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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 August 2008 (21.08.2008)

PCT

(10) International Publication Number
WO 2008/100352 A2

(51) International Patent Classification:
G06F 19/00 (2006.01)

(21) International Application Number:
PCT/US2007/083555

(22) International Filing Date:
3 November 2007 (03.11.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/856,406 3 November 2006 (03.11.2006) US

(71) Applicant *(for all designated States except US)*: **BAYLOR RESEARCH INSTITUTE** [US/US]; 3434 Live Oak Street, Suite 125, Dallas, TX 75204 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **PALUCKA, Anna, Karolina** [PL/US]; 3000 Blackburn Street #2522, Dallas, TX 75204 (US). **BANCHEREAU, Jacques, F.** [FR/US]; 6730 Northaven, Dallas, TX 75230 (US). **CHAUSSABEL, Damien** [FR/US]; 4532 Southpointe Drive, Richardson, TX 75082 (US).

(74) Agents: **FLORES, Edwin** et al.; CHALKER FLORES, LLP, 2711 LBJ Freeway, Suite 1036, Dallas, TX 75234 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: DIAGNOSIS OF METASTATIC MELANOMA AND MONITORING INDICATORS OF IMMUNOSUPPRESSION THROUGH BLOOD LEUKOCYTE MICROARRAY ANALYSIS

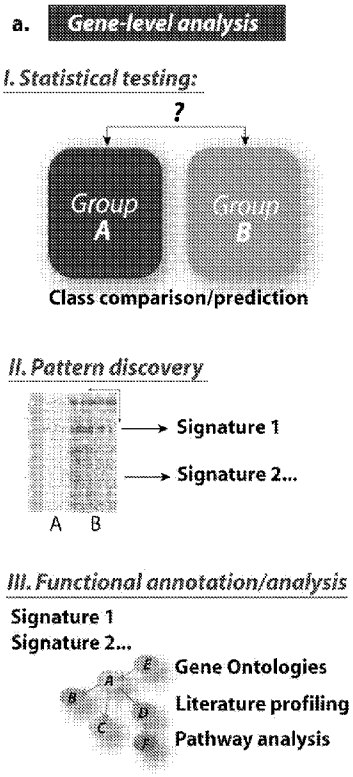


Figure 1a

(57) Abstract: The present invention includes compositions, systems and methods for the early detection and consistent determination of metastatic melanoma and/or immunosuppression using microarrays by calculating one or more expression vectors from the expression of one or more genes.

WO 2008/100352 A2

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

DIAGNOSIS OF METASTATIC MELANOMA AND MONITORING INDICATORS OF IMMUNOSUPPRESSION THROUGH BLOOD LEUKOCYTE MICROARRAY ANALYSIS

TECHNICAL FIELD OF THE INVENTION

The present invention relates in general to the field of diagnostic for monitoring indicators of metastatic melanoma and/or immunosuppression, and more particularly, to a system, method and apparatus for the diagnosis, prognosis and tracking of metastatic melanoma and monitoring indicators of immunosuppression associated with transplant recipients (e.g., liver).

LENGTHY TABLE

The patent application contains a lengthy table section. A copy of the table is available in electronic form from the USPTO web site (http://seqdata.uspto.gov/). An electronic copy of the table will also be available from the USPTO upon request and payment of the fee set forth in 37 CFR 1.19(b)(3).

BACKGROUND OF THE INVENTION

This application claims priority to United States Provisional Patent Application Serial No. 60/748,884, filed December 9, 2005, the entire contents of which are incorporated herein by reference. Without limiting the scope of the invention, its background is described in connection with diagnostic methods.

Pharmacological immunosuppression has been instrumental in transplantation, transforming a last resort experimental procedure into a routinely successful one. Cyclosporin and tacrolimus/FK506 currently constitute the mainstay for transplant recipients. Activated T cells have been considered the main cellular targets for immunosuppressive treatments, but recent reports have also noted a marked effect of these drugs on antigen presenting cells, probably contributing further to the establishment of a generalized state of immune unresponsiveness (Lee et al., 2005b; Woltman et al., 2003). While severe immunosuppression generated by pharmacological treatment is necessary for the maintenance of graft survival and function, it also exposes the recipient to life-threatening infections and malignancies. Skin cancer is a well-established complication in organ transplant recipients (Gerlini et al., 2005). A recent epidemiological study showed that nearly 20% of 1115 renal transplant patients developed skin malignancies in Europe (Bordea et al., 2004), with much higher incidences being observed in less temperate climates, e.g., as high as 28% in Australia (Carroll et al., 2003). In addition to occurring more often, skin malignancies tend to take a more aggressive clinical course in transplanted patients, with a higher propensity to metastasize distantly and lead to a fatal outcome (Barrett et al., 1993).

Tumors maintain their survival by compromising the immune system. Different mechanisms have been identified including secretion of immunosuppressive factors such as cytokines (e.g., IL-10, TGF-beta), hormones (e.g., Prostaglandin E2), and others (e.g., MIA: Melanoma inhibitory activity, Tenascin C)(Jachimczak et al., 2005; Puente Navazo et al., 2001). Furthermore, tumors might promote

development of suppressor T cells (Liyanage et al., 2002; Viguier et al., 2004), possibly via modulating dendritic cells (Gabrilovich, 2004; Lee et al., 2005a; Monti et al., 2004).

Thus, immunosuppression, whether from tumors or pharmacological treatments, has been linked to cancer progression. Therefore, a common molecular signature could be identified by profiling genome-wide transcriptional activity in peripheral blood mononuclear cell ("PBMC") samples obtained from immunosuppressed patients with either metastatic melanoma or liver allografts. The analysis of Leukocyte transcriptional profiles generated in the context of the present study supports this notion and identifies a blood signature of immunosuppression.

SUMMARY OF THE INVENTION

Genomic research is facing significant challenges with the analysis of transcriptional data that are notoriously noisy, difficult to interpret and do not compare well across laboratories and platforms. The present inventors have developed an analytical strategy emphasizing the selection of biologically relevant genes at an early stage of the analysis, which are consolidated into analytical modules that overcome the inconsistencies among microarray platforms. The transcriptional modules developed may be used for the analysis of large gene expression datasets. The results derived from this analysis are easily interpretable and particularly robust, as demonstrated by the high degree of reproducibility observed across commercial microarray platforms.

Applications for this analytical process are illustrated through the mining of a large set of PBMC transcriptional profiles. Twenty-eight transcriptional modules regrouping 4,742 genes were identified. Using the present invention is it possible to demonstrate that diseases are uniquely characterized by combinations of transcriptional changes in, e.g., blood leukocytes, measured at the modular level. Indeed, module-level changes in blood leukocytes transcriptional levels constitute the molecular fingerprint of a disease or sample.

This invention has a broad range of applications. It can be used to characterize modular transcriptional components of any biological system (e.g., peripheral blood mononuclear cells (PBMCs), blood cells, fecal cells, peritoneal cells, solid organ biopsies, resected tumors, primary cells, cells lines, cell clones, etc.). Modular PBMC transcriptional data generated through this approach can be used for molecular diagnostic, prognostic, assessment of disease severity, response to drug treatment, drug toxicity, etc. Other data processed using this approach can be employed for instance in mechanistic studies, or screening of drug compounds. In fact, the data analysis strategy and mining algorithm can be implemented in generic gene expression data analysis software and may even be used to discover, develop and test new, disease- or condition-specific modules. The present invention may also be used in conjunction with pharmacogenomics, molecular diagnostic, bioinformatics and the like, wherein in-depth expression data may be used to improve the results (e.g., by improving or sub-selecting from within the sample population) that mat be obtained during clinical trails.

More particularly, the present invention includes arrays, apparatuses, systems and method for diagnosing a disease or condition by obtaining the transcriptome of a patient; analyzing the transcriptome based on one or more transcriptional modules that are indicative of a disease or condition; and determining the patient's disease or condition based on the presence, absence or level of expression of genes within the transcriptome in the one or more transcriptional modules. The transcriptional modules may be obtained by: iteratively selecting gene expression values for one or more transcriptional modules by: selecting for the module the genes from each cluster that match in every disease or condition; removing the selected genes from the analysis; and repeating the process of gene expression value selection for genes that cluster in a sub-fraction of the diseases or conditions; and iteratively repeating the generation of modules for each clusters until all gene clusters are exhausted.

Examples of clusters selected for use with the present invention include, but are not limited to, expression value clusters, keyword clusters, metabolic clusters, disease clusters, infection clusters, transplantation clusters, signaling clusters, transcriptional clusters, replication clusters, cell-cycle clusters, siRNA clusters, miRNA clusters, mitochondrial clusters, T cell clusters, B cell clusters, cytokine clusters, lymphokine clusters, heat shock clusters and combinations thereof. Examples of diseases or conditions for analysis using the present invention include, e.g., autoimmune disease, a viral infection a bacterial infection, cancer and transplant rejection. More particularly, diseases for analysis may be selected from one or more of the following conditions: systemic juvenile idiopathic arthritis, systemic lupus erythematosus, type I diabetes, liver transplant recipients, melanoma patients, and patients bacterial infections such as *Escherichia coli, Staphylococcus aureus*, viral infections such as influenza A, and combinations thereof. Specific array may even be made that detect specific diseases or conditions associated with a bioterror agent.

Cells that may be analyzed using the present invention, include, e.g., peripheral blood mononuclear cells (PBMCs), blood cells, fetal cells, peritoneal cells, solid organ biopsies, resected tumors, primary cells, cells lines, cell clones and combinations thereof. The cells may be single cells, a collection of cells, tissue, cell culture, cells in bodily fluid, e.g., blood. Cells may be obtained from a tissue biopsy, one or more sorted cell populations, cell culture, cell clones, transformed cells, biopies or a single cell. The types of cells may be, e.g., brain, liver, heart, kidney, lung, spleen, retina, bone, neural, lymph node, endocrine gland, reproductive organ, blood, nerve, vascular tissue, and olfactory epithelium cells. After cells are isolated, these mRNA from these cells is obtained and individual gene expression level analysis is performed using, e.g., a probe array, PCR, quantitative PCR, bead-based assays and combinations thereof. The individual gene expression level analysis may even be performed using hybridization of nucleic acids on a solid support using cDNA made from mRNA collected from the cells as a template for reverse transcriptase.

