SOLID PHASE SELECTION OF DIFFERENTIALLY EXPRESSED GENES

This is a continuation-in-part of co-pending U.S. patent application Ser. No. 09/130,446 filed 6 August 1998, which is a continuation-in-part of co-pending U.S. patent application Ser. No. 09/005,222 filed 9 January 1998, which applications are incorporated by reference.

FIELD OF THE INVENTION

The invention relates generally to methods for identifying differentially expressed genes, and more particularly, to a method of competitively hybridizing differentially expressed DNAs with reference DNA sequences cloned on solid phase supports to provide a differential expression library which can be physically manipulated, *e.g.* by fluorescence-activated flow sorting.

BACKGROUND

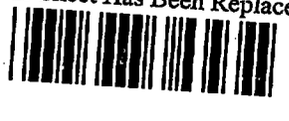
The desire to decode the human genome and to understand the genetic basis of disease and a host of other physiological states associated differential gene expression has been a key driving force in the development of improved methods for analyzing and sequencing DNA, Adams et al., Editors, Automated DNA Sequencing and Analysis (Academic Press, New York, 1994). The human genome is estimated to contain about 10^5 genes, 15-30% of which--or about 20-40 megabases--are active in any given tissue. Such large numbers of expressed genes make it difficult to track changes in expression patterns by available techniques, especially in view of the large number of genes that are expressed at relative low levels: It has been estimated that as much as 30% of mRNA consists of many thousands of distinct species each making up less than 0.5% of the total, and typically averaging less than 14 copies per cell, Sambrook et al., Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory Press, New York, 1989). Even substantial changes in expression among such low abundance mRNAs can be difficult to detect in the presence overwhelming quantities of abundant sequences.

A variety of techniques are available for analyzing gene expression that differ widely in convenience, expense, and sensitivity. Commonly used low resolution

techniques include differential display, indexing, subtraction hybridization, and numerous DNA fingerprinting techniques, e.g. Vos et al., Nucleic Acids Research, 23: 4407-4414 (1995); Hubank et al., Nucleic Acids Research, 22: 5640-5648 (1994); Lingo et al., Science, 257: 967-971 (1992); Erlander et al., International patent application PCT/US94/13041; McClelland et al., U.S. patent 5,437,975; Unrau et al., Gene, 145: 163-169 (1994); and the like. Higher resolution techniques include analysis of expressed sequence tags (ESTs), e.g. Adams et al. (cited above); analysis of concatenated fragments of expressed sequences (SAGE), e.g. Velculescu et al., Science, 270: 484-486 (1995); Zhang et al., Science, 276: 1268-1272 (1997); Velculescu et al., Cell, 88: 243-251 (1997); and the use of microarrays of oligonucleotides or polynucleotides for capturing complementary polynucleotides from expressed genes, e.g. Schena et al., Science, 270: 467-469 (1995); DeRisi et al., Science, 278: 680-686 (1997); Chee et al., Science, 274: 610-614 (1996); and the like.

The latter two high resolution techniques have shown promise as potentially robust systems for analyzing gene expression; however, there are still technical issues that need to be addressed with both approaches. In microarray systems, genes to be monitored must be known and isolated beforehand, which means different microarrays, or "DNA chips," have to be manufactured for each specialized use and for every different type of organism or species examined. With respect to microarrays constructed from fluid-delivered cDNAs, a significant degree of variability, e.g. 2-5 fold, exists in the signals generated under the same hybridization conditions, Atlas™ cDNA Expression System Users Manual (Clontech Laboratories, Palo Alto, 1998), and the systems are not readily re-usable. With respect to microarrays of synthetic oligonucleotides, a significant set-up cost for manufacturing such arrays and expensive chip-reading instruments put such systems beyond the financial capability of many potential users. In sequence tag systems, although no special instrumentation is necessary, as an extensive installed base of DNA sequencers may be used, even routine expression analysis requires a significant sequencing effort, e.g. several thousand sequencing reactions or more; the selection of type IIs tag-generating enzymes is limited; and the length (nine nucleotides) of the sequence tag in current protocols severely limits the number of cDNAs that can be uniquely labeled. It can be shown that for organisms expressing large sets of genes, such as mammalian cells, the likelihood of nine-nucleotide tags being distinct for all expressed genes is extremely

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

low, e.g. Feller, An Introduction to Probability Theory and Its Applications, Second Edition, Vol. I (John Wiley & Sons, New York, 1971).

It is clear from the above that there is a need for a convenient and sensitive technique for analyzing gene expression that permits the analysis of either known or unknown genes from any source. The availability of such a technique would find immediate application not only in medical and scientific research, but also in a host of applied fields, such as crop and livestock development, pest management, drug development, diagnostics, disease management, and the like.

SUMMARY OF THE INVENTION

Accordingly, objects of our invention include, but are not limited to, providing a method for identifying and isolating differentially expressed genes; providing a method of identifying and isolating polynucleotides on the basis of labels that generate different optical signals; providing a method for profiling gene expression of large numbers of genes simultaneously; providing a method of identifying and separating genes in accordance with whether their expression is increased or decrease under any given conditions; providing a method for identifying rare genes; and providing a method for massively parallel signature sequencing of large numbers of genes isolated according to their expression.

Our invention accomplishes these and other objects by providing differently labeled populations of polynucleotides from cell or tissue sources whose gene expression is to be compared. In comparing gene expression, differently labeled polynucleotides of a plurality of populations are competitively hybridized with reference DNA cloned on solid phase supports. Preferably, the solid phase supports are microparticles which, after such competitive hybridization, provide a differential expression library which may be manipulated by fluorescence-activated cell sorting (FACS), or other sorting means responsive to optical signals generated by labeled polynucleotides on the microparticles. Monitoring the relative signal intensity of the different labels on the microparticles permits quantification of the relative expression of particular genes in the different populations.

In one aspect of the invention, populations of microparticles having relative signal intensities of interest are isolated by FACS and the attached polynucleotides are sequenced to determine the identities of the rare or differentially expressed genes.

Preferably, the method of the invention is carried out by the following steps: a) providing a reference population of nucleic acid sequences attached to separate solid phase supports in clonal subpopulations; b) providing a population of polynucleotides of expressed genes from each of the plurality of different cells or tissues, the polynucleotides of expressed genes from different cells or tissues having a different light-generating label; c) competitively hybridizing the populations of polynucleotides of expressed genes from each of the plurality of different cells or tissues with the reference population to form duplexes between the sequences of the reference population and polynucleotides of each of the different cells or tissues such that the polynucleotides are present in duplexes on each of the solid phase supports in ratios directly related to the relative expression of their corresponding genes in the different cells or tissues; and d) detecting a relative optical signal generated by the light-generating labels of the duplexes attached thereto. In further preference, the method includes the step of sorting each solid phase support according to the relative optical signal detected. Preferably, the reference population of nucleic acids is derived from genes of the plurality of different cells or tissues being analyzed. As used herein, the phrase "polynucleotides of expressed genes" is meant to include any RNA produced by transcription, including in particular mRNA, and DNA produced by reverse transcription of any RNA, including in particular cDNA produced by reverse transcription of mRNA.

The present invention overcomes shortcoming in the art by providing compositions, methods, and kits for separating and identifying genes that are differentially expressed without requiring any previous analysis or knowledge of the sequences. The invention also permits differentially regulated genes to be separated from unregulated genes for analysis, thereby eliminating the need to analyze large numbers of unregulated genes in order to obtain information on the genes of interest.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1a and 1b illustrate FACS analysis of microparticles loaded with competitively hybridized DNA strands labeled with two different fluorescent dyes.

Figure 2 is a schematic representation of a flow chamber and detection apparatus for observing a planar array of microparticles loaded with restriction fragments for sequencing.

Figure 3a illustrates a preferred scheme for converting isolated messenger RNA (mRNA) into cDNA and insertion of the cDNA into a tag-containing vector.

Figure 3b illustrates a preferred scheme for amplifying tag-cDNA conjugates out of a vector and loading the amplified conjugates onto microparticles.

Figure 3c illustrates a preferred scheme for isolating sorted cDNAs for cloning and sequencing.

Figure 4a and 4b illustrate alternative procedures for cloning differentially expressed cDNAs isolated by FACS sorting.

Figures 5a-e illustrate flow analysis data of microparticles carrying predetermined ratios of two differently labeled cDNAs.

Figure 6 illustrates flow analysis data of microparticles carrying differently labeled cDNAs from stimulated and unstimulated THP-1 cells.

Figure 7 illustrates flow analysis data of microparticles carrying labeled cDNAs derived from mRNA of low abundance in stimulated THP-1 cells.

Figure 8 illustrates flow analysis data of microparticles carrying labeled cDNAs derived from mRNA of low abundance in human bone marrow.

Figure 9 illustrates flow analysis data of microparticles carrying differently labeled cDNAs from glucose normal and glucose starved muscle tissue.

Figure 10A illustrates an embodiment of the invention for constructing a reference nucleic acid population on microparticles.

Figure 10B illustrates an embodiment for using the reference library of Figure 10A to compare gene expression of two cell populations.

Definitions

"Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides, anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, *e.g.* 3-4, to several tens of monomeric units, *e.g.* 40-60. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the art when oligonucleotides having natural or non-natural nucleotides may be employed, *e.g.* where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which

every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, *e.g.* as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, *e.g.* described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides, usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide, *e.g.* cytosines, within the target polynucleotide "CATCGC ..." so that its sequence is represented as a binary code, *e.g.* "100101 ... " for "C-(not C)-(not C)-C-(not C)-C ... " and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of polynucleotide present in the population.

As used herein, the term "relative gene expression" or "relative expression" in reference to a gene refers to the relative abundance of the same gene expression product, usually an mRNA, in different cells or tissue types.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides compositions, methods, and kits for analyzing relative gene expression in a single or plurality of cell and/or tissue types that are of interest. The methods of the invention can be applied to polynucleotides derived from animals, plants, and microorganisms such as fungi, bacteria, mycoplasma, cyanobacteria, algae, and the like. Preferably, the polynucleotides are derived from animals, plants or microorganisms involved in fermentation process, with vertebrates and agricultural plants being most preferred. The plurality usually comprises a pair of cell or tissue types, such as a diseased tissue or cell type and a healthy tissue or cell type, or such as a cell or tissue type being subjected to a stimulus or stress, *e.g.* a change of nutrients, temperature, or the like, and the corresponding cell or tissue type in an unstressed or unstimulated state. In another embodiment, the plurality can comprise a pair of cell or tissue types having homologous genes, such as cells or tissue from different organisms. The plurality may also include more than two cell or tissue types, such as would be required in a comparison of expression patterns of the same cell or tissue over time, *e.g.* liver cells after exposure of an organism to a candidate drug, organ cells of a test animal at different developmental states, and the like. Preferably, the plurality is 2 or 3 cell or tissue types; and more preferably, it is 2 cell or tissue types.

The method of the invention typically comprises providing a reference population of nucleic acid sequences attached to separate solid phase supports in clonal subpopulations, providing at least one population of polynucleotides of expressed genes, hybridizing the population(s) of polynucleotides of expressed genes with the reference nucleic acid population, and detecting, and preferably sorting each solid phase support according to a relative optical signal generated by the duplexes attached thereto.

Figure 10A illustrates an embodiment of the invention for constructing a reference nucleic acid population on microparticles, and Figure 10B illustrates an embodiment for using such a reference library to compare gene expression of two cell populations. Messenger RNA (mRNA) is extracted (1004) from cell populations (1000) and (1002) using conventional protocols to give two populations of polynucleotides (1006) and (1008), respectively. The extraction reactions can be carried out separately or on a mixture of cell types. Preferably, the reactions are

carried out separately so that the relative quantities of mRNA from the two populations can be more readily controlled. Portions of mRNA (1006) and mRNA (1008) are combined (1010) and cDNA library (1012) is constructed in vectors carrying a repertoire of oligonucleotide tags, in accordance with the procedure described in Brenner et al., U.S. patent 5,846,719. Preferably, equal portions of mRNA, *i.e.*, equal molar quantities, are taken from each population of mRNA. A sample of vectors from library (1012) is taken and amplified, *e.g.* by polymerase chain reaction, transfection and cloning, or the like, after which the tag-cDNA conjugates (1014) carried by the vectors are excised or copied (1011) and then isolated. Loaded microparticles are then formed and prepared for use in competitive hybridization as follows (1018). The isolated tag-cDNA conjugates (1014), illustrated with oligonucleotide tags a, b, c, and d, are specifically hybridized to microparticles carrying their tag complements a', b', c', and d' (1016), respectively. The tag-cDNA conjugates are ligated to tag complements so that at least one strand of the double stranded tag-cDNA conjugate is covalently attached to the microparticle. Microparticles carrying tag-cDNA conjugates are separated from those that do not carry tag-cDNA conjugates, preferably using a fluorescence-activated cell sorter (FACS), or like instrument. The non-covalently attached strand is melted off and separated from the microparticles to yield microparticles (1020) carrying a reference nucleic acid population.

As illustrated in Figure 10b, gene expression of cells (1050) may be compared to that of cells (1052) by separately extracting (1054) mRNA (1056) and (1058) from each cell type. After construction of cDNA libraries (1062) and (1064) using conventional protocols, single stranded nucleic acid probes are generated from the respective cDNA populations (1062) and (1064), the probes preferably being labeled with optically distinguishable fluorescent dyes F (1068) and R (1066), *e.g.*, rhodamine and fluorescein. Equal amounts of the labeled polynucleotides are mixed and hybridized (1072) to the complementary strands carried by the microparticles to form duplexes (1074). After the hybridization is complete, microparticles carrying the duplexes thereby formed (1074) can be sorted (1076) in accordance to predetermined criteria, such as fluorescence ratio, fluorescence intensity, and/or the like. In such a manner, subpopulations of interest can be isolated and further analyzed, *e.g.*, those corresponding to up-regulated or down-regulated genes.

For analysis in accordance with the invention, messenger RNA (mRNA) is extracted from the cells or tissues of interest using conventional protocols, as disclosed in, for example, Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York). Preferably, the populations of mRNAs to be compared are converted into populations of labeled cDNAs by reverse transcription in the presence of a labeled nucleoside triphosphate using conventional protocols, *e.g.* Schena et al., Science 270: 467-470 (1995); DeRisi et al., Science 278: 680-686 (1997); or the like, prior to hybridization to a reference DNA population.

An important feature of the invention is that the genes whose expression levels change or are different than those of the other cells or tissues being examined may be analyzed separately from those that are not regulated or otherwise altered in response to whatever stress or condition is being studied. As described below, in the preferred embodiment gene products from the cells or tissues of interest are competitively hybridized with a reference population consisting of DNA sequences attached in clonal subpopulations to separate microparticles. As a result, microparticles carrying labeled gene products in ratios indicating differential expression may be manipulated and analyzed separately from those carrying labeled gene products in ratios indicating no change in expression, *e.g.* "house-keeping" genes, genes encoding structural proteins, or the like.

Another important feature of the invention is that the identity of the nucleic acid being analyzed, *e.g.*, genomic DNA or gene products such as cDNA, mRNA, RNA transcript, or the like, need not be known prior to analysis. After relative expression is determined, cDNAs derived from expressed genes may be identified by direct sequencing on the solid phase support, preferably a microparticle, using a number of different sequencing approaches. For identification, only a portion of the cDNAs need be sequenced. In many cases, the portion may be as small as nine or ten nucleotides, *e.g.* Velculescu et al. (cited above). Preferably, entire subpopulations of differentially expressed genes are sequenced simultaneously using MPSS, or a similar parallel analysis technique. In a preferred embodiment, this is conveniently accomplished by providing a reference population of DNA sequences such that each such sequence is attached to a separate microparticle in a clonal subpopulation. As used herein, the phrase "clonal subpopulation" refers to multiple copies of a single

kind of polynucleotide selected from a population of interest, such as a cDNA library constructed from mRNA extracted from a cell or tissue whose gene expression is being analyzed. Such clonal subpopulations may be formed in a number of ways, including by separate amplification of a poynucleotide and attacment by conventional attachment chemistries, *e.g.*, Hermansen, Bioconjugate Techniques (Academic Press, New York, 1996). As explained more fully below, clonal subpopulations are preferably formed by so-called "solid phase cloning" disclosed in Brenner, U.S. patent 5,604,097 and Brenner et al., U.S. patent 5,846,719, which are incorporated herein by reference. Briefly, such clonal subpopulations are formed by hybridizing an amplified sample of tag-DNA conjugates onto one or more solid phase support(s), *e.g.*, separate and unconnected microparticles, so that individual microparticles, or different regions of a larger support, have attached multiple copies of the same DNA.

The DNA component of the tag-DNA conjugate can be cDNA, genomic DNA, a fragment of cDNA or genomic DNA, or a synthetic DNA, such as, for example, an oligonucleotide. Preferably the tag-DNA conjugate is a cDNA or a fragment of genomic DNA ("gDNA"). The number of copies of a cDNA or gDNA in a clonal subpopulation may vary widely in different embodiments depending on several factors, including the density of tag complements on the solid phase supports, the size and composition of microparticle used, the duration of hybridization reaction, the complexity of the tag repertoire, the concentration of individual tags, the tag-DNA sample size, the labeling means for generating optical signals, the particle sorting means, signal detection system, and the like.

Guidance for making design choices relating to these factors is readily available in the literature on flow cytometry, fluorescence microscopy, molecular biology, hybridization technology, and related disciplines, as represented by the references cited herein. Preferably, the number of copies of a cDNA or a gDNA in a clonal subpopulation is sufficient to permit FACS detection and/or sorting of microparticles, wherein fluorescent signals are generated by one or more fluorescent dye molecules carried by the cDNAs attached to the microparticles. Typically, this number can be as low as a few thousand, *e.g.* 3,000-5,000, when a fluorescent molecule such as fluorescein is used, and as low as several hundred, *e.g.* 800-8000, when a rhodamine dye, such as rhodamine 6G, is used. More preferably, when loaded microparticles are detected and/or sorted by FACS or like instruments, clonal

subpopulations consist of at least 10^4 copies of a cDNA or gDNA; and most preferably, in such embodiments, clonal subpopulations consist of at least 10^5 copies of a cDNA or gDNA.

Labeled cDNAs or RNAs from the cells or tissues to be compared are competitively hybridized to the DNA sequences of the reference DNA population using conventional hybridization conditions, *e.g.* such as disclosed in Schena et al. (cited above); DeRisi et al. (cited above); or Shalon, Ph.D. Thesis entitled "DNA Microarrays," Stanford University (1995). After hybridization, an optical signal is generated by each of the two labeled species of cDNAs or RNAs so that a relative optical signal is determined for each microparticle. Preferably, such optical signals are generated and measured in a fluorescence activated cell sorter, or like instrument, which permits the microparticles to be sorted and accumulated whose relative optical signal fall with a predetermined range of values. The microparticles loaded with cDNAs or RNAs generating relative optical signals in the desired range may be isolated and identified by sequencing, such as with MPSS, as described more fully below.

