence compatible with a sticky end sequence generated by one of the first restriction enzymes. It is preferable that the members of a set of offset adaptors contain a double-stranded portion which is identical for each member of the set.

Offset adaptors can also contain or be associated with capture tags to facilitate immobilization or capture of fragments to which offset adaptors have been coupled. Offset adaptors can also contain or be associated with sorting tags to facilitate sorting or separation of fragments to which offset adaptors have been coupled. Offset adaptors can also contain or be associated with labels to facilitate detection of fragments to which offset adaptors have been coupled. Offset adaptors can also be immobilized on a substrate.

Offset adaptors can also include a protruding end at the end opposite the sticky end. Such an end can be used as, for example, a hybridization target for a label to be associated with the offset adaptor (and thus can be considered the detection portion of the offset adaptor). The two strands of an offset adaptor can be used separately in the disclosed method. For example, the two strands of the offset adaptor can be coupled to a nucleic acid fragment separately. Offset adaptors can also include one or more photocleavable nucleotides to facilitate release of adaptor-indexer sequences for detection. Photocleavable nucleotides and their use are described in WO 00/04036.

Offset adaptors need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural bases and nucleotide and oligonucleotide analogs can be used. All that is required is that the offset adaptor have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

Adaptor-Indexers

Adaptor-indexers are double-stranded nucleic acids containing a single-stranded portion and a double-stranded portion. The single-stranded portion is at one end of the adaptor-indexer and constitutes a sticky end. The sticky end is referred to as the sticky end portion of the adaptor-indexer. It is preferable that the protruding single strand (sticky end) have two, three, four, or five nucleotides. The double-stranded portion of adaptor-indexers may have any convenient sequence or length. In general, the sequence and length of the double-stranded portion is selected to be adapted to subsequent steps in the method. For example, sequences in the adaptor-indexer may be used for primer or probe hybridization. A preferred purpose of adaptor-indexers is to provide sequence for hybridization by a ligator-detector. If the fragments in the samples to which adaptors have been coupled are to be amplified, the adaptor-indexers can also provide sequence for primer hybridization (which can overlap or be contiguous with sequence for ligator-detector hybridization). Thus, preferred sequence composition and length for the double-stranded portion of adaptor-indexers will generally be those that are useful for probe and primer hybridization. Adaptor-indexers can also include a detector portion which is designed to facilitate detection of the adaptor-indexer. The detection portion can

be, for example, a sequence that is a hybridization target or it can be a label or tag.

Generally, the sequence of the double-stranded portion of an adaptor-indexer should not include the recognition sequence of any restriction enzyme to be used in a subsequent step in the method. It is preferred that adaptor-indexers not have any sequences that are self-complementary. It is considered that this condition is met if there are no complementary regions greater than six nucleotides long without a mismatch or gap.

A set of adaptor-indexers for use in the disclosed method should include different adaptor-indexers where the single-stranded portion each have a different nucleotide sequence selected from combinations and permutations of the nucleotides A, C, G, and T. Where multiple nucleic acid cleaving reagents are used in the first digest, the single-stranded portion of each adaptor-indexer can have a different nucleotide sequence compatible with a sticky end sequence generated by one of the nucleic acid cleaving reagents. While the sticky ends of adaptor-indexers in one set have different sequences, it is preferred that they be of the same length to facilitate use of the set to index fragments produced by cleavage by one nucleic acid cleaving reagent. It is preferable that the members of a set of adaptor-indexers contain a double-stranded portion which is identical for each member of the set. However, members of a set of adaptor-indexers can also have double-stranded portions that differ in some way. Similarly, some of the adaptor-indexers in a set can have identical double-stranded portions while others in the same set have different double-stranded portions. Different configurations of the set can be used to produce different types of data or probe particular relationships between tags in the disclosed method.

A preferred set of indexing linker strands comprising: (a) at least two single-stranded first oligonucleotides each having a common identical sequence, and a unique sequence of a length selected from 2, 3, 4 and 5 nucleotides selected from permutations and combinations of A, G, C and T nucleotides, at one end selected from a 3' end and a 5' end; and (b) a single-stranded second oligonucleotide whose sequence is complementary to the common sequence of the first oligonucleotides such that, when hybridized with any one of the first

oligonucleotides, a double-stranded adaptor-indexer would result which includes an end having a sticky end with a unique sequence.

Adaptor-indexers can also contain or be associated with capture tags to facilitate immobilization or capture of fragments to which adaptor-indexers have been coupled. In general, the capture tag can be one member of a binding pair such as biotin and streptavidin. Capture tags are discussed more fully elsewhere herein. Adaptor-indexers can also contain or be associated with sorting tags to facilitate sorting or separation of fragments to which adaptor-indexers have been coupled. In general, the sorting tag can be a detectable label such as a fluorescent moiety or a manipulatable moiety such as a magnetic bead. Sorting tags are discussed more fully elsewhere herein. Adaptor-indexers can also contain or be associated with labels to facilitate detection of fragments to which adaptor-indexers have been coupled. Adaptor-indexers can also be immobilized on a substrate.

Adaptor-indexers can also include a protruding end at the end opposite the sticky end. Such an end can be used as, for example, a hybridization target for a label to be associated with the adaptor-indexer (and thus can be considered the detection portion of the adaptor-indexer). The two strands of an adaptor-indexer can be used separately in the disclosed method. For example, the two strands of the adaptor-indexer can be coupled to a nucleic acid fragment separately. Adaptor-indexers can also include one or more photocleavable nucleotides to facilitate release of adaptor-indexer sequences for detection. Photocleavable nucleotides and their use are described in WO 00/04036.

Adaptor-indexers need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural bases and nucleotide and oligonucleotide analogs can be used. All that is required is that the adaptor-indexer have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

Ligator-Detectors

A ligator-detector is a nucleic acid molecule containing a single-stranded region that is complementary to a portion of a binary sequence tag generated in the disclosed method from a nucleic acid sample. The ligator-detectors generally have a specific sequence relationship to adaptor-indexers or offset

adaptors. Ligator-detectors preferably include sequence—referred to as the detector portion of the ligator-detector—matching or complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers. Thus, the ligator-detector can have sequence matching or complementary to sequence in the nucleic acid fragment adjacent to the sticky end sequence (on either or both sides), matching or complementary to the sticky end, or sequence matching or complementary to both sequence in the nucleic acid fragment adjacent to the sticky end sequence and the sticky end.

Preferably, the sequence of a ligator-detector matches or is complementary to all or part of a sticky end sequence and all or part of the adjacent sequence of the adaptor-indexer designed for use with that sticky end sequence. In this form, the sequence of a ligator-detector matches or is complementary to all or part of the recognition sequence of the first restriction enzyme(s) when cleavage is not offset from the recognition sequence. Alternatively, the ligator-detector can include sequence matching or complementary to sequence in the nucleic acid fragment adjacent to the sticky end sequence on the offset adaptor side. Whether the sequence in the ligator-detector is matching or complementary determines which strand of the adaptor-indexer and/or fragment will hybridize to the detector-ligator. It is preferred that only one type of ligator-detector—matching or complementary—is used in a given reaction of the disclosed method.

Some examples of sequence relationships between adaptor-indexers and ligator detectors are illustrated in Figures 1A-C. Ligator-detectors 1-12 in Figures 1A-C are designed to match all or part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 13-19 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 21-32 are designed to match all or part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 33-40 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Note that the part of the adjacent sequence of the adaptor-indexer embodied in the ligator-detector is contiguous

RECTIFIED SHEET (RULE 91)
ISA/EP

with the part of the sticky end sequence embodied in the ligator-detector. This is what is meant by adjacent.

Ligator-detectors 1-4 in Figures 1A-1C are designed to match all of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 5-12 are designed to match part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 2-4 and 8-12 are designed to match all or part of the sticky end sequence of adaptor-indexer I and part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 13-16 are designed to be complementary to all of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 17-19 are designed to be complementary to part of the sticky end sequence of adaptor-indexer I and all or part of the adjacent sequence of adaptor-indexer I. Ligator-detectors 14-16 are designed to be complementary to all or part of the sticky end sequence of adaptor-indexer I and part of the adjacent sequence of adaptor-indexer I. Ligator-detector 20 is designed to match all of the recognition sequence of the restriction enzyme (which generates a sticky end compatible with the sticky end of adaptor-indexer I), and all of the adjacent sequence of adaptor-indexer I. Note the extra nucleotide extending beyond the adaptor-indexer sticky end sequence. This is a flanking nucleotide in the recognition sequence.

Ligator-detectors 21-24 are designed to match all of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 25-32 are designed to match part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 22-24 and 28-31 are designed to match all or part of the sticky end sequence of adaptor-indexer II and part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 33-36 are designed to be complementary to all of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 37-40 are designed to be complementary to part of the sticky end sequence of adaptor-indexer II and all or part of the adjacent sequence of adaptor-indexer II. Ligator-detectors 34-36 and 40 are designed to be complementary to all or part

RECTIFIED SHEET (RULE 91)
ISA/EP

of the sticky end sequence of adaptor-indexer II and part of the adjacent sequence of adaptor-indexer II.

Where the nucleic acid cleaving reagents used in the first digest cleave within the recognition sequence for the nucleic acid cleaving reagent such that the recognition sequence extends beyond the sticky end sequence, the ligator-detector can also match or be complementary to all or part of the recognition sequence. Where the recognition sequence extends beyond the sticky end sequence (for example, six-base recognition sequence and four-base sticky end), the ligator-detector sequence can extend beyond the sticky end sequence of its cognate adaptor-indexer. An example of such a ligator-detector is illustrated in Figures 1A-1C (ligator-detector number 20).

While the ligator-detector can be detected using sequence-based detection systems, the ligator-detectors can also contain a label to facilitate detection of the ligator-detector. Numerous labels are known and can be used for this purpose. Ligator-detectors can also contain or be associated with capture tags to facilitate immobilization or capture of the ligator-detectors. Ligator-detectors can also contain or be associated with sorting tags to facilitate sorting or separation of the ligator-detectors. Ligator-detectors can also be immobilized on a substrate.

Ligator-detectors can also include one or more photocleavable nucleotides to facilitate release of ligator-detector sequences for detection. Photocleavable nucleotides and their use are described in WO 00/04036.

Ligator-detectors need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the ligator-detector have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

Adaptor-indexer I in Figures 1A-C is composed of nucleotides 2-25 of SEQ ID NO:1 (top strand) and nucleotides 1-20 of SEQ ID NO:2 (bottom strand). Adaptor-indexer II is composed of nucleotides 5-24 of SEQ ID NO:3 (top strand) and SEQ ID NO:4 (bottom strand). Ligator-detector 1 is nucleotides 2-25 of SEQ ID NO:1. Ligator-detector 2 is nucleotides 2-24 of SEQ ID NO:1. Ligator-detector 3 is nucleotides 2-21 of SEQ ID NO:1. Ligator-detector 4 is

RECTIFIED SHEET (RULE 91)
ISA/EP

nucleotides 2-18 of SEQ ID NO:1. Ligator-detector 5 is nucleotides 3-25 of SEQ ID NO:1. Ligator-detector 6 is nucleotides 4-25 of SEQ ID NO:1. Ligator-detector 7 is nucleotides 5-25 of SEQ ID NO:1. Ligator-detector 8 is nucleotides 3-24 of SEQ ID NO:1. Ligator-detector 9 is nucleotides 3-23 of SEQ ID NO:1. Ligator-detector 10 is nucleotides 3-20 of SEQ ID NO:1. Ligator-detector 11 is nucleotides 3-18 of SEQ ID NO:1. Ligator-detector 12 is nucleotides 4-22 of SEQ ID NO:1. Ligator-detector 13 is SEQ ID NO:2. Ligator-detector 14 is nucleotides 2-24 of SEQ ID NO:2. Ligator-detector 15 is nucleotides 4-24 of SEQ ID NO:2. Ligator-detector 16 is nucleotides 9-24 of SEQ ID NO:2. Ligator-detector 17 is nucleotides 1-23 of SEQ ID NO:2. Ligator-detector 18 is nucleotides 1-22 of SEQ ID NO:2. Ligator-detector 19 is nucleotides 1-21 of SEQ ID NO:2. Ligator-detector 20 is SEQ ID NO:1. Ligator-detector 21 is SEQ ID NO:3. Ligator-detector 22 is nucleotides 1-23 of SEQ ID NO:3. Ligator-detector 23 is nucleotides 1-21 of SEQ ID NO:3. Ligator-detector 24 is nucleotides 1-15 of SEQ ID NO:3. Ligator-detector 25 is nucleotides 2-24 of SEQ ID NO:3. Ligator-detector 26 is nucleotides 3-24 of SEQ ID NO:3. Ligator-detector 27 is nucleotides 4-24 of SEQ ID NO:3. Ligator-detector 28 is nucleotides 2-23 of SEQ ID NO:3. Ligator-detector 29 is nucleotides 2-21 of SEQ ID NO:3. Ligator-detector 30 is nucleotides 2-19 of SEQ ID NO:3. Ligator-detector 31 is nucleotides 2-16 of SEQ ID NO:3. Ligator-detector 32 is nucleotides 3-24 of SEQ ID NO:3. Ligator-detector 33 is SEQ ID NO:4. Ligator-detector 34 is nucleotides 2-24 of SEQ ID NO:4. Ligator-detector 35 is nucleotides 4-24 of SEQ ID NO:4. Ligator-detector 36 is nucleotides 9-24 of SEQ ID NO:4. Ligator-detector 37 is nucleotides 1-23 of SEQ ID NO:4. Ligator-detector 38 is nucleotides 1-22 of SEQ ID NO:4. Ligator-detector 39 is nucleotides 1-21 of SEQ ID NO:4. Ligator-detector 40 is nucleotides 5-23 of SEQ ID NO:4.

Detector Probes

Detector probes are molecules, preferably oligonucleotides, that can hybridize to nucleic acids in a sequence-specific manner. In the disclosed method, detector probes are used to capture ligator-detectors based on complementary sequences present in sample nucleic acid fragments to which the ligator-detectors are hybridized. Detector probes are preferably used in sets

having a variety of probe sequences, preferably a set of probes having every possible combination (or hybridizable to every combination) of nucleotide sequence the length of the probe. Detector probes are preferably used in sets where each probe has the same length. Preferred lengths for the probe portion of detector probes are five, six, seven, and eight nucleotides. Detector probes preferably include a probe portion (for hybridization to sample fragments) and linker portions through which the probe portion is coupled to a substrate, capture tag, sorting tag, or label. These linker portions can have any suitable structure and will generally be chosen based on the method of immobilization or synthesis of the detector probes. The linker portion can be made up of or include nucleotides. The linker portions can have any suitable length and preferably are of sufficient length to allow the probe portion to hybridize effectively. For convenience and unless otherwise indicated, reference to the length of detector probes refers to the length of the probe portion of the probes. Immobilized detector probes are detector probes immobilized on a support.

Detector probes can be, and preferably are, immobilized on a substrate. Detector probes can also contain or be associated with capture tags to facilitate immobilization or capture of the probes and ligator-detectors to which they have been coupled. Detector probes can also contain or be associated with sorting tags to facilitate sorting or separation of the probes and ligator-detectors to which they have been coupled. Detector probes can also contain or be associated with labels to facilitate detection of the probes and ligator-detectors to which they have been coupled.

Detector probes can also include one or more photocleavable nucleotides to facilitate release of probe sequences and ligator-detectors coupled to the probe. Photocleavable nucleotides and their use are described in WO 00/04036.

Detector probes need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the probe have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

Probe Arrays

Different detector probes can be used together as a set. The set can be used as a mixture of all or subsets of the probes, probes used separately in separate reactions, or immobilized in an array. Probes used separately or as mixtures can be physically separable through, for example, the use of capture tags, sorting tags, or immobilization on beads. A probe array (also referred to herein as an array) includes one or more probes or a plurality of probes immobilized at identified or predetermined locations on the array. In this context, plurality of probes refers to multiple probes each having a different sequence. Each predetermined location on the array has one type of probe (that is, all the probes at that location have the same sequence). Each location preferably will have multiple copies of the probe. The spatial separation of probes of different sequence in the array allows separate detection and identification of ligator-detectors that become coupled to the probes via hybridization of the probes to nucleic acid fragments in a nucleic acid sample. If a ligator-detector is detected at a given location in a probe array, it indicates that the sequence adjacent to the site in the nucleic acid fragment where the ligator-detector hybridized is complementary to the probe immobilized at that location in the array.

Adaptor-indexers, ligator-detectors, and offset adaptors can also be immobilized in arrays. Different modes of the disclosed method can be performed with different components immobilized, labeled, or tagged. Arrays of adaptor-indexers, ligator-detectors, and offset adaptors can be made and used as described below and elsewhere herein for the detector probes.

Solid-state substrates for use in probe arrays can include any solid material to which oligonucleotides can be coupled, directly or indirectly. This includes materials such as acrylamide, cellulose, nitrocellulose, glass, silicon, polystyrene, polyethylene vinyl acetate, polypropylene, polymethacrylate, polyethylene, polyethylene oxide, glass, polysilicates, polycarbonates, teflon, fluorocarbons, nylon, silicon rubber, polyanhydrides, polyglycolic acid, polylactic acid, polyorthoesters, polypropylfumerate, collagen, glycosaminoglycans, and polyamino acids. Solid-state substrates can have any useful form including thin films or membranes, beads, bottles, dishes, fibers,

woven fibers, shaped polymers, particles and microparticles. A preferred form for a solid-state substrate is a microtiter dish. The most preferred form of microtiter dish is the standard 96-well type.

Methods for immobilization of oligonucleotides to solid-state substrates are well established. Detector probes can be coupled to substrates using established coupling methods. For example, suitable attachment methods are described by Pease *et al.*, *Proc. Natl. Acad. Sci. USA* **91**(11):5022-5026 (1994), and Khrapko *et al.*, *Mol Biol (Mosk) (USSR)* 25:718-730 (1991). A method for immobilization of 3'-amine oligonucleotides on casein-coated slides is described by Stimpson *et al.*, *Proc. Natl. Acad. Sci. USA* **92**:6379-6383 (1995). A preferred method of attaching oligonucleotides to solid-state substrates is described by Guo *et al.*, *Nucleic Acids Res.* 22:5456-5465 (1994).

Methods for producing arrays of oligonucleotides on solid-state substrates are also known. Examples of such techniques are described in U.S. Patent No. 5,871,928 to Fodor et al., U.S. Patent No. 5,654,413 to Brenner, U.S. Patent No. 5,429,807, and U.S. Patent No. 5,599,695 to Pease et al.