Pharmacological immunosuppression promotes graft survival in transplant recipients. Endogenous immunosuppression promotes tumor survival in cancer-bearing patients. Leukocytes from patients with metastatic melanoma display an endogenous immunosuppression signature common with liver transplant

recipients under pharmacological immunosuppression. Blood microarray analyses were carried out in 25 healthy volunteers, 35 patients with metastatic melanoma, and 39 liver transplant recipients. Disease signatures were identified, and confirmed in an independent dataset, in comparison to healthy controls. Analysis of a set of 69 transcripts over-expressed preferentially in melanoma and transplant groups in comparison to six other diseases revealed remarkable functional convergence, including several repressors of interleukin-2 transcription, powerful inhibitors of NF-kappaB and MAPK pathways as well as antiproliferative molecules. Thus, patients with metastatic melanoma display an endogenous transcriptional signature of immunosuppression. This signature may now be used to identify patients at the high risk of metastatic melanoma progression.

The present invention includes a system and a method to analyze samples for the prognosis and diagnosis of metastatic melanoma and/or monitoring indicators of immunosuppression associated with transplant recipients (e.g., liver) using multiple variable gene expression analysis. The gene expression differences that remain can be attributed with a high degree of confidence to the unmatched variation. The gene expression differences thus identified can be used, for example, to diagnose disease, identify physiological state, design drugs and monitor therapies.

The sample may be screened by quantitating the mRNA, protein or both mRNA and protein level of the expression vector. When the screening is for mRNA levels, it may be quantitated by a method selected from polymerase chain reaction, real time polymerase chain reaction, reverse transcriptase polymerase chain reaction, hybridization, probe hybridization, and gene expression array. The screening method may also include detection of polymorphisms in the biomarker. Alternatively, the screening step may be accomplished using at least one technique selected from the group consisting of polymerase chain reaction, heteroduplex analysis, single stand conformational polymorphism analysis, ligase chain reaction, comparative genome hybridization, Southern blotting, Northern blotting, Western blotting, enzyme-linked immunosorbent assay, fluorescent resonance energy-transfer and sequencing. For use with the present invention the sample may be any of a number of immune cells, e.g., leukocytes or sub-components thereof.

In one embodiment, the present invention includes a method of identifying a patient with melanoma by examining the phenotype a sample for combinations of six, seven, eight, ten, fifteen, twenty, twenty-five or more genes selected from Tables 2, 8, 9, 12 and combinations thereof.

In one embodiment, the present invention includes a method of identifying gene expression response to pharmacological immunosuppression in transplant recipients by examining the phenotype a sample for combinations of six, seven, eight, ten, fifteen, twenty, twenty-five or more genes selected from Tables 10, 11, 13 and combinations thereof.

The sample may be screened by quantitating the mRNA, protein or both mRNA and protein level of the expression vector. When mRNA level is examined, it may be quantitated by a method selected from polymerase chain reaction, real time polymerase chain reaction, reverse transcriptase polymerase chain

reaction, hybridization, probe hybridization, and gene expression array. The screening method may also include detection of polymorphisms in the biomarker. Alternatively, the screening step may be accomplished using at least one technique selected from the group consisting of polymerase chain reaction, heteroduplex analysis, single stand conformational polymorphism analysis, ligase chain reaction, comparative genome hybridization, Southern blotting, Northern blotting, Western blotting, enzyme-linked immunosorbent assay, fluorescent resonance energy-transfer and sequencing. For use with the present invention the sample may be any of a number of immune cells, e.g., leukocytes or sub-components thereof.

The expression vector may be screened by quantitating the mRNA, protein or both mRNA and protein level of the expression vector. When the expression vector is mRNA level, it may be quantitated by a method selected from polymerase chain reaction, real time polymerase chain reaction, reverse transcriptase polymerase chain reaction, hybridization, probe hybridization, and gene expression array. The screening method may also include detection of polymorphisms in the biomarker. Alternatively, the screening step may be accomplished using at least one technique selected from the group consisting of polymerase chain reaction, heteroduplex analysis, single stand conformational polymorphism analysis, ligase chain reaction, comparative genome hybridization, Southern blotting, Northern blotting, Western blotting, enzyme-linked immunosorbent assay, fluorescent resonance energy-transfer and sequencing. For use with the present invention the sample may be any of a number of immune cells, e.g., leukocytes or sub-components thereof.

For example, a method of identifying a subject with melanoma by determining a database that includes the level of expression of one or more metastatic melanoma expression vectors. Another embodiment of the present invention includes a computer implemented method for determining the genotype of a sample by obtaining a plurality of sample probe intensities. Diagnosing metastatic melanoma is based upon the sample probe intensities and calculating a linear correlation coefficient between the sample probe intensities and reference probe intensities.

The present invention also includes a computer readable medium including computer-executable instructions for performing the method of determining the genotype of a sample. The method of determining the phenotype includes obtaining a plurality of sample probe intensities and diagnosing melanoma based upon the sample probe intensities for two or more metastatic melanoma expression vectors selected from those listed in Tables 2, 3, 8, 9, 12 and combinations thereof into a dataset; and calculating a linear correlation coefficient between the sample probe intensities and a reference probe intensity. The tentative phenotype is accepted as the phenotype of the sample if the linear correlation coefficient is greater than a threshold value. In addition, the present invention includes a microarray for identifying a human subject with melanoma. The microarray includes the detection of expression of two or more metastatic melanoma genes listed in Tables 2, 8, 9, 12 and combinations thereof into a dataset. The present invention provides a method of distinguishing between metastatic melanoma and immunosuppression associated with transplants by determining the level of expression of one or more

genes, and calculating one or more gene expression vectors. The melanoma-specific transcriptome-expression vectors may include values for the upregulation or downregulation of six or more genes listed in Tables 2, 3, 8, 9, 12 and combinations thereof. The present invention provides a method of identifying a subject with immunosuppression associated with transplants by determining the level of expression of one or more immunosuppression associated expression vectors. The immunosuppression-specific transcriptome-expression vectors may include values for the upregulation or downregulation of six or more genes listed in Tables 10, 11 and 13 and combinations thereof.

The present invention also includes a computer implemented method for determining the propensity for immunosuppression in a sample including obtaining a plurality of sample probe intensities and diagnosing immunosuppression based upon the sample probe intensities. A linear correlation coefficient is calculating between the plurality of sample probe intensities and a reference probe intensity. A tentative genotype is then accepted as the genotype of the sample if the linear correlation coefficient is greater than a threshold value. The melanoma-specific transcriptome-expression vectors may include values for the upregulation or downregulation of six or more genes listed in Tables 2, 8, 9, 12 and combinations thereof. The immunosuppression-specific transcriptome-expression vectors may include values for the upregulation or downregulation of six or more genes listed in Tables 10, 11 and 13 and combinations thereof.

A computer readable medium is also included that has computer-executable instructions for performing the method for determining the phenotype of a sample. The method for determining the phenotype of a sample includes obtaining a plurality of sample probe intensities and diagnosing immunosuppression based upon the sample probe intensities for two or more immunosuppression associated expression vectors. A linear correlation coefficient is calculated between the sample probe intensities and a reference probe intensity and a tentative phenotype is accepted as the phenotype of the sample if the linear correlation coefficient is greater than a threshold value. The present invention also includes a system for diagnosing immunosuppression including an expression level detector for determining the expression level of two or more immunosuppression expression vectors selected from the 1, 2, 3, 4, 5, 6, 8, 10, 15, 20, 25 or more genes. The melanoma-specific transcriptome-expression vectors and the that are used to generate expression data for each gene, which is saved into a dataset, may include values for the upregulation or downregulation of six or more genes listed in Tables 2, 8, 9, 12 and combinations thereof. The immunosuppression-specific transcriptome-expression vectors may include values in a dataset that includes the upregulation or downregulation of six or more genes listed in Tables 10, 11 and 13 and combinations thereof.

The arrays, methods and systems of the present invention may even be used to select patients for a clinical trial by obtaining the transcriptome of a prospective patient; comparing the transcriptome to one or more transcriptional modules that are indicative of a disease or condition that is to be treated in the clinical trial; and determining the likelihood that a patient is a good candidate for the clinical trial based on the presence, absence or level of one or more genes that are expressed in the patient's transcriptome

within one or more transcriptional modules that are correlated with success in a clinical trial. Generally, for each module a vector that correlates with a sum of the proportion of transcripts in a sample may be used, e.g., when each module includes a vector and wherein one or more diseases or conditions is associated with the one or more vectors. Therefore, each module may include a vector that correlates to the expression level of one or more genes within each module.

The present invention also includes arrays, e.g., custom microarrays that include nucleic acid probes immobilized on a solid support that includes sufficient probes from one or more expression vectors to provide a sufficient proportion of differentially expressed genes to distinguish between one or more diseases. For example, an array of nucleic acid probes immobilized on a solid support, in which the array includes at least two sets of probe modules, wherein the probes in the first probe set have one or more interrogation positions respectively corresponding to one or more diseases. The array may have between 100 and 100,000 probes, and each probe may be, e.g., 9-21 nucleotides long. When separated into organized prose sets, these may be interrogated separately.

The present invention also includes one or more nucleic acid probes immobilized on a solid support to form a module array that includes at least one pair of first and second probe groups, each group having one or more probes as defined by Table 1. The probe groups are selected to provide a composite transcriptional marker vector that is consistent across microarray platforms. In fact, the probe groups may even be used to provide a composite transcriptional marker vector that is consistent across microarray platforms and displayed in a summary for regulatory approval. The skilled artisan will appreciate that using the modules of the present invention it is possible to rapidly develop one or more disease specific arrays that may be used to rapidly diagnose or distinguish between different disease and/or conditions.