Preferably, clonal subpopulations of cDNAs or other DNA molecules derived from RNA are attached to microparticles using the processes illustrated in Figures 3a and 3b. First, as illustrated in Figure 3a, mRNA (300) is extracted from a cell or tissue source of interest using conventional techniques and is converted into cDNA (309) with ends appropriate for inserting into vector (316). Preferably, primer (302) having a 5' biotin (305) and poly(dT) region (306) is annealed to mRNA strands (300) so that the first strand of cDNA (309) is synthesized with a reverse transcriptase in the presence of the four deoxyribonucleoside triphosphates. Preferably, 5-methyldeoxycytidine triphosphate is used in place of deoxycytosine triphosphate in the first strand synthesis, so that cDNA (309) is hemi-methylated, except for the region corresponding to primer (302). This allows primer (302) to contain a non-methylated restriction site for releasing the cDNA from a support. The use of biotin in primer (302) is not critical to the invention and other molecular capture techniques, or moieties, can be used, *e.g.* triplex capture, or the like. Region (303) of primer (302) preferably contains a sequence of nucleotides that results in the formation of restriction site r_2 (304) upon synthesis of the second strand of cDNA (309). After isolation by binding the biotinylated cDNAs to streptavidin supports, *e.g.* Dynabeads

M-280 (Dynal, Oslo, Norway), or the like, cDNA (309) is preferably cleaved with a restriction endonuclease which is insensitive to hemimethylation (of the C's) and which recognizes site r_1 (307). Preferably, r_1 is a four-base recognition site, *e.g.* corresponding to Dpn II, or like enzyme, which ensures that substantially all of the cDNAs are cleaved and that the same defined end is produced in all of the cDNAs. After washing, the cDNAs are then cleaved with a restriction endonuclease recognizing r_2, releasing fragment (308) which is purified using standard techniques, *e.g.* ethanol precipitation, polyacrylamide gel electrophoresis, or the like. After resuspending in an appropriate buffer, fragment (308) is directionally ligated into vector (316), which carries tag (310) and a cloning site with ends (312) and (314). Preferably, vector (316) is prepared with a "stuffer" fragment in the cloning site to aid in the isolation of a fully cleaved vector for cloning.

Preparation of the tag-cDNA conjugates is not limited to the method described above and can readily be achieved in a variety of ways using conventional molecular biology techniques. For example, cDNA can be prepared by conventional methods and isolated by gel electrophoresis. This method is less preferred in part because it would bias the size distribution of the reference population. The tag can be attached by ligation of adaptors, by PCR with an oligo dT primer and a random primer, or by RACE technology (Bertling et al. (1993) *PCR Methods Appl.* 3:95-99; Frohman, M.A. (1993) *Methods Enzymol.* 218:340-356; Marathon™ CDNA Amplification Kit, Clontech Laboratories, Inc.). Attachment of the tag by cloning into a vector, as described above, is preferred for several reasons, including the ability to generate large quantities of the reference population (versus RACE, which typically yields only μg quantities), and the ability to check the sequence of the tag.

After formation of a library of tag-cDNA conjugates, a sample of host cells is usually plated to determine the number of recombinants per unit volume of culture medium. The size of sample taken for further processing preferably depends on the size of tag repertoire used in the library construction. As taught by Brenner et al., U.S. patent 5,846,719 and Brenner et al., U.S. patent 5,604,097, a sample preferably includes a number of conjugates equivalent to about one percent the size of the tag repertoire in order to minimize the selection of "doubles," *i.e.* two or more conjugates carrying the same tag and different cDNAs. Thus, for a tag repertoire consisting of a concatenation of eight 4-nucleotide "words" selected from a minimally cross-

This Sheet Has Been Replaced



hybridizing set of eight words, the size of the repertoire is 8^8, or about 1.7×10^7 tags. Accordingly, with such a repertoire, a sample of about 1.7×10^5 conjugate-containing vectors is preferably selected for amplification and further processing as illustrated in Figure 3b.

Preferably, tag-cDNA conjugates are carried in vector (330) which comprises the following sequence of elements: first primer binding site (332), restriction site r_3 (334), oligonucleotide tag (336), junction (338), cDNA (340), restriction site r_4 (342), and second primer binding site (344). After a sample is taken of the vectors containing tag-cDNA conjugates the following steps are implemented: The tag-cDNA conjugates are preferably amplified from vector (330) by use of biotinylated primer (348) and labeled primer (346) in a conventional polymerase chain reaction (PCR) in the presence of 5-methyldeoxycytidine triphosphate, after which the resulting amplicon is isolated by streptavidin capture. Restriction site r_3 preferably corresponds to a rare-cutting restriction endonuclease, such as Pac I, Not I, Fse I, Pme I, Swa I, or the like, which permits the captured amplicon to be release from a support with minimal probability of cleavage occurring at a site internal to the cDNA of the amplicon. Junction (338) which is illustrated as the sequence:

```
5'... GGGCCC ...
3'... CCCGGG ...
```

causes the DNA polymerase "stripping" reaction to be halted at the G triplet, when an appropriate DNA polymerase is used with dGTP. Briefly, in the "stripping" reaction, the 3'→5' exonuclease activity of a DNA polymerase, preferably T4 DNA polymerase, is used to render the tag of the tag-cDNA conjugate single stranded, as taught by Brenner, U.S. patent 5,604,097; and Kuijper et al., Gene, 112: 147-155 (1992). In the preferred embodiment where sorting is accomplished by formation of duplexes between tags and tag complements, tags of tag-cDNA conjugates are rendered single stranded by first selecting words that contain only three of the four natural nucleotides, and then by preferentially digesting the three nucleotide types from the tag-cDNA conjugate in the 3'→5' direction with the 3'→5' exonuclease activity of a DNA polymerase. In the preferred embodiment, oligonucleotide tags are designed to contain only A's, G's, and T's; thus, tag complements (including that in the

double stranded tag-cDNA conjugate) consist of only A's, C's, and T's. When the released tag-cDNA conjugates are treated with T4 DNA polymerase in the presence of dGTP, the complementary strands of the tags are "stripped" away to the first G. At that point, the incorporation of dG by the DNA polymerase balances the exonuclease activity of the DNA polymerase, effectively halting the "stripping" reaction. From the above description, it is clear that one of ordinary skill could make many alternative design choices for carrying out the same objective, *i.e.* rendering the tags single stranded. Such choices could include selection of different enzymes, different compositions of words making up the tags, and the like.

When the "stripping" reaction is quenched, the result is duplex (356) with single stranded tag (357). After isolation, steps (358) are implemented: the tag-cDNA conjugates are hybridized to tag complements attached to microparticles, a fill-in reaction is carried out to fill any gap between the complementary strand of the tag-cDNA conjugate and the 5' end of tag complement (362) attached to microparticle (360), and the complementary strand of the tag-cDNA conjugate is covalently bonded to the 5' end (363) of tag complement (362) by treating with a ligase. This embodiment requires, of course, that the 5' end of the tag complement be phosphorylated, *e.g.* by a kinase, such as, T4 polynucleotide kinase, or the like. The fill-in reaction is preferably carried out because the "stripping" reaction does not always halt at the first G. Preferably, the fill-in reaction uses a DNA polymerase lacking 5'→3' exonuclease activity and strand displacement activity, such as T4 DNA polymerase. Also preferably, all four dNTPs are used in the fill-in reaction, in case the "stripping" extended beyond the G triplet.

As explained further below, the tag-cDNA conjugates are hybridized to the full repertoire of tag complements. That is, among the population of microparticles, there are microparticles having every tag sequence of the entire repertoire. Thus, the tag-cDNA conjugates will hybridize to tag complements on only about one percent of the microparticles. Microparticles to which tag-cDNA have been hybridized are referred to herein as "loaded microparticles." For greater efficiency, loaded microparticles are preferably separated from unloaded microparticles for further processing. Such separation is conveniently accomplished by use of a fluorescence-activated cell sorter (FACS), or similar instrument that permits rapid manipulation and sorting of large numbers of individual microparticles. In the embodiment

illustrated in Figure 3b, a fluorescent label, *e.g.* FAM (a fluorescein derivative, Haugland, Handbook of Fluorescent Probes and Research Chemicals, Sixth Edition, (Molecular Probes, Eugene, OR, 1996)) is attached by way of primer (346).

The tag-cDNA can be attached to the tag complement on the microparticles by a procedure omitting or modifying many of the steps discussed above. For example, instead of amplifying the tag-cDNA from vector (330), the tag-cDNA can be cleaved from the vector by restriction digest, stripped, and ligated directly to the tag complement on the microparticles. This procedure omits (1) labeling the tag-cDNA with biotin and FAM, (2) amplifying the tag-cDNA, and (3) isolating the amplicon by streptavidin capture. If desired, loaded microparticles can be isolated by hybridizing with a FAM-labeled primer.

As shown in Figure 3c, after FACS, or like sorting (380), loaded microparticles (360) are isolated, treated to remove label (345), and treated to melt off the non-covalently attached strand. Label (345) is removed or inactivated so that it does not interfer with the labels of the competitively hybridized strands. Preferably, the tag-cDNA conjugates are treated with a restriction endonuclease recognizing site r_1 (342) which cleaves the tag-cDNA conjugates adjacent to primer binding site (344), thereby removing label (345) carried by the "bottom" strand, *i.e.* the strand have its 5' end distal to the microparticle. Preferably, this cleavage results in microparticle (360) with double stranded tag-cDNA conjugate (384) having protruding strand (385). 3'-labeled adaptor (386) is then annealed and ligated to protruding strand (385), after which the loaded microparticles are re-sorted by means of the 3'-label and the strand carrying the 3'-label is melted off to leave a covalently attached single strand of the cDNA (392) ready to accept denatured cDNAs or mRNAs from differentially expressed genes. Preferably, the 3'-labeled strand is melted off with sodium hydroxide treatment, or treatment with like reagent.

Clonal subpopulations of gDNAs can be attached to microparticles in a similar manner. First, genomic DNA is isolated from a cell or tissue source of interest using conventional techniques and is cleaved with at least one restriction endonuclease, which preferably cleaves at a four-base recognition, such as, for example, Dpn II, Sau3A I, Aci I, Alu I, Bfa I, BstU I, Hae III, Hha I, HinPl I, Hpa II, Mbo I, Mse I, Msp I, Nla III, Rsa I, Taq$^{\alpha}$ I, Tsp 509 I, and the like. Preferably, the cleaved fragment has an overhang of at least one base. Alternatively, genomic DNA fragments can be

prepared by shearing or sonicating the isolated genomic DNA. The tag can then be linked to the gDNA in a number of ways, including random primed PCR with primers containing the tag sequence or cloning into a vector containing a tag in a manner similar to that described above for a cDNA reference population. A label such as FAM can be attached in order to monitor the loading of the microparticles. In some instances, directional attachment onto the microparticles can be achieved by amplifying the gDNA with a primer having a consensus sequence, such as, for example, the TATA box, or a sequence complementary to a consensus sequence. When using a gDNA reference population for evaluating gene expression, it may be desirable to reduce noncoding sequence and introns in the gDNA library. For example, a large gDNA library of about 60 x 10^6 microparticles can be reduced to about 30,000-40,000 by culling, using cDNA pools as a probe.

Oligonucleotide Tags for Identification and Solid Phase Cloning

An important feature of the invention is the use of oligonucleotide tags which are members of a minimally cross-hybridizing set of oligonucleotides to construct reference DNA populations attached to solid phase supports, preferably microparticles. The sequences of oligonucleotides of a minimally cross-hybridizing set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. When oligonucleotide tags are used for sorting, as is the case for constructing a reference DNA population, tag complements are preferably attached to solid phase supports. Oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting, tracking, or labeling molecules, especially polynucleotides, such as cDNAs or mRNAs derived from expressed genes.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For example, a minimally cross-hybridizing set may consist of a set of individually

synthesized 10-mer sequences that differ from each other by at least 4 nucleotides, such set having a maximum size of 332, when constructed as disclosed in Brenner et al., U.S. patent 5,604,097. Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of 9^3, or 729, 12-mers.

When synthesized combinatorially, an oligonucleotide tag can be randomized at individual positions along its length. Preferably, however, the oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. An oligonucleotide tag can also consist of a plurality of subunits with additional nucleotides on either terminus of the oligonucleotide. The additional nucleotides can be random and/or can comprise a restriction site. Such a structure ensures the instability of a duplex or triplex having a mismatch at a terminus of the oligonucleotide. Preferably, the oligonucleotide comprises a recognition site for a rare-cutting restriction endonuclease on at least one end. In a preferred embodiment, the oligonucleotide comprises an AT-rich restriction site, such as a Pac I site, on one end. A Bsp120 site is a preferred site on the other end.

Complements of oligonucleotide tags attached to one or more solid phase supports are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Such tag complements are synthesized on the surface of a solid phase support, such as a bead, preferably microscopic, or a specific location on an array of synthesis locations on a single support, such that populations of identical, or substantially identical, sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by copies of only one type of tag complement having a particular sequence. The population of such beads or regions contains a repertoire of tag complements each with distinct sequences. As used herein in reference to

oligonucleotide tags and tag complements, the term "repertoire" means the total number of different oligonucleotide tags or tag complements that are employed for solid phase cloning (sorting) or identification. A repertoire may consist of a set of minimally cross-hybridizing set of oligonucleotides that are individually synthesized, or it may consist of a concatenation of oligonucleotides each selected from the same set of minimally cross-hybridizing oligonucleotides. In the latter case, the repertoire is preferably synthesized combinatorially.

Preferably, tag complements are synthesized combinatorially on microparticles, so that each microparticle has attached many copies of the same tag complement. A wide variety of microparticle supports may be used with the invention, including microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: Meth. Enzymol., Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046;720; and Pon, Chapter 19, in Agrawal, editor, Methods in Molecular Biology, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (*e.g.* available from PE Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (*e.g.*, TentaGelTM, Rapp Polymere, Tubingen Germany); and the like. Microparticles may also consist of dendrimeric structures, such as disclosed by Nilsen et al., U.S. patent 5,175,270. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, *e.g.* 1-2, to several hundred, *e.g.* 200-1000 μm diameter are preferable, as they facilitate the construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage. Preferably, glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN) are used as microparticles in the invention. Such microparticles are useful in a variety of sizes and are available with a variety of linkage groups for synthesizing tags and/or tag complements. More preferably, 5 μm diameter GMA beads are employed.

In a preferred embodiment, polynucleotides to be sorted, or cloned onto a solid phase support, each have an oligonucleotide tag attached, such that different polynucleotides have different tags. This condition is achieved by employing a

repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions. Of course, the sampled tag-polynucleotide conjugates are preferably amplified, *e.g.* by polymerase chain reaction, cloning in a plasmid, RNA transcription, or the like, to provide sufficient material for subsequent analysis.

Oligonucleotide tags are employed for two different purposes in certain embodiments of the invention: Oligonucleotide tags are employed to implement solid phase cloning, as described in Brenner, U.S. patent 5,604,097; and International patent application PCT/US96/09513, wherein large numbers of polynucleotides, *e.g.* several thousand to several hundred thousand, are sorted from a mixture into clonal subpopulations of identical polynucleotides on one or more solid phase supports for analysis, and they are employed to deliver (or accept) labels to identify polynucleotides, such as encoded adaptors, that number in the range of a few tens to a few thousand, *e.g.* as disclosed in Albrecht et al., International patent application PCT/US97/09472. For the former use, large numbers, or repertoires, of tags are typically required, and therefore synthesis of individual oligonucleotide tags is difficult. In these embodiments, combinatorial synthesis of the tags is preferred. On the other hand, where extremely large repertoires of tags are not required--such as for delivering labels to a plurality of kinds or subpopulations of polynucleotides in the range of 2 to a few tens, *e.g.* encoded adaptors, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized, as well as synthesized combinatorially.

Sets containing several hundred to several thousands, or even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, *e.g.* as disclosed in Frank et al., U.S. patent 4,689,405; Frank et al., Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al., Anal. Biochem., 224: 110-116 (1995); Fodor et al., International application PCT/US93/04145; Pease et al., Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern et al., J. Biotechnology,

This Sheet Has Been Replaced



35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al., Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, tag complements in mixtures, whether synthesized combinatorially or individually, are selected to have similar duplex or triplex stabilities to one another so that perfectly matched hybrids have similar or substantially identical melting temperatures. This permits mis-matched tag complements to be more readily distinguished from perfectly matched tag complements in the hybridization steps, *e.g.* by washing under stringent conditions. For combinatorially synthesized tag complements, minimally cross-hybridizing sets may be constructed from subunits that make approximately equivalent contributions to duplex stability as every other subunit in the set. Guidance for carrying out such selections is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, *e.g.* Rychlik et al., Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al., Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991); and the like. A minimally cross-hybridizing set of oligonucleotides can be screened by additional criteria, such as GC-content, distribution of mismatches, theoretical melting temperature, and the like, to form a subset which is also a minimally cross-hybridizing set.

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, *e.g.* an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, *e.g.* disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al., U.S. patent 4,980,460; Koster et al., U.S. patent 4,725,677; Caruthers et al., U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like.

Oligonucleotide tags for sorting may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers

This Sheet Has Been Replaced

of subunits, these ranges may be expressed as follows:

Numbers of Subunits in Tags in Preferred Embodiments

Monomers in Subunit	Nucleotides in Oligonucleotide Tag (12-60)	(18-40)	(25-40)
3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags for sorting are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

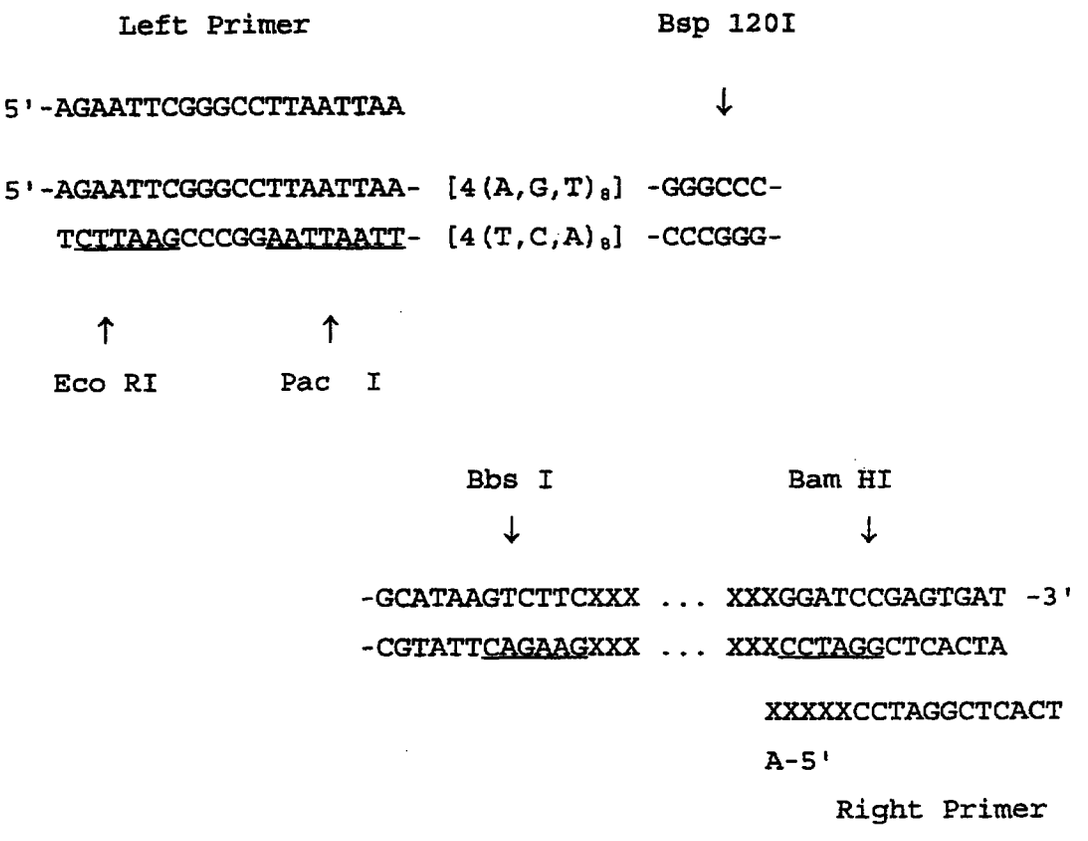
Preferably, repertoires of single stranded oligonucleotide tags for sorting contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Preferably, the length of single stranded tag complements for delivering labels is between 8 and 20. More preferably, the length is between 9 and 15.