Although preferred, it is not required that a given probe array be a single unit or structure. The set of probes may be distributed over any number of solid supports. For example, at one extreme, each probe may be immobilized in a separate reaction tube or container.

The probes in arrays can also be designed to have similar hybrid stability. This would make hybridization of fragments to detector probes more efficient and reduce the incidence of mismatch hybridization. The hybrid stability of probes can be calculated using known formulas and principles of thermodynamics (see, for example, Santa Lucia et al., *Biochemistry* 35:3555-3562 (1996); Freier et al., *Proc. Natl. Acad. Sci. USA* 83:9373-9377 (1986); Breslauer et al., *Proc. Natl. Acad. Sci. USA* 83:3746-3750 (1986)). The hybrid stability of the probes can be made more similar (a process that can be referred to as smoothing the hybrid stabilities) by, for example, chemically modifying the probes (Nguyen et al., *Nucleic Acids Res.* 25(15):3059-3065 (1997); Hohsisel, *Nucleic Acids Res.* 24(3):430-432 (1996)). Hybrid stability can also be smoothed by carrying out the hybridization under specialized conditions

(Nguyen et al., *Nucleic Acids Res.* 27(6):1492-1498 (1999); Wood et al., *Proc. Natl. Acad. Sci.* USA 82(6):1585-1588 (1985)).

Another means of smoothing hybrid stability of the probes is to vary the length of the probes. This would allow adjustment of the hybrid stability of each probe so that all of the probes had similar hybrid stabilities (to the extent possible). Since the addition or deletion of a single nucleotide from a probe will change the hybrid stability of the probe by a fixed increment, it is understood that the hybrid stabilities of the probes in a probe array will not be equal. For this reason, similarity of hybrid stability as used herein refers to any increase in the similarity of the hybrid stabilities of the probes (or, put another way, any reduction in the differences in hybrid stabilities of the probes). This is useful since any such increased similarity in hybrid stability can improve the efficiency and fidelity of hybridization and ligation of the detector probes.

The efficiency of hybridization and ligation of detector probes to sample fragments can also be improved by grouping detector probes of similar hybrid stability in sections or segments of a probe array that can be subjected to different hybridization conditions. In this way, the hybridization conditions can be optimized for particular classes of probes.

Amplification Primers

Amplification primers are oligonucleotides used to amplify binary sequence tags. Amplification primers include sequence complementary to one of the strands of a binary sequence tag. This sequence is referred to as the complementary portion of the amplification primer. Preferably, the complementary portion of an amplification primer is complementary to all or a part of one of the strands of an adaptor-indexer, all or a part of one of the strands of an offset adaptor, all or a part of the double-stranded portion of an adaptor-indexer, or all or a part of the double-stranded portion of an offset adaptor. The complementary portion of an amplification primer can be any length that supports specific and stable hybridization between the primer and the primer complement portion. Generally this is 10 to 35 nucleotides long, but is preferably 16 to 20 nucleotides long.

It is preferred that amplification primers also contain additional sequence at the 5' end of the primer that is not complementary to any part of the binary

sequence tag. This sequence is referred to as the non-complementary portion of the amplification primer. The non-complementary portion of an amplification primer may be any length, but is generally 1 to 100 nucleotides long. Amplifications primers need not be entirely single-stranded, but can contain a hairpin region formed between the 5' terminus and an internal sequence in the primer. Such amplifications primers are referred to herein as hairpin primers.

The amplification primer may also include modified nucleotides to make it resistant to exonuclease digestion or for other purposes. For example, the primer can have three or four phosphorothioate linkages between nucleotides at the 5' end of the primer. Amplification primers may contain deoxy-uridine residues that can be subsequently cleaved by an endonuclease. Amplification primers can also include one or more photocleavable nucleotides, to facilitate release of labels or mass tags in subsequent detection steps. Photocleavable nucleotides are described in WO 00/04036.

Amplification primers can also contain or be associated with capture tags to facilitate immobilization or capture of amplified sequence tags. In general, the capture tag can be one member of a binding pair such as biotin and streptavidin. Capture tags are discussed more fully elsewhere herein. Amplification primers can also contain or be associated with sorting tags to facilitate sorting or separation of amplified sequence tags. In general, the sorting tag can be a detectable label such as a fluorescent moiety or a manipulatable moiety such as a magnetic bead. Sorting tags are discussed more fully elsewhere herein. Amplification primers can also contain or be associated with labels to facilitate detection of amplified sequence tags. Amplification primers can also be immobilized on a substrate.

Amplification primers need not be composed of naturally occurring nucleotides. Modified nucleotides, unnatural based and nucleotide and oligonucleotide analogs can be used. All that is required is that the amplification primer have the general structure described herein and be capable of the interactions and reactions required in the disclosed method.

Anchored Primers

Anchored primers are oligonucleotides including an oligo dT portion and an anchor portion. The oligo dT portion is a string of dT residues at the 5' end of

the primer. The anchor portion is one or more nucleotides at the 3' end of the primer that are not all dT. Anchored primers are useful for cDNA synthesis. Some forms of anchored primers are described by Liang et al., *Nucleic Acids Res,* 21(14): 3269-75 (1993), and Liang and Pardee, *Science* 257:967-971 (1992). A preferred form of anchored primer comprises 16 dT residues and 2 residues at the 3'-end other than TN. Examples of anchored primers are the sequences TTTTTTTTTTTTTTTGC (SEQ ID NO:5), TTTTTTTTTTTTTTTGT (SEQ ID NO:6), and TTTTTTTTTTTTTTTCA (SEQ ID NO:7). There are 12 different primers of this type, each distinguished by the nucleotide sequence of the 3' terminal dinucleotide. Anchored primers are useful for generating a cDNA preparations of lower complexity. This is accomplished by performing reverse transcription in the presence of one anchored primer or in the presence of a set of anchored primers collectively having only a subset of the sequences possible. For example, six different cDNA preparations of reduced complexity may be generated by using six distinct sets of anchored primers, each set comprising only two of all the possible anchored primers ending in a unique non-TN dinucleotide as described above.

Labels

To aid in detection and quantitation of ligator-detectors coupled to detector probes, labels can be incorporated into, coupled to, or associated with, ligator-detectors, offset adaptors, detector probes, and/or adaptor-indexers. It is preferred that the ligator-detector be labeled. A label is any molecule that can be associated with ligator-detectors, directly or indirectly, and which results in a measurable, detectable signal, either directly or indirectly. A label is associated with a component when it is coupled or bound, either covalently or non-covalently, to the component. A label is coupled to a component when it is covalently coupled to the component. Many suitable labels for incorporation into, coupling to, or association with nucleic acid are known. Examples of labels suitable for use in the disclosed method are radioactive isotopes, fluorescent molecules, phosphorescent molecules, bioluminescent molecules, enzymes, antibodies, and ligands.

Examples of suitable fluorescent labels include fluorescein (FITC), 5,6-carboxymethyl fluorescein, Texas red, nitrobenz-2-oxa-1,3-diazol-4-yl (NBD), coumarin, dansyl chloride, rhodamine, 4'-6-diamidino-2-phenylinodole (DAPI), and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. Preferred fluorescent labels are fluorescein (5-carboxyfluorescein-N-hydroxysuccinimide ester) and rhodamine (5,6-tetramethyl rhodamine). Preferred fluorescent labels for simultaneous detection are FITC and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. The absorption and emission maxima, respectively, for these fluors are: FITC (490 nm; 520 nm), Cy3 (554 nm; 568 nm), Cy3.5 (581 nm; 588 nm), Cy5 (652 nm: 672 nm), Cy5.5 (682 nm; 703 nm) and Cy7 (755 nm; 778 nm), thus allowing their simultaneous detection. The fluorescent labels can be obtained from a variety of commercial sources, including Molecular Probes, Eugene, OR and Research Organics, Cleveland, Ohio.

Labeled nucleotides are the preferred form of label since they can be directly incorporated into ligator-detectors during synthesis. Examples of labels that can be incorporated into DNA or RNA include nucleotide analogs such as BrdUrd (Hoy and Schimke, *Mutation Research* 290:217-230 (1993)), BrUTP (Wansick *et al.*, J. *Cell Biology* 122:283-293 (1993)) and nucleotides modified with biotin (Langer *et al.*, *Proc. Natl. Acad. Sci. USA* 78:6633 (1981)) or with suitable haptens such as digoxygenin (Kerkhof, *Anal. Biochem.* 205:359-364 (1992)). Suitable fluorescence-labeled nucleotides are Fluorescein-isothiocyanate-dUTP, Cyanine-3-dUTP and Cyanine-5-dUTP (Yu *et al.*, *Nucleic Acids Res.*, 22:3226-3232 (1994)). A preferred nucleotide analog detection label for DNA is BrdUrd (BUDR triphosphate, Sigma), and a preferred nucleotide analog detection label for RNA is Biotin-16-uridine-5'-triphosphate (Biotin-16-dUTP, Boehringher Mannheim). Fluorescein, Cy3, and Cy5 can be linked to dUTP for direct labeling. Cy3.5 and Cy7 are available as avidin or anti-digoxygenin conjugates for secondary detection of biotin- or digoxygenin-labeled probes.

Labels that are incorporated into nucleic acid, such as biotin, can be subsequently detected using sensitive methods well-known in the art. For example, biotin can be detected using streptavidin-alkaline phosphatase conjugate (Tropix, Inc.), which is bound to the biotin and subsequently detected

by chemiluminescence of suitable substrates (for example, chemiluminescent substrate CSPD: disodium, 3-(4-methoxyspiro-[1,2,-dioxetane-3-2'-(5'-chloro)tricyclo [3.3.1.1^{3,7}]decane]-4-yl) phenyl phosphate; Tropix, Inc.).

Other labels include molecular or metal barcodes, mass labels, and labels detectable by nuclear magnetic resonance, electron paramagnetic resonance, surface enhanced raman scattering, surface plasmon resonance, fluorescence, phosphorescence, chemiluminescence, resonance raman, microwave, or a combination. Mass labels are compounds or moieties that have, or which give the labeled component, a distinctive mass signature in mass spectroscopy. Mass labels are useful when mass spectroscopy is used for detection. Preferred mass labels are peptide nucleic acids and carbohydrates. Combinations of labels can also be useful. For example, color-encoded microbeads having, for example, 256 unique combinations of labels, are useful for distinguishing numerous components. For example, 256 different ligator-detectors can be uniquely labeled and detected allowing multiplexing and automation of the disclosed method.

Useful labels are described in de Haas et al., "Platinum porphyrins as phosphorescent label for time-resolved microscopy," *J. Histochem. Cytochem.* 45(9):1279-92 (1997); Karger and Gesteland, "Digital chemiluminescence imaging of DNA sequencing blots using a charge-coupled device camera," *Nucleic Acids Res.* 20(24):6657-65 (1992); Keyes et al., "Overall and internal dynamics of DNA as monitored by five-atom-tethered spin labels," *Biophys. J.* 72(1):282-90 (1997); Kirschstein et al., "Detection of the DeltaF508 mutation in the CFTR gene by means of time- resolved fluorescence methods," *Bioelectrochem. Bioenerg.* 48(2):415-21 (1999); Kricka, "Selected strategies for improving sensitivity and reliability of immunoassays," *Clin. Chem.* 40(3):347-57 (1994); Kricka, "Chemiluminescent and bioluminescent techniques," *Clin. Chem.* 37(9):1472-81 (1991); Kumke et al., "Temperature and quenching studies of fluorescence polarization detection of DNA hybridization," *Anal. Chem.* 69(3):500-6 (1997); McCreery, "Digoxigenin labeling," *Mol. Biotechnol.* 7(2):121-4 (1997); Mansfield et al., "Nucleic acid detection using non-radioactive labeling methods," *Mol. Cell Probes* 9(3):145-56 (1995); Nurmi et al., "A new label technology for the detection of specific polymerase chain

reaction products in a closed tube," *Nucleic Acids Res.* 28(8):28 (2000); Oetting et al. "Multiplexed short tandem repeat polymorphisms of the Weber 8A set of markers using tailed primers and infrared fluorescence detection," *Electrophoresis* 19(18):3079-83(1998); Roda et al., "Chemiluminescent imaging of enzyme-labeled probes using an optical microscope-videocamera luminograph," *Anal. Biochem.* 257(1):53-62 (1998); Siddiqi et al., "Evaluation of electrochemiluminescence- and bioluminescence-based assays for quantitating specific DNA," *J. Clin. Lab. Anal.* 10(6):423-31 (1996); Stevenson et al., "Synchronous luminescence: a new detection technique for multiple fluorescent probes used for DNA sequencing," *Biotechniques* 16(6):1104-11 (1994); Vo-Dinh et al., "Surface-enhanced Raman gene probes," *Anal. Chem.* 66(20):3379-83 (1994); Volkers et al., "Microwave label detection technique for DNA in situ hybridization," *Eur. J. Morphol.* 29(1):59-62 (1991).

Metal barcodes, a form of molecular barcode, are 30-300 nm diameter by 400-4000 nm multilayer multi metal rods. These rods are constructed by electrodeposition into an alumina mold, then the alumina is removed leaving these small multilayer objects behind. The system can have up to 12 zones encoded, in up to 7 different metals, where the metals have different reflectivity and thus appear lighter or darker in an optical microscope depending on the metal; this leads to practically unlimited identification codes. The metal bars can be coated with glass or other material, and probes attached to the glass using methods commonly known in the art; assay readout is by fluorescence from the target, and the identity of the probe is from the light dark pattern of the barcode.

Methods for detecting and measuring signals generated by labels are known. For example, radioactive isotopes can be detected by scintillation counting or direct visualization; fluorescent molecules can be detected with fluorescent spectrophotometers; phosphorescent molecules can be detected with a spectrophotometer or directly visualized with a camera; enzymes can be detected by measurement or visualization of the product of a reaction catalyzed by the enzyme; antibodies can be detected by detecting a secondary detection label coupled to the antibody. Such methods can be used directly in the disclosed method of amplification and detection. As used herein, detection molecules are molecules which interact with amplified nucleic acid and to which

one or more detection labels are coupled. In another form of detection, labels can be distinguished temporally via different fluorescent, phosphorescent, or chemiluminescent emission lifetimes. Multiplexed time-dependent detection is described in Squire et al., J. Microscopy 197(2):136-149 (2000), and WO 00/08443.

Quantitative measurement of the amount or intensity of a label can be used. For example, quantitation can be used to determine if a given label, and thus the labeled component, is present at a threshold level or amount. A threshold level or amount is any desired level or amount of signal and can be chosen to suit the needs of the particular form of the method being performed.

Capture Tags

A capture tag is any compound that can be used to separate compounds or complexes having the capture tag from those that do not. Preferably, a capture tag is a compound, such as a ligand or hapten, that binds to or interacts with another compound, such as ligand-binding molecule or an antibody. It is also preferred that such interaction between the capture tag and the capturing component be a specific interaction, such as between a hapten and an antibody or a ligand and a ligand-binding molecule. Capture tags preferably are antibodies, ligands, binding proteins, receptor proteins, haptens, aptamers, carbohydrates, synthetic polyamides, or oligonucleotides. Preferred binding proteins are DNA binding proteins. Preferred binding proteins are DNA binding proteins. Preferred DNA binding proteins are zinc finger motifs, leucine zipper motifs, helix-turn-helix motifs. These motifs can be combined in the same specific binding molecule.

Preferred capture tags, described in the context of nucleic acid probes, are described by Syvnen *et al., Nucleic Acids Res.,* 14:5037 (1986). Preferred capture tags include biotin, which can be incorporated into nucleic acids. In the disclosed method, capture tags incorporated into adaptor-indexers or offset adaptors can allow sample fragments (to which the adaptors have been coupled) to be captured by, adhered to, or coupled to a substrate. Such capture allows simplified washing and handling of the fragments, and allows automation of all or part of the method.

Properties of zinc fingers, zinc finger motifs, and their interactions, are described by Nardelli et al., *Zinc finger-DNA recognition: analysis of base specificity by site- directed mutagenesis.* Nucleic Acids Res, **20**(16):4137-44 (1992), Jamieson et al., *In vitro selection of zinc fingers with altered DNA-binding specificity.* Biochemistry, **33**(19):5689-95 (1994), Chandrasegaran, S. and J. Smith, *Chimeric restriction enzymes: what is next?* Biol Chem, **380**(7-8):841-8 (1999), and Smith et al., *A detailed study of the substrate specificity of a chimeric restriction enzyme.* Nucleic Acids Res, **27**(2):674-81 (1999).

Capturing sample fragments on a substrate may be accomplished in several ways. In one embodiment, capture docks are adhered or coupled to the substrate. Capture docks are compounds or moieties that mediate adherence of a sample fragment by binding to, or interacting with, a capture tag on the fragment. Capture docks immobilized on a substrate allow capture of the fragment on the substrate. Such capture provides a convenient means of washing away reaction components that might interfere with subsequent steps.

Substrates for use in the disclosed method can include any solid material to which components of the assay can be adhered or coupled. Examples of substrates include, but are not limited to, materials such as acrylamide, cellulose, nitrocellulose, glass, silicon, polystyrene, polyethylene vinyl acetate, polypropylene, polymethacrylate, polyethylene, polyethylene oxide, polysilicates, polycarbonates, teflon, fluorocarbons, nylon, silicon rubber, polyanhydrides, polyglycolic acid, polylactic acid, polyorthoesters, polypropylfumerate, collagen, glycosaminoglycans, and polyamino acids. Substrates can have any useful form including thin films or membranes, beads, bottles, dishes, fibers, woven fibers, shaped polymers, particles and microparticles. Preferred forms of substrates are plates and beads. The most preferred form of beads are magnetic beads.

In one embodiment, the capture dock is an oligonucleotide. Methods for immobilizing and coupling oligonucleotides to substrates are well established. For example, suitable attachment methods are described by Pease *et al.*, *Proc. Natl. Acad. Sci. USA* 91(11):5022-5026 (1994), and Khrapko *et al.*, *Mol Biol (Mosk) (USSR)* 25:718-730 (1991). A method for immobilization of 3'-amine oligonucleotides on casein-coated slides is described by Stimpson *et al.*, *Proc.*

Natl. Acad. Sci. USA 92:6379-6383 (1995). A preferred method of attaching oligonucleotides to solid-state substrates is described by Guo *et al.*, *Nucleic Acids Res.* 22:5456-5465 (1994).