The present invention also include a method for displaying transcriptome vector data by separating one or more genes into one or more modules to visually display an aggregate gene expression vector value for each of the modules; and displaying the aggregate gene expression vector value for overexpression, underexpression or equal expression of the aggregate gene expression vector value in each module. In one example, overexpression is identified with a first identifier and underexpression is identified with a second identifier. Examples of identifiers include colors, shapes, patterns, light/dark, on/off, symbols and combinations thereof. For example, overexpression is identified with a first identifier and underexpression is identified with a second identifier, wherein the first identifier is a first color and the second identifier is a second color, and wherein first and second identifiers are superimposed to provide a combined color.

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the features and advantages of the present invention, reference is now made to the detailed description of the invention along with the accompanying figures and in which:

FIGURES 1A to 1C show the basic microarray data mining strategy steps involved in accepted gene-level microarray data analysis (FIGURE 1A), the modular mining strategy of the present invention FIGURE 1B and a full size representation of the module extraction algorithm FIGURE 1C to generate on or more datasets used to create the expression vectors;

FIGURE 2 is a graph representing transcriptional profiles showing levels of modular gene expression profiles across an independent group of samples;

FIGURE 3 is a distribution of keyword occurrence in the literature obtained for four sets of coordinately expressed genes;

FIGURE 4 illustrates a modular microarray analysis strategy for characterization of the transcriptional system;

FIGURE 5 is an analysis of patient blood leukocyte transcriptional profiles;

FIGURE 6 illustrates module maps of transcriptional changes caused by disease;

FIGURE 7 illustrates the identification of a blood leukocyte transcriptional signature in patients with metastatic melanoma;

FIGURE 8 illustrates the validation of microarray results in an independent set of samples;

FIGURE 9 illustrates the identification of a blood leukocyte transcriptional signature in transplant recipients under immunosuppressive drug therapy

FIGURE 10-13 illustrate detailed results of the module-level analysis;

FIGURE 14 illustrates the module-level analysis for distinctive transcriptional signatures in blood from patients with metastatic melanoma and from liver transplant recipients;

FIGURE 15 illustrates the mapping transcriptional changes in patient blood leukocytes at the module level;

FIGURE 16 illustrates the module-level analysis for common transcriptional signatures in blood from patients with metastatic melanoma and from liver transplant recipients;

FIGURE 17 illustrates the analysis of significance patterns with genes expressed at higher levels in both melanoma and liver transplant patients compared to healthy volunteers;

FIGURE 18 illustrates the modular distribution of ubiquitous and specific gene signatures common to melanoma and transplant groups;

FIGURE 19 illustrates the transcriptional signature of immunosuppression;

FIGURE 20 shows a statistical group comparison between patients and their respective controls;

FIGURE 21 shows the analysis of significance patterns for genes over-expressed in SLE patients but not in patients with acute Influenza A infection;

FIGURE 22 shows the patterns of significance for genes common to Influenza A and SLE; and

FIGURE 23 is a functional analysis of genes shared by patients with Influenza infection and Lupus grouped according to significance patterns.

DETAILED DESCRIPTION OF THE INVENTION

While the making and using of various embodiments of the present invention are discussed in detail below, it should be appreciated that the present invention provides many applicable inventive concepts that can be embodied in a wide variety of specific contexts. The specific embodiments discussed herein are merely illustrative of specific ways to make and use the invention and do not delimit the scope of the invention.

To facilitate the understanding of this invention, a number of terms are defined below. Terms defined herein have meanings as commonly understood by a person of ordinary skill in the areas relevant to the present invention. Terms such as "a", "an" and "the" are not intended to refer to only a singular entity, but include the general class of which a specific example may be used for illustration. The terminology herein is used to describe specific embodiments of the invention, but their usage does not delimit the invention, except as outlined in the claims. Unless defined otherwise, all technical and scientific terms used herein have the meaning commonly understood by a person skilled in the art to which this invention belongs. The following references provide one of skill with a general definition of many of the terms used in this invention: Singleton, et al., Dictionary Of Microbiology And Molecular Biology (2d ed. 1994); The Cambridge Dictionary Of Science And Technology (Walker ed., 1988); The Glossary Of Genetics, 5th Ed., R. Rieger et al. (eds.), Springer Verlag (1991); and Hale & Marham, The Harper Collins Dictionary Of Biology (1991).

Various biochemical and molecular biology methods are well known in the art. For example, methods of isolation and purification of nucleic acids are described in detail in WO 97/10365, WO 97/27317, Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, (P. Tijssen, ed.) Elsevier, N.Y. (1993); Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part 1. Theory and Nucleic Acid Preparation, (P. Tijssen, ed.) Elsevier, N.Y. (1993); and Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, N.Y., (1989); and Current Protocols in Molecular Biology, (Ausubel, F. M. et al., eds.) John Wiley & Sons, Inc., New York (1987-1999), including supplements such as supplement 46 (April 1999).

BIOINFORMATICS DEFINITIONS

As used herein, an "object" refers to any item or information of interest (generally textual, including noun, verb, adjective, adverb, phrase, sentence, symbol, numeric characters, etc.). Therefore, an object is anything that can form a relationship and anything that can be obtained, identified, and/or searched from

a source. "Objects" include, but are not limited to, an entity of interest such as gene, protein, disease, phenotype, mechanism, drug, etc. In some aspects, an object may be data, as further described below.

As used herein, a "relationship" refers to the co-occurrence of objects within the same unit (e.g., a phrase, sentence, two or more lines of text, a paragraph, a section of a webpage, a page, a magazine, paper, book, etc.). It may be text, symbols, numbers and combinations, thereof

As used herein, "meta data content" refers to information as to the organization of text in a data source. Meta data can comprise standard metadata such as Dublin Core metadata or can be collection-specific. Examples of metadata formats include, but are not limited to, Machine Readable Catalog (MARC) records used for library catalogs, Resource Description Format (RDF) and the Extensible Markup Language (XML). Meta objects may be generated manually or through automated information extraction algorithms.

As used herein, an "engine" refers to a program that performs a core or essential function for other programs. For example, an engine may be a central program in an operating system or application program that coordinates the overall operation of other programs. The term "engine" may also refer to a program containing an algorithm that can be changed. For example, a knowledge discovery engine may be designed so that its approach to identifying relationships can be changed to reflect new rules of identifying and ranking relationships.

As used herein, "semantic analysis" refers to the identification of relationships between words that represent similar concepts, e.g., though suffix removal or stemming or by employing a thesaurus. "Statistical analysis" refers to a technique based on counting the number of occurrences of each term (word, word root, word stem, n-gram, phrase, etc.). In collections unrestricted as to subject, the same phrase used in different contexts may represent different concepts. Statistical analysis of phrase co-occurrence can help to resolve word sense ambiguity. "Syntactic analysis" can be used to further decrease ambiguity by part-of-speech analysis. As used herein, one or more of such analyses are referred to more generally as "lexical analysis." "Artificial intelligence (AI)" refers to methods by which a non-human device, such as a computer, performs tasks that humans would deem noteworthy or "intelligent." Examples include identifying pictures, understanding spoken words or written text, and solving problems.

As used herein, the term "database" or "dataset" refer to repositories for raw or compiled data, even if various informational facets can be found within the data fields. A database is typically organized so its contents can be accessed, managed, and updated (e.g., the database is dynamic). The term "database" and "source" are also used interchangeably in the present invention, because primary sources of data and information are databases. However, a "source database" or "source data" refers in general to data, e.g., unstructured text and/or structured data, which are input into the system for identifying objects and determining relationships. A source database may or may not be a relational database. However, a

system database usually includes a relational database or some equivalent type of database or dataset which stores or includes stored values relating to relationships between objects.

As used herein, a "system database" and "relational database" are used interchangeably and refer to one or more collections of data organized as a set of tables containing data fitted into predefined categories. For example, a database table may comprise one or more categories defined by columns (e.g. attributes), while rows of the database may contain a unique object for the categories defined by the columns. Thus, an object such as the identity of a gene might have columns for its presence, absence and/or level of expression of the gene. A row of a relational database may also be referred to as a "set" and is generally defined by the values of its columns. A "domain" in the context of a relational database is a range of valid values a field such as a column may include.

As used herein, a "domain of knowledge" refers to an area of study over which the system is operative, for example, all biomedical data. It should be pointed out that there is advantage to combining data from several domains, for example, biomedical data and engineering data, for this diverse data can sometimes link things that cannot be put together for a normal person that is only familiar with one area or research/study (one domain). A "distributed database" refers to a database that may be dispersed or replicated among different points in a network.

Terms such "data" and "information" are often used interchangeably, as are "information" and "knowledge." As used herein, "data" is the most fundamental unit that is an empirical measurement or set of measurements. Data is compiled to contribute to information, but it is fundamentally independent of it. Information, by contrast, is derived from interests, e.g., data (the unit) may be gathered on ethnicity, gender, height, weight and diet for the purpose of finding variables correlated with risk of cardiovascular disease. However, the same data could be used to develop a formula or to create "information" about dietary preferences, i.e., likelihood that certain products in a supermarket have a higher likelihood of selling.

As used herein, "information" refers to a data set that may include numbers, letters, sets of numbers, sets of letters, or conclusions resulting or derived from a set of data. "Data" is then a measurement or statistic and the fundamental unit of information. "Information" may also include other types of data such as words, symbols, text, such as unstructured free text, code, etc. "Knowledge" is loosely defined as a set of information that gives sufficient understanding of a system to model cause and effect. To extend the previous example, information on demographics, gender and prior purchases may be used to develop a regional marketing strategy for food sales while information on nationality could be used by buyers as a guideline for importation of products. It is important to note that there are no strict boundaries between data, information, and knowledge; the three terms are, at times, considered to be equivalent. In general, data comes from examining, information comes from correlating, and knowledge comes from modeling.