In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (*e.g.* whether ribose or deoxyribose nucleosides are employed), base modifications (*e.g.* methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments. Conditions for annealing single-stranded or duplex tags to

their single-stranded or duplex complements are well known, *e.g.* Ji et al., Anal. Chem. 65: 1323-1328 (1993); Cantor et al., U.S. patent 5,482,836; and the like. Use of triplex tags in sorting has the advantage of not requiring a "stripping" reaction with polymerase to expose the tag for annealing to its complement.

An exemplary tag library for sorting is shown below (SEQ ID NO: 1).



Formula I

The flanking regions of the oligonucleotide tag may be engineered to contain restriction sites, as exemplified above, for convenient insertion into and excision from cloning vectors. Optionally, the right or left primers may be synthesized with a biotin attached (using conventional reagents, *e.g.* available from Clontech Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or cleavage. Preferably, for making tag-fragment conjugates, the above library is inserted into a conventional cloning vector, such a pUC19, or the like. Optionally, the vector containing the tag library may contain a "stuffer" region, "XXX ... XXX," which facilitates isolation of fragments fully digested with, for example, Bam HI and Bbs I.

An important aspect of the invention is the sorting and attachment of populations of DNA sequences, *e.g.* from a cDNA library, to microparticles or to

separate regions on a solid phase support such that each microparticle or region has substantially only one kind of sequence attached; that is, such that the DNA sequences are present in clonal subpopulations. This objective is accomplished by insuring that substantially all different DNA sequences have different tags attached. This condition, in turn, is brought about by taking only a sample of the full ensemble of tag-DNA sequence conjugates for analysis. (It is acceptable that identical DNA sequences have different tags, as it merely results in the same DNA sequence being operated on or analyzed twice.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the DNA sequences; it can be carried out inherently as a secondary effect of the techniques used to process the DNA sequences and tags; or sampling can be carried out both overtly and as an inherent part of processing steps.

If a sample of n tag-DNA sequence conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r)=e^{-\lambda}(\lambda)^r/_r$, where r is the number of conjugates having the same tag and $\lambda=np$, where p is the probability of a given tag being selected. If $n=10^6$ and $p=1/(1.67 \times 10^7)$ (for example, if eight 4-base words described in Brenner et al. were employed as tags), then $\lambda=.0149$ and $P(2)=1.13 \times 10^{-4}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained by serial dilutions of a mixture containing tag-fragment conjugates.

As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. Preferably, at least ninety-five percent of the DNA sequences have unique tags attached.

Preferably, DNA sequences are conjugated to oligonucleotide tags by inserting the sequences into a conventional cloning vector carrying a tag library. For example, cDNAs may be constructed having a Bsp 120 I site at their 5' ends and after digestion with Bsp 120 I and another enzyme such as Sau 3A or Dpn II may be directionally inserted into a pUC19 carrying the tags of Formula I to form a tag-cDNA library, which includes every possible tag-cDNA pairing. A sample is taken from this library

for amplification and sorting. Sampling may be accomplished by serial dilutions of the library, or by simply picking plasmid-containing bacterial hosts from colonies. After amplification, the tag-cDNA conjugates may be excised from the plasmid.

After the oligonucleotide tags are prepared for specific hybridization, *e.g.* by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

Specificity of the hybridizations of tag to their complements may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words a repertoire of 8^8, or about 1.67×10^7, tags and tag complements are produced. In a library of tag-DNA sequence conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of 1.67×10^7 microparticles, then only a sparse subset of the sampled microparticles would be loaded. Preferably, loaded microparticles may be separated from unloaded microparticles by a fluorescence activated cell sorting (FACS) instrument using conventional protocols after DNA sequences have been fluorescently labeled and denatured. After loading and FACS sorting, the label may be cleaved prior use or other analysis of the attached DNA sequences.

A reference DNA population may consist of any set of DNA sequences whose frequencies in different test populations is sought to be compared. Preferably, a reference DNA population for use in the analysis of gene expression in a plurality of cells or tissues is constructed by generating a cDNA library from each of the cells or tissues whose gene expression is being compared. This may be accomplished either by pooling the mRNA extracted from the various cells and/or tissues, or it may be accomplished by pooling the cDNAs of separately constructed cDNA libraries. Alternatively, a reference DNA population may be constructed from genomic DNA. The objective is to obtain a set of DNA sequences that will include all of the sequences that could possibly be expressed in any of the cells or tissues being analyzed. Once the DNA sequences making up a reference DNA population are obtained, they must be conjugated with oligonucleotide tags for solid phase cloning. Preferably, the DNA sequences are prepared so that they can be inserted into a vector carrying an appropriate tag repertoire, as described above, to form a library of tag-DNA sequence conjugates. A sample of conjugates is taken from this library, amplified, and loaded onto microparticles. It is important that the sample be large enough so that there is a high probability that all of the different types of DNA sequences are represented on the loaded microparticles. For example, if among a plurality of cells being compared a total of about 25,000 genes are expressed, then a sample of about five-fold this number, or about 125,000 tag-DNA sequence conjugates, should be taken to ensure that all possible DNA sequences will be represented among the loaded microparticles with about a 99% probability, *e.g.* Sambrook et al. (cited above).

In another embodiment, the reference population can comprise a set of polynucleotides encoding a specific set or sets of proteins selected from the group consisting of cell cycle proteins, signal transduction pathway proteins, oncogene gene products, tumor suppressors, kinases, phosphatases, transcription factors, growth factor receptors, growth factors, extracellular matrix proteins, proteases, cytoskeletal proteins, membrane receptors, Rb pathway proteins, p53 pathway proteins, proteins involved in metabolism, proteins involved in cellular responses to stress, cytokines, proteins involved in DNA damage and repair, and proteins involved in apoptosis. Such polynucleotides are typically attached to the solid phase supports through oligonucleotides having a unique sequence per solid support, but such polynucleotides

can also be attached to the solid phase supports through an oligonucleotide with a sequence common for each solid phase support, such as, for example a polyadenylated oligonucleotide.

Preferably, after the tag-DNA sequence conjugates are sampled, they are amplified by PCR using a fluorescently labeled primer to provide sufficient material to load onto the tag complements of the microparticles and to provide a means for distinguishing loaded from unloaded microparticles, as disclosed in Brenner et al. (cited above). Preferably, the PCR primer also contains a sequence which allows the generation of a restriction site of a rare-cutting restriction endonuclease, such as Pac I, in the double stranded product so that the fluorescent label may be cleave from the end of the cDNA prior to the competitive hybridization of labeled DNA strands derived from cells or tissue being studied. After such loading, the specifically hybridized tag-DNA sequence conjugates are ligated to the tag complements and the loaded microparticles are separated from the unloaded microparticles by FACS. The fluorescent label is cleaved from the DNA strands of the loaded microparticles and the non-covalently attached strand is removed by denaturing with heat, formamide, NaOH, and/or with like means, using conventional protocols. The microparticles are then ready for competitive hybridization.

Competitive Hybridization and Light-Generating Labels

Gene expression products, *e.g.* mRNA or cDNA, from the cells and/or tissues being analyzed are isolated. The expression products are labeled so as to distinguish the source. Preferably, the products from each source comprise a label different from the label comprised by the products of any other source, *e.g.*, each having a unique and distinguishable emission frequency. Alternatively, the product of one source can be left unlabeled. The expression products can be labeled by conventional techniques, *e.g.* DeRisi et al. (cited above), or the like. Preferably, a light-generating label is incorporated into cDNAs reverse transcribed from the extracted mRNA, or an oligonucleotide tag is attached for providing a labeled tag complement for identification. A large number of light-generating labels are available, including fluorescent, colorimetric, chemiluminescent, and electroluminescent labels. Generally, such labels produce an optical signal which may comprise an absorption frequency, an emission frequency, an intensity, a signal lifetime, or a combination of

such characteristics. Preferably, fluorescent labels are employed, either by direct incorporation of fluorescently labeled nucleoside triphosphates or by indirect application by incorporation of a capture moiety, such as biotinylated nucleoside triphosphates or an oligonucleotide tag, followed by complexing with a moiety capable of generating a fluorescent signal, such as a streptavidin-fluorescent dye conjugate or a fluorescently labeled tag complement. Preferably, the optical signal detected from a fluorescent label is an intensity at one or more characteristic emission frequencies. Selection of fluorescent dyes and means for attaching or incorporating them into DNA strands is well known, *e.g.* DeRisi et al. (cited above), Matthews et al., Anal. Biochem., Vol 169, pgs. 1-25 (1988); Haugland, Handbook of Fluorescent Probes and Research Chemicals (Molecular Probes, Inc., Eugene, 1992); Keller and Manak, DNA Probes, 2nd Edition (Stockton Press, New York, 1993); and Eckstein, editor, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991); Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Ju et al., Proc. Natl. Acad. Sci., 92: 4347-4351 (1995) and Ju et al., Nature Medicine, 2: 246-249 (1996); and the like.

Preferably, light-generating labels are selected so that their respective optical signals can be related to the quantity of labeled DNA strands present and so that the optical signals generated by different light-generating labels can be compared. Measurement of the emission intensities of fluorescent labels is the preferred means of meeting this design objective. For a given selection of fluorescent dyes, relating their emission intensities to the respective quantities of labeled DNA strands requires consideration of several factors, including fluorescent emission maxima of the different dyes, quantum yields, emission bandwidths, absorption maxima, absorption bandwidths, nature of excitation light source(s), and the like. Guidance for making fluorescent intensity measurements and for relating them to quantities of analytes is available in the literature relating to chemical and molecular analysis, *e.g.* Guilbault, editor, Practical Fluorescence, Second Edition (Marcel Dekker, New York, 1990); Pesce et al., editors, Fluorescence Spectroscopy (Marcel Dekker, New York, 1971); White et al., Fluorescence Analysis: A Practical Approach (Marcel Dekker, New York, 1970); and the like. As used herein, the term "relative optical signal" means a ratio of signals from different light-generating labels that can be related to a ratio of differently labeled DNA strands of identical, or substantially identical, sequences that

form duplexes with a complementary reference DNA strand. Preferably, a relative optical signal is a ratio of fluorescence intensities of two or more different fluorescent dyes.

Competitive hybridization between the labeled DNA strands derived from the plurality of cells or tissues is carried out by applying equal quantities of the labeled DNA strands from each such source to the microparticles loaded with the reference DNA population in a conventional hybridization reaction. The particular amounts of labeled DNA strands added to the competitive hybridization reaction vary widely depending on the embodiment of the invention. Factors influencing the selection of such amounts include the quantity of microparticles used, the type of microparticles used, the loading of reference DNA strands on the microparticles, the complexity of the populations of labeled DNA strands, and the like. Hybridization is competitive in that differently labeled DNA strands with identical, or substantially identical, sequences compete to hybridize to the same complementary reference DNA strands. The competitive hybridization conditions are selected so that the proportion of labeled DNA strands forming duplexes with complementary reference DNA strands reflects, and preferably is directly proportional to, the amount of that DNA strand in its population in comparison with the amount of the competing DNA strands of identical sequence in their respective populations. Thus, if a first and second differently labeled DNA strands with identical sequence are competing for hybridization with a complementary reference DNA strand such that the first labeled DNA strand is at a concentration of 1 ng/μl and the second labeled DNA strand is at a concentration of 2 ng/μl, then at equilibrium it is expected that one third of the duplexes formed with the reference DNA would include first labeled DNA strands and two thirds of the duplexes would include second labeled DNA strands. Guidance for selecting hybridization conditions is provided in many references, including Keller and Manak, (cited above); Wetmur, (cited above); Hames et al., editors, Nucleic Acid Hybridization: A Practical Approach (IRL Press, Oxford, 1985); and the like.

Another aspect of the invention is a kit for analyzing differentially expressed genes, comprising a mixture of microparticles, each microparticle having a population of identical single stranded nucleic acid molecules attached thereto, the single stranded nucleic acid molecules being different on each microparticle and comprising a polynucleotide derived from an mRNA of at least one cell or tissue source.

Preferably, each of said nucleic acid molecules further comprises an oligonucleotide tag in juxtaposition with said polynucleotide and positioned between said microparticle and said polynucleotide. The kit can further comprise a population of cDNA molecules from at least one of said cell or tissue sources, reagents for labeling the cDNA populations, reagents for performing competitive hybridization, and the like. If desired, the cDNA molecules in the kit are provided in fluorescently labeled form. The kit can contain additional components for performing competitive hybridization, such as, for example, hybridization buffers, PCR buffers and standards, and the like. The kit can further comprise at least one container or several containers for each of the components and can comprise printed instructions for use in analyzing differentially expressed genes.

The invention also provides a kit for preparing a reference population, comprising a plurality of microparticles having oligonucleotide tag complements attached thereto, the oligonucleotide tag complement sequence being different on each microparticle. The kit can further comprise a plurality of vectors comprising a library of tags having sequences complementary to the tag complements. The kit can further comprise a population of polynucleotides from at least one cell or tissue source, preferably cDNAs. When a population of polynucleotides is included, preferably the population of polynucleotides is contained in a container separate from said plurality of microparticles. The kit can also contain reagents for preparing the reference population, such as, for example, adaptors, labels, polymerase, dNTP's, labelled dNTP's, PCR buffers, and the like, as well as printed instructions for preparing the reference population.

Flow Sorting of Microparticles with Up-Regulated and/or Down-Regulated Gene Products

After labeled polynucleotides are competitively hybridized to a reference population on microparticles, the microparticles may be analyzed and/or sorted in a number of ways depending on the chemical and/or physical properties of the microparticles and the attached sequences. For example, microparticles of interest may be mechanically separated by micro-manipulators, magnetic microparticles may be sorted by adjusting or manipulating magnetic fields, charged microparticles may be manipulated by electrophoresis, or the like. The following references provide

guidance for selecting means for analyzing and/or sorting microparticles: Pace, U.S. Patent 4,908,112; Saur et al., U.S. Patent 4,710,472; Senyei et al., U.S. Patent 4,230,685; Wilding et al., U.S. Patent 5,637,469; Penniman et al., U.S. Patent 4,661,225; Karnaukhov et al., U.S. Patent 4,354,114; Abbott et al., U.S. Patent 5,104,791; Gavin et al., PCT publication WO 97/40383; and the like. Preferably, microparticles containing fluorescently labeled DNA strands are conveniently classified and sorted by a commercially available FACS instrument, *e.g.* Van Dilla et al., Flow Cytometry: Instrumentation and Data Analysis (Academic Press, New York, 1985); Fulwyler et al., U.S. Patent 3,710,933; Gray et al., U.S. Patent 4,361,400; Dolbeare et al., U.S. Patent 4,812,394; and the like. For fluorescently labeled DNA strands competitively hybridized to a reference strand, preferably the FACS instrument has multiple fluorescent channel capabilities. Preferably, upon excitation with one or more high intensity light sources, such as a laser, a mercury arc lamp, or the like, each microparticle will generate fluorescent signals, usually fluorescence intensities, related to the quantity of labeled DNA strands from each cell or tissue types carried by the microparticle. As shown in Figure 1a of Example 1, when fluorescent intensities of each microparticle are plotted on a two-dimensional graph, microparticles indicating equal expression levels will be on or near the diagonal (100) of the graph. Up-regulated and down-regulated genes will appear in the off-diagonal regions (112). Such microparticles are readily sorted by commercial FACS instruments by graphically defining sorting parameters to enclose one or both off-diagonal regions (112) as shown in Figure 1b. Thus, microparticles can be sorted according to their relative optical signal, and if desired, collected for further analysis by accumulating those microparticles generating a signal within a predetermined range of values corresponding to a difference in gene expression among the different cell or tissue sources.

Flow Sorting of Microparticles According to the Abundance of Nucleic Acid Sequences from which the Polynucleotides are Derived

Microparticles containing fluorescently labeled DNA strands can also be classified and sorted according to the abundance of the gene products from which they are derived. The abundance of a nucleic acid sequence can be determined by the methods described above for determining relative gene expression and can be

correlated with the level of intensity of the optical signal generated by the polynucleotides bound to the microparticles. A lower intensity is indicative of a rarer nucleic acid sequence, such as a rare gene product. Rare genes are genes encoding an mRNA which is present in about 100 copies per cell or less, with increasing preference for less than about 50 copies to less than about 25 copies, with less than about 10 copies per cell being most preferred. Rare genes can be isolated by collecting microparticles with low fluorescent intensities as shown in Examples 9 and 10. The collected microparticles typically comprise less than about 5% of the total microparticles, with increasing preference for less than about 2.5%, 1%, to 0.5% with less than about 0.1% being most preferred.

Alternatively, since hybridization rates are proportionate to the abundance of a nucleic acid sequence, less abundant nucleic acid sequences can be isolated by setting the hybridization conditions such that nucleic acid sequences present in a lower abundance in a cell or tissue source remain unhybridized. Suitable hybridization conditions include those conditions used for producing normalized cDNA libraries (Patanjali et al., *Proc. Natl. Acad. Sci. USA*, 88:1943-1947 (1991)). For example, rare genes can be isolated by collecting unhybridized DNA after allowing a maximum period of time for hybridization of the abundant DNA species.

Repetitive sequences can often complicate the mapping and analysis of polymorphisms. Repetitive sequences exist due to the presence in the genome of transposons, retrotransposons, retroviruses, short interspersed repetitive elements (SINEs) such as Alu sequences, satellite DNA, minisatellite DNA, megasatellite DNA, and the like. Repetitive sequences can be removed from a DNA population as described above by sorting rapidly hybridizing DNA species away from DNA species that are slower to hybridize. Preferably, the unhybridized population is substantially enriched in polynucleotides derived from non-repetitive nucleic acid sequences.

Another aspect of the invention is a kit for analyzing and/or isolating nucleic acid sequences with respect to their abundance comprising microparticles prepared as described above and printed instructions for use.

Identification of Sorted Genes by Massively Parallel Signature Sequencing (MPSS)

Expressed genes may be identified in parallel by MPSS, which is a combination of two techniques: one for tagging and sorting fragments of DNA for parallel processing (*e.g.* Brenner et al., International application PCT/US96/09513), and another for the stepwise sequencing the end of a DNA fragment (*e.g.* Brenner, U.S. patent 5,599,675 and Albrecht et al., International patent application PCT/US97/09472). After an initial digestion of a target polynucleotide with a first restriction endonuclease, restriction fragments are ligated to oligonucleotide tags as described below, and in Brenner et al., International application PCT/US96/09513, so that the resulting tag-fragment conjugates may be sampled, amplified, and sorted onto separate solid phase supports by specific hybridization of the oligonucleotide tags with their tag complements.