In another embodiment, the capture dock is the anti-hybrid antibody. Methods for immobilizing antibodies to substrates are well established. Immobilization can be accomplished by attachment, for example, to aminated surfaces, carboxylated surfaces or hydroxylated surfaces using standard immobilization chemistries. Examples of attachment agents are cyanogen bromide, succinimide, aldehydes, tosyl chloride, avidin-biotin, photocrosslinkable agents, epoxides and maleimides. A preferred attachment agent is glutaraldehyde. These and other attachment agents, as well as methods for their use in attachment, are described in *Protein immobilization: fundamentals and applications*, Richard F. Taylor, ed. (M. Dekker, New York, 1991), Johnstone and Thorpe, *Immunochemistry In Practice* (Blackwell Scientific Publications, Oxford, England, 1987) pages 209-216 and 241-242, and *Immobilized Affinity Ligands*, Craig T. Hermanson *et al.*, eds. (Academic Press, New York, 1992). Antibodies can be attached to a substrate by chemically cross-linking a free amino group on the antibody to reactive side groups present within the substrate. For example, antibodies may be chemically cross-linked to a substrate that contains free amino or carboxyl groups using glutaraldehyde or carbodiimides as cross-linker agents. In this method, aqueous solutions containing free antibodies are incubated with the solid-state substrate in the presence of glutaraldehyde or carbodiimide. For crosslinking with glutaraldehyde the reactants can be incubated with 2% glutaraldehyde by volume in a buffered solution such as 0.1 M sodium cacodylate at pH 7.4. Other standard immobilization chemistries are known by those of skill in the art.

Sorting Tags

A sorting tag is any compound that can be used to sort or separate compounds or complexes having the sorting tag from those that do not. In general, all capture tags can be sorting tags. Sorting tags also include compounds and moieties that can be detected and which can mediate the sorting of tagged components. Such forms of sorting tags are generally not also capture tags. For example, a fluorescent moiety can allow sorting of components tagged

with the moiety from those that are not (or those with a different tag). However, such a fluorescent moiety does not necessarily have a suitable capture dock with which it can interact and be captured. Preferably, a sorting tag is a label, such as a fluorescent label, that can mediate sorting.

Amplification Target Circle

An amplification target circle (ATC) is a circular single-stranded DNA molecule, generally containing between 40 to 1000 nucleotides, preferably between about 50 to 150 nucleotides, and most preferably between about 50 to 100 nucleotides. Portions of ATCs have specific functions making the ATC useful for rolling circle amplification (RCA). These portions are referred to as the primer complement portion and the reporter tag portions. The primer complement portion and the reporter tag portion are required elements of an amplification target circle. Those segments of the ATC that do not correspond to a specific portion of the ATC can be arbitrarily chosen sequences. It is preferred that ATCs do not have any sequences that are self-complementary. It is considered that this condition is met if there are no complementary regions greater than six nucleotides long without a mismatch or gap.

An amplification target circle, when replicated, gives rise to a long DNA molecule containing multiple repeats of sequences complementary to the amplification target circle. This long DNA molecule is referred to herein as tandem sequences DNA (TS-DNA). TS-DNA contains sequences complementary to the primer complement portion and the reporter tag portions. These sequences in the TS-DNA are referred to as primer sequences (which match the sequence of the rolling circle replication primer) and reporter tags. Amplification target circles and their use are further described in U.S. Patent No. 5,854,033.

Method

The disclosed method involves the following basic steps. A nucleic acid sample is incubated with one or more nucleic acid cleaving reagents, preferably restriction endonucleases, that results in a set of DNA fragments cleaved at particular sites. The sample is then mixed with one or more offset adaptors, each of which has a recognition sequence for a nucleic acid cleaving reagent that cleaves at a site offset from the recognition sequence. The offset adaptors are

then covalently coupled, preferably by ligation, onto the DNA fragments. The offset adapters should have ends compatible with the ends of the nucleic acid fragments.

The nucleic acid sample is incubated with one or more nucleic acid cleaving reagents, preferably restriction endonucleases, such that a set of DNA fragments having sticky ends with a variety of sequences is generated. Preferred for this purpose is the use of a single Type IIS restriction endonuclease having an offset cleavage site. Since such Type IIS restriction endonucleases cleave at a site different from the recognition sequence, this results in a set of DNA fragments having sticky ends with a variety of sequences. A similar effect can be obtained by digesting the nucleic acid sample with a mixture of restriction endonucleases or other nucleic acid cleaving reagents which cleave at their recognition site.

For a four base sticky end, there are 256 possible sequences. The general formula is $N = 4^X$ where X is the length of the sticky end and N is the number of possible sequences. In a sufficiently complex nucleic acid sample, all of these sequences will be represented in the ends of the set of DNA fragments. The nucleic acid sample is also divided into aliquots (referred to as index samples); preferably as many aliquots as there are sticky end sequences (that is, $N = 4^X$ aliquots). Where multiple restriction endonucleases are used, the nucleic acid sample is preferably divided into index samples before digestion. Where a single restriction endonuclease is used, the nucleic acid sample is preferably divided into index samples following digestion. Each index sample is then mixed with a different adaptor-indexer, each of which has a sticky end compatible with one of the possible sticky ends on the DNA fragments in that index sample. The adaptor-indexers are then coupled onto compatible DNA fragments. This results in the formation of binary sequence tags. The binary sequence tags have adaptors ligated to each end. The binary sequence tags can then be amplified, if desired, using any suitable method, such as PCR. Sequences in the adaptors can be used as primer binding sites for this amplification.

The binary sequence tags can then be analyzed. Preferably, the binary sequence tags are hybridized with ligator-detectors. One portion of each ligator-

detector matches or is complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers. Preferably, the ligator-detector has sequence matching or complementary to one of the possible sticky end sequences generated by digestion with the restriction enzymes. The ligator-detector can also be complementary to nucleotides in the recognition sequence of the restriction enzymes if restriction enzymes that cleave in their recognition sequence are used. Ligator-detectors can also have sequence matching or complementary to common sequences in the adaptor-indexers. In this case, the appropriate ligator-detector is used with each of the index samples. That is, the ligator-detector oligonucleotide used in each index sample matches or is complementary to sequence, including sticky end sequence, in the adaptor-indexer sequence used in that index sample. Alternatively, the ligator-detector can have sequence matching or complementary to sequence of the nucleic acid fragment (to which an adaptor-indexer has been coupled) adjacent to the sticky end sequence and on opposite the side of the fragment from the adaptor-indexer (that is, the offset adaptor side of the binary sequence tag).

Each index sample then can be mixed with detector probes and the probes are coupled to the ligator-detectors. Preferably, the set of probes used include every possible sequence of a given length (for example, every possible six base sequence). The detector probes can be immobilized in an array.

The ends of the detector probes and the ligator-detectors are coupled together only if the probe hybridizes adjacent to the end of the ligator-adaptor. Thus, a ligator-detector is coupled to a detector probe only when a sequence complementary to the probe is present immediately adjacent to the region in a binary sequence tag derived from the original sample to which the end of the ligator-detector hybridizes (preferably the sticky end sequences).

Binary sequence tag will result in an association of a ligator-detector (and an adaptor-indexer and offset adaptor) with a detector probe. This association will be detected through a signal generated from one or several of the associated components. In a preferred form of the disclosed method, the set of detector probes (if multiple sets are used) in which the signal for a given fragment is determined by the sequence of the original sticky end sequence (or recognition sequence). Each different sticky end or recognition sequence is

processed in a separate index sample; a separate set of detector probes is used for each index sample or derivative index sample. The probe in the set of probes to which the signal for a given fragment is associated and detected is determined by the sequence in the binary sequence tag adjacent to the sticky end sequence (or recognition sequence) since the detector probe must hybridize to this sequence in order to be coupled to the ligator-detector hybridized to the binary sequence tag. A complex nucleic acid sample will produce a unique pattern of signals in the probe sets. It is this pattern that allows unique cataloging of nucleic acid samples and sensitive and powerful comparisons of the patterns of signals produced from different nucleic acid samples.

The set of detector probes, and particular probe in the set, in which a signal is associated identifies the sequence of the sticky end of the DNA fragment that gave rise to the signal and of the sequence adjacent to the sticky end. This is a ten base sequence when a four base sticky end and six base probes are used. The set of detector probes identifies the sticky end sequence (the four bases) because each set of detector probes used a different adaptor-indexer having a different sticky end sequence. The particular probe identifies the sequence adjacent to the sticky end (the six bases) because each probe has a different probe with a different sequence. Only the probe with sequence complementary to the adjacent sequence will hybridize and thus become associated with a signal.

The information generated is similar when probe arrays are used. The array (if multiple arrays are used) in which the signal for a given fragment is detected is determined by the sequence of the original sticky end sequence (or recognition sequence). Each different sticky end or recognition sequence is processed in a separate index sample; a separate array is used for each index sample or derivative index sample. The location in the array in which the signal for a given fragment is detected is determined by the sequence in the binary sequence tag adjacent to the sticky end sequence (or recognition sequence) since the probe must hybridize to this sequence in order to be coupled to the ligator-detector hybridized to the binary sequence tag. A complex nucleic acid sample will produce a unique pattern of signals on the arrays.

The array, and location in the array, where a DNA fragment generates a signal identifies the sequence of the sticky end of the DNA fragment and of the sequence adjacent to the sticky end. This is a ten base sequence when a four base sticky end and six base probes are used. The array identifies the sticky end sequence (the four bases) because each array used a different adaptor-indexer having a different sticky end sequence. The location in the array identifies the sequence adjacent to the sticky end (the six bases) because each location in the array has a different probe with a different sequence. Only the probe with sequence complementary to the adjacent sequence will hybridize and thus become associated with a signal.

The disclosed method is performed using one or more nucleic acid cleaving reagents that cleave at a site offset from their recognition sequence. Preferred nucleic acid cleaving reagents for use in the disclosed method are Type IIS restriction endonucleases, which are enzymes that cleave DNA at locations outside of (or offset from) the recognition site and which generate sticky ends. Examples of Type IIS restriction endonucleases are FokI, BbvI, HgaI, BspMI and SfaNI.

Nucleic acid cleaving reagents for use in the disclosed method produce sticky ends encompassing permutations and combinations of the four nucleotides, A, C, G, and T. The larger the number of protruding bases, the greater the number of possible permutations and combinations of terminal nucleotide sequences, and the more specific the indexing is likely to be. For example, a restriction endonuclease such as FokI, which releases fragments with four base, 5'-protruding sticky ends, will generate fragments having 4^4 or 256 possible protruding tetranucleotide ends. The length of the recognition sequence, the length of the sticky end generated, and the length of the probes used together determine the number of data bins (that is, probe identities) into which the binary sequence tag are sorted. By using sticky ends and detector probes of sufficient length, the sorting of fragments can be matched to the complexity of the sample being analyzed.

The use of a comprehensive panel of adaptor-indexers provides a means for attaching specific functional modifications to selected subsets of a complex mixture of nucleic acid fragments and identifying the molecules so modified.

Such a defined subset of molecules may be further resolved by additional cleavage and indexing, or by any of the established techniques such as cloning, PCR amplification, or gel electrophoresis. Individual members of the class may be distinguished by identifying characteristics such as length, sequence, or restriction endonuclease maps. The sequence of the sticky ends of the adaptor-indexers provides a means of indexing a large number of nucleic acid fragments.

Detector probes of different sequence can be immobilized at different locations on a probe array. In this way, the sequence of the probes on the probe array and the sequence of the binary sequence tags determine where on the array ligator-detectors become coupled. The presence, amount, presence and amount, or absence of ligator-detector at different locations in the probe arrays thus forms a pattern of signals that provides a signature or fingerprint of the binary sequence tags, and thus of the nucleic acid sample based on the presence or absence of specific nucleic acid sequences in the sample. For this reason, cataloging of this pattern of signals (that is, the pattern of the presence, amount, presence and amount, or absence of ligator-detectors) is an embodiment of the disclosed method that is of particular interest. As discussed elsewhere herein, the probability of probe mismatch can be used to create more complex catalogs based on differential hybridization of particular fragments to different detector probes.

Catalogs can be made up of, or be referred to, as, for example, a pattern of ligator-detectors on probe arrays, a pattern of the presence of ligator-detectors on probe arrays, a catalog of binary sequence tags, a catalog of nucleic acid fragments in a sample, or a catalog of nucleic acid sequences in a sample. The information in the catalog is preferably in the form of positional information (that is, location in the detector array) or, more preferably, in the form of sequences. Preferred sequence information for catalogs include sequences of detector probes to which a ligator-detector was coupled and sequences of nucleic acid fragments present in the sample (derived from the locations in the detector array where ligator-detectors were coupled). Catalogs can also contain or be made up of other information derived from the information generated in the disclosed method, and can be combined with information obtained or generated from any other source. The informational nature of catalogs produced using the

disclosed method lends itself to combination and/or analysis using known bioinformatics systems and methods.

Such catalogs of nucleic acid samples can be compared to a similar catalog derived from any other sample to detect similarities and differences in the samples (which is indicative of similarities and differences in the nucleic acids in the samples). For example, a catalog of a first nucleic acid sample can be compared to a catalog of a sample from the same type of organism as the first nucleic acid sample, a sample from the same type of tissue as the first nucleic acid sample, a sample from the same organism as the first nucleic acid sample, a sample obtained from the same source but at time different from that of the first nucleic acid sample, a sample from an organism different from that of the first nucleic acid sample, a sample from a type of tissue different from that of the first nucleic acid sample, a sample from a strain of organism different from that of the first nucleic acid sample, a sample from a species of organism different from that of the first nucleic acid sample, or a sample from a type of organism different from that of the first nucleic acid sample.

The same type of tissue is tissue of the same type such as liver tissue, muscle tissue, or skin (which may be from the same or a different organism or type of organism). The same organism refers to the same individual, animal, or cell. For example, two samples taken from a patient are from the same organism. The same source is similar but broader, referring to samples from, for example, the same organism, the same tissue from the same organism, the same cDNA, or the same cDNA library. Samples from the same source that are to be compared can be collected at different times (thus allowing for potential changes over time to be detected). This is especially useful when the effect of a treatment or change in condition is to be assessed. Samples from the same source that have undergone different treatments can also be collected and compared using the disclosed method. A different organism refers to a different individual organism, such as a different patient, a different individual animal. Different organism includes a different organism of the same type or organisms of different types. A different type of organism refers to organisms of different types such as a dog and cat, a human and a mouse, or *E. coli* and *Salmonella*. A different type of tissue refers to tissues of different types such as liver and

kidney, or skin and brain. A different strain or species of organism refers to organisms differing in their species or strain designation as those terms are understood in the art.

Offset adaptors and adaptor-indexers are preferably coupled to nucleic acid fragments in their double-stranded form since ligation will be more efficient. However, the two strands of an offset adaptor or adaptor-indexer can be used separately in the disclosed method. For example, the two strands of the offset adaptor or adaptor-indexer can be coupled to a nucleic acid fragment separately. Thus, the steps of mixing and coupling offset adaptors or adaptor-indexers to nucleic acid fragments encompasses both mixing and coupling of double-stranded forms of offset adaptor or adaptor-indexer and separately mixing and coupling the strands of offset adaptors or adaptor-indexers.

When a first offset adaptor strand is coupled to a nucleic acid fragment, the double-stranded recognition site for cleavage can be formed other than by covalently coupling the second offset adaptor strand to the nucleic acid fragment. For example, the second offset adaptor strand can be hybridized to the first offset adaptor strand, but not coupled to the nucleic acid fragment. The hybrid need only stay together through cleavage of the fragment with second nucleic acid cleaving reagent. The offset adaptor region of the nucleic acid fragment need not be fully double-stranded so long as the recognition site is functional. Alternatively, the other strand of the offset adaptor can be formed by filling in the single-stranded portion of the first offset adaptor coupled to the nucleic acid fragment.

When a first adaptor-indexer strand is coupled to a nucleic acid fragment, the second strand need not be used or added. Coupling of the first adaptor-indexer strand alone can complete the formation of one strand of a binary sequence tag. This is all that is required for many forms of manipulation and analysis of binary sequence tags. A coupled, double-stranded adaptor-indexer can be formed other than by covalently coupling the second adaptor-indexer strand to the nucleic acid fragment. For example, the other strand of the adaptor-indexer can be formed by filling in the single-stranded portion of the first adaptor-indexer coupled to the nucleic acid fragment. As with the second

offset adaptor strand, the second adaptor-indexer strand can be hybridized to the first adaptor-indexer strand without coupling it to the nucleic acid fragment.

When comparing catalogs of binary sequence tags obtained from related samples, it is possible to identify the presence of a subset of correlated pairs. Binary sequence tags produced using the disclosed method generally occur in correlated pairs, except in a few cases where the initial cleavage site occurs near the terminus of a DNA substrate. Correlated pairs of binary sequence tags are the two tags that result from the initial cleavage at a particular site. An offset adaptor is coupled to each end of the cleavage site, eventually resulting in the two correlated binary sequence tags.

cDNA generated from long mRNA molecules is likely to generate a complex pattern of binary sequence tags. On the other hand, shorter cDNAs are likely to generate relatively fewer binary sequence tags. If a small concentration of a di-deoxy nucleotide triphosphate is included in the first strand cDNA synthesis reaction, first strand synthesis will result in a distribution of strands with a relatively lower representation of longer molecules. In the specific case where ddCTP is used as a terminator, the probability P of chain extension at every position where C is present is given by:

P=([dCTP])/([dCTP]+q[ddCTP]), where q is the incorporation efficiency of ddCTP relative to dCTP. The probability of extension for a chain containing C at n positions is P^n.

By generating the second strand with random primers, a skewed distribution of double-stranded DNA fragments is preserved in which sequences near the 3'-end of the cDNA are over-represented. As a result, binary sequence tags originating from sequences near the 3' end will also be over-represented relative to binary sequence tags originating from sequences closer to the 5' end. Among the binary sequence tags generated from such a cDNA, it will be possible to identify a pattern of correlated pairs, where the abundance (that is, the signal intensity) of each correlated pair is nearly the same. Different correlated pairs arise from cleavage sites distant from each other and can be ordered according to their abundance. For genomes of known sequence, the data obtained from several assays using cDNAs with a skewed 3'-end representation, can be used to generate large datasets of correlated tags and the corresponding

intensity signal gradients. These gradients can be calibrated with the distance of each binary sequence tag from the 3'-end of the cDNA. Signal gradients of correlated tags can thus serve as calibrators in other assays involving the analysis of mRNA transcripts from genomes of unknown sequence.