As used herein, "a program" or "computer program" refers generally to a syntactic unit that conforms to the rules of a particular programming language and that is composed of declarations and statements or

instructions, divisible into, "code segments" needed to solve or execute a certain function, task, or problem. A programming language is generally an artificial language for expressing programs.

As used herein, a "system" or a "computer system" generally refers to one or more computers, peripheral equipment, and software that perform data processing. A "user" or "system operator" in general includes a person, that uses a computer network accessed through a "user device" (e.g., a computer, a wireless device, etc) for the purpose of data processing and information exchange. A "computer" is generally a functional unit that can perform substantial computations, including numerous arithmetic operations and logic operations without human intervention.

As used herein, "application software" or an "application program" refers generally to software or a program that is specific to the solution of an application problem. An "application problem" is generally a problem submitted by an end user and requiring information processing for its solution.

As used herein, a "natural language" refers to a language whose rules are based on current usage without being specifically prescribed, e.g., English, Spanish or Chinese. As used herein, an "artificial language" refers to a language whose rules are explicitly established prior to its use, e.g., computer-programming languages such as C, C++, Java, BASIC, FORTRAN, or COBOL.

As used herein, "statistical relevance" refers to using one or more of the ranking schemes (O/E ratio, strength, etc.), where a relationship is determined to be statistically relevant if it occurs significantly more frequently than would be expected by random chance.

As used herein, the terms "coordinately regulated genes" or "transcriptional modules" are used interchangeably to refer to grouped, gene expression profiles (e.g., signal values associated with a specific gene sequence) of specific genes. Each transcriptional module correlates two key pieces of data, a literature search portion and actual empirical gene expression value data obtained from a gene microarray. The set of genes that is selected into a transcriptional module based on the analysis of gene expression data (using the module extraction algorithm described above). Additional steps are taught by Chaussabel, D. & Sher, A. Mining microarray expression data by literature profiling. Genome Biol 3, RESEARCH0055 (2002), (http://genomebiology.com/2002/3/10/research/0055) relevant portions incorporated herein by reference and expression data obtained from a disease or condition of interest, e.g., Systemic Lupus erythematosus, arthritis, lymphoma, carcinoma, melanoma, acute infection, autoimmune disorders, autoinflammatory disorders, etc.).

The Table below lists examples of keywords that were used to develop the literature search portion or contribution to the transcription modules. The skilled artisan will recognize that other terms may easily be selected for other conditions, e.g., specific cancers, specific infectious disease, transplantation, etc. For example, genes and signals for those genes associated with T cell activation are described hereinbelow as Module ID "M 2.8" in which certain keywords (e.g., Lymphoma, T-cell, CD4, CD8, TCR, Thymus, Lymphoid, IL2) were used to identify key T-cell associated genes, e.g., T-cell surface markers (CD5, CD6, CD7, CD26, CD28, CD96); molecules expressed by lymphoid lineage cells

(lymphotoxin beta, IL2-inducible T-cell kinase, TCF7; and T-cell differentiation protein mal, GATA3, STAT5B). Next, the complete module is developed by correlating data from a patient population for these genes (regardless of platform, presence/absence and/or up or downregulation) to generate the transcriptional module. In some cases, the gene profile does not match (at this time) any particular clustering of genes for these disease conditions and data, however, certain physiological pathways (e.g., cAMP signaling, zinc-finger proteins, cell surface markers, etc.) are found within the "Underdetermined" modules. In fact, the gene expression data set may be used to extract genes that have coordinated expression prior to matching to the keyword search, i.e., either data set may be correlated prior to cross-referencing with the second data set.

Table 1. Examples of Genes within Distinct Modules

Module I.D.	Number of probe sets	Keyword selection	Assessment
M 1.1	76	Ig, Immunoglobulin, Bone, Marrow, PreB, IgM,Mu.	Plasma cells. Includes genes coding for Immunoglobulin chains (e.g. IGHM, IGJ, IGLL1, IGKC, IGHD) and the plasma cell marker CD38.
M 1.2	130	Platelet, Adhesion, Aggregation, Endothelial, Vascular	Platelets. Includes genes coding for platelet glycoproteins (ITGA2B, ITGB3, GP6, GP1A/B), and platelet-derived immune mediators such as PPPB (pro-platelet basic protein) and PF4 (platelet factor 4).
M 1.3	80	Immunoreceptor, BCR, B-cell, IgG	B-cells. Includes genes coding for B-cell surface markers (CD72, CD79A/B, CD19, CD22) and other B-cell associated molecules: Early B-cell factor (EBF), B-cell linker (BLNK) and B lymphoid tyrosine kinase (BLK).
M 1.4	132	Replication, Repression, Repair, CREB, Lymphoid, TNF-alpha	Undetermined. This set includes regulators and targets of cAMP signaling pathway (JUND, ATF4, CREM, PDE4, NR4A2, VIL2), as well as repressors of TNF-alpha mediated NF-KB activation (CYLD, ASK, TNFAIP3).
M 1.5	142	Monocytes, Dendritic, MHC, Costimulatory, TLR4, MYD88	Myeloid lineage. Includes molecules expressed by cells of the myeloid lineage (CD86, CD163, FCGR2A), some of which being involved in pathogen recognition (CD14, TLR2, MYD88). This set also includes TNF family members (TNFR2, BAFF).
M 1.6	141	Zinc, Finger, P53, RAS	Undetermined. This set includes genes coding for signaling molecules, e.g. the zinc finger containing inhibitor of activated STAT (PIAS1 and PIAS2), or the nuclear factor of activated T-cells NFATC3.
M 1.7	129	Ribosome, Translational, 40S, 60S, HLA	MHC/Ribosomal proteins. Almost exclusively formed by genes coding MHC class I molecules (HLA-A,B,C,G,E)+ Beta 2-microglobulin (B2M) or Ribosomal proteins (RPLs, RPSs).
M 1.8	154	Metabolism, Biosynthesis, Replication, Helicase	Undetermined. Includes genes encoding metabolic enzymes (GLS, NSF1, NAT1) and factors involved in DNA replication (PURA, TERF2, EIF2S1).
M 2.1	95	NK, Killer, Cytolytic, CD8, Cell-mediated, T-cell, CTL, IFN-g	Cytotoxic cells. Includes cytotoxic T-cells amd NK-cells surface markers (CD8A, CD2, CD160, NKG7, KLRs), cytolytic molecules (granzyme, perforin, granulysin), chemokines (CCL5, XCL1) and CTL/NK-cell associated molecules (CTSW).
M 2.2	49	Granulocytes, Neutrophils, Defense, Myeloid, Marrow	Neutrophils. This set includes innate molecules that are found in neutrophil granules (Lactotransferrin: LTF, defensin: DEAF1, Bacterial Permeability Increasing protein: BPI, Cathelicidin antimicrobial protein: CAMP…).
M 2.3	148	Erythrocytes, Red, Anemia, Globin,	Erythrocytes. Includes hemoglobin genes (HGBs) and other erythrocyte-associated genes (erythrocytic alkirin:ANK1,

Module I.D.	Number of probe sets	Keyword selection	Assessment
		Hemoglobin	Glycophorin C: GYPC, hydroxymethylbilane synthase: HMBS, erythroid associated factor: ERAF).
M 2.4	133	Ribonucleoprotein, 60S, nucleolus, Assembly, Elongation	Ribosomal proteins. Including genes encoding ribosomal proteins (RPLs, RPSs), Eukaryotic Translation Elongation factor family members (EEFs) and Nucleolar proteins (NPM1, NOAL2, NAP1L1).
M 2.5	315	Adenoma, Interstitial, Mesenchyme, Dendrite, Motor	Undetermined. This module includes genes encoding immune-related (CD40, CD80, CXCL12, IFNA5, IL4R) as well as cytoskeleton-related molecules (Myosin, Dedicator of Cytokenesis, Syndecan 2, Plexin C1, Distrobrevin).
M 2.6	165	Granulocytes, Monocytes, Myeloid, ERK, Necrosis	Myeloid lineage. Includes genes expressed in myeloid lineage cells (IGTB2/CD18, Lymphotoxin beta receptor, Myeloid related proteins 8/14 Formyl peptide receptor 1), such as Monocytes and Neutrophils.
M 2.7	71	No keywords extracted.	Undetermined. This module is largely composed of transcripts with no known function. Only 20 genes associated with literature, including a member of the chemokine-like factor superfamily (CKLFSF8).
M 2.8	141	Lymphoma, T-cell, CD4, CD8, TCR, Thymus, Lymphoid, IL2	T-cells. Includes T-cell surface markers (CD5, CD6, CD7, CD26, CD28, CD96) and molecules expressed by lymphoid lineage cells (lymphotoxin beta, IL2-inducible T-cell kinase, TCF7, T-cell differentiation protein mal, GATA3, STAT5B).
M 2.9	159	ERK, Transactivation, Cytoskeletal, MAPK, JNK	Undetermined. Includes genes encoding molecules that associate to the cytoskeleton (Actin related protein 2/3, MAPK1, MAP3K1, RAB5A). Also present are T-cell expressed genes (FAS, ITGA4/CD49D, ZNF1A1).
M 2.10	106	Myeloid, Macrophage, Dendritic, Inflammatory, Interleukin	Undetermined. Includes genes encoding for Immune-related cell surface molecules (CD36, CD86, LILRB), cytokines (IL15) and molecules involved in signaling pathways (FYB, TICAM2-Toll-like receptor pathway).
M 2.11	176	Replication, Repress, RAS, Autophosphorylation, Oncogenic	Undetermined. Includes kinases (UHMK1, CSNK1G1, CDK6, WNK1, TAOK1, CALM2, PRKCI, ITPKB, SRPK2, STK17B, DYRK2, PIK3R1, STK4, CLK4, PKN2) and RAS family members (G3BP, RAB14, RASA2, RAP2A, KRAS).
M 3.1	122	ISRE, Influenza, Antiviral, IFN-gamma, IFN-alpha, Interferon	Interferon-inducible. This set includes interferon-inducible genes: antiviral molecules (OAS1/2/3/L, GBP1, G1P2, EIF2AK2/PKR, MX1, PML), chemokines (CXCL10/IP-10), signaling molecules (STAT1, STAt2, IRF7, ISGF3G).
M 3.2	322	TGF-beta, TNF, Inflammatory, Apoptotic, Lipopolysaccharide	Inflammation I. Includes genes encoding molecules involved in inflammatory processes (e.g. IL8, ICAM1, C5R1, CD44, PLAUR, IL1A, CXCL16), and regulators of apoptosis (MCL1, FOXO3A, RARA, BCL3/6/2A1, GADD45B).
M 3.3	276	Inflammatory, Defense, Lysosomal, Oxidative, LPS	Inflammation II. Includes molecules inducing or inducible by inflammation (IL18, ALOX5, ANPEP, AOAH, HMOX1, SERPINB1), as well as lysosomal enzymes (PPT1, CTSB/S, NEU1, ASAH1, LAMP2, CAST).
M 3.4	325	Ligase, Kinase, KIP1, Ubiquitin, Chaperone	Undetermined. Includes protein phosphatases (PPP1R12A, PTPRC, PPP1CB, PPM1B) and phosphoinositide 3-kinase (PI3K) family members (PIK3CA, PIK32A, PIP5K3).
M 3.5	22	No keyword extracted	Undetermined. Composed of only a small number of transcripts. Includes hemoglobin genes (HBA1, HBA2, HBB).
M 3.6	288	Ribosomal, T-cell, Beta-catenin	Undetermined. This set includes mitochondrial ribosomal proteins (MRPLs, MRPs), mitochondrial elongations factors (GFM1/2), Sortin Nexins (SN1/6/14) as well as lysosomal ATPases (ATP6V1C/D).
M 3.7	301	Spliceosome,	Undetermined. Includes genes encoding proteasome subunits