Once an amplified sample of DNA fragments is sorted onto solid phase supports to form homogeneous populations of substantially identical fragments, the ends of the fragments are preferably sequenced with an adaptor-based method of DNA sequencing that includes repeated cycles of ligation, identification, and cleavage, such as the method described in Brenner, U.S. patent 5,599,675. In further preference, adaptors used in the sequencing method each have a protruding strand and an oligonucleotide tag selected from a minimally cross-hybridizing set of oligonucleotides, as taught by Albrecht et al., International patent application PCT/US97/09472. Such adaptors are referred to herein as "encoded adaptors." Encoded adaptors whose protruding strands form perfectly matched duplexes with the complementary protruding strands of a fragment are ligated. After ligation, the identity and ordering of the nucleotides in the protruding strand is determined, or "decoded," by specifically hybridizing a labeled tag complement, or "de-coder" to its corresponding tag on the ligated adaptor.

The preferred sequencing method is carried out with the following steps: (a) ligating an encoded adaptor to an end of a fragment, the encoded adaptor having a nuclease recognition site of a nuclease whose cleavage site is separate from its recognition site; (b) identifying one or more nucleotides at the end of the fragment by the identity of the encoded adaptor ligated thereto; (c) cleaving the fragment with a

This Sheet Has Been Replaced



nuclease recognizing the nuclease recognition site of the encoded adaptor such that the fragment is shortened by one or more nucleotides; and (d) repeating said steps (a) through (c) until said nucleotide sequence of the end of the fragment is determined. In the identification step, successive sets of tag complements, or "de-coders," are specifically hybridized to the respective tags carried by encoded adaptors ligated to the ends of the fragments. The type and sequence of nucleotides in the protruding strands of the polynucleotides are identified by the label carried by the specifically hybridized de-coder and the set from which the de-coder came, as described below.

Identification of Sorted Genes by Conventional Sequencing

Gene products carried by microparticles may be identified after sorting, *e.g.* by FACS, using conventional DNA sequencing protocols. Suitable templates for such sequencing may be generated in several different ways starting from the sorted microparticles carrying differentially expressed gene products. For example, the reference DNA attached to an isolated microparticle may be used to generate labeled extension products by cycle sequencing, *e.g.* as taught by Brenner, International application PCT/US95/12678. In this embodiment, primer binding site (400) is engineered into the reference DNA (402) distal to tag complement (406), as shown in Figure 4a. After isolating a microparticle, *e.g.* by sorting into separate microtiter well, or the like, the differentially expressed strands are melted off, primer (404) is added, and a conventional Sanger sequencing reaction is carried out so that labeled extension products are formed. These products are then separated by electrophoresis, or like techniques, for sequence determination. In a similar embodiment, sequencing templates may be produced without sorting individual microparticles. Primer binding sites (400) and (420) may be used to generate templates by PCR using primers (404) and (422). The resulting amplicons containing the templates are then cloned into a conventional sequencing vector, such as M13. After transfection, hosts are plated and individual clones are selected for sequencing.

In another embodiment, illustrated in Figure 4b, primer binding site (412) may be engineered into the competitively hybridized strands (410). This site need not have a complementary strand in the reference DNA (402). After sorting, competitively hybridized strands (410) are melted off of reference DNA (402) and amplified, *e.g.* by PCR, using primers (414) and (416), which may be labeled and/or derivatized with

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

biotin for easier manipulation. The melted and amplified strands are then cloned into a conventional sequencing vector, such as M13, which is used to transfect a host which, in turn, is plated. Individual colonies are picked for sequencing.

Example 1

Construction of a Tagged cDNA Library, Sampling, and Loading Tagged cDNAs onto Microparticles

In this example, a preferred protocol for preparing tagged reference DNA for loading onto microparticles is described. Briefly, cDNA from each of the cell or tissue types of interest is prepared and directionally cloned into a vector containing the tag element of Formula I. Preferably, the mRNA extracted from such cells or tissues is combined, usually in equal proportions, prior to first strand synthesis. mRNA is obtained using standard protocols, after which first and second strand synthesis is carried out as exemplified and the resulting cDNAs are inserted into a vector containing a tag element of Formula I, or like tag element. The vectors containing the tag-cDNA conjugates are then used to transform a suitable host, typically a conventional bacterial host, after which a sample of cells from the host culture is further expanded and vector DNA is extracted. The tag-cDNA conjugates are preferably amplified from the vectors by PCR and processed as described below for loading onto microparticles derivatized with tag complements. After the non-covalently attached strand is melted off, the cDNA-containing microparticles are ready to accept competitively hybridized gene products in accordance with the invention. Specific guidance relating to the indicated steps is available in Sambrook et al. (cited above); Ausbel et al., editors, Current Protocols in Molecular Biology (John Wiley & Sons, New York, 1995); and like guides on molecular biology techniques.

A pellet of approximately 5 μg of mRNA is resuspended in 45 μl (final volume) of a first strand pre-mix consisting of 10 μl 5x SuperScript buffer (250 mM Tris-Cl, pH 8.3, 375 mM KCl, and 15 mM $MgCl_2$) (GIBCO/BRL) (or like reverse transcriptase buffer), 5 μl 0.1 M dithiothreitol (DTT), 2.5 μl 3dNTP/methyl-dCTP mix (10 μM each of dATP, dGTP, dTTP, and 5-methyl-dCTP, e.g available from Pharmacia Biotech), 1 μl RNasin, 12 μl 0.25 μg/μl of reverse transcription primer shown below, and 14.5 μl H2O.

This Sheet Has Been Replaced



5'-biotin-GACATGCTGCATTGAGACGATTCTTTTTTTTTTTTTTTTTV

Reverse Transcription Primer (SEQ ID NO: 2)

After incubation for 15 min at room temperature, 5 ml of 200 U/μl SuperScript is added and the mixture is incubated for 1 hr at 42°C. After the 1 hr incubation, the above mixture (about 50 μl total) is added to a second-strand premix on ice (volume 336 μl) consisting of 80 μl 5x second-strand buffer (94 mM Tris-Cl, pH 6.9, 453 mM KCl, 23 mM $MgCl_2$, and 50 mM $(NH_4)2SO_4$ to give a total reaction volume of about 386 μl. Separately, 4 μl of 0.8 U/μl RNase H (3.2 units) and 10 μl of 10 unit/μl E. coli DNA polymerase I (100 units) are combined and the combined enzyme mixture is added to the above second-strand reaction mixture, after which the total reaction volume is microfuged 5 sec and then incubated for 1 hr at 16°C and for 1 hr at room temperature to give the following double stranded cDNA (SEQ ID NO: 3):

```
5'-biotin-GACATGCTGCATTGAGACGATTCTTTTTTTTTTTTTTTTTVXXX ... XGATCXXX-3'
          CTGTACGACGTAACTCTGCTAAGAAAAAAAAAAAAAAAAAABXXX ... XCTAGXXX-5'
                              ↑                            ↑
                            Bsm BI                       Dpn II
```

where the X's indicated nucleotides in the cDNAs, V represents A, C, or G, and B represents C, G, or T. Note that the reverse transcription primer sequence has been selected to give a Bsm BI site in the cDNAs which results in a 5'-GCAT overhang upon digestion with Bsm BI.

After phenol/chloroform extraction and ethanol precipitation, the cDNA is resuspended in the manufacturer's recommended buffer for digestion with Dpn II (New England Biolabs, Beverely, MA), which is followed by capture of the biotinylated fragment on avidinated beads (Dynal, Oslo, Norway). After washing, the captured fragments are digested with Bsm BI to release the following cDNAs (SEQ ID NO: 4) which are precipitated in ethanol:

```
GCATTGAGACGATTCTTTTTTTTTTTTTTTTTTVXXX ... X      -3'
```


This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

ACTCTGCTAAGAAAAAAAAAAAAAAAAABXXX ... XCTAG -5'

A conventional cloning vector, such as BlueScript II, pBC, or the like (Stratagene Cloning Systems, La Jolla, CA), is engineered to have the following sequence of elements (SEQ ID NO: 5)(which are those shown in Formula I):

```
5'-...TTAATTAAGGA [TAG] GGGCCCGCATAAGTCTTC [STUFFER]
GGATCC...-3'

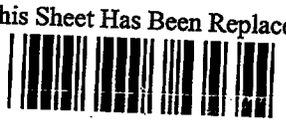
3'-...AATTAATTCCT [TAG] CCCGGGCGTATTCAGAAG [STUFFER] CCTAGG...-5'
      ↑                                ↑              ↑
    Pac I                            Bbs I          Bam HI
```

After digestion with Bbs I and Bam HI, the vector is purified by gel electrophoresis and combined with the cDNAs for ligation. Note that the vector has been engineered so that the Bbs I digestion results in an end compatible with the Bsm BI-digested end of the cDNAs. After ligation, a suitable host bacteria is transformed and a culture is expanded for subsequent use.

From the expanded culture, a sample of host cells are plated to determine the fraction that carry vectors with inserted cDNAs, after which an aliquot of culture corresponding to about 1.7×10^5 insert-containing cells is withdrawn and separately expanded in culture. This represents about one percent of the repertoire of tags of the type illustrated in Formula I.

Preferably, the tag-cDNA conjugates are amplified out of the vectors by PCR using a conventional protocol, such as the following. For each of 8 replicate PCRs, the following reaction components are combined: 1 µl vector DNA (125 ng/µl for a library, 10^9 copies for a single clone); 10 µl 10x Klentaq Buffer (Clontech Laboratories, Palo Alto, CA); 0.25 µl biotinylated 20-mer "forward" PCR primer (1 nmol/µl); 0.25 µl FAM-labeled 20-mer "reverse" PCR primer (1 nmol/µl); 1 µl 25 mM dATP, dGTP, dTTP, and 5-methyl-dCTP (total dNTP concentration 100 mM); 5 µl DMSO; 2 µl 50x Klentaq enzyme; and 80.5 µl H_2O (for a total volume of 100 µl).

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

The PCR is run in an MJR DNA Engine (MJ Research), or like thermal cycler, with the following protocol: 1) 94°C for 4 min; 2) 94°C 30 sec; 3) 67°C 3 min; 4) 8 cycles of steps 2 and 3; 5) 94°C 30 sec, 6) 64°C 3 min, 7) 22 cycles of steps 5 and 6; 8) 67°C for 3 min; and 9) hold at 4°C.

The 8 PCR mixtures are pooled and 700 µl phenol is added at room temperature, after which the combined mixture is vortexed for 20-30 sec and then centrifuged at high speed (*e.g.* 14,000 rpm in an Eppendorf bench top centrifuge, or like instrument) for 3 min. The supernatant is removed and combined with 700 µl chloroform (24:1 mixture of chloroform:iso-amyl alcohol) in a new tube, vortexed for 20-30 sec, and centrifuged for 1 min, after which the supernatant is transferred to a new tube and combined with 80 µl 3M sodium acetate and 580 µl isopropanol. After centrifuging for 20 min, the supernatant is removed and 1 ml 70% ethanol is added. The mixture is centrifuged for 5-10 min, after which the ethanol is removed and the precipitated DNA is dried in a speedvac.

After resuspension, the cDNA is purified on avidinated magnetic beads (Dynal) using the manufacturer's recommended protocol and digested with Pac I (1 unit of enzyme per µg of DNA), also using the manufacturer's recommended protocol (New England Biolabs, Beverly, MA). The cleaved DNA is extracted with phenol/chloroform followed by ethanol precipitation. The tags of the tag-cDNA conjugates are rendered single stranded by combining 2 units of T4 DNA polymerase (New England Biolabs) per µg of streptavidin-purified DNA. 150 µg of streptavidin-purified DNA is resuspended in 200 µl H_2O and combined with the following reaction components: 30 µl 10 NEB Buffer No. 2 (New England Biolabs); 9 µl 100 mM dGTP; 30 µl T4 DNA polymerase (10 units/µl); and 31 µl H_2O; to give a final reaction volume of 300 µl. After incubation for 1 hr at 37°C, the reaction is stopped by adding 20 µl 0.5 M EDTA, and the T4 DNA polymerase is inactivated by incubating the reaction mixture for 20 min at 75°C. The tag-cDNA conjugates are purified by phenol/chloroform extraction and ethanol precipitation.

5 µm GMA beads with tag complements are prepared by combinatorial synthesis on an automated DNA synthesizer (Gene Assembler Special /4 Primers, Pharmacia Biotech, Bjorkgatan, Sweden, or like instrument) using conventional phosphoramidite chemistry, wherein nucleotides are condensed in the 3'→5' direction. In a preferred embodiment, a 28-nucleotide "spacer" sequence is synthesized,

followed by the tag complement sequence (8 "words" of 4 nucleotides each for a total of 32 nucleotides in the tag complement), and a sequence of three C's. Thus, the beads are derivatized with a 63-mer oligonucleotide. The length of the "spacer" sequence is not critical; however, the proximity of the bead surface may affect the activity of enzymes that are use to treat tag complements or captured sequences. Therefore, if such processing is employed, a spacer long enough to avoid such surface effects is desirable. Preferably, the spacer is between 10 and 30 nucleotides, inclusive. The following sequence (SEQ ID NO: 6), containing a Pac I site, is employed in the present embodiment:

5'-CCC-[Tag Complement]-TCCTTAATTAACTGGTCTCACTGTCGCA-bead

↑

Pac I

Preferably, the tag-cDNA conjugates are hybridized to tag compliments on beads of a number corresponding to at least a full repertoire of tag complements, which in the case of the present embodiment is 8^8, or about 1.6×10^7 beads. The number of beads in a given volume is readily estimated with a hemocytometer.

Prior to hybridization of the tag-cDNA conjugates, the 5' ends of the tag complements are phosphorylated, preferably by treatment with a polynucleotide kinase. Briefly, 2.5×10^8 beads suspended in 100 µl H_2O are combined with 100 µl 10x NEB buffer No. 2 (New England Biolabs, Beverly, MA), 10 µl 100 mM ATP, 1 µl 10% Tween 20, 17 µl T4 polynucleotide kinase (10 units/µl), and 772 µl H_2O for a final volume of 1000 µl. After incubating for 2 hr at 37°C with vortexing, the temperature is increased to 65°C for 20 min to inactivate the kinase, with continued vortexing. After incubation, the beads are washed twice by spinning down the beads and resuspending them in 1 ml TE (Sambrook et al., Molecular Cloning, Second Edition, Cold Spring Harbor Laboratory) containing .01% Tween 20.

For hybridization of tag-cDNA conjugates to tag complements, the tag-cDNA conjugates as prepared above are suspended in 50 µl H_2O and the resulting mixture is combined with 40 µl 2.5x hybridization buffer, after which the combined mixture is filtered through a Spin-X spin column (0.22 µm) using a conventional protocol to give a filtrate containing the tag-cDNA conjugates. (5 ml of the 2.5x hybridization

This Sheet Has Been Replaced



buffer consists of 1.25 ml 0.1 M $NaPO_4$ (pH 7.2), 1.25 ml 5 M NaCl, 0.25 ml 0.5% Tween 20, 1.50 ml 25% dextran sulfate, and 0.75 ml H_2O.) Approximately 1.8 x 10^7 beads in 10 µl TE/Tween buffer (TE with .01% Tween 20) is centrifuged so that the beads form a pellet and the TE/Tween is removed. To the beads, 25 µl of 1x hybridization buffer (10 mM NaPO4 (pH 7.2), 500 mM NaCl, 0.01% Tween 20, 3% dextran sulfate) is added and the mixture is vortexed to fully resuspend the beads, after which the mixture is centrifuged so that the beads form a pellet and the supernatant is removed.

The tag-cDNA conjugates in the above filtrate are incubated at 75°C for 3 min and combined with the beads, after which the mixture is vortexed to fully resuspend the beads. The resulting mixture is further incubated at 75°C with vortexing for approximately three days (60 hours). After hybridization, the mixture is centrifuged for 2 min and the supernatant is removed, after which the beads are washed twice with 500 µl TE/Tween and resuspended in 500 µl 1x NEB buffer No. 2 with .01% Tween 20. The beads are incubated at 64°C in this solution for 30 min, after which the mixture is centrifuged so that the beads form a pellet, the supernatant is removed, and the beads are resuspended in 500 µl TE/Tween.

Loaded beads are sorted from unloaded beads using a high speed cell sorter, preferably a MoFlo flow cytometer equipped with an argon ion laser operating at 488 nm (Cytomation, Inc., Ft. Collins, CO), or like instrument. After sorting, the loaded beads are subjected to a fill-in reaction by combining them with the following reaction components: 10 µl 10x NEB buffer No. 2, 0.4 µl 25 mM dNTPs, 1 µl 1% Tween 20, 2 µl T4 DNA polymerase (10 units/ml), and 86.6 µl H_2O, for a final reaction volume of 100 µl. After incubation at 12°C for 30 min with vortexing, the reaction mixture is centrifuged so that the beads form a pellet and the supernatant is removed. The pelleted beads are resuspended in a ligation buffer consisting of 15 µl 10x NEB buffer No. 2, 1.5 µl 1% Tween 20, 1.5 µl 100 mM ATP, 1 µl T4 DNA ligase (400 units/ ml), and 131 µl H_2O, to give a final volume of 150 µl. The ligation reaction mixture is incubated at 37°C for 1 hr with vortexing, after which the beads are pelleted and washed once with 1x phosphate buffered saline (PBS) with 1 mM $CaCl_2$. The beads are resuspended in 45 µl PBS (with 1 mM $CaCl_2$) and combined with 6 µl Pronase solution (10 mg/ml, Boehringer Mannheim, Indianapolis, IN), after which the mixture is incubated at 37°C for 1 hr with vortexing. After centrifugation,

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

the loaded beads are washed twice with TE/Tween and then once with 1x NEB Dpn II buffer (New England Biolabs, Beverly, MA).

The tag-cDNA conjugates loaded onto beads are cleaved with Dpn II to produce a four-nucleotide protruding strand to which a complementary adaptor carrying a 3'-label is ligated. Accordingly, the loaded beads are added to a reaction mixture consisting of the following components: 10 µl 10x NEB Dpn II buffer, 1 µl 1% Tween, 4 µl Dpn II (50 units/ml), and 85 µl H_2O, to give a final reaction volume of 100 µl. The mixture is incubated at 37°C overnight with vortexing, after which the beads are pelleted, the supernatant is removed, and the beads are washed once with 1x NEB buffer No. 3. To prevent self-ligation, the protruding strands of the tag-cDNA conjugates are treated with a phosphatase, *e.g.* calf intestine phosphatase (CIP), to remove the 5' phosphates. Accordingly, the loaded beads are added to a reaction mixture consisting of the following components: 10 µl 10x NEB buffer No. 3, 1 µl 1% Tween 20, 5 µl CIP (10 units/µl), and 84 µl H_2O, to give a final reaction volume of 100 µl. The resulting mixture is incubated at 37°C for 1 hr with vortexing, after which the beads are pelleted, washed once in PBS containing 1 mM $CaCl_2$, treated with Pronase as described above, washed twice with TE/Tween, and once with 1x NEB buffer No. 2.