This can be illustrated with the following example of a catalog with three candidate correlated tags as might be obtained from an experiment performed in an organism whose genomic sequence is known.

Tag address	Test/control ratio	Test level	Control level
A	15.5	322	21
B	14.8	319	22
C	16.1	244	15
D	15.3	239	16
F	14.9	171	11
G	16.0	189	12

In this example, the top pair of correlated tags is located near the 3'-end of a unique cDNA, where the cDNA has been up-regulated by a factor of approximately 15 in the test sample. The other two pairs of tags have in common a very similar test/control expression ratio, and occur at positions of increasing distance from the 3'-end of the cDNA. The absolute levels of expression also occur as binary pairs, and show a correlation with the distance in nucleotides relative to the 3'-end of the cDNA. That is, the absolute level of the tags goes down the further source sequence of the tag is from the 3' end. The level of both binary sequence tags in a pair go down together, in a correlated fashion (thus, keeping their ratio the same).

If a similar assay were performed in an organism whose genome has not been sequenced, correlated binary sequence tags that share approximately the same expression ratio, and that display graded levels of absolute expression values, may be used to infer the possible order of the binary sequence tags derived from a specific cDNA. Repeating the assay using cDNA generating in the presence of a different level of the same ddNTP, or the same level of a different ddNTP, will generate additional data that can confirm or strengthen the

identification, ordering, and separation distance of the putative correlated binary sequence tags believed to be derived from the same gene. The analysis of correlated binary sequence tags can be simplified by using anchored primers. As described elsewhere herein, anchored primers can be used to produce a less complex nucleic acid sample. The reduced complexity will result in fewer tags to be analyzed. It should also be noted that alternative splicing events may lead to different distance maps for correlated binary sequence tags derived from genes that have such splicing patterns.

Since correlated tags are derived from adjacent sequences in the original nucleic acid molecule, once correlated pairs are identified, their sequences, when put together, represent a longer sequence in the original nucleic acid fragment. For example, when the disclosed method is performed using MboI and FokI, the information content of the correlated binary tags would be 24 (4+6+4+6+4) nucleotides. These longer sequences can be added to the catalog of tags.

The presence, amount, presence and amount, or absence of ligator-detectors coupled to detector probes can be accomplished by detection of labels incorporated into, coupled to, or associated with the ligator-detectors. Alternatively, the ligator-detectors can be detected based on detection of their sequence. These detections are generally referred to as direct detection of coupling of ligator-detectors. Any of the numerous sequence-specific detection techniques can be used for this purpose, including, for example, hybridization of labeled probes. The presence, amount, presence and amount, or absence of ligator detectors can also be detected by generating a signal mediated by the ligator-detector. Use of the ligator-detector as a primer for rolling circle replication, described below, is a preferred example of this. The presence, amount, presence and amount, or absence of ligator detectors can also be detected by detecting the detector probe to which the ligator-detector is coupled, the adaptor-indexer associated with the coupled ligator-detector, the offset adaptor associated with the coupled ligator-detector, or a combination. These detections are generally referred to as indirect detection of coupling of ligator-detectors.

The signal to be detected for the binary sequence tags can be increased by nucleic acid amplification during the method. It is preferred either that the

binary sequence tags be amplified or that the ligator-detectors that have been coupled to detector probes be amplified or mediate amplification of another nucleic acid. In the first case, the binary sequence tags can be amplified using any suitable method. These include polymerase chain reaction (PCR), ligase chain reaction (LCR), self-sustained sequence replication (3SR), nucleic acid sequence based amplification (NASBA), strand displacement amplification (SDA), amplification with Qβ replicase, and rolling circle amplification (RCA) (Birkenmeyer and Mushahwar, *J. Virological Methods*, 35:117-126 (1991); Landegren, *Trends Genetics*, 9:199-202 (1993); Lizardi *et al.*, *Nature Genetics* 19(3):225-232 (1998)). In the case of ligator-detector amplification, a preferred form of amplification is rolling circle amplification of a single-stranded circular DNA molecule primed by the ligator-detector. In this way, a long tandem repeat of the DNA circle is generated with the amplified strand anchored to the detector array via the ligator-detector. This technique of amplification is described in PCT application WO 97/19193. If the ligator-detector is used as a rolling circle replication primer, there is no need to incorporate a label in the ligator-detector since the amplified DNA can be detected (either directly or via an incorporated label).

Amplification of the binary sequence tags is facilitated by the presence of offset adaptor and adaptor-indexer sequence at the ends of the binary sequence tags. For example, the offset adaptor sequences and adaptor-indexer sequences can be used for amplification of primer sequences. The offset adaptor and adaptor-indexer sequences can also be used to circularize the binary sequence tags for subsequent amplification by rolling circle replication. Rolling circle amplification is described in U.S. Patent No. 5,854,033 and PCT application WO 97/19193.

In another embodiment, the strands of binary sequence tags can be separated prior to hybridization to the ligator-detectors. Such strand separation can improve the efficiency of ligator-detector hybridization. This separation can be accomplished using any suitable technique. Strand separation is preferably accomplished by including a capture tag or sorting tag on one of the strands of the offset adaptors or adaptor-indexers. Such a capture tag can then be used to

immobilize one strand of the binary sequence tags while the other strands are washed away. Either the immobilized or washed strand can be carried forward in the method. A sorting tag can allow separation of the strands by a sorting technique.

In another embodiment, the concentrations of the various nucleic acid fragments in the index samples are normalized. Preferably, the concentrations of binary sequence tags are normalized. Normalization can be performed either before or after any amplification step that may be used. A preferred technique for fragment normalization involves immobilizing one strand of the nucleic acid fragments (preferably the binary sequence tags), denaturing the nucleic acid fragments, renaturing the nucleic acid fragments for a time greater than the $c_0t_{1/2}$ for abundant nucleic acid fragments and less than the $c_0t_{1/2}$ for rare nucleic acid fragments, and collecting the un-renatured nucleic acid fragments.

A number of varied probe sets and arrays are known in the art and can be used with the disclosed method. Terstappen et al. (Tibbe *et al.*, *Optical tracking and detection of immunomagnetically selected and aligned cells.* Nat Biotechnol, **17**(12):1210-3 (1999); U.S. Patent No. 5,985,153 (Dolan and Terstappen, *Magnetic separation apparatus and methods employing an internal magnetic capture gradient and an external transport force*); U.S. Patent No. 5,993,665 (Terstappen and Liberti, *Quantitative cell analysis methods employing magnetic separation*); U.S. Patent No. 6,013,188 (Terstappen and Liberti, *Methods for biological substance analysis employing internal magnetic gradients separation and an externally-applied transport force*)) have demonstrated immunomagnetically selected and fluorescently labeled probes for detection of cells of interest. In their technique the cells are labeled using immunospecific binding probes and the resulting labeled cells are induced to move into detector range by an externally applied magnetic field. Lithographic processing of one wall of the sample vessel leads to the improvement of locating the tagged cells along well determined spatial patterns.

Thorp et al. (Napier *et al.*, *Probing biomolecule recognition with electron transfer: electrochemical sensors for DNA hybridization.* Bioconjug Chem, **8**(6):906-13 (1997); U.S. Patent No. 5,968,745 (Thorp et al., *Polymer-electrodes for detecting nucleic acid hybridization and method of use thereof*);

U.S. Patent No. 5,871,918 (Thorp *et al.*, *Electrochemical detection of nucleic acid hybridization*); WO 99/64847 (Welch, *Electochemical probes for detection of molecular interactions and drug discovery*)) have developed an electrochemical detection method. Their method directly detects target nucleic acids without sample amplification or use of fluorescent labels. Detection is accomplished by following the oxidation-reduction cycle of $Ru(bpy)_3^{2+}$ by electrochemical measurement. The measurement elements have synthetically prepared oligonucleotide probes attached to electrodes, the electrodes have been fabricated into a variety of formats including biochips, microtiter plates and hand-held devices.

Spectral labeling coupled with a counting technique, such as flow cytometry, has been exploited for analysis of DNA samples (U.S. Patent No. 5,736,330 (Fulton, *Method and compositions for flow cytometric determination of DNA sequences*); WO 99/19515 (Phipps *et al.*, *Precision fluorescently dyed particles and methods of making and using same*); WO 99/37814 (Chandler and Chandler, *Microparticles with multiple fluorescent signals*)). In such analyses, micron scale beads are color coded according to the probes attached to their surface, targets are labeled with an analytical fluorescent tag, and the coincidence events containing the bead color and the analytical color are counted. In this manner a probe array of many colors can be read out very quickly and easily.

In another technique utilizing microspheres, Walt et al. (Walt, *Techview: molecular biology. Bead-based fiber-optic arrays.* Science, **287**(5452):451-2 (2000); WO 98/50782 (Ferguson et al., *Fiber optic biosensor for selectively detecting oligonucleotide species in a mixed fluid sample*); U.S. Patent No. 6,023,540 (Walt and Michael, *Fiber optic sensor with encoded microspheres*); Michael *et al.*, *Randomly ordered addressable high-density optical sensor arrays.* Anal Chem, **70**(7):1242-8 (1988)) have developed a system where the probes are attached to the microspheres and the microspheres subsequently self assemble in a random spatial pattern into the distal end of a fiber optic array. The "optical bar code" signature of each microsphere provides the identity of the attached probe, and signal of the labeled target indicates the concentration of the target.

One of the ways in which sequence tags identified using the disclosed method can be used is to create arrays of oligomers based on the sequence tags. Thus, new oligomers, with appropriate coupling chemistry and spacers, are synthesized which correspond to the sequence tag. These tags are arrayed and coupled appropriately on a microarray slide to yield a new microarray with expressed sequence tag probes for the organism of interest. Such arrays can be used in any of a variety of methods that involve arrays of oligonucleotides, including the disclosed method. In particular, the array can be used to detect similarities and differences between sequences present in the original sample and any future sample.

If genomic sequence information is available, a BLAST search using sequence derived from binary sequence tags may indicate the presence of a promoter if the sequence lies 5' of a characterized gene. *In silico* analysis of predicted methylation sites should facilitate such assumptions. If genomic sequence is unavailable, the sequence tag can serve as a starting point to facilitate characterization of previously undefined genomic DNA regions of interest.

Modification Assisted Analysis of Binary Sequence Tags (MAABST)

Modification assisted analysis of binary sequence tags (MAABST) is a form of BEST that assesses modification of sequences in nucleic acid molecules by detecting differential cleavage based on the presence or absence of modification in the molecules. For example, a site that is methylated in a nucleic acid molecule will not be cut by a restriction enzyme that is sensitive to methylation at that site. A restriction enzyme that is insensitive to methylation will cleave at that site, thus producing a different pattern of sequence tags. Comparison of the results with different samples of nucleic acids can establish differences in the modification levels or patterns in the different samples. As used herein, a nucleic acid cleaving reagent or restriction enzyme that is sensitive to modification in its recognition site is a nucleic acid cleaving reagent or restriction enzyme that will either cleave only when the site is unmodified or will cleave only when the site is modified (that is, the nucleic acid cleaving reagent or restriction enzyme requires a particular modification state for cleavage). A nucleic acid cleaving reagent or restriction enzyme that is

insensitive to modification in its recognition site is a nucleic acid cleaving reagent or restriction enzyme that will cleave regardless of whether the site is modified or unmodified.

MAABST is useful for assessing the state, level, and condition of modifications in nucleic acid molecules. Many nucleic acid modifications are known that have biological effects and significance. Methylation, for example, is a universal mechanism for regulating gene expression in animals and plants (Bird and Wolffe, *Methylation-induced repression-belts, braces, and chromatin*, Cell **99**:451-454 (1999); Finnegan *et al.*, *DNA Methylation in Plants*, Annual Rev Physiol **49**:223-247 (1998); Bird, *DNA Methylation de Novo*, Science **286**:2287-2288 (1999)). Understanding the physiological consequences of methylation has utility in a number of fields. It is well documented that methylation of promoter regions can repress transcription, both *in vitro* and *in vivo* (Baylin *et al.*, *Alterations in DNA methylation: a fundamental aspect of neoplasia*, Adv Cancer Res **72**:141-96 (1998)). For instance, the promoters of several genes implicated in neoplasia and tumor suppression are subject to hypermethylation (Melki et al., *Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia*, Cancer Res **59**(15):3730-40 (1999)). Methylation also performs important functions in plant development and flowering.

The use of differential methylation to study gene function traditionally required prior knowledge of DNA sequences subject to methylation, obtained only after substantial effort to clone, sequence and verify the methylation sensitivity of the region of interest. MAABST expedites the identification of differentially methylated sequences by, in its preferred forms, combining the power of high throughput microarray technology and *in silico* analysis with the sensitivity and quantitation of differential display.

MAABST has several advantages over other methods currently used to identify promoters on a genomic scale. Current approaches to identify promoter and other regulatory elements in a high throughput manner include: *in silico* analysis of nucleotide sequence for transcription factor binding sites, β-lactamase insertion (Whitney *et al.*, *A genome-wide functional assay of signal*

transduction in living mammalian cells, Nat Biotechnol **16**(13):1329-33 (1998)), COBRA (Xiong and Laird, *COBRA: a sensitive and quantitative DNA methylation assay*, Nuc Acid Res **25**(12):2532-2534 (1997)), and restriction landmark genomic scanning (Costell *et al.*, *aberrant CpG-island methylation has non-random and tumour-type-specific patterns*, Nature Genetics **25**:132-138 (2000)).

Other forms of modification are indicative of certain types of DNA damage caused by particular agents. These include alkylation, dimerization, derivatization, depurination, or ADP-ribosylation. Examples of modifications and their source are described in Lodovici et al., Levels of 8-hydroxydeoxyguanosine as a marker of DNA damage in human leukocytes, Free Radic Biol Med 28(1):13-7 (2000); Maehira et al., Alterations of protein kinase C, 8-hydroxydeoxyguanosine, and K-ras oncogene in rat lungs exposed to passive smoking, Clin Chim Acta 289(1-2):133-44 (1999); Gamboa Da Costa et al., Characterization of the Major DNA Adduct Formed by alpha-Hydroxy-N-desmethyltamoxifen in Vitro and in Vivo, Chem Res Toxicol 13(3):200-207 (2000); Phillips et al., Methods of DNA adduct determination and their application to testing compounds for genotoxicity, Environ Mol Mutagen 35(3):222-233 (2000); Airoldi et al., Carcinogen-DNA adducts as tools in risk assessment, Adv Exp Med Biol 472:231-40 (1999); Purewal et al., Association between acetylator genotype and 2-amino-1-methyl-6- phenylimidazo[4,5-b]pyridine (PhIP) DNA adduct formation in colon and prostate of inbred Fischer 344 and Wistar Kyoto rats, Cancer Lett 149(1-2):53-60 (2000).

MAABST uses the same basic steps as BEST and includes cleavage with a nucleic acid cleaving reagent and coupling of an offset adaptor to the nucleic acid fragments. The difference is that MAABST uses a nucleic acid cleaving reagent that is sensitive to modification of its recognition site. Thus, the nucleic acid cleaving reagent will either not cleave at a site that is modified or cleave only at a site that is modified. In nucleic acid fragments that have a modification at the recognition site, the fragment will not be cleaved by the nucleic acid cleavage reagent and no offset adaptor will be added to the end.

Uncleaved nucleic acid fragments can be culled from the procedure in a number of ways. For example, if the nucleic acid fragments are to be amplified

following addition of the offset adaptors, amplification can be made dependent on the presence of an offset adaptor on the end. This can be accomplished, for example by using a PCR primer complementary to sequence in the offset adaptor. The uncleaved nucleic acid fragments can also be culled by, for example, including a capture tag, sorting tag, or label in the offset adaptor. By capturing or sorting the fragments based on the presence or absence of the capture or sorting tag, only those fragments containing an offset adaptor are carried forward in the procedure.

If a label is associated with the offset adaptor, all of the fragments can be carried forward in the procedure but only those having an offset adaptor (and thus a label) will produce a detectable signal (in this scenario, the label on the offset adaptor would have to be detected in association with a detector probe and/or a ligator-detector). Use of a label in this manner in the offset adaptor can also allow discrimination of cleaved and uncleaved fragments (that is, fragments cleaved or not cleaved by the nucleic acid cleaving reagent). This can be accomplished by detecting coupling of ligator-detectors to detector probes via labels, capture tags, or sorting tags on the detector probes, ligator-detectors, or adaptor-indexers and also detecting the presence, amount, presence and amount, or absence of an offset adaptor on the fragment via its label.

MAABST can also be used to determine how one type of cell or organism influences gene expression or other biological pathways in another type of cell or organism. For example, suppose that a mouse is genetically altered via gene targeting to inactivate a particular methyltransferase gene (there are several known for mouse but for illustration purposes, assume there is only one). Cells of interest (for instance, B cells) are taken from a mouse genetically altered to inactivate the methyl transferase mechanism and mixed with other cells of interest (for instance, T cells) obtained from a normal mouse. The B and T cells are mixed together. The T cell methylation pattern between B cell mixed and non-B cell mixed could then be compared. It can be concluded that the observed changes in methylation occurred within the T cells since the B cells harbored an inactivated methyltransferase.

Using standard differential gene expression techniques, it would be very difficult for many genes, and impossible for others, to distinguish which transcripts originated in the B cells and which transcripts originated in the T

cells (since any two cell types will share expression of a number of genes). However, due to the prior genetic alterations, that is, deleting the methyltransferase gene, MAABST allows examination of gene expression in a mixed cell population. An example of MAABST is described in Illustration 7.

Mass Spectroscopy Detection

Mass spectrometry techniques can be utilized for detection in BEST. These techniques include matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy. Such techniques allow automation and rapid throughput of multiple samples and assays.

Mass spectrometry detection works better with smaller molecules so it is preferable to cut some BEST components prior to, or as part of mass spectrometry detection. A number of methods are contemplated where an oligonucleotide molecule to be detected is cut to a shorter length prior to detection by mass spectrometry. The BEST protocol would proceed as normal and, in the preferred embodiment, the surface that has the hexamer probes attached would be compatible with the source region of a matrix assisted laser desorption ionization, time of flight, mass spectrometer (MALDI-TOF-MS). The resultant fragment of the BEST process for this particular case would look something like:

```
          surface----HHHHHHIIIIZZZZ...-L-3'
                     ||||||||||||||
3'-...YYYYCCTACTAGXXXXXXXXXXXZZZZ...-5'
```

Where:

H are the hexamer probe;

3'-CTAG-5' in lower strand indicates the initial Type II restriction site (MboI);

3'-CCTAC-5'in lower strand indicates the Type IIS restriction site (FokI);

Y are the remainder of the offset adapter;

I are the quadramer from the indexing step;

Z are the remainder of the adapter-indexer;

- are linker, universal bases, mimics or other analogs;

X are complementary bases, not germane;

L is a label.