Module I.D.	Number of probe sets	Keyword selection	Assessment
		Methylation, Ubiquitin	(PSMA2/5, PSMB5/8); ubiquitin protein ligases HIP2, STUB1, as well as components of ubiqutin ligase complexes (SUGT1).
M 3.8	284	CDC, TCR, CREB, Glycosylase	Undetermined. Includes genes encoding enzymes: aminomethyltransferase, arginyltransferase, asparagines synthetase, diacylglycerol kinase, inositol phosphatases, methyltransferases, helicases…
M 3.9	260	Chromatin, Checkpoint, Replication, Transactivation	Undetermined. Includes genes encoding kinases (IBTK, PRKRIR, PRKDC, PRKCI) and phosphatases (e.g. PTPLB, PPP2CB/3CB, PTPRC, MTM1, MTMR2).

As used herein, the term "array" refers to a solid support or substrate with one or more peptides or nucleic acid probes attached to the support. Arrays typically have one or more different nucleic acid or peptide probes that are coupled to a surface of a substrate in different, known locations. These arrays, also described as "microarrays", "gene-chips" or DNA chips that may have 10,000; 20,000, 30,000; or 40,000 different identifiable genes based on the known genome, e.g., the human genome. These pan-arrays are used to detect the entire "transcriptome" or transcriptional pool of genes that are expressed or found in a sample, e.g., nucleic acids that are expressed as RNA, mRNA and the like that may be subjected to RT and/or RT-PCR to made a complementary set of DNA replicons. Arrays may be produced using mechanical synthesis methods, light directed synthesis methods and the like that incorporate a combination of non-lithographic and/or photolithographic methods and solid phase synthesis methods. Bead arrays that include 50-mer oligonucleotide probes attached to 3 micrometer beads may be used that are, e.g., lodged into microwells at the surface of a glass slide or are part of a liquid phase suspension arrays (e.g., Luminex or Illumina) that are digital beadarrays in liquid phase and uses "barcoded" glass rods for detection and identification.

Various techniques for the synthesis of these nucleic acid arrays have been described, e.g., fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may be peptides or nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate. Arrays may be packaged in such a manner as to allow for diagnostics or other manipulation of an all inclusive device, see for example, U.S. Pat. No. 6,955,788, relevant portions incorporated herein by reference.

BIOLOGICAL DEFINITIONS

As used herein, the term "disease" refers to a physiological state of an organism with any abnormal biological state of a cell. Disease includes, but is not limited to, an interruption, cessation or disorder of cells, tissues, body functions, systems or organs that may be inherent, inherited, caused by an infection, caused by abnormal cell function, abnormal cell division and the like. A disease that leads to a "disease state" is generally detrimental to the biological system, that is, the host of the disease. With respect to the present invention, any biological state, such as an infection (e.g., viral, bacterial, fungal, helminthic, etc.), inflammation, autoinflammation, autoimmunity, anaphylaxis, allergies, premalignancy, malignancy,

surgical, transplantation, physiological, and the like that is associated with a disease or disorder is considered to be a disease state. A pathological state is generally the equivalent of a disease state.

Disease states may also be categorized into different levels of disease state. As used herein, the level of a disease or disease state is an arbitrary measure reflecting the progression of a disease or disease state as well as the physiological response upon, during and after treatment. Generally, a disease or disease state will progress through levels or stages, wherein the affects of the disease become increasingly severe. The level of a disease state may be impacted by the physiological state of cells in the sample.

As used herein, the terms "therapy" or "therapeutic regimen" refer to those medical steps taken to alleviate or alter a disease state, e.g., a course of treatment intended to reduce or eliminate the affects or symptoms of a disease using pharmacological, surgical, dietary and/or other techniques. A therapeutic regimen may include a prescribed dosage of one or more drugs or surgery. Therapies will most often be beneficial and reduce the disease state but in many instances the effect of a therapy will have non-desirable or side-effects. The effect of therapy will also be impacted by the physiological state of the host, e.g., age, gender, genetics, weight, other disease conditions, etc.

As used herein, the term "pharmacological state" or "pharmacological status" refers to those samples that will be, are and/or were treated with one or more drugs, surgery and the like that may affect the pharmacological state of one or more nucleic acids in a sample, e.g., newly transcribed, stabilized and/or destabilized as a result of the pharmacological intervention. The pharmacological state of a sample relates to changes in the biological status before, during and/or after drug treatment and may serve a diagnostic or prognostic function, as taught herein. Some changes following drug treatment or surgery may be relevant to the disease state and/or may be unrelated side-effects of the therapy. Changes in the pharmacological state are the likely results of the duration of therapy, types and doses of drugs prescribed, degree of compliance with a given course of therapy, and/or un-prescribed drugs ingested.

As used herein, the term "biological state" refers to the state of the transcriptome (that is the entire collection of RNA transcripts) of the cellular sample isolated and purified for the analysis of changes in expression. The biological state reflects the physiological state of the cells in the sample by measuring the abundance and/or activity of cellular constituents, characterizing according to morphological phenotype or a combination of the methods for the detection of transcripts.

As used herein, the term "expression profile" refers to the relative abundance of RNA, DNA or protein abundances or activity levels. The expression profile can be a measurement for example of the transcriptional state or the translational state by any number of methods and using any of a number of gene-chips, gene arrays, beads, multiplex PCR, quantitiative PCR, run-on assays, Northern blot analysis, Western blot analysis, protein expression, fluorescence activated cell sorting (FACS), enzyme linked immunosorbent assays (ELISA), chemiluminescence studies, enzymatic assays, proliferation studies or any other method, apparatus and system for the determination and/or analysis of gene expression that are readily commercially available.

As used herein, the term "transcriptional state" of a sample includes the identities and relative abundances of the RNA species, especially mRNAs present in the sample. The entire transcriptional state of a sample, that is the combination of identity and abundance of RNA, is also referred to herein as the transcriptome. Generally, a substantial fraction of all the relative constituents of the entire set of RNA species in the sample are measured.

As used herein, the term "transcriptional vectors," "expression vectors," "genomic vectors" (used interchangeably) refers to transcriptional expression data that reflects the "proportion of differentially expressed genes." For example, for each module the proportion of transcripts differentially expressed between at least two groups (e.g., healthy subjects versus patients). This vector is derived from the comparison of two groups of samples. The first analytical step is used for the selection of disease-specific sets of transcripts within each module. Next, there is the "expression level." The group comparison for a given disease provides the list of differentially expressed transcripts for each module. It was found that different diseases yield different subsets of modular transcripts. With this expression level it is then possible to calculate vectors for each module(s) for a single sample by averaging expression values of disease-specific subsets of genes identified as being differentially expressed. This approach permits the generation of maps of modular expression vectors for a single sample, e.g., those described in the module maps disclosed herein. These vector module maps represent an averaged expression level for each module (instead of a proportion of differentially expressed genes) that can be derived for each sample. These composite "expression vectors" are formed through successive rounds of selection: 1) of the modules that were significantly changed across study groups and 2) of the genes within these modules which are significantly changed across study groups (Figure 2, *step II*). Expression levels are subsequently derived by averaging the values obtained for the subset of transcripts forming each vector (Figure 2, *step III*). Patient profiles can then be represented by plotting expression levels obtained for each of these vectors on a graph (e.g. on a radar plot). Therefore a set of vectors results from two round of selection, first at the module level, and then at the gene level. Vector expression values are composite by construction as they derive from the average expression values of the transcript forming the vector.

Using the present invention it is possible to identify and distinguish diseases not only at the module-level, but also at the gene-level; i.e., two diseases can have the same vector (identical proportion of differentially expressed transcripts, identical "polarity"), but the gene composition of the expression vector can still be disease-specific. This disease-specific customization permits the user to optimize the performance of a given set of markers by increasing its specificity.