The following 3'-labeled adaptor (SEQ ID NO: 7) is prepared using conventional reagents, *e.g.* Clontech Laboratories (Palo Alto, CA):

```
5'-pGATCACGAGCTGCCAGTC-FAM
   TGCTCGACGGTCAG
```

where "p" is a 5' phosphate group and "FAM" is a fluorescein dye attached to the 3' carbon of the last nucleotide of the top strand by a commercially available 3' linker group (Clontech Laboratories). The ligation is carried out in the following reaction mixture: 5 µl 10x NEB buffer No. 2, 0.5 µl 1% Tween 20, 0.5 µl 100 mM ATP, 5 ml 3'-labeled adaptor (100 pmol/µl), 2.5 µl T4 DNA ligase (400 units/µl) and 36.5 µl H_2O, to give a final reaction volume of 50 µl. The reaction mixture is incubated at 16°C overnight with vortexing, after which the beads are washed once with PBS containing 1 mM $CaCl_2$ and treated with Pronase as described above. After this initial ligation, the nick remaining between the adaptor and tag-cDNA conjugate is sealed by

This Sheet Has Been Replaced

simultaneously treating with both a kinase and a ligase as follows. Loaded beads are resuspended in a reaction mixture consisting of the following components: 15 μ0x NEB buffer No. 2, 1.5 μl 1% Tween 20, 1.5 μl 100 mM ATP, 2 μl T4 polynucleotide kinase (10 units/μl), 1 μl T4 DNA ligase (400 units/μl), and 129 μl H_2O, for a final reaction volume of 150 μl.. The reaction mixture is incubated at 37°C for 1 hr with vortexing, after which the beads are washed once with PBS containing 1 mM $CaCl_2$, treated with Pronase as described above, and washed twice with TE/Tween.

After the labeled strand is melted off, preferably by treatment with 150 mM NaOH, the reference DNA on the beads is ready for competitive hybridization of differentially expressed gene products.

Example 2

Preparation of a Yeast Reference DNA Population Attached to Microparticles

In this example, Saccharomyces cerevisiae cells of strain YJM920 MATa Gal+ SUC2 CUP1 are grown in separate rich and minimal media cultures essentially as describe by Wodicka et al. (cited above). mRNA extracted from cells grown under both conditions are used to establish a reference cDNA population which is tagged, sampled, amplified, labeled, and loaded onto microparticles. Loaded microparticles are isolated by FACS, labels are removed, and the non-covalently bound strands of the loaded DNA are melted off and removed.

Yeast cells are grown at 30°C either in rich medium consisting of YPD (yeast extract/peptone/glucose, Bufferad, Newark, NJ) or in minimal medium (yeast nitrogen base without amino acids, plus glucose, Bufferad). Cell density is measured by counting cells from duplicate dilutions, and the number of viable cells per milliliter is estimated by plating dilutions of the cultures on YPD agar immediately before collecting cells for mRNA extraction. Cells is mid-log phase (1-5 x 10^7 cells/ml) are pelleted, washed twice with AE buffer solution (50 mM NaAc, pH 5.2, 10 mM EDTA), frozen in a dry ice-ethanol bath, and stored at -80°C.

mRNA is extracted as follows for both the construction of the reference DNA library and for preparation of DNA for competitive hybridization. Total RNA is extracted from frozen cell pellets using a hot phenol method, described by Schmitt et al., Nucleic Acids Research, 18: 3091-3092 (1990), with the addition of a chloroform-

This Sheet Has Been Replaced



isoamyl alcohol extraction just befor precipitation of the total RNA. Phase-Lock Gel (5 Prime-3 Prime, Inc., Boulder, CO) is used for all organic extractions to increase RNA recovery and decrease the potential for contamination of the RNA with material from the organic interface. Poly(A)$^+$ RNA is purified from the total RNA with an oligo-dT selection step (Oligotex, Qiagen, Chatsworth, CA).

5 μg each of mRNA from cells grown on rich medium and minimal medium are mixed for construction of a cDNA library in a pUC19 containing the tag repertoire of Formula I. The tag repertoire of Formula I is digested with Eco RI and Bam HI and inserted into a similarly digested pUC19. The mRNA is reverse transcribed with a commercially available kit (Strategene, La Jolla, CA) using an olgio-dT primer containing a sequence which generates a Bsm BI site identical to that of Formula I upon second strand synthesis. The resulting cDNAs are cleaved with Bsm BI and Dpn II and inserted into the tag-containing pUC19 after digestion with Bsm BI and Bam HI. After transfection and colony formation, the density of pUC19 tranformants is determined so that a sample containing approximately thirty thousand tag-cDNA conjugates may be obtained and expanded in culture. Alternatively, a sample of tag-cDNA conjugates are obtained by picking approximately 30 thousand clones, which are then mixed and expanded in culture.

From a standard miniprep of plasmid, the tag-cDNA conjugates are amplified by PCR with 5-methyldeoxycytosine triphosphate substituted for deoxycytosine triphosphate. The following 19-mer forward and reverse primers (SEQ ID NO: 8 and SEQ ID NO: 9), specific for flanking sequences in pUC19, are used in the reaction:

```
forward primer:        5'-biotin-AGTGAATTCGGGCCTTAATTAA

reverse primer:        5'-FAM-GTACCCGCGGCCGCGGTCGACTCTAGAGGATC
```

where "FAM" is an NHS ester of fluorescein (Clontech Laboratories, Palo Alto, CA) coupled to the 5' end of the reverse primer via an amino linkage, *e.g.* Aminolinker II (Perkin-Elmer, Applied Biosystems Division, Foster City, CA). The reverse primer is selected so that a Not I site is reconstituted in the double stranded product. After PCR amplification, the tag-cDNA conjugates are isolated on avidinated beads, *e.g.* M-280 Dynabeads (Dynal, Oslo, Norway).

This Sheet Has Been Replaced



After washing, the cDNAs bound to the beads are digested with Pac I releasing the tag-cDNA conjugates and a stripping reaction is carried out to render the oligonucleotide tags single stranded. After the reaction is quenched, the tag-cDNA conjugate is purified by phenol-chloroform extraction and combined with 5.5 Om GMA beads carrying tag complements, each tag complement having a 5' phosphate. Hybridization is conducted under stringent conditions in the presence of a thermal stable ligase so that only tags forming perfectly matched duplexes with their complements are ligated. The GMA beads are washed and the loaded beads are concentrated by FACS sorting, using the fluorescently labeled cDNAs to identify loaded GMA beads. The isolated beads are treated with Pac I to remove the fluorescent label, after which the beads are heated in an NaOH solution using conventional protocols to remove the non-covalently bound strand. After several washes the GMA beads are ready for competitive hybridization.

Example 3

Isolation and Identification of Up-Regulated and Down-Regulated Genes in Yeast Exposed to Different Growth Conditions

In this example, mRNA is extracted from cells of each culture and two populations of labeled polynucleotides are produced by a single round of poly(dT) primer extension by a reverse transcriptase in the presence of fluorescently label nucleoside triphosphates. Equal amounts of each of the labeled polynucleotides are then combined with the GMA beads of Example 1 carrying the reference DNA population for competitive hybridization, after which the beads are analyzed by FACS and those in the off-diagonal regions are accumulated for MPSS analysis.

Fluorescent nucleoside triphosphates Cy3-dUTP or CY5-dUTP (Amersham) are incorporated into cDNAs during reverse transcription of 1. μg of poly(A)$^+$ RNA obtained as described in Example 1 using a poly(dT)$_{16}$ primer in separate reactions. After heating the primer and RNA to 70°C for 10 min, the reaction mixture is transferred to ice and a premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleoside triphosphates are added to give the following concentrations: 500 μM for dATP, dCTP, and dGTP; 200 μM for dTTP; and 100 mM each for Cy3-dUTP or CY5-dUTP. After incubation at 42°C for 2 hours, unincorporated fluorescent nucleotides are removed by first

This Sheet Has Been Replaced



diluting the reaction mixture with 470 µl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating to about 5 µl using a Centricon-30 concentrator (Amicon). Purified labeled cDNA from both reactions is combined and resuspended in 11 µl of 3.5 x SSC containing 10 µg poly (dA) and 0.3 µl of 10% SDS. Prior to hybridization the solution is boiled for 2 min and allowed to cool to room temperature, after which it is applied to the GMA beads and incubated for about 8-12 hours at 62°C. After washing twice in 2 x SSC and 0.2% SDS, the GMA beads are resuspended in NEB-2 buffer (New England Biolabs, Beverly, MA) and loaded in a Coulter EPICS Elite ESP flow cytometer for analysis and sorting. In a two dimensional fluorescence intensity contour plot, the GMA beads generate a pattern as shown in Figure 1a. Sorting parameters are set as shown in Figure 1b so that GMA beads in the off-diagonal regions (112) are sorted and collected for MPSS analysis.

The labeled cDNA strands are melted from the GMA beads and removed by centrifugation. After several washes, a primer is annealed to the primer binding site shown in Formula I and extended in a conventional polymerization reaction to reconstitute the double stranded DNAs on the GMA beads which include the Dpn II site, described above. After digestion with Dpn II, beads loaded with tag-cDNA conjugates are placed in an instrument for MPSS analysis, as described in Albrecht et al. (cited above).

The top strands of the following 16 sets of 64 encoded adaptors (SEQ ID NO: 10 through SEQ ID NO: 25) are each separately synthesized on an automated DNA synthesizer (model 392 Applied Biosystems, Foster City) using standard methods. The bottom strand, which is the same for all adaptors, is synthesized separately then hybridized to the respective top strands:

SEQ ID NO.	Encoded Adaptor
10	5'-pANNNTACAGCTGCATCCCttggcgctgagg pATGCACGCGTAGGG-5'
11	5'-pNANNTACAGCTGCATCCCtgggcctgtaag pATGCACGCGTAGGG-5'
12	5'-pCNNNTACAGCTGCATCCCttgacgggtctc pATGCACGCGTAGGG-5'
13	5'-pNCNNTACAGCTGCATCCCtgcccgcacagt pATGCACGCGTAGGG-5'
14	5'-pGNNNTACAGCTGCATCCCttcgcctcggac

This Sheet Has Been Replaced



```
                pATGCACGCGTAGGG-5'
15          5'-pNGNNTACAGCTGCATCCCtgatccgctagc
                pATGCACGCGTAGGG-5'
16          5'-pTNNNTACAGCTGCATCCCttccgaacccgc
                pATGCACGCGTAGGG-5'
17          5'-pNTNNTACAGCTGCATCCCtgaggggatag
                pATGCACGCGTAGGG-5'
18          5'-pNNANTACAGCTGCATCCCttcccgctacac
                pATGCACGCGTAGGG-5'
19          5'-pNNNATACAGCTGCATCCCtgactccccgag
                pATGCACGCGTAGGG-5'
20          5'-pNNCNTACAGCTGCATCCCtgtgttgcgcgg
                pATGCACGCGTAGGG-5'
21          5'-pNNNCTACAGCTGCATCCCtctacagcagcg
                pATGCACGCGTAGGG-5'
22          5'-pNNGNTACAGCTGCATCCCtgtcgcgtcgtt
                pATGCACGCGTAGGG-5'
23          5'-pNNNGTACAGCTGCATCCCtcggagcaacct
                pATGCACGCGTAGGG-5'
24          5'-pNNTNTACAGCTGCATCCCtggtgaccgtag
                pATGCACGCGTAGGG-5'
25          5'-pNNNTTACAGCTGCATCCCtcccctgtcgga
                pATGCACGCGTAGGG-5'
```

where N is any of dA, dC, dG, or dT; p is a phosphate group; and the nucleotides indicated in lower case letters are the 12-mer oligonucleotide tags. Each tag differs from every other by 6 nucleotides. Equal molar quantities of each adaptor are combined in NEB #2 restriction buffer (New England Biolabs, Beverly, MA) to form a mixture at a concentration of 1000 pmol/μL.

Each of the 16 tag complements are separately synthesized as amino-derivatized oligonucleotides and are each labeled with a fluorescein molecule (using an NHS-ester of fluorescein, available from Molecular Probes, Eugene, OR) which is attached to the 5' end of the tag complement through a polyethylene glycol linker (Clonetech Laboratories, Palo Alto, CA). The sequences of the tag complements are simply the 12-mer complements of the tags listed above.

Ligation of the adaptors to the target polynucleotide is carried out in a mixture consisting of 5 μl beads (20 mg), 3 μL NEB 10x ligase buffer, 5 μL adaptor mix (25 nM), 2.5 μL NEB T4 DNA ligase (2000 units/μL), and 14.5 μL distilled water. The

This Sheet Has Been Replaced



mixture is incubated at 16°C for 30 minutes, after which the beads are washed 3 times in TE (pH 8.0).

After centrifugation and removal of TE, the 3' phosphates of the ligated adaptors are removed by treating the polynucleotide-bead mixture with calf intestinal alkaline phosphatase (CIP) (New England Biolabs, Beverly, MA), using the manufacturer's protocol. After removal of the 3' phosphates, the CIP may be inactivated by proteolytic digestion, *e.g.* using PronaseTM (available form Boeringer Mannhiem, Indianapolis, IN), or an equivalent protease, with the manufacturer's protocol. The polynucleotide-bead mixture is then washed, treated with a mixture of T4 polynucleotide kinase and T4 DNA ligase (New England Biolabs, Beverly, MA) to add a 5' phosphate at the gap between the target polynucleotide and the adaptor, and to complete the ligation of the adaptors to the target polynucleotide. The bead-polynucleotide mixture is then washed in TE.

Separately, each of the labeled tag complements is applied to the polynucleotide-bead mixture under conditions which permit the formation of perfectly matched duplexes only between the oligonucleotide tags and their respective complements, after which the mixture is washed under stringent conditions, and the presence or absence of a fluorescent signal is measured. Tag complements are applied in a solution consisting of 25 nM tag complement 50 mM NaCl, 3 mM Mg, 10 mM Tris-HCl (pH 8.5), at 20°C, incubated for 10 minutes, then washed in the same solution (without tag complement) for 10 minute at 55°C.

After the four nucleotides are identified as described above, the encoded adaptors are cleaved from the polynucleotides with Bbv I using the manufacturer's protocol. After an initial ligation and identification, the cycle of ligation, identification, and cleavage is repeated three times to give the sequence of the 16 terminal nucleotides of the target polynucleotide.

Preferably, analysis of the hybridized encoded adaptors takes place in an instrument which i) constrains the loaded microparticles to be disposed in a planar array in a flow chamber, ii) permits the programmed delivery of process reagents to the flow chamber, and iii) detects simultaneously optical signals from the array of microparticles. Such a preferred instrument is shown diagrammatically in Figure 2, and more fully disclosed in Bridgham et al., International patent application PCT/US98/11224. Briefly, flow chamber (500) is prepared by etching a cavity

This Sheet Has Been Replaced



having a fluid inlet (502) and outlet (504) in a glass plate (506) using standard micromachining techniques, *e.g.* Ekstrom et al., International patent application PCT/SE91/00327; Brown, U.S. patent 4,911,782; Harrison et al., Anal. Chem. 64: 1926-1932 (1992); and the like. The dimension of flow chamber (500) are such that loaded microparticles (508), *e.g.* GMA beads, may be disposed in cavity (510) in a closely packed planar monolayer of 100-200 thousand beads. Cavity (510) is made into a closed chamber with inlet and outlet by anodic bonding of a glass cover slip (512) onto the etched glass plate (506), *e.g.* Pomerantz, U.S. patent 3,397,279. Reagents are metered into the flow chamber from syringe pumps (514 through 520) through valve block (522) controlled by a microprocessor as is commonly used on automated DNA and peptide synthesizers, *e.g.* Bridgham et al., U.S. patent 4,668,479; Hood et al., U.S. patent 4,252,769; Barstow et al., U.S. patent 5,203,368; Hunkapiller, U.S. patent 4,703,913; or the like.

Three cycles of ligation, identification, and cleavage are carried out in flow chamber (500) to give the sequences of 12 nucleotides at the termini of each of approximately 100,000 fragments. Nucleotides of the fragments are identified by hybridizing tag complements to the encoded adaptors as described above. Specifically hybridized tag complements are detected by exciting their fluorescent labels with illumination beam (524) from light source (526), which may be a laser, mercury arc lamp, or the like. Illumination beam (524) passes through filter (528) and excites the fluorescent labels on tag complements specifically hybridized to encoded adaptors in flow chamber (500). Resulting fluorescence (530) is collected by confocal microscope (532), passed through filter (534), and directed to CCD camera (536), which creates an electronic image of the bead array for processing and analysis by workstation (538). Preferably, after each ligation and cleavage step, the cDNAs are treated with PronaseTM or like enzyme. Encoded adaptors and T4 DNA ligase (Promega, Madison, WI) at about 0.75 units per μL are passed through the flow chamber at a flow rate of about 1-2 μL per minute for about 20-30 minutes at 16°C, after which 3' phosphates are removed from the adaptors and the cDNAs prepared for second strand ligation by passing a mixture of alkaline phosphatase (New England Bioscience, Beverly, MA) at 0.02 units per μL and T4 DNA kinase (New England Bioscience, Beverly, MA) at 7 units per μL through the flow chamber at 37°C with a flow rate of 1-2 μL per minute for 15-20 minutes. Ligation is accomplished by T4

This Sheet Has Been Replaced



DNA ligase (.75 units per mL, Promega) through the flow chamber for 20-30 minutes. Tag complements at 25 nM concentration are passed through the flow chamber at a flow rate of 1-2 μL per minute for 10 minutes at 20°C, after which fluorescent labels carried by the tag complements are illuminated and fluorescence is collected. The tag complements are melted from the encoded adaptors by passing hybridization buffer through the flow chamber at a flow rate of 1-2 μL per minute at 55°C for 10 minutes. Encoded adaptors are cleaved from the cDNAs by passing Bbv I (New England Biosciences, Beverly, MA) at 1 unit/μL at a flow rate of 1-2 μL per minute for 20 minutes at 37°C.

Example 4

FACS Analysis of Microparticles Loaded with Different Ratios of DNAs Labeled with Fluorescein and CY5

In this example, the sensitivity of detecting different ratios of differently labeled cDNAs was tested by constructing a reference DNA population consisting of a single clone and then competitively hybridizing to the reference DNA population different ratios of complementary strands labeled with different fluorescent dyes. The reference DNA population consisted of a cDNA clone, designated "88.11," which is an 87-basepair fragment of an expressed gene of the human monocyte cell line THP-1, available from the American Type Culture Collection (Rockville, Maryland) under accession number TIB 202. The nucleotide sequence of 88.11 has a high degree of homology to many entries in the GenBank Expressed Sequence Tag library, *e.g.* gb AA830602 (98%). The reference DNA population, which consisted of only 88.11 cDNA, was prepared as described in Example 1, with the exception that a special population of microparticles was prepared in which all microparticles had the same tag complement attached. The corresponding oligonucleotide tag was attached to the 88.11 cDNA. Thus, only monospecific populations of tags and tag complements were involved in the experiment. After competitive hybridization, the loaded microparticles were analyzed on a Cytomation, Inc. (Ft. Collins, CO) FACS instrument as described above.