The bottom strand is SEQ ID NO:8.

For DNA samples of greater than approximately 50 bases the performance of mass spectrometry techniques is reduced . Chemical, biological, physical (thermal), and other cleaving reagents can be used to generate smaller, more optimal, sub-fragments to be analyzed in the mass spectrometer. The degree of fragmentation is somewhat tunable in instruments like the Q-TOF systems (Micromass, US head office at Suite 407N, 100 Cummings Center, Beverly, MA 01915-6101, USA.) where one can look at the parent ion, then increase the fragmentation to see the decomposition fragments and thus the sequence; such a technique is contemplated to determine the full sized sub-fragment, and infer the sequence (which is longer sequence information than for the basic BEST method) of the sub-fragment through these known tools. The detectable fragment can be top strand, bottom strand, or both strands depending upon the scheme. The label may be a cleavable mass tag or the strand need not be labeled.

There are several useful cleaving reagents for this purpose. For example, one technique is that of Szybalski (described elsewhere herein) where FokI is used to cut at a fixed distance from an arbitrary, specific, recognition site. This technique can be extended to other restriction enzymes of Type IIS or Type III. This technique can also be used twice, once to trim off the end nearer the surface, once to trim off the end further from the surface; preferably a Type II enzyme would be used to cut the end furthest from the surface.

Use of McrBC (New England Biolabs), can be used to cut at methylcytosine sites adjacent to G/A. The cut site is not well defined (approximately 30 bases) which may be used to advantage to generate the parent as well as the fragmentation set. Metal containing porphyrins attached to oligonucleotides have been shown to cut DNA very near the porphyrin when exposed to light (Texaphyrins, US5607924). One could denature and use a hybridization texaphryin and light to cleave the remaining strand. Another cleavage technology is that of Dervan (Cartwright *et al.*, *Cleavage of chromatin with methidiumpropyl-EDTA . iron(II).* Proc Natl Acad Sci U S A, **80**(11):3213-7 (1983); Schultz and Dervan, *Sequence-specific double-strand cleavage of DNA by penta-N- methylpyrrolecarboxamide-EDTA X Fe(II).* Proc Natl Acad Sci U S A, **80**(22):6834-7 (1983)). Techniques using photocleavable linkages are

described by Olejnik et al. (Olejnik *et al.*, *Photocleavable peptide-DNA conjugates: synthesis and applications to DNA analysis using MALDI-MS.* Nucleic Acids Res, **27**(23):4626-31 (1999); Olejnik et al., *Photocleavable affinity tags for isolation and detection of biomolecules.* Methods Enzymol, **291**:135-54 (1998); Olejnik et al., *Photocleavable aminotag phosphoramidites for 5'-termini DNA/RNA labeling.* Nucleic Acids Res, **26**(15):3572-6 (1998); Olejnik *et al.*, *Photocleavable biotin derivatives: a versatile approach for the isolation of biomolecules.* Proc Natl Acad Sci U S A, **92**(16):7590-4 (1995)) These linkages can be cleaved using light to release the fragment from the surface, thus providing a more gentle desorption. WO 0004036 describes photocleavable nucleotides and methods for their use.

In one embodiment, a mass label such as peptide nucleic acid (PNA) molecules (Hanvey et al., Science 258:1481-1485 (1992)) of different sequence and molecular weight can be used as labels that bind specifically to sequence in ligator-detectors or adaptor-indexers. Laser desorption of the samples is used to generate MALDI-TOF mass spectra of the PNA labels, which are released into the spectrometer and resolved by mass. The intensity of each PNA label reveals the relative amount of different components (e.g. ligator-detectors or adaptor-indexers. In other words, the PNA spectra generate scalar values that are indirect indicators of the relative abundance of the labeled component at specific locations in an array.

The mass of the fragment itself and/or its fragmentation pattern, generated, for example, by collisionally induced dissociation (CID), can be used to verify that the hexamer was correctly hybridized and will provide additional control/tester ratio information. A preferred instrument for this CID option would make use of a tandem mass spectrometer of the class of a MALDI-qQTOF as described by Loboda et al (Loboda *et al. Design and Performance of a MALDI-QqTOF Mass Spectrometer.* in *47th ASMS Conference.* 1999. Dallas, Texas.) where the first quadrapole is used to select the mass of interest and the collision cell is used to generate the fragment spectrum. It has been reported by Shaw-Smith et al (Biotechniques, **28**:958-964 (2000)) that the subdivision of a cDNA population by indexing (Unrau and Deugau, Gene **145**(2):163-9 (1994)) is subject to the generation of redundant subsets due to mismatched ligation of

adapters. In the majority of cases, the mismatch ligation involves a single mispaired base. Ligation conditions may be modified, by using thermostable ligases at high temperature, to reduce the frequency of mismatch ligation, but often the conditions that lead to reduction of mismatches also result in reduced ligation efficiency for AT-rich overhangs. One form of the disclosed method provides an alternative to indexing, which may be described as "capture of redundant subsets." Conditions for ligation of restriction enzyme fragments to adapters may be modified for high-yield ligation of perfectly matched sequences, with concurrent ligation of a few sequences with a single base mismatch. This high-yield method results in the generation of redundant subsets, instead of unique subsets. The cDNA fragments present in each redundant subset may then be further analyzed, for example, as follows:

1. A pair of adaptor-indexer-specific oligonucleotides, one of which contains a biotin, are used to amplify the cDNA fragment by PCR. For assays involving a tester and a control sample, one of the primers used to amplify one of these samples will preferably contain one or more additional 5'-terminal bases, in order to make the masses of the control and tester tags easily resolvable by the mass spectrometer.

2. Single stranded amplicons are isolated by binding to streptavidin beads, followed by release of the unbound strand.

3. The single stranded cDNA tags are hybridized with a ligator-detector oligonucleotide, and then contacted with a microarray comprising all hexamers. Ligation is performed in a solution in contact with the array surface, under conditions generating maximum specificity of hexamer base pairing. A different hexamer microarray is used for each of the 256 possible adaptor-indexers.

4. The microarray is washed to remove weakly bound cDNA tags.

5. The microarray spots are covered with a suitable matrix for performing analysis of DNA by mass spectrometry.

6. Mass analysis is performed using a MALDI source, tandem quadrupole, quadrupole, time-of-fight mass spectrometer by tuning the first quadrupole filter for transmission of a single mass-to-charge. The microarray consists of all possible hexamers; a correctly hybridized cDNA tag bound at a unique address has a single, well defined mass. In addition, for each of the 256

possible adapters, the sequence of the adapter cohesive end is known, and also corresponds to a single mass. Hence, it is possible to determine in advance which precise cDNA mass tag window to use for each of the 4096 array addresses, and for each of the 256 microarrays. For example, there exist a total of 84 possible different mass combinations for the unique cDNA tags that bind to specific addresses on a microarray of 4096 hexamers, using a single adaptor sequence. The general formula is a binomial coefficient of the form: Combinations = [(n+r-1)!]/[n!(r-1)!]. In this expression "n" is the number of bases in a detector sequence in the microarray, while "r" is the number of possible values for the mass of a base. Resolving for n=6 and r=4, r-1 = 3; [(6+3)*(6+2)*(6+1)]/3! = 9*8*7/6 = 84.

The signal measured in the mass spectrometer corresponds to the number of DNA ions that have the mass for which a specific array address has been tuned. Most DNA molecules that are present at an incorrect address will have a single base mismatch, either in the hexamer sequence, or in the adaptor-indexer sequence; all such molecules will have a different mass, and will not be detected. Molecules with two mismatches are most likely to also have a different mass, except in the special case of molecules with two mismatched bases that exactly compensate each other. Such molecules may be scored by collecting a fragmentation pattern in the collision cell of the tandem mass spectrometer their incorrect fragment masses in the fragmentation (see below).

An additional, optional step in the mass spectrometry analysis is fragmentation in the subsequent collision cell—a quadrupole acting as an ion guide and a region of relatively high pressure chemically inert gas—followed by TOF analysis of the DNA fragments. Fragmentation and TOF analysis will resolve the fragments originating from correct adapter-indexer ligation and correct hexamer ligation, from other fragments of equal mass, but different sequence, resulting from possible mutually compensatory mismatches at the adaptor-indexer ligation and the hexamer ligation steps.

Further, photocleavable nucleotides or linkers can be used which will yield a shorter, well defined, fragment from the binary sequence tag upon laser irradiation in the MALDI source. Photocleavable linkers may be used in the

attachment of the mass labels wherein the linkage will cleave and release the mass tag when the samples are subjected to the UV source of the MALDI.

Still further, use of uracil rather than thymine in the synthetic adapters (phosphoramidite chemicals available from Glenn Research) may be used in conjunction with uracil-DNA glycosylase, UDG, (available from New England Biolabs) to introduce specific strand breaks which can be designed to release particular fragments upon treatment with UDG. Such strand breaks may be engineered to be offset in the sense and antisense strands such that the complex remains substantially intact near room temperature until subjected to the laser of the MALDI source.

Probability Detection

Sequencing by hybridization is known to produce mismatch errors (Lipshutz, *Likelihood DNA sequencing by hybridization.* J Biomol Struct Dyn, 11(3):637-53 (1993)). Database searching for sequence information currently is regular expression based and requires matched "letters" between the database entry and the search sequence. BEST allows replacement of regular expression matching (match versus no-match per base) with a probability function to determine a confidence in the assignment of the identity of a binary sequence tag.

The disclosed method uses ligation to improve the specificity of the hybridization near the ligation site. Despite this improvement, there will remain a finite probability of a mismatch, particularly for nucleotides more removed from the ligation site. The error rate depends on at least two mismatch properties: base pairing (that is, A with G) and distance from the ligation site.

As an illustration of the process to determine the confidence value, consider the two bases in a hexamer probe furthest from the ligation site, numbering the bases as shown here.

```
                  <hexamer>
    surface-spacer-NNNNNNnnnn-ligator-detector-label-3' <probe>
                    ||||||||||||||||||||||||
3'-offset-adapter..NNNNNNnnnn-adapter-indexer-5'        <target>
                    123456 <position>
```

```
                        <hexamer>
surface-linker-spacer-NNNNNNnnnn-ligator-detect-signal-3'<probe>
                        ||||||||||||||||||||||||
   3'-offset-adapter..NNNNNNNNNnnnn-adapter-indexer-5'  <target>
                        123456 <position>
```

where for this particular case the structure is surface---linker-spacer------ ATXXXX, focusing on the AT (positions 1 and 2) bases for purpose of the immediate illustration.

To evaluate the possible set of sequences represented, weight matrices are used, following Dayhoff (Dayhoff et al., *A model of evolutionary changes in proteins*, in *Atlas of Protein Sequence and Structure*, Dayhoff, ed. 1978, National Biomedical Research Foundation: Washington DC) and Venezia (Venezia and O'Hara, *Rapid motif compliance scoring with match weight sets.* Comput Appl Biosci, **9**(1):65-9 (1993)). The coefficient in these matrices will be determined experimentally for the BEST system. Below is an example of matrices (with illustrative coefficients) representing position 1 and 2, where the columns represent the upper strand nucleotide and the rows represent the lower strand nucleotide. The actual coefficients can be determined empirically.

```
      Position 1                  Position 2
   A    T    C    G            A    T    C    G
A[.02,.90,.03,.05]          A[.01,.97,.01,.01]
T[.90,.02,.03,.05]          T[.97,.01,.01,.01]
C[.02,.03,.05,.90]          C[.01,.01,.01,.97]
G[.03,.02,.90,.05]          G[.01,.01,.97,.01]
```

For the case of a perfect match detection on the hexamer ATXXXX the score is determined to be the product of the coefficients of the matrices, shown below here in bold; 0.90 x 0.97 = 0.87.

```
      Position 1                  Position 2
   A    T    C    G            A    T    C    G
A[.02,.90,.03,.05]          A[.01,.97,.01,.01]
T[.90,.02,.03,.05]          T[.97,.01,.01,.01]
C[.02,.03,.05,.90]          C[.01,.01,.01,.97]
G[.03,.02,.90,.05]          G[.01,.01,.97,.01]
```

A case where a singe base mismatch in one strand occurs, for example A→G in position 1 on the hexamer side, the score is determined in a similar fashion, to be 0.05 x 0.97 = 0.05

Position 1				Position 2			
A	T	C	G	A	T	C	G
A[.02,	.90,	.03,	.05]	A[.01,	**.97**,	.01,	.01]
T[.90,	.02,	.03,	**.05**]	T[.97,	.01,	.01,	.01]
C[.02,	.03,	.05,	.90]	C[.01,	.01,	.01,	.97]
G[.03,	.02,	.90,	.05]	G[.01,	.01,	.97,	.01]

This procedure can be extended to an arbitrary number of bases in a similar manner. For a given number of nucleotides the score can be computed for all possible mismatches and rank ordered to reveal the most probable identity. A cut-off score can be used to reduce the number of possible identities from the matrix estimation. For example using the example matrices above, sequences with a threshold score above 0.50 would yield only one sequence, that being a sequence which matches the probe.

This method of estimating sequences and their respective probability scores from the universe of mismatch events for a said probe can from extended from 1 to n, where n is the number of free bases available for hybridization.

In an organism that has not been completely characterized (i.e. at least sequenced and consensus sequence assembled) a confidence value for uniqueness can be computed if a random distribution of bases is assumed. For example, if there is a candidate of 15 bases in length, in an organism which has an estimated 10^8 base genome, the 15 base fragment is expected to be unique because $10^8/4^{15}= 0.1$ is much less than 1. The genome would have to be 10 times larger before an occurrence of two instances of the particular 15 base fragment would be expected.

The distributions, in known genomes, are known not to be completely random and the initial assumption of a random distribution can be improved as information is gathered. This new information can be used to assign and use confidence values.

As an example, consider a fictitious gene family ABCD, whose members are ABCD1, ABCD2 and ABCD3. The three members were discovered

following some event such as heat shock, and they are thus putatively assigned to belong to the heat shock family of genes and happen to have significant stretches of conserved sequence among the family of genes. Also consider the organism to be a plant, where ABCD1 was isolated from the plant root, ABCD2 was isolated from the plant leaf, and ABCD3 was isolated from the plant flower. The estimation matrix may look like

```
        1    2    3
ABCD1[.60,.15,.05]
ABCD2[.25,.60,.15]
ABCD3[.05,.15,.60]
```

where the column 1 represents root, column 2 represents leaf and column 3 represents flower.

In a single experiment where there is a high confidence in the sequence but the sequence may belong to one of the three known members of the family, the source of the sample (i.e. root, leaf or flower) allow estimation of the identity of the gene. For the fully mathematically closed treatment the matrix must contain all elements of the family, here to allow for a still to be found gene in this family, the rows and columns do not add to 1; all the other members are assigned a sum of 0.05, the values to be updated as the amount of information known about the organism increases.

One can extend this estimation to include organism homology. That is, a search of a database of all organisms for a binary sequence tag from gene ABCD1 of Plant 1 may turn up matches to Plant 2, Plant 3, Mammal 1, etc. The estimation matrix would be constructed from the known organism data in the database.

The calculations and analysis described above can be illustrated using the following example of construction of a catalog. Consider a two probe detector array, a control sample, and a tester sample. Consider the two probes to have the known sequences: A, <substrate--linker--AGGGAG-3'>, and, B, <substrate--linker--ATGGAG>. These probes will capture their cognate sequence: AA, <...TCCCTC...>, and, BB, <...TACCTC...> from the control and tester samples, as well as some mismatched species with lower probability as

described herein. Utilizing the estimation matrix technique as discussed above the probabilities of the correct matching can be calculated.

The BEST procedure is conducted on the control and tester, resulting signals are collected from the probe detector array, and a catalog is made which contains the four signals:

	control			tester	
	AA	BB		AA	BB
A	.30	.03	A	.80	.10
B	.03	.50	B	.03	.50

The catalog also contains the probabilities, and/or entries derived from the probabilities, for each probe/target combination, as discussed above. For purpose of illustration, let us assume that the probability of having probe sequence A paired with target sequence AA is 0.80, and the probability of having probe sequence A paired with sequence BB is 0.10, probe sequence B paired with target sequence AA is 0.05, and the probability of having probe sequence B paired with sequence BB is 0.75, or

	estimation	
	AA	BB
A	.80	.10
B	.05	.75

It is a simple matter of application of linear algebra to determine the signals corresponding to each target. Here, for example, multiplying the corresponding entries together to convert the control and tester to the pattern corresponding to the probabilistic pattern of the target of interest. For example, the total signal ascribed, in the control sample, to AA target is 0.30 x 0.80 (on A probe site, perfect match) + 0.03 x 0.05 (on B probe site, imperfect match) = approximately 0.24. On the tester sample, the AA target signal is 0.80 x 0.80 + 0.03 x 0.05 = approximately 0.64. Comparison of the pattern for the control and tester, for the sequence corresponding to AA, exhibits an increase in the relative amount of AA from 0.24 to 0.64 for control to tester respectively. All other entries in the pattern are calculated in the same fashion.

Illustrations

The disclosed method can be further understood by way of the following illustrations which involve examples of the disclosed method. The illustrations are not intended to limit the scope of the method in any way.

Illustration 1: Mass Spectrometry Detection.

This illustration is an example of the disclosed method using known enzymes and mass spectrometry for detection. The signals of the control and tester samples are distinguished in the mass spectrometer in that the labeling moieties have different masses. The ratio of the appropriate peaks represents the ratio of the control and tester materials. For the control and tester samples, steps 1-8 are conducted in parallel, with the labels of step 8 being a heavy and a light mass tag for control and tester, respectively. The resulting mixtures for the control and tester are pooled and then contacted, simultaneously, with the array of step 9.