Using modules as a foundation grounds expression vectors to coherent functional and transcriptional units containing minimized amounts of noise. Furthermore, the present invention takes advantage of composite transcriptional markers. As used herein, the term "composite transcriptional markers" refers to the average expression values of multiple genes (subsets of modules) as compared to using individual genes as markers (and the composition of these markers can be disease-specific). The composite

transcriptional markers approach is unique because the user can develop multivariate microarray scores to assess disease severity in patients with, e.g., SLE, or to derive expression vectors disclosed herein. The fact that expression vectors are composite (i.e. formed by a combination of transcripts) further contributes to the stability of these markers. Most importantly, it has been found that using the composite modular transcriptional markers of the present invention the results found herein are reproducible across microarray platform, thereby providing greater reliability for regulatory approval. Indeed, vector expression values proved remarkably robust, as indicated by the excellent reproducibility obtained across microarray platforms; as well as the validation results obtained in an independent set of pediatric lupus patients. These results are of importance since improving the reliability of microarray data is a prerequisite for the widespread use of this technology in clinical practice.

Gene expression monitoring systems for use with the present invention may include customized gene arrays with a limited and/or basic number of genes that are specific and/or customized for the one or more target diseases. Unlike the general, pan-genome arrays that are in customary use, the present invention provides for not only the use of these general pan-arrays for retrospective gene and genome analysis without the need to use a specific platform, but more importantly, it provides for the development of customized arrays that provide an optimal gene set for analysis without the need for the thousands of other, non-relevant genes. One distinct advantage of the optimized arrays and modules of the present invention over the existing art is a reduction in the financial costs (e.g., cost per assay, materials, equipment, time, personnel, training, etc.), and more importantly, the environmental cost of manufacturing pan-arrays where the vast majority of the data is irrelevant. The modules of the present invention allow for the first time the design of simple, custom arrays that provide optimal data with the least number of probes while maximizing the signal to noise ratio. By eliminating the total number of genes for analysis, it is possible to, e.g., eliminate the need to manufacture thousands of expensive platinum masks for photolithography during the manufacture of pan-genetic chips that provide vast amounts of irrelevant data. Using the present invention it is possible to completely avoid the need for microarrays if the limited probe set(s) of the present invention are used with, e.g., digital optical chemistry arrays, ball bead arrays, beads (e.g., Luminex), multiplex PCR, quantitiative PCR, run-on assays, Northern blot analysis, or even, for protein analysis, e.g., Western blot analysis, 2-D and 3-D gel protein expression, MALDI, MALDI-TOF, fluorescence activated cell sorting (FACS) (cell surface or intracellular), enzyme linked immunosorbent assays (ELISA), chemiluminescence studies, enzymatic assays, proliferation studies or any other method, apparatus and system for the determination and/or analysis of gene expression that are readily commercially available.

The "molecular fingerprinting system" of the present invention may be used to facilitate and conduct a comparative analysis of expression in different cells or tissues, different subpopulations of the same cells or tissues, different physiological states of the same cells or tissue, different developmental stages of the same cells or tissue, or different cell populations of the same tissue against other diseases and/or normal cell controls. In some cases, the normal or wild-type expression data may be from samples analyzed at or

about the same time or it may be expression data obtained or culled from existing gene array expression databases, e.g., public databases such as the NCBI Gene Expression Omnibus database.

As used herein, the term "differentially expressed" refers to the measurement of a cellular constituent (e.g., nucleic acid, protein, enzymatic activity and the like) that varies in two or more samples, e.g., between a disease sample and a normal sample. The cellular constituent may be on or off (present or absent), upregulated relative to a reference or downregulated relative to the reference. For use with gene-chips or gene-arrays, differential gene expression of nucleic acids, e.g., mRNA or other RNAs (miRNA, siRNA, hnRNA, rRNA, tRNA, etc.) may be used to distinguish between cell types or nucleic acids. Most commonly, the measurement of the transcriptional state of a cell is accomplished by quantitative reverse transcriptase (RT) and/or quantitative reverse transcriptase-polymerase chain reaction (RT-PCR), genomic expression analysis, post-translational analysis, modifications to genomic DNA, translocations, in situ hybridization and the like.

For some disease states it is possible to identify cellular or morphological differences, especially at early levels of the disease state. The present invention avoids the need to identify those specific mutations or one or more genes by looking at modules of genes of the cells themselves or, more importantly, of the cellular RNA expression of genes from immune effector cells that are acting within their regular physiologic context, that is, during immune activation, immune tolerance or even immune anergy. While a genetic mutation may result in a dramatic change in the expression levels of a group of genes, biological systems often compensate for changes by altering the expression of other genes. As a result of these internal compensation responses, many perturbations may have minimal effects on observable phenotypes of the system but profound effects to the composition of cellular constituents. Likewise, the actual copies of a gene transcript may not increase or decrease, however, the longevity or half-life of the transcript may be affected leading to greatly increases protein production. The present invention eliminates the need of detecting the actual message by, in one embodiment, looking at effector cells (e.g., leukocytes, lymphocytes and/or sub-populations thereof) rather than single messages and/or mutations.

The skilled artisan will appreciate readily that samples may be obtained from a variety of sources including, e.g., single cells, a collection of cells, tissue, cell culture and the like. In certain cases, it may even be possible to isolate sufficient RNA from cells found in, e.g., urine, blood, saliva, tissue or biopsy samples and the like. In certain circumstances, enough cells and/or RNA may be obtained from: mucosal secretion, feces, tears, blood plasma, peritoneal fluid, interstitial fluid, intradural, cerebrospinal fluid, sweat or other bodily fluids. The nucleic acid source, e.g., from tissue or cell sources, may include a tissue biopsy sample, one or more sorted cell populations, cell culture, cell clones, transformed cells, biopies or a single cell. The tissue source may include, e.g., brain, liver, heart, kidney, lung, spleen, retina, bone, neural, lymph node, endocrine gland, reproductive organ, blood, nerve, vascular tissue, and olfactory epithelium.

The present invention includes the following basic components, which may be used alone or in combination, namely, one or more data mining algorithms; one or more module-level analytical processes; the characterization of blood leukocyte transcriptional modules; the use of aggregated modular data in multivariate analyses for the molecular diagnostic/prognostic of human diseases; and/or visualization of module-level data and results. Using the present invention it is also possible to develop and analyze composite transcriptional markers, which may be further aggregated into a single multivariate score.

The present inventors have recognized that current microarray-based research is facing significant challenges with the analysis of data that are notoriously "noisy," that is, data that is difficult to interpret and does not compare well across laboratories and platforms. A widely accepted approach for the analysis of microarray data begins with the identification of subsets of genes differentially expressed between study groups. Next, the users try subsequently to "make sense" out of resulting gene lists using pattern discovery algorithms and existing scientific knowledge.

Rather than deal with the great variability across platforms, the present inventors have developed a strategy that emphasized the selection of biologically relevant genes at an early stage of the analysis. Briefly, the method includes the identification of the transcriptional components characterizing a given biological system for which an improved data mining algorithm was developed to analyze and extract groups of coordinately expressed genes, or transcriptional modules, from large collections of data.

The biomarker discovery strategy described herein is particularly well adapted for the exploitation of microarray data acquired on a global scale. Starting from ~44,000 transcripts a set of 28 modules was defined that are composed of nearly 5000 transcripts. Sets of disease-specific composite expression vectors were then derived. Vector expression values (expression vectors) proved remarkably robust, as indicated by the excellent reproducibility obtained across microarray platforms. This finding is notable, since improving the reliability of microarray data is a prerequisite for the widespread use of this technology in clinical practice. Finally, expression vectors can in turn be combined to obtain unique multivariate scores, therefore delivering results in a form that is compatible with mainstream clinical practice. Interestingly, multivariate scores recapitulate global patterns of change rather than changes in individual markers. The development of such "global biomarkers" can be used for both diagnostic and pharmacogenomics fields.

In one example, twenty-eight transcriptional modules regrouping 4742 probe sets were obtained from 239 blood leukocyte transcriptional profiles. Functional convergence among genes forming these modules was demonstrated through literature profiling. The second step consisted of studying perturbations of transcriptional systems on a modular basis. To illustrate this concept, leukocyte transcriptional profiles obtained from healthy volunteers and patients were obtained, compared and analyzed. Further validation of this gene fingerprinting strategy was obtained through the analysis of a published microarray dataset.

Remarkably, the modular transcriptional apparatus, system and methods of the present invention using pre-existing data showed a high degree of reproducibility across two commercial microarray platforms.

The present invention includes the implementation of a widely applicable, two-step microarray data mining strategy designed for the modular analysis of transcriptional systems. This novel approach was used to characterize transcriptional signatures of blood leukocytes, which constitutes the most accessible source of clinically relevant information.

As demonstrated herein, it is possible to determine, differential and/or distinguish between two disease based on two vectors even if the vector is identical (+/+) for two diseases – e.g. M1.3 = 53% down for both SLE and FLU because the composition of each vector can still be used to differentiate them. For example, even though the proportion and polarity of differentially expressed transcripts is identical between the two diseases for M1.3, the gene composition can still be disease-specific. The combination of gene-level and module-level analysis considerably increases resolution. Furthermore, it is possible to use 2, 3, 4, 5, 10, 15, 20, 25, 28 or more modules to differentiate diseases.