88.11 cDNA was also cloned into a vector identical to that of Example 1 (330 of Figure 3b), except that it did not contain tag 336. 10 μg of vector DNA was linearized by cleaving to completion with Sau 3A, an isoschizomer of Dpn II (342 of

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

Figure 3b), after which two 1 μg aliquots of the purified linear DNA were taken. From each 1 μg aliquot, about 20 μg of labeled single stranded DNA product was produced by repeated cycles of linear amplification using primers specific for primer binding site 332. In one aliquot, product was labeled by incorporation of rhodamine R110-labeled dUTP (PE Applied Biosystems, Foster City, CA); and in the other aliquot, product was labeled by incorporation of CY5-labeled dUTP (Amersham Corporation, Arlington Heights, IL). Quantities of the labeled products were combined to form seven 5 μg amounts of the two products in ratios of 1:1, 2:1, 1:2, 4:1, 1:4, 8:1, and 1:8. The 5 μg quantities of labeled product were separately hybridized to 1.6×10^5 microparticles (GMA beads with 88.11 cDNA attached) overnight at 65°C in 50 μl 4x SSC with 0.2% SDS, after which the reaction was quenched by diluting to 10 ml with ice-cold TE/Tween buffer (defined above). The loaded microparticles were centrifuged, washed by suspending in 0.5 ml 1x SSC with 0.2% SDS for 15 min at 65°C, centrifuged, and washed again by suspending in 0.5 ml 0.1x SSC with 0.2% SDS for 15 min at 55°C. After the second washing, the microparticles were centrifuged and resuspended in 0.5 ml TE/Tween solution for FACS analysis.

The results are shown in Figures 5a-5e, where in each Figure the vertical axis corresponds to CY5 fluorescence and the horizontal axis corresponds to rhodamine R110 fluorescence. In Figure 5a, a population of microparticles were combined that had either all R110-labeled DNA or all CY5-labeled DNA hybridized to the complementary reference strands. Contours 550 and 552 are clearly distinguished by the detection system of the FACS instrument and microparticles of both populations produce readily detectable signals. Figure 5b illustrates the case where the R110- and CY5-labeled strands are hybridized in equal proportions. As expected, the resulting contour is located on the diagonal of the graph and corresponds to the position expected for non-regulated genes. Figures 5c through 5e show the analysis of three pairs of competitive hybridizations: i) R110- and CY5-labeled strands hybridized in a 2:1 concentration ratio and a 1:2 concentration ratio, ii) R110- and CY5-labeled strands hybridized in a 4:1 concentration ratio and a 1:4 concentration ratio, and iii) R110- and CY5-labeled strands hybridized in an 8:1 concentration ratio and a 1:8 concentration ratio. The data of Figure 5c suggest that genes up-regulated or down-regulated by a factor of two are detectable in the present embodiment, but that

This Sheet Has Been Replaced



SUBSTITUTE SHEET (RULE 26)

significant overlap may exist between signals generated by regulated and non-regulated genes. Figures 5d and 5e suggest that genes up-regulated or down-regulated by a factor of four or higher are readily detectable over non-regulated genes.

Example 5

FACS Analysis of Differentially Expressed Genes from Stimulated and Unstimulated THP-1 Cells

In this example, a reference DNA population attached to microparticles was constructed from cDNA derived from THP-1 cells stimulated as indicated below. Equal concentrations of labeled cDNAs from both stimulated and unstimulated THP-1 cells were then competitively hybridized to the reference DNA population, as described in Example 1, and the microparticles carrying the labeled cDNAs were analyzed by a FACS instrument. THP-1 cells were stimulated by treatment with phorbol 12-myristate 13-acetate (PMA) and lipopolysaccharide (LPS).

THP-1 cells were grown in T-165 flasks (Costar, No. 3151) containing 50 ml DMEM/F12 media (Gibco, No. 11320-033) supplemented with 10% fetal bovine serum (FBS)(Gibco, No. 26140-038), 100 units/ml penicillin, 100 μg/ml streptomycin (Gibco, No. 15140-122), and 0.5 μM β-mercaptoethanol (Sigma, No. M3148). Cultures were seeded with 1 x 10^5 cells/ml and grown to a maximal density of 1 x 10^6. Doubling time of the cell populations in culture was about 36 hours. Cells were treated with PMA as follows: Cells from a flask (about 5 x 10^7 cells) were centrifuged (Beckman model GS-6R) at 1200 rpm for 5 minutes and resuspended in 50 ml of fresh culture media (without antibiotics) containing 5 μl of 1.0 mM PMA (Sigma, No. P-8139) in DMSO (Gibco No. 21985-023) or 5 μl DMSO (for the unstimulated population), after which the cells were cultured for 48 hours. Following the 48 hour incubation, media and non-adherent cells were aspirated from the experimental flask (*i.e.* containing stimulated cells) and fresh media (without antibiotics) was added, the fresh media containing 10 μl of 5 mg/ml LPS (Sigma, No. L-4130) in phosphate buffered saline (PBS). The culture of unstimulated cells was centrifuged (Beckman model GS-6R) at 1200 rpm for 5 minutes at 4°C so that a pellet formed which was then resuspended in 50 ml of fresh growth media containing 10 μl PBS. Both the cultures of stimulated and unstimulated cells were incubated at 37°C for four hours, after which cells were harvested as follows: Media was aspirated from

This Sheet Has Been Replaced



the cultures and adherent cells were washed twice with warm PBS, after which 10 ml PBS was added and the cells were dislodged with a cell scaper. The dislodged cells were collected and their concentration was determined with a hemocytometer, after which they were centrifuged (Beckman model GS-6R) at 1200 rpm for 5 minutes to form a pellet which was used immediately for RNA extraction.

mRNA was extracted from about 5 x 10^6 cells using a FastTrack 2.0 kit (No. K1593-02, Invitrogen, Inc. San Diego, CA) for isolating mRNA. The manufacturer's protocol was followed without significant alterations. A reference DNA population attached to microparticles was constructed from mRNA extracted from stimulated cells, as described in Example 1. Separate cDNA libraries were constructed from mRNA extracted from stimulated and unstimulated cells. The vectors used for the libraries were identical to that of Example 1, except that they did not contain oligonucleotide tags (336 of Figure 3b). Following the protocol of Example 4, approximately 2.5 μg of rhodamine R110-labeled single stranded DNA was produced from the cDNA library derived from stimulated cells, and approximately 2.5 μg of CY5-labeled single stranded DNA was produced from the cDNA library derived from unstimulated cells. The two 2.5 μg aliquots were mixed and competitively hybridized to the reference DNA on 9.34 x 10^5 microparticles. The reaction conditions and protocol was as described in Example 4.

After hybridization, the microparticles were sorted by a Cytomation, Inc. MoFlo FACS instrument as described above. Figure 6 contains a conventional FACS contour plot 600 of the frequencies of microparticles with different fluorescent intensity values for the two fluorescent dyes. Approximately 10,000 microparticles corresponding to up-regulated genes (sort window 602 of Figure 6) were isolated, and approximately 12,000 microparticles corresponding to down-regulated genes (sort window 604 of Figure 6) were isolated. After melting off the labeled strands, as described above, the cDNAs carried by the microparticles were amplified using a commercial PCR cloning kit (Clontech Laboratories, Palo Alto, CA), and cloned into the manufacturer's recommended cloning vector. After transformation, expansion of a host culture, and plating, 87 colonies of up-regulated cDNAs were picked and 73 colonies of down-regulated cDNAs were picked. cDNAs carried by plasmids extracted from these colonies were sequenced using conventional protocols on a PE

This Sheet Has Been Replaced

Applied Biosystems model 373 automated DNA sequencer. The identified sequences are listed in Tables 1 and 2.

Table 1

Up-Regulated Genes

No. Copies	Description	GenBank Indentifier
19	LD78/MIP-1	HUMCKLD78
16	TNF-inducible (TSF-6) mRNA	HUMTSG6A
15	GRO-γ (MIP-2β)	HUMGROG5
6	GRO-β (MIP-2α)	HUMGROB
6	act-2	HUMACT2A
4	guanylate binding protein isoform I (GBP-2)	HUMGBP1
4	spermidine/spermin N1-acetyltransferase	HUMSPERMNA
4	adipocyte lipid-binding protein	HUMALBP
3	Fibronectin	HSFIB1
3	interleukin-8	HSMDNCF
1	insulin-like growth factor binding protein 3	HSIGFBP3M
1	interferon-γ inducible early response gene	HSINFGER
1	type IV collangenase	
1	cathepsin L	HSCATHL
1	EST	
1	EST	
1	Genomic/EST	HSAC002079

Table 2

Down-Regulated Genes

No. Copies	Description	GenBank Indentifier
16	Elongation factor 1	HSEF1AC
4	Ribosomal protein S3a/v-fos tranf. Effector	HUMFTE1A
6	Ribosomal protein S7	HUMRPS17
2	Translationally controlled tumor protein	HSTUMP
3	23 kD highly basic protein	HS23KDHBP
2	Laminin receptor	HUMLAMR
2	Cytoskeletal gamma-actin	HSACTCGR
2	Ribosomal protein L6	HSRPL6AA
2	Ribosomal protein L10	HUMRP10A
2	Ribosomal protein L21	HSU14967
2	Ribosomal protein S27	HSU57847
1	Ribosomal protein L5	HSU14966
1	Ribosomal protein L9	HSU09953
1	Ribosomal protein L17	HSRPL17
1	Ribosomal protein L30	HSRPL30
1	Ribosomal protein L38	HSRPL38
1	Ribosomal protein S8	HSRPS8
1	Ribosomal protein S13	HSRPS13
1	Ribosomal protein S18	HSRPS18
1	Ribosomal protein S20	HUMRPS20
1	Acidic ribosomal phosphoprotein P0	HUMPPARP0
1	26S proteasome subunit p97	HUM26SPSP
1	DNA-binding protein B	HUMAAE
1	T-cell cyclophilin	HSCYCR
1	Interferon inducible 6-26 mRNA	HSIFNIN4
1	Hematopoetic proteoglycan core protein	HSHPCP
1	Fau	HSFAU
1	beta-actin	HSACTB
1	Nuclear enc. mito. serine hydroxymethyltrans.	HUMSHMTB
1	Mito. Cytochrome c oxidase subunit II	HUMMTCDK
1	Genomic	W92931
1	EST	W84529
1	EST	AA933890
1	EST	AA206288
1	EST	AA649735
1	EST	N34678
1	EST	AA166702
1	EST	AA630799
3	Genomic	

Example 6

FACS Analysis of Differentially Expressed Genes from Stimulated and Unstimulated THP-1 Cells

(Experiment: Comp 11)

A reference DNA population attached to microparticles was constructed from cDNA derived from stimulated THP-1 cells. cDNA from stimulated and unstimulated THP-1 cells was prepared for competitive hybridization as follows. 20 μg each of the THP-1 unstimulated probe library (U3A-TL) and the THP-1 stimulated probe library (S3A-TL) were digested with 50 units of Sau3A to prepare the vector for linear PCR. The DNA was purified by phenol/chloroform extraction and fluorescently labelled by PCR. For calibration purposes, both CY5 and R110 were used to label each condition.

The U3A-TL DNA was labeled with CY5 and the S3A-TL DNA was labeled with R110. Briefly, a reaction mixture containing 80 μl 10X PCR Buffer; 16 μl biotinylated primer (B-Primer, 125 pmole/:l); 16 μl dNTPs (6.25 mM); 4 μg template; 16 μl Klentaq enzyme; 64 μl R110 dUTP or 6.4 μl of CY5 dUTP; and water to bring the total volume to 800 μl. This mixture was dispensed into 8 aliquots, which then underwent 34 cycles of PCR according to the following protocol: 1) 94°C 3 min; 2) 94°C 30 sec; 3) 62°C 30 sec; 4) 72°C 1 min; and 5) 72°C 10 min. The PCR reaction was purified and the colored nucleotides were removed by precipitation.

Reference Population

The Comp 11 bead library consisted of 2,667,369 beads, with a complexity of 1 million clones from the THP-1 stimulated library. The beads were prepared as described above as outlined in Figure 3. The starting PMT2 mean for the FITC signal was 19.5. The duplexed DNA on the beads was denatured with 2.5 ml 150mM NaOH washes at RT for 15min with mild vortexing. The efficiency of the denaturization was determined by measuring the remaining FITC signal mean, which was 2.2, i.e., 11.3% residual fluorescence. The beads were washed twice in .5 ml of 4X SSC .1% SDS.

Competitive Hybridization

100,000 beads were hybridized with 10 μg of each linear PCR product of the stimulated probe library (S3A-TL) labeled with CY5 and the same library labeled with R110. 936,542 beads were hybridized with 10 μg of CY5 stimulated probe and 10 μg of R110 unstimulated probe. The beads were assembled in 50 μl with a final

buffer composition of 4X SSC/.1% SDS. The samples were heated to 80°C for 3 minutes, the probes were added and the temperature was moved to 65°C. Hybridization continued for 16 hrs. with vortexing. The beads were ice quenched in 10 ml of TE Tween. The recovered samples were rinsed 2 times with 1X SSC /.1% SDS, resuspended in .5 ml of 1X SSC /.1%SDS, and washed at 65°C for 15 min. The beads were rinsed in .1X SSC /.1%SDS and washed at 55°C in .1X SSC /.1% SDS for 15 min. The samples were rinsed with TE Tween and 10,000 events of both samples were analyzed on the BD FacsCaliber. 10,163 beads (1.15%), the brightest CY5 off the 1:1 diagonal, were sorted. 11,977 beads (1.35%), the brightest R110 off the 1:1 diagonal, were sorted. The beads were pooled in a PCR reaction, TA cloned, and sequenced. The identified sequences are listed in Tables 3 and 4.

Table 3

Comp 11: Downregulated Genes

No. Copies	Description	GenBank Identifier
99	23 kD highly basic protein	HS23KDHBP
1	26S proteasome subunit p55	AB003103
7	26S proteasome subunit p97	HUM26SPSP
1	28kD heat shock protein	HSHSP28
3	90 kD HSP	HSHSP90R
1	α NAC	HSANAC
2	α-enolase	HSAEP
3	α1 acid glycoprotein	HUMAGP1A
21	Acidic ribosomal phosphoprotein P0	HUMPPARP0
4	Acidic ribosomal phosphoprotein P1	HUMPPARP1
3	Acidic ribosomal phosphoprotein P2	HUMPPARP2
1	activin β-C chain	HSACTNBC
3	Adenylyl cyclase-associated protein (CAP)	HUMADCY
2	ADP/ATP translocase	HUMTLCA
3	Allograft-inflammatory factor 1	HSU19713
13	Antioxidant enzyme AOE37-2	HSU25182
1	Arp2/3 protein complex subunit p21Arc	AF006084
2	Arp2/3 protein complex subunit p41Arc	AF06086
1	ATP-dependent RNA helicase	AB001636
1	B94	HUMB94
7	basic transcription factor 3a (BTF3a)	HSBTF3
3	BBC1	HSBBC1
3	beta-actin	HSACTB
1	brain-expressed HHCPA78 homolog	S73591
1	c-myc transcription factor puf	HUMPUF
3	Calmodulin	HUMCAMA
1	cAMP response element regulatory protein	HUMCREB2A

This Sheet Has Been Replaced



No. Copies	Description	GenBank Identifier
1	cis-acting sequence	HUMCIS
1	Cks1 protein homolog	HSCKSNS2
3	clathrin assembly protein 50	HSU36188
1	CLP	HUMCLPB
5	Cu/Zn SOD-1	HSSODR1
1	Cyclophilin	HUMCYCLO
3	Cytochrome c oxidase cox VIIa-L	HSCOX7AL
1	Cytochrome c oxidase subunit Vb	HUMCOXCA
3	Cytochrome c oxidase subunit Vic	HSCOVIC
4	Cytoskeletal gamma-actin	HSACTCGR
3	Cytoskeletal tropomycin TM30	HSTROPCR
4	DNA-binding protein B	HUMAAE
2	EBV small RNAs associated protein (EAP)	HSEAP
30	Elongation factor 1 α	HSEF1AC
1	Elongation factor 1 δ	HSEF1DELA
1	Elongation factor 1 γ	HSEF1GMR
1	ERp28 protein	HSERP28
9	Fau	HSFAU
1	ferritin L chain	HUMFERL
1	Fibronectin receptor	HSFNRA
4	Fus	HSFUSA
2	G-β-like protein	HUMMHBA123
4	Glutaminyl tRNA synthetase	HSGTS
2	H+ ATP synthase subunit b	HSATPSYN
2	H3.3 histone, class B	HUMHISH3B
10	Heat shock factor binding protein 1	AF068754
5	Heat shock protein 86	HSHSP86
4	Hematopoetic lineage cell specific protein	HSHEAM
2	Hematopoetic proteoglycan core protein	HSHPCP
1	HLA-DR associated protein	HSPHAPII
3	HMG-17	HUMHMG17
2	Icln chlorine channel regulatory protein	HSU17899
2	IL-8	HSMDNCF
1	IMP dehydrogenase	HUMIMP
3	Initiation factor 4B	HSINTFA4B
1	Insulinoma rig analog	HUMIDB
2	Interferon inducible 6-26 mRNA	HSIFNIN4
1	KIAA0116	HUMORFA10
1	KIAA0164	D79986
5	KIAA0571	AB01114B
11	Lactate dehydrogenase B	HSLDHBR
12	Laminin receptor	HUMLAMR
7	LD78/MIP-1	HUMCKLD78
1	Leucine-rich protein	HUM130LEU
1	LLRep3	HSLLREP3
1	low Mr GTP-binding protein (RAB32)	HSU71127

No. Copies	Description	GenBank Identifier
1	LST1	HSLST1G
2	MAPKAP kinase (3pK)	HSU09578
5	MHC protein hom. to chicken B complex protein	HUMMHBA123
1	Mitochondrial cytochrome c oxidase subunit II	HUMMTCDK
2	Mitochondrial phosphate carrier protein	SSMPCP
1	Mitochondrial serine hydroxymethyl transferase	HUMSHMTB
1	Mitochondrial tRNAs	MIT1HS
2	Mitochondrial ubiquinone-binding protein	HSUBPQPC
1	Mn SOD-2	HUMSUDIS
2	Myeloid progenitor inhibitory factor (MPIF-1)	HSU85767
1	Myosin regulatory light chain	HSMRLCM
1	Nuclear-encoded mito. serine hydroxymethyltransferase	HUMSHMTB
1	P2U nucleotide receptor	S74902
8	Palmitoyl-protein thioesterase	HSU44772
1	Phosphate carrier	HSPHOSC
2	Prothymosin α	HUMTHYMA
21	Ribosomal protein L10	HUMRP10A
4	ribosomal protein L11	HSRPL11
8	ribosomal protein L14	D87735
13	ribosomal protein L17	HSRPL17
4	ribosomal protein L18a	HUMRIBPROD
27	ribosomal protein L21	HSU14967
4	ribosomal protein L23 (putative)	HSL23MR
4	ribosomal protein L25	HSU12465
4	ribosomal protein L26	HSRP26AA
6	ribosomal protein L27a	HSU14968
1	ribosomal protein L28	HSU14969
15	ribosomal protein L29	HSRPL29
10	ribosomal protein L3	HUMRRL3A
2	ribosomal protein L30	HUMRPL30A
1	ribosomal protein L30	HSRPL30
12	ribosomal protein L32	HSRPL32
1	ribosomal protein L34	HUMRPL34A
1	ribosomal protein L35	HSU12465
3	ribosomal protein L37a	HSRPL37A
11	ribosomal protein L38	HSRPL38
8	ribosomal protein L4	HSHRPL4
3	ribosomal protein L41	AF026844
19	ribosomal protein L5	HSU14966
22	ribosomal protein L6	HSRPL6AA
17	ribosomal protein L7	HSRBPRL7A
13	ribosomal protein L7a	HUMRPL7A
19	ribosomal protein L9	HSU09953