1. Make double stranded cDNA using reverse transcriptase following standard procedures.

2. Digest with restriction endonuclease, MboI which has a ^GATC_ recognition site. Depicted cDNA is SEQ ID NO:9.

```
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||     + MboI
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5'
```



```
          (A)                    (B)
5'-NNNNNNNNNN-3'        5'-GATCNNNNNNNNNN-3'
   ||||||||||                   ||||||||||
3'-NNNNNNNNNNCTAG-5'    3'-    NNNNNNNNNN-5'
```

3. Attach offset adapters. These offset adapters contain a Type IIS endonuclease recognition site. The offset adapters are hybridized and ligated to the correct fragments. The example Type IIS enzyme used here is FokI which has recognition and cut positions of GGATG (9/13).

```
   5'-NNNNNNNNNN-3'          5'-GATCATCC..off-adap-3'
(A)  ||||||||||                    ||||||||||||
   3'-NNNNNNNNNNCTAG-5'        3'-TAGG..off-adap-5'
```

```
    5'-off-adap..GGAT-3'          5'-GATCNNNNNNNNN-3'
(B)    ||||||||||||||       +          ||||||||||
    3'-off-adap..CCTACTAG-5'          3'-NNNNNNNNN-5'
```

4. Second digest. Cut with Type IIS restriction endonuclease. For clarity, only the 'B' fragment (GGATGATCNNNNNNNNN; SEQ ID NO:10) is followed below.

```
     5'-off-adap..GGATGATCNNNNNNNNN-3'
(B)     |||||||||||||||||||||||||||    + FokI (GGATG)
     3'-off-adap..CCTACTAGNNNNNNNNN-5'
```



```
     5'-off-adap..GGATGATCNNNNNN-3'
(B)     ||||||||||||||||||||||||
     3'-off-adap..CCTACTAGNNNNNNnnnn-5'
```

5. Divide resulting digest into 256 wells on a 384 microtiter plate. Add adapter-indexer containing a capture tag (biotin in this case), hybridize and ligate to the construct.

```
  5'-off-adap..GGATGATCNNNNNN-3'  5'-nnnn-adapter-indexer-3'
(B)  ||||||||||||||||||||||||       +     ||||||||||||||
  3'-off-adap..CCTACTAGNNNNNNnnnn-5'      adapter-indexer-Biotin
```



```
   5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)   ||||||||||||||||||||||||||||||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-Biotin
```

5a. Cleanup using biotin-advidin interaction to retain only ligated constructs.

6. Amplify. Two amplification primers are used, one designed to be complementary to a strand of the offset adapter, and another designed to be complementary to the adapter-indexer. A number of phosphorothioate linkages are included in the 3' end of the lower strand; these are used to protect against the exonuclease digest which follows.

Amplicons generated from PCR:

```
   5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)  ||||||||||||||||||||||||||||||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

7. Exonuclease digestion step to produce single stranded amplicon.

```
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

8. Add ligator-detector with a mass label, using different labels for tester and control.

```
                                5'-nnnn-ligator-detect-label
                                   ||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

9. Localize in array. An array may be constructed which contains up to, in this illustration, all 4096 hexamers. For this illustration consider the array to be on a glass slide, and to contain all 4096 hexamers covalently attached to the slide through a PEG linker-spacer.

Combine the control and tester solutions, and contact the mixture with the array under hybridization conditions. Once hybridized, ligate.

```
                                   hexamer
surface----linker-spacer------NNNNNN-3' +

                                  5'-nnnn-ligator-detect-label
                                     ||||||||||||||||||
          3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
                                     |
                                     |
                                     v
 *surface--linker-spacer- ---NNNNNNnnnn-ligator-detect-label
                             ||||||||||||||||||||||||
          3'-adapter..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

10. Detection by MALDI-TOF-MS.

Coat the glass slide with an appropriate matrix material, such as 2,5-dihydroxy benzoic acid or others as known in the art. Place the glass slide in the source region of a MALDI-TOF-MS. Individual hexamer regions are sampled by the laser, which will release the fragments complementary to the strands tethered to the surface, and detected in the time-of-fight spectrometer. Because the control and tester will have different masses in the spectrum, the ratio of the heavy to light mass signals will represent the ratio of control and tester DNA inputs. Notice that this detection method will discriminate against array

mismatched samples because a single base mismatch will cause the parent peak to shift by the mass of the base change; such mass accuracy for short oligonucleotides is known in the art.

11. Statistics and error detection.

Recognizing that each restriction cut of step 2 results in two tags, data from correlated tags, in the case of known genomes, are further utilized to error detect and intensity correct. Other options for mass spectroscopy detection are described below.

Illustration 2: Fluorescence Detection on Probe Arrays.

This illustration is an example of the disclosed method using known enzymes, probe arrays on glass slides, and fluorescence readout. The signals of the control and tester samples are distinguished by use of differing fluorescence labels, where the ratio of the appropriate peaks represents the ratio of the control and tester materials. For the control and tester samples, steps 1-8 are conducted in parallel, with the labels of step 8 being a two differing fluorescence tags for control and tester. The mixtures for the control and tester are pooled and then contacted, simultaneously, with the array of step 9.

1. Make double stranded cDNA using reverse transcriptase following standard procedures.

2. Digest with restriction endonuclease, MboI which has a ^GATC_ recognition site. Depicted cDNA is SEQ ID NO:9.

```
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||     + MboI
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5'
```



```
          (A)                    (B)
5'-NNNNNNNNNN-3'        5'-GATCNNNNNNNNNN-3'
   ||||||||||                   ||||||||||
3'-NNNNNNNNNNCTAG-5'    3'-    NNNNNNNNNN-5'
```

3. Attach offset adapters. These offset adapters contain a Type IIS endonuclease recognition site. The offset adapters are hybridized and ligated to the correct fragments. The example Type IIS enzyme used here is FokI which has recognition and cut positions of GGATG(9/13).

```
      5'-NNNNNNNNNN-3'          5'-GATCATCC..off-adap-3'
(A)     ||||||||||                    ||||||||||||
      3'-NNNNNNNNNNCTAG-5'          3'-TAGG..off-adap-5'
```

```
      5'-off-adap..GGAT-3'        5'-GATCNNNNNNNNNN-3'
(B)     ||||||||||||||      +           ||||||||||
      3'-off-adap..CCTACTAG-5'        3'-NNNNNNNNNN-5'
```

4. Second digest. Cut with Type IIS restriction endonuclease. For clarity, only the 'B' fragment (GGATGATCNNNNNNNNNN; SEQ ID NO:10) is followed below.

```
      5'-off-adap..GGATGATCNNNNNNNNNN-3'
(B)     ||||||||||||||||||||||||||||    + FokI (GGATG)
      3'-off-adap..CCTACTAGNNNNNNNNNN-5'
```



```
      5'-off-adap..GGATGATCNNNNNN-3'
(B)     ||||||||||||||||||||||
      3'-off-adap..CCTACTAGNNNNNNnnnn-5'
```

5. Divide resulting digest into 256 wells on a 384 microtiter plate. Add adapter-indexer containing a capture moiety (biotin in this case), hybridize and ligate to the construct.

```
      5'-off-adap..GGATGATCNNNNNN-3' 5'-nnnn-adapter-indexer-3'
(B)     ||||||||||||||||||||||||     +      ||||||||||||||
      3'-off-adap..CCTACTAGNNNNNNnnnn-5'    adapter-indexer-Biotin
```



```
      5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)     ||||||||||||||||||||||||||||||||||||||||||||||
      3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-Biotin
```

5a. Cleanup using biotin-advidin interaction to retain only ligated constructs.

6. Amplify. Two amplification primers are used, one designed to be complementary to a strand of the offset adapter, and another designed to be complementary to the adapter-indexer. A number of phosphorothioate linkages

are included in the 3' end of the lower strand; these are useful to protect against the exonuclease digest which follows.

Amplicons generated from PCR:

```
    5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)  ||||||||||||||||||||||||||||||||||||||||||||||||
    3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

7. Exonuclease digestion step to produce single stranded amplicon.

```
    3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

8. Add ligator-detector.

```
                                5'-nnnn-ligator-detect-label
                                   ||||||||||||||||||
    3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

9. Localize in array. An array may be constructed which contains, for example, up to all 4096 hexamers. For this illustration consider the array to be on a glass slide, and to contain all 4096 hexamers covalently attached to the slide through a PEG linker-spacer.

Combine the control and tester solutions, and contact the mixture with the array under hybridization conditions. Once hybridized, ligate.

```
                                    hexamer
surface----linker-spacer------NNNNNN-3' +

                                    5'-nnnn-ligator-detect-label
                                       ||||||||||||||||||
          3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
                                    |
                                    |
                                    v
 *surface--linker-spacer- ---NNNNNNnnnn-ligator-detect-label
                             ||||||||||||||||||||||||
          3'-adapter..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

10. Detection by fluorescence.

Place the glass slide into a slide scanner capable of at least two color detection (several such instruments are commercially available, for example

products from GSI Lumonics, Axon, Virtek, and others). Because the control and tester will have different colors, the ratio of the signals in the two colors will represent the ratio of control and tester DNA inputs. Other options for label detection are described below.

Illustration 3: Labels and sorting.

The following illustration makes use of labels, sorting, and microbeads in the disclosed method. In this illustration, 256 types of ligator-detector are each coupled to the surface of 256 different color encoded microbeads, thus each ligator-detector is identified by a single color. After annealing and ligating these new "microbead ligator-detectors" to the target sequences, the 256 sets of microbead ligator-detectors are loaded simultaneously in a 4096-well microtiter plate containing 4096 hexamer probes, one probe per well. Only a perfect match during hybridization will ligate a fluorochrome-labeled hexamer to a specific microbead labeled complex. The Luminex100 (Luminex Corporation) flow analyzer can distinguish color encoded microbeads and measure their fluorescence simultaneously, the six bases adjacent to the ligator-detector can be identified by knowing the specific address (and the hexamer contained at that address) of the wells on the microtiter plate. The design is illustrated below.

The steps up to generation of the single-stranded amplicon are the same as basic BEST, including enzyme cuts, offset adaptor ligation, Type IIS enzyme cuts, adaptor-indexer ligation, PCR amplification, and capturing and denaturing the fragments to generate single-strands of binary sequence tags.

The 256 single stranded binary sequence tags are annealed to 256 ligator-detectors. There are 256 different sequences of ligator-detectors, which comprise the 256 different ligator-detectors complementary to each of the 4-base sticky ends generated. The ligator-detector is labeled with a signaling moiety, such as a fluorescent dye or a fluorescent bead. Each of the 256 annealed single stranded preparations is split into 4096 aliquots and hybridized to one of 4096 hexamer probes. For example, the 4096 hexamers are contained in a 4096 well microtiter plate, with each well containing a single hexamer probe. Each of the 4096 hexamer probes is fluorescently labeled at the 5'-end and contains a free 3'-hydroxyl end.

Following hybridization and ligation of the hexamer probe with the annealed single-stranded amplicons, the 256 preparations are transferred to another well containing streptavidin and unbound material is washed away (the hexamer probe also contains a biotin group). Two signals are then measured.

One signal corresponds to the sample (either tester or control, for instance). Tester and control samples are distinguished by fluorescently encoded beads (Luminex). In one embodiment, 512 colors can be used; 256 colors for the tester and 256 colors for the control. However, 256 color encoded beads could also be used where the beads are "offset" between the tester and control. For instance, color 1 corresponds to hexamer 2 for the tester but hexamer 3 for the control, and so on. The second signal derives from the labeled hexamer probe and measures the level of the single-stranded DNA annealed to the ligator-detector. The two signals measured here could be read simultaneously on an instrument such as the Luminex100.

Illustration 4: BEST with Early Amplification.

In some cases, particularly for the detection of very low abundance mRNA, an early amplification is beneficial. In illustrations 1 and 2 above, the sample is divided into 256 aliquots prior to an amplification step. In an initial sample with a plurality of cDNA molecules, some species are expected to be present at concentrations below the detectable limit of the system after being divided into the 256 aliquots. To overcome such a concentration limitation, the following form of the method introduces an amplification step prior to dividing the sample into 256 equivalent pools.

1. Make double stranded cDNA.

2. Select a set of restriction enzymes that includes both Type IIS and Type II restriction enzymes. The sets can be designed to address different experimental outcomes through the selection of recognition sites, number of bases in the recognition site, reach of the Type IIS enzyme, etc., with the design such that all or most of the unknown bases between the two restriction sites are determined. Here FokI (IIS) and MboI (II) are utilized.

3. First digest cDNA with Type II restriction endonuclease, 4 base overhang on the 5' end is preferred. MboI has the recognition site ^GATC. Depicted cDNA is SEQ ID NO:9.

```
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||     + MboI
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5'
```

yields two fragments, (A) and (B)

```
            (A)                      (B)
5'-NNNNNNNNNN-3'          5'-GATCNNNNNNNNNN-3'
   ||||||||||                       ||||||||||
3'-NNNNNNNNNNCTAG-5'      3'-    NNNNNNNNNN-5'
```

4. Add an offset adapter that contains a Type IIS recognition site by hybridization and ligation. Here the Type IIS is FokI with the recognition site GGATG(9/13). This adapter also contains a region to be used as a universal primer for a PCR in a later step, call the region UP1.

```
   5'-NNNNNNNNNN-3'          5'-GATCATCC..UP1-adapter-3'
(A)   ||||||||||                    ||||||||||||||||||
   3'-NNNNNNNNNNCTAG-5'          3'-TAGG..UP1-adapter-5'

    5'-adapter-UP1..GGAT-3'       5'-GATCNNNNNNNNNN-3'
(B)    ||||||||||||          +          ||||||||||
    3'-adapter-UP1..CCTACTAG-5'       3'-NNNNNNNNNN-5'
```

5. Second digest. Type IIS restriction endonuclease.

The nascent constructs are digested using FokI. For clarity, only the 'B' fragment (GGATGATCNNNNNNNNNN; SEQ ID NO:10) is followed below.

```
     5'-adapter-UP1..GGATGATCNNNNNNNNNN-3'
(B)     ||||||||||||||||||||||||||||||||    + FokI (GGATG)
     3'-adapter-UP1..CCTACTAGNNNNNNNNNN-5'
```

yields

```
     5'-adapter-UP1..GGATGATCNNNNNN-3'
(B)     ||||||||||||||||||||||||||
     3'-adapter-UP1..CCTACTAGNNNNNNnnnn-5'
```

6. Add a plurality of adapter-indexers. There are 256 different types of adapter-indexers, each adapter is constructed with a common section encoding a

universal primer (UP2), an adapter-specific PCR section (AS1) followed by a adapter-specific 4 base 5' overhang. A possible AS1 would be the complement of the 4 base 5'-overhang.

```
5'-adapter-UP1..GGATGATCNNNNNN-3'     5'-nnnn-AS1-UP2-adapter-3'
   ||||||||||||||||||||||||||||||  +          ||||||||||||||||
3'-adapter-UP1..CCTACTAGNNNNNNnnnn-5'     3'-AS1-UP2-adapter-5'
```

To yield

```
5'-adapter-UP1..GGATGATCNNNNNNnnnn-AS1-UP2-adapter-3'
   ||||||||||||||||||||||||||||||||||||||||||||||||
3'-adapter-UP1..CCTACTAGNNNNNNnnnn-AS1-UP2-adapter-5'
```

7. PCR amplify using universal primers complementary to UP1 and UP2. This round of amplifications would yield amplification on the order of 500 times.

8. Divide amplicons into plurality of wells. Here consider all 256 possibilities and transfer to 256 wells of a 384 well microtiter plate.

9. Perform further PCR. Now use universal primers for UP1 with biotin attached and the second primer is specific for a particular AS1, thus specific (1 of 256) biotinylated adapter specific amplicons are generated.

```
   biotin-5'-adapter-UP1..GGATGATCNNNNNNnnnn-AS1-UP2-adapter-3'
(B)          ||||||||||||||||||||||||||||||||||||||||||||||||
          3'-adapter-UP1..CCTACTAGNNNNNNnnnn-AS1-UP2-adapter-5'
```

10. The remaining steps mirror those disclosed in Illustration 1.

Cleanup using biotin-advidin interaction to retain constructs, create single stranded fragments:

```
3'-adapter-UP1..CCTACTAGNNNNNNnnnn-AS1-UP2-adapter-5'
```

Add ligator-detector:

```
                              5'-nnnn-AS1-UP2-ligator-detect-signal
                                 ||||||||||||||||||||||||||||||
3'-adapter-UP1..CCTACTAGNNNNNNnnnn-AS1-UP2-adapter-indexer-5'
```

Hybridize to probe array of 4096 hexamers:

```
surface-linker-spacer--NNNNNNnnnn-AS1-UP2-ligator-detect-signal
                       ||||||||||||||||||||||||||||||||||||||
   3'-adapter..CCTACTAGNNNNNNnnnn-AS1-UP2-adapter-indexer-5'
```

Detect by mass spectrometry (or fluorescence, if the label is fluorometric as in Illustration 2).

Illustration 5: Pyrosequencing.

Detection of binary sequence tags may be made by any sequencing technique known in the art. A preferred technique is pyrosequencing. A preferred instrument for this is available from Pyrosequencing AB, Vallongatan 1, SE-752 28 Uppsala, Sweden. Using this instrument the unknown bases between the Type II and Type IIS recognition sites may be sequenced in the following manner.

1. Make double stranded cDNA

2. Cleave with Type II enzyme, preferably one with a 4 base recognition and 4 base 5' overhang.

3. Prepare offset adapters having the Type IIS recognition site and the compatible overhang to the overhang of the Type II cleavage in the manner described in illustration 1.

4. Hybridize and ligate the offset adapter to the cleaved fragments from step 2.

5. Cleave the cDNA with Type IIS enzyme which cleaves at a site offset from the recognition sequence introduced by the addition of offset adaptors.

6. Distribute the solution from step 5 into 256 replicate instances in 256 wells of a 384 well microtiter plate. These are index samples.

7. Ligate indexing adapters. Each index sample is incubated with a different adapter-indexer each of which has a sticky end compatible with one of the possible sticky ends on the DNA fragments in that index sample. The adapter-indexers are then coupled onto compatible DNA fragments by ligation, resulting in the formation of binary sequence tags with adapters covalently attached to each end.

8. Dilute. Each index sample of binary sequence tags is diluted to achieve a concentration of less than one molecule per well when the sample is

distributed to 96, 384, or 1536 well microtiter plates in the following step. The dilution will depend upon a number of factors including the number of wells into which the sample subsequently is distributed, the Type II cleavage frequency, and the initial input amount of DNA.

9. Distribute. The contents of each of the wells of step 8, which contain one type of adapter indexer each, are distributed to a number of wells. In this illustration, each well is transferred to 96 wells, for a total of 256 microtiter plates.