The term "gene" refers to a nucleic acid (e.g., DNA) sequence that includes coding sequences necessary for the production of a polypeptide (e.g.,), precursor, or RNA (e.g., mRNA). The polypeptide may be encoded by a full length coding sequence or by any portion of the coding sequence so long as the desired activity or functional property (e.g., enzymatic activity, ligand binding, signal transduction, immunogenicity, etc.) of the full-length or fragment is retained. The term also encompasses the coding region of a structural gene and the sequences located adjacent to the coding region on both the 5' and 3' ends for a distance of about 2 kb or more on either end such that the gene corresponds to the length of the full-length mRNA and 5' regulatory sequences which influence the transcriptional properties of the gene. Sequences located 5' of the coding region and present on the mRNA are referred to as 5'-untranslated sequences. The 5'-untranslated sequences usually contain the regulatory sequences. Sequences located 3' or downstream of the coding region and present on the mRNA are referred to as 3'-untranslated sequences. The term "gene" encompasses both cDNA and genomic forms of a gene. A genomic form or clone of a gene contains the coding region interrupted with non-coding sequences termed "introns" or "intervening regions" or "intervening sequences." Introns are segments of a gene that are transcribed into nuclear RNA (hnRNA); introns may contain regulatory elements such as enhancers. Introns are removed or "spliced out" from the nuclear or primary transcript; introns therefore are absent in the messenger RNA (mRNA) transcript. The mRNA functions during translation to specify the sequence or order of amino acids in a nascent polypeptide.

As used herein, the term "nucleic acid" refers to any nucleic acid containing molecule, including but not limited to, DNA, cDNA and RNA. In particular, the terms "a gene in Table X" refers to at least a portion or the full-length sequence listed in a particular table, as found hereinbelow. The gene may even be found or detected a genomic form, that is, it includes one or more intron(s). Genomic forms of a gene may also include sequences located on both the 5' and 3' end of the coding sequences that are present on

the RNA transcript. These sequences are referred to as "flanking" sequences or regions. The 5' flanking region may contain regulatory sequences such as promoters and enhancers that control or influence the transcription of the gene. The 3' flanking region may contain sequences that influence the transcription termination, post-transcriptional cleavage, mRNA stability and polyadenylation.

As used herein, the term "wild-type" refers to a gene or gene product isolated from a naturally occurring source. A wild-type gene is that which is most frequently observed in a population and is thus arbitrarily designed the "normal" or "wild-type" form of the gene. In contrast, the term "modified" or "mutant" refers to a gene or gene product that displays modifications in sequence and/or functional properties (i.e., altered characteristics) when compared to the wild-type gene or gene product. It is noted that naturally occurring mutants can be isolated; these are identified by the fact that they have altered characteristics (including altered nucleic acid sequences) when compared to the wild-type gene or gene product.

As used herein, the term "polymorphism" refers to the regular and simultaneous occurrence in a single interbreeding population of two or more alleles of a gene, where the frequency of the rarer alleles is greater than can be explained by recurrent mutation alone (typically greater than 1%).

As used herein, the terms "nucleic acid molecule encoding," "DNA sequence encoding," and "DNA encoding" refer to the order or sequence of deoxyribonucleotides along a strand of deoxyribonucleic acid. The order of these deoxyribonucleotides determines the order of amino acids along the polypeptide protein) chain. The DNA sequence thus codes for the amino acid sequence.

As used herein, the terms "complementary" or "complementarity" are used in reference to polynucleotides (i.e., a sequence of nucleotides) related by the base-pairing rules. For example, the sequence "A-G-T," is complementary to the sequence "T-C-A." Complementarity may be "partial," in which only some of the nucleic acids' bases are matched according to the base pairing rules. Or, there may be "complete" or "total" complementarity between the nucleic acids. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands. This is of particular importance in amplification reactions, as well as detection methods that depend upon binding between nucleic acids.

As used herein, the term "hybridization" is used in reference to the pairing of complementary nucleic acids. Hybridization and the strength of hybridization (i.e., the strength of the association between the nucleic acids) is impacted by such factors as the degree of complementarity between the nucleic acids, stringency of the conditions involved, the T_m of the formed hybrid, and the G:C ratio within the nucleic acids. A single molecule that contains pairing of complementary nucleic acids within its structure is said to be "self-hybridized."

As used herein the term "stringency" is used in reference to the conditions of temperature, ionic strength, and the presence of other compounds such as organic solvents, under which nucleic acid hybridizations are conducted. Under "low stringency conditions" a nucleic acid sequence of interest will hybridize to its exact complement, sequences with single base mismatches, closely related sequences (e.g., sequences

with 90% or greater homology), and sequences having only partial homology (e.g., sequences with 50-90% homology). Under "medium stringency conditions," a nucleic acid sequence of interest will hybridize only to its exact complement, sequences with single base mismatches, and closely related sequences (e.g., 90% or greater homology). Under "high stringency conditions," a nucleic acid sequence of interest will hybridize only to its exact complement, and (depending on conditions such a temperature) sequences with single base mismatches. In other words, under conditions of high stringency the temperature can be raised so as to exclude hybridization to sequences with single base mismatches.

As used herein, the term "probe" refers to an oligonucleotide (i.e., a sequence of nucleotides), whether occurring naturally as in a purified restriction digest or produced synthetically, recombinantly or by PCR amplification, that is capable of hybridizing to another oligonucleotide of interest. A probe may be single-stranded or double-stranded. Probes are useful in the detection, identification and isolation of particular gene sequences. Any probe used in the present invention may be labeled with any "reporter molecule," so that it is detectable in any detection system, including, but not limited to enzyme (e.g., ELISA, as well as enzyme-based histochemical assays), fluorescent, radioactive, luminescent systems and the like. It is not intended that the present invention be limited to any particular detection system or label.

As used herein, the term "target," refers to the region of nucleic acid bounded by the primers. Thus, the "target" is sought to be sorted out from other nucleic acid sequences. A "segment" is defined as a region of nucleic acid within the target sequence.

As used herein, the term "Southern blot" refers to the analysis of DNA on agarose or acrylamide gels to fractionate the DNA according to size followed by transfer of the DNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized DNA is then probed with a labeled probe to detect DNA species complementary to the probe used. The DNA may be cleaved with restriction enzymes prior to electrophoresis. Following electrophoresis, the DNA may be partially depurinated and denatured prior to or during transfer to the solid support. Southern blots are a standard tool of molecular biologists (Sambrook et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Press, NY, pp 9.31-9.58, 1989).

As used herein, the term "Northern blot" refers to the analysis of RNA by electrophoresis of RNA on agarose gels, to fractionate the RNA according to size followed by transfer of the RNA from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized RNA is then probed with a labeled probe to detect RNA species complementary to the probe used. Northern blots are a standard tool of molecular biologists (Sambrook, et al., supra, pp 7.39-7.52, 1989).

As used herein, the term "Western blot" refers to the analysis of protein(s) (or polypeptides) immobilized onto a support such as nitrocellulose or a membrane. The proteins are run on acrylamide gels to separate the proteins, followed by transfer of the protein from the gel to a solid support, such as nitrocellulose or a nylon membrane. The immobilized proteins are then exposed to antibodies with reactivity against an

antigen of interest. The binding of the antibodies may be detected by various methods, including the use of radiolabeled antibodies.

As used herein, the term "polymerase chain reaction" ("PCR") refers to the method of K. B. Mullis (U.S. Pat. Nos. 4,683,195 4,683,202, and 4,965,188, hereby incorporated by reference), which describe a method for increasing the concentration of a segment of a target sequence in a mixture of genomic DNA without cloning or purification. This process for amplifying the target sequence consists of introducing a large excess of two oligonucleotide primers to the DNA mixture containing the desired target sequence, followed by a precise sequence of thermal cycling in the presence of a DNA polymerase. The two primers are complementary to their respective strands of the double stranded target sequence. To effect amplification, the mixture is denatured and the primers then annealed to their complementary sequences within the target molecule. Following annealing, the primers are extended with a polymerase so as to form a new pair of complementary strands. The steps of denaturation, primer annealing and polymerase extension can be repeated many times (i.e., denaturation, annealing and extension constitute one "cycle"; there can be numerous "cycles") to obtain a high concentration of an amplified segment of the desired target sequence. The length of the amplified segment of the desired target sequence is determined by the relative positions of the primers with respect to each other, and therefore, this length is a controllable parameter. By virtue of the repeating aspect of the process, the method is referred to as the "polymerase chain reaction" (hereinafter "PCR"). Because the desired amplified segments of the target sequence become the predominant sequences (in terms of concentration) in the mixture, they are said to be "PCR amplified".

As used herein, the terms "PCR product," "PCR fragment," and "amplification product" refer to the resultant mixture of compounds after two or more cycles of the PCR steps of denaturation, annealing and extension are complete. These terms encompass the case where there has been amplification of one or more segments of one or more target sequences.

As used herein, the term "real time PCR" as used herein, refers to various PCR applications in which amplification is measured during as opposed to after completion of the reaction. Reagents suitable for use in real time PCR embodiments of the present invention include but are not limited to TaqMan probes, molecular beacons, Scorpions primers or double-stranded DNA binding dyes.

As used herein, the term "transcriptional upregulation" as used herein refers to an increase in synthesis of RNA, by RNA polymerases using a DNA template. For example, when used in reference to the methods of the present invention, the term "transcriptional upregulation" refers to an increase of least 1 to 2 fold, 2 to 3 fold, 3 to 10 fold, and even greater than 10 fold, in the quantity of mRNA corresponding to a gene of interest detected in a sample derived from an individual predisposed to SLE as compared to that detected in a sample derived from an individual who is not predisposed to SLE. However, the system and evaluation is sufficiently specific to require less that a 2 fold change in expression to be detected. Furthermore, the change in expression may be at the cellular level (change in expression within a single

cell or cell populations) or may even be evaluated at a tissue level, where there is a change in the number of cells that are expressing the gene. Particularly useful differences are those that are statistically significant.

Conversely, the term "transcriptional downregulation" refers to a decrease in synthesis of RNA, by RNA polymerases using a DNA template. For example, when used in reference to the methods of the present invention, the term "transcriptional downregulation" refers to a decrease of least 2 fold, 2 to 3 fold, 3 to 10 fold, and even greater than 10 fold, in the quantity of mRNA corresponding to a gene of interest detected in a sample derived from an individual predisposed to SLE as compared to that detected in a sample derived from an individual who is not predisposed to such a condition or to a database of information for wild-type and/or normal control, e.g., fibromyalgia. Again, the system and evaluation is sufficiently specific to require less that a 2 fold change in expression to be detected. Particularly useful differences are those that are statistically significant.