No. Copies	Description	GenBank Identifier
25	ribosomal protein S11	HSRPS11
5	ribosomal protein S13	HSRPS13
11	ribosomal protein S15a	HSRPS15A
5	ribosomal protein S16	HUMSRAA
28	ribosomal protein S17	HUMRPS17
35	ribosomal protein S18	HSRPS18
2	ribosomal protein S19	HUMS19RP
6	ribosomal protein S20	HUMRPS20
11	ribosomal protein S27	HSU57847
17	ribosomal protein S28(hu homolog of yeast)	HUMRSPT
3	ribosomal protein S3	HSHUMS3
4	ribosomal protein S3a/v-fos transf. effector protein	HUMFTE1A
6	ribosomal protein S4	HUMRPS4X
1	ribosomal protein S7	HUMRPS7A
20	ribosomal protein S8	HSRPS8
1	RNAse/angiogenin inhibitor	HSRAI
1	small nuclear RNA U2	HSU25766
5	T-cell cyclophilin	HSCYCR
1	T-cell surface glycoprotein	HSE2
1	TI-227H	HUMTI227HC
1	transcriptional coactivator PC4	HSU12979
1	translation initiation factor 2 β subunit	HUMELF2
1	translation initiation factor eIF3 p40 subunit	HSU54559
2	translationally controlled tumor protein	HSTUMP
1	U1 small nuclear RNP-specific C protein	HSU1RNPC
2	ubiquinol-cytochrome c oxidase smallest subunit	D55636
1	ubiquinone binding protein	HUMQBPCA
4	Ubiquitin	HSU49869
2	Ubiquitin	HUMUBI13
7	ubiquitin Uba52	HSUBA52P
11	ubiquitin Uba80	HSIBA80R
8	EST	AA149853
6	EST	AA759306
3	EST	AI053510
2	EST	AA630799
2	EST	N26031
2	EST	AA843411
2	EST	AA234913
2	EST	AI034446
2	EST	AI054090
1	EST	AI054090
1	EST	AA828574
1	EST	AI087086
1	EST	AI031866

No. Copies	Description	GenBank Identifier
1	EST	AI040041
1	EST	AA542832
1	EST	N73319
1	EST	AA464447
1	EST	AA993034
1	EST	AA933890
1	EST	AI095923
1	EST	AA166703
1	EST	AA573139
1	EST	AA938293
1	EST	H43222
1	EST	AA576961
1	EST	AI015173
1	EST	AI015700
1	EST	W95680
1	EST	AA934688
1	EST	AA204731
2	NO MATCH	
1	NO MATCH	
1	NO MATCH	
1	NO MATCH	
918	**total sequenced (downregulated)**	

Table 4

Comp 11: Upregulated Genes

No. Copies	Description	GenBank Identifier
6	23 kD highly basic protein	HS23KDHBP
103	Act-2	HUMACT2A
1	activated B cell factor 1	AF060154
1	activating transcription factor 3	HUMATF3X
2	adenylyl cyclase-associated protein (CAP)	HUMADCY
47	adipocyte lipid-binding protein	HUMALBP
3	aquaporin 9	AB008775
17	ATPase	HUMH01A
22	B94	HUMB94
1	Cathepsin B	HUMCATHB
10	Cathepsin L	HSCATHL
5	EBV-induced protein	HSU19261
1	Elongation factor 1	HSEF1AC
46	Fibronectin	HSFIB1
57	Guanylate binding protein isoform I (GBP-2)	HUMGBP1
2	IFN-γ inducible early response gene	HSINFGER
33	IGF binding protein 3	HSIGFBP3M

No. Copies	Description	GenBank Identifier
1	IL-1 receptor antagonist	HSI1RA
3	IL-1β	HUMIL1BA
20	IL-8	HSMDNCF
4	Insulin-like growth factor binding protein 3	HSIGFBP3M
3	JKA3 mRNA induced upon T cell stimulation	HSU38443
2	KIAA0251	D87438
3	Macrophage scavenger receptor type I	HUMRMSR1
184	MIP-1α (LD78)	HUMCKLD78
218	MIP-2α (GRO-β)	HUMGROB
50	MIP-2β (GRO-γ)	HUMGROG5
58	Mn SOD	HSMNSOD
4	Musculin	AF087036
1	Paraplegin	HSPARAPLE
3	Prostaglandin endoperoxide synthase-2	HUMPTGS2
1	RANTES	HUMRANTES
1	Reticulocalbin	HUMRCN
1	Ribosomal protein L21	HSU14967
1	Ribosomal protein L7	HSRBPRL7A
1	Ribosomal protein S28	HUMRSPT
34	Spermidine/spermine N1-acetyltransferase	HUMSPERMNA
1	Striated muscle contraction reg. Protein	HUMID2B
95	TNF-inducible (TSG-6) mRNA	HUMTSG6A
10	TNFα	HSTNFR
2	Translation initiation factor 2β	HUMELF2
1	TRNA-Ala	HSCR6ALAT
17	Type IV collagenase	HUM4COLA
19	EST	AA916304
15	EST	AA873350
10	EST	AA011639
8	EST	AA346072
3	EST	AA284427
2	EST, IL-1/TNF-inducible	HSEST222
2	EST	W88513
1	EST	AA904231
1	EST	AA767777
1	EST	AA969937
1	EST	AA528703
4	Genomic	HSAC000119
4	Genomic	AC000403
1	Genomic	AC004130
6	NO MATCH	
2	NO MATCH	
2	NO MATCH	
1	NO MATCH	
1	NO MATCH	
1	NO MATCH	

No. Copies	Description	GenBank Identifier
1	NO MATCH	
1	NO MATCH	
1157	**Total sequenced (upregulated)**	

Example 7

FACS Analysis of Differentially Expressed Genes from Stimulated and Unstimulated THP-1 Cells

(Experiment: Comp 14)

In a separate experiment, reference DNA population preparation and competitive hybridization were done as described in Example 6. 9150 beads (0.89%), the brightest CY5 off the 1:1 diagonal, were sorted. 11085 beads (1.15%), the brightest R110 off the 1:1 diagonal, were sorted. The identified sequences are listed in Tables 5 and 6.

Table 5

Comp 14: Downregulated Genes

No. Copies	Description
29	H.sapiens mRNA for 23 kD highly basic protein
13	H.sapiens mRNA for ribosomal protein S18
12	Laminin receptor homolog mRNA
12	H.sapiens mRNA for ribosomal protein L26
12	Human ribosomal L5 protein mRNA
9	Human mRNA for elongation factor 1-alpha
9	H.sapiens mRNA for large subunit of ribosomal protein L21
7	H.sapiens gene for ribosomal protein L38
6	Homo sapiens cDNA, 3' end /clone=IMAGE
6	H.sapiens rpS8 gene for ribosomal protein S8
5	Human ribosomal protein L3 mRNA
5	Human Ki nuclear autoantigen mRNA
5	Human ribosomal protein L7a mRNA
5	Novel
5	Human mRNA for ribosomal protein S11
5	Neuroblastoma RAS viral (v-ras) oncogene homolog
5	Human mitochondrial DNA
5	H.sapiens initiation factor 4B cDNA
5	Human endothelial-monocyte activating polypeptide II mRNA
5	Novel
5	Human monocytic leukaemia zinc finger protein (MOZ) mRNA
4	Human platelet activating factor acetylhydrolase, brain isoform, 45 kDa subunit (LIS1) gene

No. Copies	Description
4	Human ferritin L chain mRNA
4	Human DNA sequence from cosmid cN37F10 on chromosome 22q11.2-qter
4	Human mRNA for core I protein
4	H.sapiens mRNA homologous to mouse P21 mRNA
4	H.sapiens mRNA for ribosomal protein L6
4	Human ribosomal protein L9 pseudogene
4	Homo sapiens cDNA, 3' end /clone=486654
4	Human MHC protein homologous to chicken B complex protein mRNA
4	Human elongation factor EF-1-alpha gene
3	H.sapiens Sub1.5 mRNA
3	Human mRNA for Apo1_Human (MER5(Aop1-Mouse)-like protein)
3	Homo sapiens chromosome 5, P1 clone 702A10 (LBNL H56)
3	Human fumarase precursor (FH) mRNA
3	Homo sapiens cDNA, 3' end /clone=IMAGE:1695780
3	Human GST1-Hs mRNA for GTP-binding protein
3	H.sapiens mRNA for RNA polymerase II 140 kDa subunit
3	Homo sapiens ribosomal protein L30 mRNA
3	Human ribosomal protein S17 mRNA
3	HSEST222 Homo sapiens cDNA /clone=MEC-222 /gb=X84721 /gi=673398 /ug=Hs.115716 /len=558
3	Homo sapiens Arp2/3 protein complex subunit p21-Arc (ARC21) mRNA
3	Human cytoplasmic dynein light chain 1 (hdlc1) mRNA
3	Human ribosomal protein S3a mRNA
3	Human mRNA for heat shock protein hsp86
2	Homo sapiens Munc13 mRNA
2	Human translational initiation factor 2 beta subunit (eIF-2-beta) mRNA
2	Human mRNA for potential laminin-binding protein
2	Human cyclophilin-related processed pseudogene
2	Homo sapiens ribosomal protein S20 (RPS20) mRNA
2	Human acidic ribosomal phosphoprotein P1 mRNA
2	Human ribosomal protein S13 (RPS13) mRNA
2	Novel
2	Homo sapiens cDNA /clone=IMAGE:979232
2	Homo sapiens cDNA, 3' end /clone=81477
2	Human intercellular adhesion molecule-1 (ICAM-1) mRNA
2	Human mRNA for ribosomal protein L17
2	Human mRNA for carboxyl methyltransferase
2	Human mRNA for cytoskeletal gamma-actin
2	Homo sapiens cDNA, 3' end /clone=626635
2	Human nucleophosmin mRNA
2	Human ribosomal protein L10 mRna
2	Novel
2	Y box binding protein-1 (YB-1) mRNA
2	Human guanylate binding protein isoform I (GBP-2) mRNA
2	Homo sapiens cyclin D3 (CCND3) mRNA
2	Novel
2	Homo sapiens cDNA, 3' end /clone=IMAGE:1474218

No. Copies	Description
2	Homo sapiens Ubiquitin mRNA sequence
2	Human ribosomal protein L7
2	Human phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
2	HSC3EF102 Homo sapiens cDNA, 3' end /clone
	Total singlets: 189
	Total contigs: 72
	Total seq reads in contigs: 306
	Total seqs to be searched: 610

Table 6

Comp 14: Upregulated Genes

No. Copies	Description
77	Human mRNA for putative cytokine 21 (HC21)
31	Human gene for tumor necrosis factor (TNF-alpha)
27	Human insulin-like growth factor-binding protein-3 gene
26	Human LD78 alpha gene
23	Human mRNA for macrophage inflammatory protein-2beta (MIP2beta)
20	Human cytokine LD78 gene
20	H.sapiens gene for spermidine/spermine N1-acetyltransferase
20	Human gene for melanoma growth stimulatory activity (MGSA)
17	Human ferritin H chain mRNA
13	Novel
13	Human adipocyte lipid-binding protein
12	Human interleukin 8 (IL8) gene
9	Homo sapiens cDNA, 3' end /clone=73864
8	Human ATL-derived PMA-responsive (APR) peptide mRNA
8	Human ATL-derived PMA-responsive (APR) peptide mRNA
7	Human cell surface glycoprotein CD44 mRNA
7	H.sapiens SOD-2 gene
7	Human hypoxanthine phosphoribosyltransferase (HPRT) gene
6	Human tumor necrosis factor-inducible (TSG-6) mRNA fragment
6	Human adenosine receptor (A2) gene
5	Human phosphatidylinositol 3-kinase catalytic subunit p110delta Mrna
4	Human BAC clone RG104I04 no function
4	Homo sapiens adenosine triphosphatase mRNA
4	Human mRNA (3'-fragment) for (2'-5') oligo A synthetase E
4	Genomic sequence no function
2	Human type IV collagenase mRNA
2	Human ribosomal protein S17 mRNA
2	Homo sapiens cDNA, 3' end /clone=IMAGE:1459553
2	Human interleukin 1-beta (IL1B) gene

Example 8

FACS Analysis of Differentially Expressed Genes from Stimulated and Unstimulated THP-1 Cells
(Experiment: Comp 15)

In a separate experiment, cDNA from stimulated and unstimulated THP-1 cells was prepared for competitive hybridization as described in Example 6. The reference DNA population was prepared as described in Example 6, except that the Comp 15 bead library consisted of 2,570,000 beads, with a complexity of 1 million clones from the THP-1 stimulated library and the THP-1 unstimulated library (50% of each). 13,988 beads (.87%), the brightest CY5 off the 1:1 diagonal, were sorted. 17,393 beads (1.08%), the brightest R110 off the 1:1 diagonal, were sorted. The identified sequences are listed in Tables 7 and 8.

Table 7

Comp 15: Downregulated Genes

No. Copies	Description
25	H.sapiens mRNA for 23 kD highly basic protein
17	Homo sapiens ribosomal protein L30 Mrna
16	H.sapiens mRNA for ribosomal protein S18
15	H.sapiens mRNA for ribosomal protein L6
14	L44-like ribosomal protein (L44L) and FTP3 (FTP3) genes
11	Homo sapiens PYRIN (MEFV) mRNA
9	Human cathepsin G mRNA
8	Human mRNA for ribosomal protein S11
8	Novel
8	H.sapiens mRNA for ribosomal protein L37a
8	Novel
8	H.sapiens mRNA for ribosomal protein L26
8	H.sapiens mRNA for translationally controlled tumor protein p21 Homology
8	Human deoxyuridine triphosphatase (DUT) Mrna
8	Human growth factor independence-1 (Gfi-1) mRNA
7	Human mRNA for ribosomal protein L39
7	Human ribosomal protein L10 mRNA
6	Human ribosomal protein L9 mRNA, complete cds. 5/96
6	Homo sapiens cDNA, 3' end /clone=IMAGE:1862607 /clone_end=3' /gb=AI053436 /ug=Hs.135355 /len=138
6	Human gene for catalase Weak Homology
5	H.sapiens mRNA for ribosomal protein L7
5	Homo sapiens (clone cori-1cl5) S29 ribosomal protein mRNA
5	Human mRNA for HBp15/L22
5	H.sapiens mRNA for NEFA protein
5	Novel
5	Human mRNA for potential laminin-binding protein
5	HSEST222
4	Human ribosomal protein S16 mRNA
4	Homo sapiens cDNA /clone=IMAGE:1118473 /gb=AA603101 /gi=2436962 /ug=Hs.14214 /len=621
4	H.sapiens mRNA for large subunit of ribosomal protein L21
4	Human HMG-17 gene for non-histone chromosomal protein HMG-17
4	Human ribosomal protein L5 mRNA
4	H.sapiens Uba80 mRNA for ubiquitin. 2/97
4	Human interferon-inducible mRNA
3	Homo sapiens mRNA for ribosomal protein L14
3	H.sapiens rpS8 gene for ribosomal protein S8
3	Homo sapiens monocyte/macrophage Ig-related receptor MIR-7 (MIR cl-7) mRNA
3	Homo sapiens U2 snRNP auxiliary factor small subunit
3	Homo sapiens 3-phosphoglycerate dehydrogenase mRNA

No. Copies	Description
3	Novel
3	Homo sapiens aflatoxin aldehyde reductase AFAR mRNA
3	Homo sapiens ribosomal protein L18a mRNA
3	Homo sapiens histone H2A.F/Z variant (H2AV) mRNA
3	Human ribosomal protein L27a mRNA
3	H.sapiens gene for ribosomal protein L38
3	Homo sapiens cDNA /clone=IMAGE:1089890 /gb=AA584384 /gi=2368993 /ug=Hs.100437 /len=434
3	Human ribosomal protein S17 mRNA
3	Human cyclophilin-related processed pseudogene
3	H.sapiens MUC5B gene, rearranged DNA fragment
2	Homo sapiens gene for ribosomal protein L41
2	Homo sapiens glia maturation factor beta mRNA
2	Novel
2	Human ribosomal protein L7a
2	Human ribosomal protein S13 (RPS13) mRNA
2	Homo sapiens cDNA /clone=IMAGE:979448 /gb=AA523303 /gi=2264015 /ug=Hs.15476 /len=640
2	Human profilin mRNA
2	Homo sapiens cDNA, 3' end /clone=1391189 /clone_end=3' /gb=AA781132 /ug=Hs.110803 /len=658
2	Human mRNA for mitochondrial ATP synthase (F1-ATPase) alpha subunit
2	Human mRNA for cytoskeletal gamma-actin
2	Human mRNA for ribosomal protein L32
2	H.sapiens beta-sarcoglycan gene
2	Human mRNA for 26S proteasome subunit p31
2	H.sapiens mRNA for ribosomal protein S15a
2	Novel
2	Novel
2	Homo sapiens IgE receptor beta chain (HTm4) mRNA
2	Human HuR RNA binding protein (HuR) mRNA
2	human alpha-tubulin mRNA
2	H.sapiens mRNA for elongations factor Tu-mitochondrial
2	Homo sapiens cDNA, 3' end /clone=550365 /clone_end=3' /gb=AA098869 /gi=1644973 /ug=Hs.103088 /len=526
2	Homo sapiens cDNA, 3' end /clone=448402 /clone_end=3' /gb=AA777529 /ug=Hs.11355 /len=529
2	Human mRNA for proteasome subunit HsC10-II
2	Homo sapiens RCL (Rcl) mRNA
2	Homo sapiens clone DT1P1A10 mRNA, CAG
2	Novel
2	Human prothymosin-alpha gene
	Total singlets: 213
	Total contigs: 76
	Total seq reads in contigs: 366
	Total seqs to be searched: 717

Table 8

Comp 15: Upregulated Genes

No. Copies	Description
188	Human gene for tumor necrosis factor (TNF-alpha)
61	Cytochrome P450
38	H.sapiens mRNA for uridine phosphorylase
27	HuEST
14	Human tumor necrosis factor-inducible (TSG-6) mRNA fragment
11	Homo Sapiens Chromosone 21 clone
8	Novel
6	Novel
6	Novel
5	SOD-2 Gene
5	Human LD78 beta gene
4	Adenosine receptor A2
3	Human Mitocondrial DNA
3	Human spermidine/spermine N1-acetyltransferase (SSAT) gene
3	H.sapiens mRNA for 23 kD highly basic protein
3	HuEST
3	Human tumor necrosis factor alpha inducible protein A20 mRNA
2	Human spermidine/spermine N1-acetyltransferase (SSAT) gene
2	Human plasma membrane Ca2+ pumping ATPase mRNA
2	Cathepsin L
2	GRO3 oncogene MIP2-beta
2	Small inducible cytokine A4 (homologous to mouse Mip-1b) ACT2
2	GRO2 oncogene MIP2-alpha
2	Human LD78 alpha gene
2	Interleukin 8
	Total singlets: 91
	Total contigs: 25
	Total seq reads in contigs: 404
	Total seqs to be searched: 726

Example 9

Isolation of Rare Genes From Stimulated THP-1 Cells

(Experiment: Cot 3)

In this example, rare genes are isolated from stimulated THP-1 cells by collecting beads of lower relative intensity. Bead and probe libraries were constructed from mRNA prepared from phorbol ester treated THP-1 cultured cells. Six bead libraries (160K complexity) were loaded twice to BP 11 combitagged beads. A total of 1,260,000 beads were sorted. The beads were filled in and ligated. The top

strand of the beads was stripped with 2.5 ml 150 mM NaOH washes at room temperature for 15 minutes with mild vortexing. The beads were washed twice in 0.5 ml of 4X SSC/0.1% SDS. 100,000 beads were hybridized overnight with 50 ng of CY5 labelled probe from stimulated THP-1 cells in 4X SSC/0.1% SDS at 65°. The recovered samples were rinsed 2 times with 1X SSC/0.1% SDS, resuspended in 0.5 ml of 1X SSC/0.1% SDS, and washed at 65°C for 15 minutes. The beads were then rinsed in 0.1X SSC/0.1% SDS and washed at 55°C in 0.1X SSC/0.1% SDS for 15 minutes. 98,880 clones were analyzed and sorted by flow cytometry. Sample CT003E contained 126 clones which barely hybridized any CY5 probe. Sample CT003F contained 1557 clones that did not find enough probe to migrate to the diagonal. These beads contained the least frequent copies in our probe library. 50 clones from each gate (see Figure 7) were picked for sequence analysis. The identified sequences are listed in Table 9.