10. Amplify. Each aliquot of the binary sequence tags is amplified using a suitable amplification method, such as PCR. Two PCR primers are used, one designed to be complementary to a sequence of the offset-adapter and the other designed to be complementary to the adapter-indexer. Because there should only be one molecule in the well which can be amplified, there should be one predominant molecular species following the amplification.

11. Cleanup. The amplified binary sequence tags may be cleaned up using any number of methods known in the art. The preferred method is to have a capture tag incorporated in the PCR primer, for example biotin, such that single stranded amplicons are retained for further steps.

12. Detection by Pyrosequencing. Using a sequencing primer complementary to one of the adapter sequences the unknown, adjacent bases can be directly determined through the technique of pyrosequencing. Because there is predominantly one single stranded DNA fragment from each well, there should be one predominant sequence corresponding to that fragment. The commercial instrument is capable of sequencing ten bases per well, for 96 wells, in about 10 minutes. The expression level of a given binary sequence tag is proportional to the number of times the sequence occurs.

Illustration 6: Hairpin Primers.

This illustration describes the use of hairpin primers. The method described in illustration 1 can be adapted to make use of a cleavable hairpin primer and release the tag in the following fashion:

1. Produce binary sequence tags as in Illustration 1 and then perform the PCR step using a hairpin primer, the hairpin primer containing a uracil in the hairpin sequence, or near the hairpin sequence. The hairpin primer is a nucleic

acid molecule that contains a primer sequence and that can form a stem-loop or hairpin structure. A hairpin structure forms at the end of the amplified fragments and facilitates coupling of the fragments to the probes (see next step). The hairpin structure serves the function of the ligator-detector hybridized to a binary sequence tag by forming a duplex region next to a single-stranded region. This allows the end of the amplified fragment to be ligated to a probe array (see next step).

2. Hybridize and ligate onto the probe hexamer array.

```
    xxxxxxNNNNNNNNNNNNNNN
    |||||||||||||||||    U   Labile Hairpin (control)
...nnnnnnnnnnNNNNNNNNNN

    xxxxxxNNNNNNNNNNNNNNN
    |||||||||||||||||    U   Labile Hairpin (tester)
...nnnnnnnnnnNNNNNNNNNNM
```

where x is the hexamer probe, N is the hairpin, M is an additional base or bases, n is the binary sequence tag, | indicates base pairing.

3. Wash with alkali to remove non-ligated tag-hairpins.

4. Cleave with uracil-DNA glycosylase.

The released fragment to be analyzed will be:

```
...nnnnnnnnnnNNNNNNNNNN    (control)
...nnnnnnnnnnNNNNNNNNNNM    (tester)
```

5. Detect the cleaved tags, resolving the two different masses, using MALDI-TOF. Use of a tandem mass spectrometer, as described above, to fragment the cleaved tags will determine some or all of the tag sequence, and improve the signal to noise.

Hairpin primers may be utilized to multiplex the readout of the control and tester of a tag from the same address of an surface array.

1. Produce binary sequence tags from cDNA as in Illustration 1 and then generate PCR products using hairpin primers. Use different hairpin primers for the tester and control, a uracil in the synthetic adapters for the testers and a thymine in the synthetic adapters for the controls. A fluorescence label my be incorporated into the hairpin using standard fluorescent labeled nucleotides.

2. Hybridize and ligate to probe array.

```
    xxxxxxNNNNNNNNNNNNNNN
    ||||||||||||||||||    dT (Stable Hairpin, Control)
...nnnnnnnnnnNNNNNNNNN*NN

    xxxxxxNNNNNNNNNNNNNNN
    ||||||||||||||||||     U (Labile Hairpin, Tester)
...nnnnnnnnnnNNNNNNNNN*NN
```

where x is the hexamer probe, N is the hairpin, n is the binary sequence tag, | indicates base pairing, * indicates a fluorescently labeled nucleotide.

3. Read signal before uracil-DNA glycosylase, this corresponds to control plus tester total signal.

4. Cleave hairpins containing uracil using uracil-DNA glycosylase. The thymine containing hairpin will remain intact.

5. Wash the slide.

6. Read signal AFTER uracil-DNA glycosylase, this corresponds to control signal only.

Illustration 7: MAABST with Mass Spectrometry Detection.

This illustration is a detailed look using known enzymes and mass spectrometry for detection. The signals of the control and tester samples are distinguished in the mass spectrometer in that the signaling moieties have different masses, where the ratio of the appropriate peaks represents the ratio of the control and tester materials.

1. Make double stranded cDNA.

2. Digest with a methylation sensitive restriction endonuclease. Depicted cDNA is SEQ ID NO:9.

```
                  CH3
                  |
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||   + methyl sensitive RE
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5'

                  ↓

                  CH3
                  |
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||    if nucleic acid is methylated
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5'
```

```
            (A)                    (B)
5'-NNNNNNNNNN-3'          5'-GATCNNNNNNNNN-3'
   ||||||||||                     |||||||||
3'-NNNNNNNNNCTAG-5'       3'-    NNNNNNNNN-5'
```

if nucleic acid is unmethylated

3. Create Type IIS constructs. Add offset adapter encoding a Type IIS endonuclease by hybridization and ligation. FokI = GGATG(9/13).

```
                  CH3
                  |
5'-NNNNNNNNNNGATCNNNNNNNNNN-3'
   ||||||||||||||||||||||||||    if nucleic acid is methylated,
3'-NNNNNNNNNNCTAGNNNNNNNNNN-5' sequence is not cut and
                                ligation is not possible,
                                and no binary sequence tags are
                                generated
```

if DNA is cut, offset adapter is ligated:

```
   5'-NNNNNNNNNN-3'          5'-GATCATCC..adapter-3'
(A)   ||||||||||                    ||||||||||||
   3'-NNNNNNNNNCTAG-5'          3'-TAGG..adapter-5'

    5'-off-adap..GGAT-3'        5'-GATCNNNNNNNNN-3'
(B)    ||||||||||||||      +          ||||||||||
    3'-off-adap..CCTACTAG-5'       3'-NNNNNNNNN-5'
```

4. Second digest, Type IIS restriction endonuclease. For clarity, only the 'B' fragment (GGATGATCNNNNNNNNN; SEQ ID NO:10) is followed below.

```
     5'-off-adap..GGATGATCNNNNNNNNN-3'
(B)    |||||||||||||||||||||||||||    + FokI (GGATG)
     3'-off-adap..CCTACTAGNNNNNNNNN-5'
                      |
                      ▼
     5'-off-adap..GGATGATCNNNNNN-3'
(B)    ||||||||||||||||||||||||
     3'-off-adap..CCTACTAGNNNNNNnnnn-5'
```

5. Divide cDNA digest into 256 wells on a 384 microtiter plate. Add immobilizing indexer adapter and ligate to construct.

```
   5'-off-adap..GGATGATCNNNNNN-3'     5'-nnnn-adapter-indexer-3'
(B)   ||||||||||||||||||||||||        +        ||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-5'         adapter-indexer-
                                                        Biotin
```



```
   5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)   ||||||||||||||||||||||||||||||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-Biotin
```

5a. Cleanup using biotin-advidin interaction to retain constructs.

6. Amplification. Two amplification primers are used, one designed to be complementary to a strand of the offset adapter, and another designed to be complementary to the adapter-indexer. A number of phosphorothioate linkages are included in the 3' end of the lower strand; these are necessary to protect against the exonuclease digest which follows.

Amplicons generated from PCR

```
   5'-off-adap..GGATGATCNNNNNNnnnn-adapter-indexer-3'
(B)   ||||||||||||||||||||||||||||||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

7. Exonuclease digestion step to produce single stranded amplicon.

```
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

8. Add ligator-detector.

```
                              5'-nnnn-ligator-detect-signal
                                 ||||||||||||||||||
   3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

9. Hybridize to solid surface index of 4096 hexamers. Ligate.

```
        hexamer
surface----linker-spacer------NNNNNN-3' +

                                    5'-nnnn-ligator-detect-signal
                                       ||||||||||||||||||
      3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```



```
*surface--linker-spacer----NNNNNNnnnn-ligator-detect-signal
                           ||||||||||||||||||||||||
      3'-off-adap..CCTACTAGNNNNNNnnnn-adapter-indexer-5'
```

10. Place the surface in the source region of a MALDI- TOF-MS. Individual hexamer regions are sampled by the laser and detected in the time-of-fight spectrometer. Ratio of the heavy to light mass signals will represent the ratio of control and tester DNA inputs. As an option, the mass of the DNA strand itself and/or it's fragmentation pattern can be used to verify that the hexamer was correctly hybridized and will provide additional control/tester ratio information.

11. Recognizing that each restriction cut of step 2 results in two tags, data from correlated binary sequence tags, in the case of known genomes, are further utilized to error detect and correct.

CLAIMS

We claim:

1. A method of producing binary sequence tags from nucleic acid fragments in a nucleic acid sample, the method comprising

(a) incubating a nucleic acid sample with one or more first nucleic acid cleaving reagents to produce nucleic acid fragments,

(b) mixing one or more offset adaptors with the nucleic acid sample and covalently coupling the offset adaptors to the nucleic acid fragments,

(c) incubating the nucleic acid sample with one or more second nucleic acid cleaving reagents to produce nucleic acid fragments with sticky ends, wherein the second nucleic acid cleaving reagents cleave at a site offset from their recognition sequence, wherein each offset adaptor has a recognition sequence for at least one of the second nucleic acid cleaving reagents,

(d) mixing one or more adaptor-indexers with the nucleic acid sample and covalently coupling the adaptor-indexers to the nucleic acid fragments, wherein each adaptor-indexer has a different sticky end, wherein each sticky end of the adaptor-indexers is compatible with a sticky end generated by the second nucleic acid cleaving reagents,

wherein the nucleic acid fragments to which offset adaptors and adaptor-indexers have been coupled are binary sequence tags.

2. The method of claim 1 wherein the binary sequence tags are amplified, detected, identified, sequenced, cataloged, or a combination.

3. The method of claim 2 wherein the binary sequence tags are detected, wherein detection comprises determining, directly or indirectly, the presence, amount, presence and amount, or absence of one or more binary sequence tags.

4. The method of claim 1 wherein the offset adaptors and the adaptor-indexers are covalently coupled to the nucleic acid fragments by ligation.

5. The method of claim 1 wherein the first and second nucleic acid cleaving reagents are restriction enzymes.

6. The method of claim 5 wherein the recognition sequence of the restriction enzymes are from four to thirty nucleotides in length.

7. The method of claim 5 wherein the recognition sequence of the restriction enzymes are from four to ten nucleotides in length.

8. The method of claim 5 wherein the recognition sequence of the restriction enzymes are from four to eight nucleotides in length.

9. The method of claim 5 wherein the first restriction enzymes have a four base recognition sequence, wherein one second restriction enzyme is used, and wherein the second restriction enzyme is a Type IIS restriction enzyme that cleaves at a site different from its recognition sequence.

10. The method of claim 1 wherein binary sequence tags are amplified.

11. The method of claim 10 wherein the binary sequence tags are labeled during amplification.

12. The method of claim 10 wherein binary sequence tags are amplified by the polymerase chain reaction.

13. The method of claim 1 further comprising

hybridizing the binary sequence tags to detector probes.

14. The method of claim 13 wherein the detector probes are coupled to beads.

15. The method of claim 14 wherein the beads contain a label.

16. The method of claim 15 wherein the label is a molecular barcode.

17. The method of claim 15 wherein the label is a mass label.

18. The method of claim 15 further comprising

sorting or separating the binary sequence tags via the labels.

19. The method of claim 14 further comprising

sorting or separating the binary sequence tags via the beads.

20. The method of claim 13 wherein the detector probes are immobilized on a substrate in an array.

21. The method of claim 13 wherein the detector probes contain a label.

22. The method of claim 21 further comprising

sorting or separating the binary sequence tags via the labels.

23. The method of claim 1 further comprising

detecting the binary sequence tags by mass spectroscopy.

24. The method of claim 23 wherein the binary sequence tags are detected by matrix-assisted laser desorption/ionization time-of-flight mass spectroscopy.

25. The method of claim 23 wherein, prior to, or during detection the binary sequence tags are fragmented by collisionally induced dissociation.

26. The method of claim 23 wherein one or both strands of the binary sequence tags are cleaved or partially degraded prior to detection.

27. The method of claim 26 wherein the binary sequence tags are partially degraded using Texaphyrin technology.

28. The method of claim 23 wherein the binary sequence tags are hybridized to detector probes prior to detection.

29. The method of claim 28 wherein the detector probes are immobilized on a substrate in an array.

30. The method of claim 1 further comprising

sequencing all or a portion of one or more of the binary sequence tags.

31. The method of claim 30 wherein the portion of the binary sequence tag that does not correspond to sequence in either the offset adaptor or the adaptor-indexer is sequenced.

32. The method of claim 30 further comprising, prior to sequencing

diluting each nucleic acid sample and dividing into aliquots containing, on average, one binary sequence tag molecule per aliquot, and

amplifying the binary sequence tag in each aliquot.

33. The method of claim 1 wherein each adaptor-indexer has a double-stranded portion, wherein the double-stranded portion of each adaptor-indexers is different.

34. The method of claim 1 wherein each adaptor-indexer has a double-stranded portion, wherein the double-stranded portion of each adaptor-indexers is the same.

35. The method of claim 1 wherein each adaptor-indexer has a double-stranded portion, wherein the double-stranded portions of at least two of adaptor-indexers are different.

36. The method of claim 1 wherein each adaptor-indexer has a double-stranded portion, wherein the double-stranded portions of at least two of adaptor-indexers are the same.

37. The method of claim 1 wherein each adaptor-indexer contains a label.

38. The method of claim 37 further comprising

sorting or separating the binary sequence tags via the labels.

39. The method of claim 1 further comprising

separating the strands of the binary sequence tags.

40. The method of claim 39 wherein each binary sequence tag comprises a top strand and a bottom strand,

wherein the top strands have offset adaptor sequence at their 5' ends, wherein the bottom strands have adaptor-indexer sequence at their 5' ends,

wherein the top strands of the binary sequence tags are separated from the bottom strands of the binary sequence tags.

41. The method of claim 40 further comprising

hybridizing one or more ligator-detectors with either the top strands of the binary sequence tags or the bottom strands of the binary sequence tags, wherein each ligator-detector comprises sequence matching or complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers,

hybridizing the nucleic acid sample with a detector array comprising one or more probes and covalently coupling the ligator-detectors to the probes, wherein each probe has a different sequence, and

detecting, directly or indirectly, coupling of ligator-detectors to the detector array probes.

42. The method of claim 1 further comprising

hybridizing one or more ligator-detectors with the binary sequence tags, wherein each ligator-detector comprises sequence matching or complementary to all or part of sequence including, and adjacent to, the sticky end of at least one of the adaptor-indexers,

hybridizing the nucleic acid sample with a detector array comprising one or more probes and covalently coupling the ligator-detectors to the probes, wherein each probe has a different sequence, and

detecting, directly or indirectly, coupling of ligator-detectors to the detector array probes.

43. The method of claim 42 wherein coupling of the ligator-detector is detected by rolling circle replication of an amplification target circle wherein replication is primed by the ligator-detector.

44. The method of claim 43 wherein uncoupled ligator-detectors do not prime rolling circle replication of an amplification target circle.

45. The method of claim 42 wherein the nucleic acid cleaving reagents generate sticky ends having N different sequences, and wherein the sample is divided into N index samples.

46. The method of claim 42 wherein the detector array probes are all of the same length.

47. The method of claim 46 wherein the detector array probes are six, seven, or eight nucleotides long.

48. The method of claim 42 wherein the detector array probes all have similar hybrid stability.

49. The method of claim 42 wherein each offset adaptor, adaptor-indexer, ligator-detector, or detector array probe contains a label, wherein coupling of the ligator-detectors to the probes is detected via the label.

50. The method of claim 1 wherein the collection of binary sequence tags produced from the nucleic acid sample constitutes a catalog of binary sequence tags for the nucleic acid sample.

51. The method of claim 1 further comprising

performing steps (a) through (d) on a plurality of nucleic acid samples.

52. The method of claim 51 further comprising

performing steps (a) through (d) on a control nucleic acid sample,

identifying differences between the collection of binary sequence tags produced from the nucleic acid samples and the control nucleic acid sample.

53. The method of claim 52 wherein the differences are differences in the presence, amount, presence and amount, or absence of binary sequence tags produced from the nucleic acid samples and the control nucleic acid sample.

54. The method of claim 51 wherein the steps (a) through (d) are performed on a control nucleic acid sample and a tester nucleic acid sample,

wherein the tester nucleic acid sample, or the source of the tester nucleic acid sample, is treated, prior to step (a), so as to destroy, disrupt or eliminate one or more nucleic acid molecules in the tester nucleic acid sample,

wherein the binary sequence tags corresponding to the destroyed, disrupted, or eliminated nucleic acid molecules will be produced from the control nucleic acid sample but not the tester nucleic acid sample.

55. The method of claim 54 wherein the tester nucleic acid sample is treated so as to destroy, disrupt or eliminate one or more nucleic acid molecules in the tester nucleic acid sample.

56. The method of claim 55 wherein the tester and control nucleic acid samples are samples of messenger RNA, the method further comprising, prior to step (a)

reverse transcribing the messenger RNA molecules to produce first cDNA strands of the messenger RNA molecules,

destroying or disrupting one or more first cDNA strands by sequence-specific cleavage in the tester nucleic acid sample but not the control nucleic acid sample,

synthesizing second cDNA strands from the first DNA strands.

57. The method of claim 55 wherein the tester and control nucleic acid samples are samples of messenger RNA, the method further comprising, prior to step (a)

reverse transcribing the messenger RNA molecules to produce first and second cDNA strands of the messenger RNA molecules,

destroying or disrupting one or more second cDNA strands by sequence-specific cleavage in the tester nucleic acid sample but not the control nucleic acid sample.

58. The method of claim 54 wherein the source of the tester nucleic acid sample is treated so as to destroy, disrupt or eliminate one or more nucleic acid molecules in the tester nucleic acid sample.

59. The method of claim 58 wherein the treatment of the source is accomplished by exposing cells from which the tester sample will be derived with a compound, composition, or condition that will reduce or eliminate expression of one or more genes.

60. The method of claim 54 further comprising identifying differences in the binary sequence tags produced from the control nucleic acid sample and tester nucleic acid sample.

61. The method of claim 51 further comprising

identifying differences between the collection of binary sequence tags produced from the nucleic acid samples.

62. The method of claim 61 further comprising identifying or preparing nucleic acid fragments corresponding the binary sequence tags produced from one or more of the nucleic acid samples but not from one or more of the other nucleic acid samples.