Table 9

THP-1 Rare Genes

No. Copies	Descripton	GenBank Identifier
CT003E		
2	Alu primary transcript	U67828
1	AMP deaminase	HSAMPD3B
1	BBC1	HSBBC1
14	CD44	HUMCD44B
1	clone 23933 mRNA	HSU79273
7	EST	AA905212
1	EST	AA975736
1	EST	N53143
1	EST	AA808221
1	EST	AA826047
1	EST	AA736779
1	EST	AA994497
1	EST	AI049999
1	EST (88% homology)	AA626040
8	EST (contains Alu repeat)	AA129219
9	EST (contains Alu repeat)	AI085719
1	EST (contains Alu repeat)	W07654
1	EST (Sau3A not present)	AA553627
3	ferritin H chain	HUMFERH
1	IL1 β	HUMIL1BA

No. Copies	Descripton	GenBank Identifier
1	KIAA0098 gene mRNA	HUMKG1DD
1	mito. cyt oxidase subunit I pseudogene	AF035429
1	Mitochondrial genome	MIHSGENOM
1	NADH:ubiquinone oxidoreductase NDUF56 subunit	AF044959
2	no match	
1	no match	
1	only 12 bases	
1	only 12 bases	
1	only 13 bases	
1	Pyruvate kinase, M gene for M1-type & M2-type	HSPKM12
4	TNF α	HSTNFR
7	TNF type I recept. assoc. prot./DNAse I/HSP75	HSU12595/D831 95/AF043254
1	type IV collagenase	HUM4COLA
1	Ubiquitin hydrolyzing enzyme I (UBHI)	AF022789
1	VASP gene	HSVASP413
CT003F		
1	Apolipoprotein C-II	HSAPOC2G
3	BBC1	HSBBC1
1	clone s153 mRNA fragment	HUMFRCC
5	cytoskeletαλ γ actin	HSACTCGR
1	elongation factor 1 α	HSEF1AC
1	EST	AA905212
2	EST	AA977353
1	EST	AA135810
1	EST (contains Alu repeat)	H08741
1	EST	AA282788
1	EST	AA226660
1	EST (85% homology; contains Alu, CACA tract)	AA704393
1	EST (86% homology; contains Alu)	H60533
1	EST (88% homology)	AA228701
2	EST (contains Alu repeat)	AI085719
1	EST (contains Alu repeat)	AA713891
1	EST (rat)	AI136745
1	ferritin H chain	HUMFERH
1	genomic (72 bp; 88% homology)	HSAC002082
1	ICAM-1	HUMICAMA1M
1	IL1 β	HUMIL1BA
1	Interferon γ receptor accessory factor 1	HSU05877
2	mito. cyt oxidase subunit I pseudogene	AF035429
1	no match	
1	no match	

No. Copies	Descripton	GenBank Identifier
1	no match (40 bp)	
1	OTK27	D50420
1	p65, rat (partial match to hu synaptotagmin I)	RRP65
1	rat α tubulin (100% hom to rat; 78% to human)	RNATUBZ
1	Ribophorin II	HSRIBIIR
3	Ribosomal protein S3	HSU14991
1	RING4	HSRING4
12	sec61-complex β subunit	HUMSEC61B
1	TNF α	HSTNFR
1	type IV collagenase	HUM4COLA

Example 10

Isolation of Rare Genes From Human Bone Marrow

Bead and probe libraries were constructed from commercially available mRNA from bone marrow. Six bead libraries (160K complexity) were loaded twice to BP 12 combitagged beads. They formed mixes 216, 217, 218, and 219. A total of 3,150,000 beads were sorted. The beads were filled in and ligated. The top strand of mix 217 was stripped off with NaOH. The CT1 bone marrow probe was linearly amplified with CY5 nucleotides and then purified. 200,000 beads were hybridized with 5 and 50 ng of probe overnight at 65°. 180,000 clones from the 5 nG hybridization were interrogated and sorted. Sample CT001 contained 996 clones which barely hybridized any CY5 probe. CT002 sample contained 1988 clones that did not find enough probe to migrate to the diagonal. These beads contained the least frequent copies in our probe library. 200 clones from each gate (see Figure 8) were picked for sequence analysis.

Example 11

FACS Analysis of Differentially Expressed Genes from Normal and Glucose Starved Human Muscle Tissue

Bead and probe libraries were constructed from mRNA prepared from muscle tissue in two states: glucose normal (basal) and glucose starved (clamp). Six bead libraries (160K complexity) from the glucose normal state were loaded to BP 12 combitagged beads to form mix 237. A total of 810,000 beads were sorted. The beads were filled in and ligated. The beads were digested with DpnII enzyme and ligated to an adapter with FITC on the strand opposite to the covalently attached DNA

strand. The top strand of mix 217 was stripped off with NaOH. The CT1 glucose normal probe (13,510,000 complexity) was linearly amplified with CY5 nucleotides and then purified. The CT2 glucose starred probe (7,132,000 complexity) was linearly amplified with R110 nucleotides and then purified. 250,000 beads were hybridized with 5ug of each probe overnight at 65°. 230,000 clones were interrogated and sorted. Sample UP001 contained 968 clones which were upregulated. Sample DN001 contained 1652 clones which were down regulated. 1000 clones from each gate (see Figure 9) were picked for sequence analysis. The identified sequences are listed in Tables 10 and 11.

Table 10

Downregulated Genes in Starved Human Muscle

No. Copies	Description
23	Human mRNA for slow skeletal troponin C
27	Human alkali myosin light chain 1 Mrna
14	Human messenger RNA for beta-globin
13	Human lymphocytic antigen CD59/MEM43 mRNA
10	P6=cytochrome c oxidase subunit Vlc homolog
8	H.sapiens mRNA homologous to mouse P21 mRNA
5	Human SPARC/osteonectin mRNA
5	3', mRNA sequence
4	Pan troglodytes beta-2-microglobulin mRNA
4	reductase
4	Homo sapiens gene for ribosomal protein L41
3	Homo sapiens ribosomal protein L30 mRNA
3	IMAGE: 1388067
2	ni65c01.s1 NCI_CGAP_Pr12 Homo sapiens cDNA clone IMAGE:981696
2	Homo sapiens mRNA for KIAA0454 protein
2	Human gene for cardiac beta myosin heavy chain

Table 11

Upregulated Genes in Starved Human Muscle

No. Copies	Description
4	Human mitochondrion cytochrome b gene
4	Homo sapiens sarcosin mRNA
4	laminin receptor homolog
3	H.sapiens mRNA for 23 kD highly basic protein
3	Human ENO3 mRNA for beta-enolase
3	alpha-tropomyosin
3	alpha B-crystallin
3	Human mRNA for muscle phosphofructokinase
2	Baboon beta-myosin heavy-chain mRNA
2	Human mRNA 3'-fragment for glycogen phosphorylase
2	Human ribosomal L5 protein mRNA
2	H.sapiens mRNA for ribosomal protein L37a
2	Human cytochrome c oxidase subunit VII (COX8) mRNA

All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

The invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the appended claims.



We claim:

1. A method of analyzing differential gene expression, comprising:

providing a reference population of nucleic acid sequences attached to separate solid phase supports in clonal subpopulations;

providing a population of polynucleotides of expressed genes from a first cell or tissue source and at least one population of polynucleotides of expressed genes from a different cell or tissue source, the polynucleotides of expressed genes from each source comprising a light-generating label different from the label comprised by polynucleotides of any other source;

competitively hybridizing the populations of polynucleotides of expressed genes from each source with the reference nucleic acid population to form duplexes between the nucleic acid sequences of the reference nucleic acid population and the polynucleotides of each source such that the polynucleotides are present in duplexes on each of the solid phase supports in ratios directly related to the relative expression of their corresponding genes in the sources; and

detecting a relative optical signal generated by the light-generating labels of the duplexes attached thereto.

2. The method of Claim 1, wherein said nucleic acid sequences are DNA sequences.

3. The method of Claim 2, wherein said step of providing said reference population further includes:

forming at least one population of tag-cDNA conjugates from mRNA extracted from at least one of said sources and a repertoire of oligonucleotide tag;

removing a sample of the tag-cDNA conjugates; and

amplifying the tag-cDNA conjugates of the sample.

4. The method of Claim 3, wherein said populations of tag-cDNA conjugates are formed from mRNA extracted from each of said sources, the method further comprising combining said populations of tag-cDNA conjugates from each of said sources prior to removing said sample.



5. The method of Claim 4, wherein said sample is sufficiently small relative to said total tag-cDNA conjugates that substantially all different cDNAs have different oligonucleotide tags.

6. The method of Claim 5, wherein said step of providing said reference population further includes attaching said tag-cDNA conjugates of said sample to said separate solid phase supports by specifically hybridizing said oligonucleotide tags of said tag-cDNA conjugates to their respective complements.

7. The method of Claim 6, wherein said step of amplifying comprises replicating said tag-cDNA conjugates of said sample in a polymerase chain reaction.

8. The method of Claim 6, wherein said step of amplifying comprises replicating said tag-cDNA conjugates of said sample by inserting said tag-cDNA conjugates into a cloning vector and transfecting a host cell therewith.

9. The method of Claim 6, wherein said sample includes a number of oligonucleotide tags less than or equal to one percent of said oligonucleotide tags in said repertoire.

10. The method of Claim 2, wherein said reference DNA population is derived from said expressed genes of all of said sources being analyzed.

11. The method of Claim 2, further comprising sorting each solid phase support according to said relative optical signal.

12. The method of Claim 2, wherein said different light-generating labels are different fluorescent labels.

13. The method of Claim 12, wherein said population of polynucleotides of expressed genes are populations of cDNAs.

14. The method of Claim 13, further comprising the steps of:

accumulating each said solid phase support having said relative optical signal with a value within one or more predetermined ranges of values corresponding to a difference in gene expression among said sources; and

identifying said polynucleotides on each of said solid supports by determining a nucleotide sequence of a portion of each of said polynucleotides.

15. The method of Claim 14, wherein said relative optical signal is a ratio of fluorescence intensities and wherein said populations of polynucleotides are from two sources.

16. The method of Claim 15, wherein said portion of said polynucleotides is a sequence of at least ten nucleotides.

17. The method of Claim 15, wherein said step of identifying includes simultaneous sequencing of at least ten thousand of said polynucleotides by massively parallel signature sequencing.

18. A method of isolating polynucleotides derived from genes differentially expressed in a plurality of different cells or tissues, the method comprising the steps of:

providing a reference DNA population of DNA sequences attached to separate microparticles in clonal subpopulations;

providing a population of polynucleotides derived from genes expressed in each of the plurality of different cells or tissues, each polynucleotide having a light-generating label capable of generating an optical signal indicative of the cells or tissues from which it is derived;

competitively hybridizing the populations of polynucleotides of genes expressed in each of the plurality of different cells or tissues with the reference DNA population to form duplexes between the DNA sequences of the reference DNA population and polynucleotides from each of the different cells or tissues such that the

polynucleotides are present in duplexes on each of the microparticles in ratios directly related to the relative expression of their corresponding genes in the different cells or tissues; and

isolating polynucleotides corresponding to genes differentially expressed in the different cells or tissues by sorting microparticles in accordance with the optical signals generated by the populations of polynucleotides hybridized thereto.

19. The method of Claim 18, wherein said reference DNA population is derived from genes expressed in the plurality of different cells or tissues being analyzed.

20. The method of Claim 19, wherein said plurality of different cells or tissues is two and wherein said optical signal is a fluorescent signal.

21. The method of Claim 20, wherein said populations of polynucleotides are labeled with different fluorescent labels.

22. The method of Claim 21, wherein said populations of polynucleotides are populations of cDNAs.

23. The method of Claim 22, wherein said step of competitively hybridizing includes providing hybridization conditions which result in substantially all of said duplexes being perfectly matched duplexes.

24. The method of Claim 23, wherein said step of isolating includes sorting said microparticles in accordance with the ratio of fluorescence intensities generated by said populations of cDNAs hybridized thereto.

25. The method of Claim 24, wherein said step of isolating includes sorting said microparticle with a fluorescence-activated cell sorter.

26. The method of Claim 25, further including the step of identifying said isolated cDNAs by determining a nucleotide sequence of a portion of each said isolated cDNA.

27. A method of determining relative abundance of gene products, comprising:

providing a reference DNA population of DNA sequences attached to separate solid phase supports in clonal subpopulations;

providing a population of polynucleotides derived from genes expressed in at least one cell or tissue source, the polynucleotides having a light-generating label;

hybridizing the polynucleotides with the reference DNA population to form duplexes between the DNA sequences of the reference DNA population and the polynucleotides; and

sorting each solid phase support according to the optical signal generated by the light-generating labels of the duplexes attached thereto,

wherein relative abundance of the gene products is correlated with the relative level of intensity of the optical signals obtained from the duplexes, wherein a lower intensity is indicative of a rarer gene product.

28. The method of Claim 27, further comprising isolating solid phase supports having lower relative intensities, wherein said isolated solid phase supports comprise at most about 5% of the total solid phase supports provided.

29. The method of Claim 28, wherein said isolated solid phase supports comprise at most about 0.5% of the total supports provided.

30. A method of isolating polynucleotides according to the abundance of the nucleic acid sequences from which they are derived, comprising:

providing a reference DNA population of DNA sequences attached to separate microparticles in clonal subpopulations;

providing a population of polynucleotides derived from nucleic acid sequences present in the cells of at least one cell or tissue source, each polynucleotide having a light-generating label capable of generating an optical signal;

competitively hybridizing the population of polynucleotides with the reference DNA population to form duplexes between the DNA sequences of the reference DNA population and the polynucleotides, the hybridizing being conducted under conditions which provide a hybridization rate proportionate to the abundance of the polynucleotide wherein less abundant polynucleotides would remain unhybridized;

sorting the polynucleotides into a hybridized population and an unhybridized population.

31. The method of Claim 30, wherein said polynucleotides are hybridized with said reference DNA population under conditions such that said unhybridized population comprises polynucleotides derived from rare gene products.

32. The method of Claim 30, wherein said polynucleotides are hybridized with said reference DNA population under conditions such that said unhybridized population is substantially enriched in polynucleotides derived from nonrepetitive nucleic acid sequences.

33. A composition comprising a mixture of microparticles, each microparticle having a population of identical single stranded nucleic acid molecules attached thereto, the single stranded nucleic acid molecules being different on each microparticle and comprising an oligonucleotide tag in juxtaposition with a polynucleotide derived from an mRNA of at least one cell or tissue source.

34. The composition of Claim 33, wherein said nucleic acid molecules are DNA.

35. The composition of Claim 34, wherein said polynucleotides are derived from a plurality of cell or tissue sources.

36. The composition of Claim 35, wherein said mixture comprises at least 100 different microparticles.

37. The composition of Claim 35, wherein said mixture comprises at least 1000 different microparticles.

38. The composition of Claim 35, wherein said mixture comprises at least 10^4 different microparticles.

39. The composition of Claim 35, wherein said oligonucleotide tag is about 12 to about 60 nucleotides in length.

40. The composition of Claim 35, wherein said oligonucleotide tag is about 18 to about 40 nucleotides in length.

41. The composition of Claim 35, wherein said oligonucleotide tag is about 25 to about 40 nucleotides in length.

42. A composition comprising a mixture of microparticles, each microparticle having a population of identical single stranded nucleic acid molecules attached thereto, the single stranded nucleic acid molecules being different on each microparticle and each of the different nucleic acid molecules comprising a polynucleotide encoding a protein selected from the group consisting of cell cycle proteins, signal transduction pathway proteins, oncogene gene products, tumor suppressors, kinases, phosphatases, transcription factors, growth factor receptors, growth factors, extracellular matrix proteins, proteases, cytoskeletal proteins, membrane receptors, Rb pathway proteins, p53 pathway proteins, proteins involved in metabolism, proteins involved in cellular responses to stress, cytokines, proteins involved in DNA damage and repair, and proteins involved in apoptosis.

43. The composition of Claim 42, wherein each of said nucleic acid molecules further comprises an oligonucleotide tag in juxtaposition with said polynucleotide and positioned between said microparticle and said polynucleotide.

44. The composition of claim 43, wherein each of said microparticles comprises a set of oligonucleotide tags having a sequence different from the oligonucleotide tags of any other microparticle in said composition.

45. The composition of Claim 42, wherein said polynucleotides encode kinases.

46. The composition of Claim 42, wherein said polynucleotides encode cell-cycle proteins.

47. The composition of Claim 42, wherein said polynucleotides encode signal transduction pathway proteins.

48. The composition of Claim 42, wherein said polynucleotides encode proteins involved in apoptosis.

49. The composition of Claim 42, wherein said polynucleotides encode proteins involved in metabolism.

50. A kit for preparing a reference population, comprising:
a plurality of microparticles having oligonucleotide tag complements attached thereto, the oligonucleotide tag complement sequence being different on each microparticle.

51. The kit of Claim 50, further comprising a plurality of vectors comprising a library of tags, the tags having sequences complementary to said tag complements.

52. The kit of Claim 51, further comprising a population of polynucleotides from at least one cell or tissue source.

53. The kit of Claim 52, wherein said polynucleotides are cDNAs.

54. The kit of Claim 52, wherein said population of polynucleotides is contained in a container separate from said plurality of microparticles.

55. The kit of Claim 51, further comprising at least one reagent for preparing said reference population.

56. A kit for analyzing differentially expressed genes, comprising:

a mixture of microparticles, each microparticle having a population of identical single stranded nucleic acid molecules attached thereto, the single stranded nucleic acid molecules being different on each microparticle and comprising polynucleotide derived from an mRNA of at least one cell or tissue source.

57. The kit of Claim 56, wherein each of said nucleic acid molecules further comprises an oligonucleotide tag in juxtaposition with said polynucleotide and positioned between said microparticle and said polynucleotide.

58. The kit of Claim 56, further comprising printed instructions for use in analyzing differentially expressed genes.

59. The kit of Claim 56, further comprising a container.

60. The kit of Claim 56, further comprising a population of cDNA molecules from at least one of said cell or tissue sources.