63. The method of claim 1 further comprising identifying or preparing nucleic acid fragments corresponding the binary sequence tags produced from the nucleic acid sample.

64. The method of claim 63 further comprising using the prepared nucleic acid fragments as probes for analyzing a different nucleic acid sample.

65. The method of claim 64 wherein analysis of the different nucleic acid sample involves detection, quantitation, identification, comparison, screening, sequencing, culling, destruction, sorting, capturing, or a combination, of nucleic acid molecules in the different nucleic acid sample.

66. The method of claim 1 wherein a single type of offset adaptor is used.

67. The method of claim 1 wherein one second nucleic acid cleaving reagent is used, wherein the second nucleic acid cleaving reagent is a Type IIS restriction enzyme.

68. The method of claim 1 wherein at least one of the first nucleic acid cleaving reagents is sensitive to modification of its recognition site.

69. The method of claim 68 wherein the first and second nucleic acid cleaving reagents are restriction enzymes.

70. The method of claim 69 wherein the modification to the recognition site is methylation, alkylation, dimerization, derivatization, depurination, or ADP-ribosylation.

71. The method of claim 69 wherein the modification is present in the nucleic acid fragments when isolated or is introduced to the nucleic acid fragments after isolation.

72. The method of claim 69 further comprising, following coupling of the adaptor-indexers to the nucleic acid fragments,

amplifying the nucleic acid fragments in the index samples to which offset adaptors and adaptor-indexers have been coupled.

73. The method of claim 72 further comprising determining the sequence of a portion of at least one of the nucleic acid fragments in the nucleic acid sample.

74. The method of claim 69 further comprising, following coupling of the offset adaptors to the nucleic acid fragments,

separating nucleic acid fragments coupled to offset adaptors from nucleic acid fragments not coupled to offset adaptors, wherein only nucleic acid fragments coupled to offset adaptors are used in step (c).

75. The method of claim 69 wherein at least one of the first restriction enzymes (1) is insensitive to modification of its recognition site and (2) has the same recognition site as the first restriction enzyme that is sensitive to modification of its recognition site,

the method further comprising, prior to digestion with the first restriction enzymes,

dividing each index sample into a set of two or more of index samples,

wherein each secondary index sample in each set of index samples is digested with a different first restriction enzyme,

wherein steps (a) through (d) are performed with each of the secondary index samples.

76. The method of claim 75 further comprising,

comparing the pattern of the presence or absence of binary sequence tags made using the first restriction enzyme that is sensitive to modification of its recognition site with the pattern of the presence or absence of binary sequence tags made using the first restriction enzyme that is insensitive to modification of its recognition site and that has the same recognition site as the first restriction enzyme that is sensitive to modification of its recognition site,

wherein differences in the patterns indicate modification of nucleic acids in the nucleic acid sample.

77. The method of claim 69 wherein the pattern of the presence, amount, presence and amount, or absence of binary sequence tags constitutes a catalog of nucleic acid fragments in the nucleic acid sample.

78. The method of claim 77 further comprising preparing a second catalog of nucleic acid fragments in a second nucleic acid sample and comparing the first catalog and second catalog, wherein differences in the first and second catalogs indicate differences in modification of the first and second nucleic acid samples.

79. The method of claim 77 wherein the second nucleic acid sample is a sample from the same type of cells as the first nucleic acid sample except that the cells from which the first nucleic acid sample is derived are modification-deficient relative to the cells from which the second nucleic acid sample is derived.

80. The method of claim 77 wherein the second nucleic acid sample is a sample from a different type of cells than the first nucleic acid sample, and wherein the cells from which the first nucleic acid sample is derived are modification-deficient relative to the cells from which the second nucleic acid sample is derived.

81. The method of claim 1 wherein the nucleic acid sample is a cDNA sample, wherein the cDNA is synthesized in the presence of one or more dideoxy nucleoside triphosphates,

the method further comprising

identifying correlated binary sequence tags.

82. The method of claim 81 wherein the correlated binary sequence tags are identified by

calculating the expression levels and the expression ratios of a plurality of the binary sequence tags,

identifying binary sequence tags that have similar expression ratios,

grouping the binary sequence tags that have similar expression ratios in binary pairs, to generate a list of correlated binary sequence tags,

ordering the binary pairs in the list of correlated binary sequence tags according to expression level of the binary sequence tags, and

fitting the values of the expression levels to a standard curve based on an inverse exponential function.

83. The method of claim 82 further comprising

identifying correlated binary sequence tags using a different a cDNA sample, wherein the cDNA is synthesized in the presence of a different concentration of the one or more dideoxy nucleoside triphosphates or a different set of one or more dideoxy nucleoside triphosphates,

correlating changes in expression of each list of correlated binary sequence tags with the predicted slope changes predicted by the inverse exponential function corresponding to each of the dideoxy terminator levels, and

predicting the order and position within the cDNA of each list of binary tags.

84. The method of claim 81 wherein the cDNA is synthesized using one or more anchored primers.

85. The method of claim 1 further comprising

performing steps (a) through (d) on a second nucleic acid sample.

86. The method of claim 85 wherein the second nucleic acid sample is a sample from the same type of organism as the first nucleic acid sample.

87. The method of claim 85 wherein the second nucleic acid sample is a sample from the same type of tissue as the first nucleic acid sample.

88. The method of claim 85 wherein the second nucleic acid sample is a sample from the same organism as the first nucleic acid sample.

89. The method of claim 88 wherein the second nucleic acid sample is obtained at a different time than the first nucleic acid sample.

90. The method of claim 85 wherein the second nucleic acid sample is a sample from a different organism than the first nucleic acid sample.

91. The method of claim 85 wherein the second nucleic acid sample is a sample from a different type of tissue than the first nucleic acid sample.

92. The method of claim 85 wherein the second nucleic acid sample is a sample from a different species of organism than the first nucleic acid sample.

93. The method of claim 85 wherein the second nucleic acid sample is a sample from a different strain of organism than the first nucleic acid sample.

94. The method of claim 85 wherein the second nucleic acid sample is a sample from a different cellular compartment than the first nucleic acid sample.

95. The method of claim 1 further comprising,

dividing the nucleic acid sample into a plurality of index samples.

96. The method of claim 95 wherein a different adaptor-indexer is mixed with each index sample.

97. The method of claim 95 wherein a different second nucleic acid cleaving reagent is mixed with each index sample.

98. The method of claim 95 wherein a different offset adaptor is mixed with each index sample.

99. The method of claim 95 wherein a different first nucleic acid cleaving reagent is mixed with each index sample.

100. The method of claim 95 wherein the nucleic acid sample is divided into a plurality of index samples prior to step (d).

101. The method of claim 95 wherein the nucleic acid sample is divided into a plurality of index samples prior to step (c).

102. The method of claim 95 wherein the nucleic acid sample is divided into a plurality of index samples prior to step (b).

103. The method of claim 95 wherein the nucleic acid sample is divided into a plurality of index samples prior to step (a).

104. The method of claim 95 wherein one or more of the index samples are divided into a plurality of secondary index samples.

105. The method of claim 104 wherein a different adaptor-indexer is mixed with each secondary index sample.

106. The method of claim 104 wherein a different second nucleic acid cleaving reagent is mixed with each secondary index sample.

107. The method of claim 104 wherein a different offset adaptor is mixed with each secondary index sample.

108. The method of claim 104 wherein one or more of the secondary index samples are divided into a plurality of tertiary index samples.

109. The method of claim 108 wherein a different adaptor-indexer is mixed with each tertiary index sample.

110. The method of claim 108 wherein a different second nucleic acid cleaving reagent is mixed with each tertiary index sample.

111. The method of claim 95 wherein the binary sequence tags are amplified prior to dividing the nucleic acid sample into a plurality of index samples.

112. The method of claim 95 wherein the binary sequence tags are amplified following dividing the nucleic acid sample into a plurality of index samples.

113. The method of claim 112 further comprising

prior to amplifying the binary sequence tags, diluting each index sample and dividing into aliquots containing, on average, one binary sequence tag molecule per aliquot, and

following amplifying the binary sequence tags, sequencing all or a portion of one or more of the binary sequence tags are sequenced.

114. The method of claim 1 further comprising, following step (d),

mixing the nucleic acid sample with one or more different hairpin primers, wherein each hairpin primer comprises a different primer sequence, wherein each primer sequence is complementary to all or part of the sequence of at least one of the adaptor-indexers,

incubating the nucleic acid sample under conditions that promote amplification of binary sequence tags, wherein amplified binary sequence tags are formed which have hairpin primer sequences at one or both ends,

incubating the nucleic acid sample under conditions that promote formation of hairpin structures by the hairpin primer sequences at the ends of the amplified binary sequence tags,

hybridizing the nucleic acid sample with a plurality of detector probes and covalently coupling the hairpin structures to the probes, wherein each probe has a different sequence, and

detecting coupling of the amplified fragments to different detector probes.

115. The method of claim 1 wherein the concentration of the binary sequence tags is normalized.

116. The method of claim 115 wherein the concentration of the binary sequence tags is normalized by immobilizing one strand of the binary sequence tags, denaturing the binary sequence tags, renaturing the binary sequence tags for a time greater than the $c_0t_{1/2}$ for abundant binary sequence tags and less than the $c_0t_{1/2}$ for rare binary sequence tags, and collecting the un-renatured binary sequence tags.

117. A method of producing binary sequence tags from nucleic acid fragments in a nucleic acid sample, the method comprising

(a) incubating a nucleic acid sample with one or more first nucleic acid cleaving reagents to produce nucleic acid fragments,

(b) mixing one or more first offset adaptor strands with the nucleic acid sample and covalently coupling the first offset adaptor strands to the nucleic acid fragments, wherein, after coupling, the first offset adaptor strands are fully or partially single-stranded,

(c) treating the nucleic acid sample to result in full or partial complementary sequences hybridized to the first offset adaptor strands,

(d) incubating the nucleic acid sample with one or more second nucleic acid cleaving reagents to produce nucleic acid fragments with sticky ends, wherein the

second nucleic acid cleaving reagents cleave at a site offset from their recognition sequence, wherein each first offset adaptor strand has a recognition sequence for at least one of the second nucleic acid cleaving reagents,

(e) mixing one or more adaptor-indexers with the nucleic acid sample and covalently coupling the adaptor-indexers to the nucleic acid fragments, wherein each adaptor-indexer has a different sticky end, wherein each sticky end of the adaptor-indexers is compatible with a sticky end generated by the second nucleic acid cleaving reagents,

wherein the nucleic acid fragments to which offset adaptors and adaptor-indexers have been coupled are binary sequence tags.

118. The method of claim 117 wherein treatment of the nucleic acid sample to result in full or partial complementary sequences hybridized to the first offset adaptor strands is accomplished by hybridizing second offset adaptor strands to the coupled first offset adaptor strands.

119. The method of claim 118 further comprising covalently coupling the second offset adaptor strands to the nucleic acid fragments.

120. The method of claim 117 wherein treatment of the nucleic acid sample to result in full or partial complementary sequences hybridized to the first offset adaptor strands is accomplished by filling in the single-stranded portion of the first offset adaptor strands.

121. A method of producing binary sequence tags from nucleic acid fragments in a nucleic acid sample, the method comprising

(a) incubating a nucleic acid sample with one or more first nucleic acid cleaving reagents to produce nucleic acid fragments,

(b) mixing one or more offset adaptors with the nucleic acid sample and covalently coupling the offset adaptors to the nucleic acid fragments,

(d) incubating the nucleic acid sample with one or more second nucleic acid cleaving reagents to produce nucleic acid fragments with sticky ends, wherein the second nucleic acid cleaving reagents cleave at a site offset from their recognition sequence, wherein each first offset adaptor strand has a recognition sequence for at least one of the second nucleic acid cleaving reagents,

(e) mixing one or more first adaptor-indexer strands with the nucleic acid sample and covalently coupling the first adaptor-indexer strands to the nucleic acid

fragments, wherein each first adaptor-indexer strand has a different end sequence, wherein each end sequence of the first adaptor-indexer strands is compatible with a sticky end generated by the second nucleic acid cleaving reagents, wherein, after coupling, the first adaptor-indexer strands are fully or partially single-stranded,

wherein the nucleic acid fragments to which offset adaptors and first adaptor-indexer strands have been coupled are binary sequence tags.

122. The method of claim 121 further comprising treating the nucleic acid sample to result in full or partial complementary sequences hybridized to the first adaptor-indexer strands.

123. The method of claim 122 wherein treatment of the nucleic acid sample to result in full or partial complementary sequences hybridized to the first adaptor-indexer strands is accomplished by hybridizing second adaptor-indexer strands to the coupled first adaptor-indexer strands.

124. The method of claim 123 further comprising covalently coupling the second adaptor-indexer strands to the nucleic acid fragments.

125. The method of claim 122 wherein treatment of the nucleic acid sample to result in full or partial complementary sequences hybridized to the first adaptor-indexer strands is accomplished by filling in the single-stranded portion of the first adaptor-indexer strands.

126. A method of producing binary sequence tags from nucleic acid fragments in a nucleic acid sample, the method comprising

(a) incubating a nucleic acid sample with one or more first nucleic acid cleaving reagents to produce nucleic acid fragments,

(b) mixing one or more first offset adaptor strands with the nucleic acid sample and covalently coupling the first offset adaptor strands to the nucleic acid fragments, wherein, after coupling, the first offset adaptor strands are fully or partially single-stranded,

(c) treating the nucleic acid sample to result in full or partial complementary sequences hybridized to the first offset adaptor strands,

(d) incubating the nucleic acid sample with one or more second nucleic acid cleaving reagents to produce nucleic acid fragments with sticky ends, wherein the second nucleic acid cleaving reagents cleave at a site offset from their recognition

sequence, wherein each first offset adaptor strand has a recognition sequence for at least one of the second nucleic acid cleaving reagents,

(e) mixing one or more first adaptor-indexer strands with the nucleic acid sample and covalently coupling the first adaptor-indexer strands to the nucleic acid fragments, wherein each first adaptor-indexer strand has a different end sequence, wherein each end sequence of the first adaptor-indexer strands is compatible with a sticky end generated by the second nucleic acid cleaving reagents, wherein, after coupling, the first adaptor-indexer strands are fully or partially single-stranded,

wherein the nucleic acid fragments to which offset adaptors and adaptor-indexers have been coupled are binary sequence tags.

FIG. 1A

Adaptor-indexer I

CATGCGGATCCTAAGGCTTACGCC
GCCTAGGATTCCGAATGCGG

Adaptor-indexer II

TAATGGAAGCTGGATTCGCG
GTTCATTACCTTCGACCTAAGCGC

Ligator-detectors

1. **CATG**CGGATCCTAAGGCTTACGCC
2. **CATG**CGGATCCTAAGGCTTACGC
3. **CATG**CGGATCCTAAGGCTTA
4. **CATG**CGGATCCTAAGGC
5. **ATG**CGGATCCTAAGGCTTACGCC
6. **TG**CGGATCCTAAGGCTTACGCC
7. **G**CGGATCCTAAGGCTTACGCC

Ligator-detectors

21. **CAAG**TAATGGAAGCTGGATTCGCG
22. **CAAG**TAATGGAAGCTGGATTCGC
23. **CAAG**TAATGGAAGCTGGATTC
24. **CAAG**TAATGGAAGCT
25. **AAG**TAATGGAAGCTGGATTCGCG
26. **AG**TAATGGAAGCTGGATTCGCG
27. **G**TAATGGAAGCTGGATTCGCG

RECTIFIED SHEET (RULE 91)
ISA/EP

FIG. 1B

8. **ATG**CGGATCCTAAGGCTTACGC
9. **ATG**CGGATCCTAAGGCTTACG
10. **ATG**CGGATCCTAAGGCTT
11. **ATG**CGGATCCTAAGGC
12. **TG**CGGATCCTAAGGCTTAC
13. GGCGTAAGCCTTAGGATCCG**CATC**
14. GCGTAAGCCTTAGGATCCG**CATC**
15. GTAAGCCTTAGGATCCG**CATC**
16. CCTTAGGATCCG**CATC**
17. GGCGTAAGCCTTAGGATCCG**CAT**
28. **AAG**TAATGGAAGCTGGATTCGC
29. **AAG**TAATGGAAGCTGGATTC
30. **AAG**TAATGGAAGCTGGAT
31. **AAG**TAATGGAAGCTG
32. **AG**TAATGGAAGCTGGATTCGCG
33. CGCGAATCCAGCTTCCATTA**CTTG**
34. GCGAATCCAGCTTCCATTA**CTTG**
35. GAATCCAGCTTCCATTA**CTTG**
36. CAGCTTCCATTA**CTTG**
37. CGCGAATCCAGCTTCCATTA**CTT**

FIG. 1C

18. GGCGTAAGCCTTAGGATCCG**CA**

19. GGCGTAAGCCTTAGGATCCG**C**

20. G**CATG**CGGATCCTAAGGCTTACGCC

38. CGCGAATCCAGCTTCCATTA**CT**

39. CGCGAATCCAGCTTCCATTA**C**

40. GAATCCAGCTTCCATTA**CTT**

SEQUENCE LISTING

<110> Yale University

<120> Binary Encoded Sequence Tags

<130> AGL 100 PCT

<140> Not Yet Assigned
<141> 2000-08-11

<160> 10

<170> PatentIn Ver. 2.1

<210> 1
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
Ligator-detector

<400> 1
gcatgcggat cctaaggctt acgcc 25

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

<220>
<223> Description of Artificial Sequence:
Ligator-detector

<400> 2
ggcgtaagcc ttaggatccg catc 24

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

<220>
<223> Description of Artificial Sequence:

Ligator-detector

<400> 3
caagtaatgg aagctggatt cgcg 24

<210> 4
<211> 24
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence:
Ligator-detector

<400> 4
cgcgaatcca gcttccatta cttg 24

<210> 5
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer

<400> 5
ttttttttttt ttttttgc 18

<210> 6
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer

<400> 6
ttttttttttt ttttttgt 18

<210> 7
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Primer

<400> 7
ttttttttttt ttttttca 18

<210> 8
<211> 26
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Nucleic Acid Fragment

<400> 8
nnnnnnnnnn nnnngatcat ccnnnn 26

<210> 9
<211> 24
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Depicted cDNA

<400> 9
nnnnnnnnnn gatcnnnnnn nnnn 24

<210> 10
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Nucleic Acid Fragment

<400> 10
ggatgatcnn nnnnnnnn 18