Both transcriptional "upregulation"/overexpression and transcriptional "downregulation"/underexpression may also be indirectly monitored through measurement of the translation product or protein level corresponding to the gene of interest. The present invention is not limited to any given mechanism related to upregulation or downregulation of transcription.

The term "eukaryotic cell" as used herein refers to a cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and bluegreen algae.

As used herein, the term "in vitro transcription" refers to a transcription reaction comprising a purified DNA template containing a promoter, ribonucleotide triphosphates, a buffer system that includes a reducing agent and cations, e.g., DTT and magnesium ions, and an appropriate RNA polymerase, which is performed outside of a living cell or organism.

As used herein, the term "amplification reagents" refers to those reagents (deoxyribonucleotide triphosphates, buffer, etc.), needed for amplification except for primers, nucleic acid template and the amplification enzyme. Typically, amplification reagents along with other reaction components are placed and contained in a reaction vessel (test tube, microwell, etc.).

As used herein, the term "diagnosis" refers to the determination of the nature of a case of disease. In some embodiments of the present invention, methods for making a diagnosis are provided which permit determination of SLE.

The present invention may be used alone or in combination with disease therapy to monitor disease progression and/or patient management. For example, a patient may be tested one or more times to determine the best course of treatment, determine if the treatment is having the intended medical effect, if the patient is not a candidate for that particular therapy and combinations thereof. The skilled artisan will

recognize that one or more of the expression vectors may be indicative of one or more diseases and may be affected by other conditions, be they acute or chronic.

As used herein, the term "pharmacogenetic test" refers to an assay intended to study interindividual variations in DNA sequence related to, e.g., drug absorption and disposition (pharmacokinetics) or drug action (pharmacodynamics), which may include polymorphic variations in one or more genes that encode the functions of, e.g., transporters, metabolizing enzymes, receptors and other proteins.

As used herein, the term "pharmacogenomic test" refers to an assay used to study interindividual variations in whole-genome or candidate genes, e.g., single-nucleotide polymorphism (SNP) maps or haplotype markers, and the alteration of gene expression or inactivation that may be correlated with pharmacological function and therapeutic response.

As used herein, an "expression profile" refers to the measurement of the relative abundance of a plurality of cellular constituents. Such measurements may include, e.g., RNA or protein abundances or activity levels. The expression profile can be a measurement for example of the transcriptional state or the translational state. See U.S. Pat. Nos. 6,040,138, 5,800,992, 6,020135, 6,033,860, relevant portions incorporated herein by reference. The gene expression monitoring system, include nucleic acid probe arrays, membrane blot (such as used in hybridization analysis such as Northern, Southern, dot, and the like), or microwells, sample tubes, gels, beads or fibers (or any solid support comprising bound nucleic acids). See, e.g., U.S. Pat. Nos. 5,770,722, 5,874,219, 5,744,305, 5,677,195 and 5,445,934, relevant portions incorporated herein by reference. The gene expression monitoring system may also comprise nucleic acid probes in solution.

The gene expression monitoring system according to the present invention may be used to facilitate a comparative analysis of expression in different cells or tissues, different subpopulations of the same cells or tissues, different physiological states of the same cells or tissue, different developmental stages of the same cells or tissue, or different cell populations of the same tissue.

As used herein, the term "differentially expressed: refers to the measurement of a cellular constituent varies in two or more samples. The cellular constituent can be either up-regulated in the test sample relative to the reference or down-regulated in the test sample relative to one or more references. Differential gene expression can also be used to distinguish between cell types or nucleic acids. See U.S. Pat. No. 5,800,992, relevant portions incorporated herein by reference.

Therapy or Therapeutic Regimen: In order to alleviate or alter a disease state, a therapy or therapeutic regimen is often undertaken. A therapy or therapeutic regimen, as used herein, refers to a course of treatment intended to reduce or eliminate the affects or symptoms of a disease. A therapeutic regimen will typically comprise, but is not limited to, a prescribed dosage of one or more drugs or surgery. Therapies, ideally, will be beneficial and reduce the disease state but in many instances the effect of a therapy will have non-desirable effects as well. The effect of therapy will also be impacted by the physiological state of the sample.

Modules display distinct "transcriptional behavior". It is widely assumed that co-expressed genes are functionally linked. This concept of "guilt by association" is particularly compelling in cases where genes follow complex expression patterns across many samples. The present inventors discovered that transcriptional modules form coherent biological units and, therefore, predicted that the co-expression properties identified in the initial dataset would be conserved in an independent set of samples. Data were obtained for PBMCs isolated from the blood of twenty-one healthy volunteers. These samples were not used in the module selection process described above.

The present invention includes the following basic components, which may be used alone or in combination, namely, one or more data mining algorithms; one or more module-level analytical processes; the characterization of blood leukocyte transcriptional modules; the use of aggregated modular data in multivariate analyses for the molecular diagnostic/prognostic of human diseases; and/or visualization of module-level data and results. Using the present invention it is also possible to develop and analyze composite transcriptional markers, which may be further aggregated into a single multivariate score.

The present inventors have recognized that current microarray-based research is facing significant challenges with the analysis of data that are notoriously "noisy," that is, data that is difficult to interpret and does not compare well across laboratories and platforms. A widely accepted approach for the analysis of microarray data begins with the identification of subsets of genes differentially expressed between study groups. Next, the users try subsequently to "make sense" out of resulting gene lists using pattern discovery algorithms and existing scientific knowledge.

Rather than deal with the great variability across platforms, the present inventors have developed a strategy that emphasized the selection of biologically relevant genes at an early stage of the analysis. Briefly, the method includes the identification of the transcriptional components characterizing a given biological system for which an improved data mining algorithm was developed to analyze and extract groups of coordinately expressed genes, or transcriptional modules, from large collections of data.

The biomarker discovery strategy that we have developed is particularly well adapted for the exploitation of microarray data acquired on a global scale. Starting from ~44,000 transcripts the inventors defined 28 modules composed of nearly 5000 transcripts. Sets of disease-specific composite expression vectors were then derived. Vector expression values proved remarkably robust, as indicated by the excellent reproducibility obtained across microarray platforms. This finding is notable, since improving the reliability of microarray data is a prerequisite for the widespread use of this technology in clinical practice. Finally, vectors can in turn be combined to obtain unique multivariate scores, therefore delivering results in a form that is compatible with mainstream clinical practice. Interestingly, multivariate scores recapitulate global patterns of change rather than changes in individual markers. The development of such "global biomarkers" constitutes therefore a promising prospect for both diagnostic and pharmacogenomics fields.

In one example, twenty-eight transcriptional modules regrouping 4742 probe sets were obtained from 239 blood leukocyte transcriptional profiles. Functional convergence among genes forming these modules was demonstrated through literature profiling. The second step consisted of studying perturbations of transcriptional systems on a modular basis. To illustrate this concept, leukocyte transcriptional profiles obtained from healthy volunteers and patients were obtained, compared and analyzed. Further validation of this gene fingerprinting strategy was obtained through the analysis of a published microarray dataset. Remarkably, the modular transcriptional apparatus, system and methods of the present invention using pre-existing data showed a high degree of reproducibility across two commercial microarray platforms.

The present invention includes the implementation of a widely applicable, two-step microarray data mining strategy designed for the modular analysis of transcriptional systems. This novel approach was used to characterize transcriptional signatures of blood leukocytes, which constitutes the most accessible source of clinically relevant information.

As demonstrated herein, it is possible to determine, differential and/or distinguish between two disease based on two vectors even if the vector is identical (+/+) for two diseases – e.g. M1.3 = 53% down for both SLE and FLU because the composition of each vector can still be used to differentiate them. For example, even though the proportion and polarity of differentially expressed transcripts is identical between the two diseases for M1.3, the gene composition can still be disease-specific. The combination of gene-level and module-level analysis considerably increases resolution. Furthermore, it is possible to use 2, 3, 4, 5, 10, 15, 20, 25, 28 or more modules to differentiate diseases.

Material and methods. Processing of blood samples. All blood samples were collected in acid citrate dextrose tubes (BD Vacutainer) and immediately delivered at room temperature to the Baylor Institute for Immunology Research, Dallas, TX, for processing. Peripheral blood mononuclear cells (PBMCs) from 3-4 ml of blood were isolated *via* Ficoll gradient and immediately lysed in RLT reagent (Qiagen, Valencia, CA) with beta-mercaptoethanol (BME) and stored at –80°C prior to the RNA extraction step.

Microarray analysis. Total RNA was isolated using the RNeasy kit (Qiagen) according to the manufacturer's instructions and RNA integrity was assessed using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA).

Affymetrix GeneChips: These microarrays include short oligonucleotide probe sets synthesized *in situ* on a quartz wafer. Target labeling was performed according to the manufacturer's standard protocol (Affymetrix Inc., Santa Clara, CA). Biotinylated cRNA targets were purified and subsequently hybridized to Affymetrix HG-U133A and U133B GeneChips (>44,000 probe sets). Arrays were scanned using an Affymetrix confocal laser scanner. Microarray Suite, Version 5.0 (MAS 5.0; Affymetrix) software was used to assess fluorescent hybridization signals, to normalize signals, and to evaluate signal detection calls. Normalization of signal values per chip was achieved using the MAS 5.0 global method of scaling to the target intensity value of 500 per GeneChip. A gene expression analysis software

program, GeneSpring, Version 7.1 (Agilent), was used to perform statistical analysis and hierarchical clustering.

Illumina BeadChips: These microarrays include 50mer oligonucleotide probes attached to 3μm beads, which are lodged into microwells at the surface of a glass slide. Samples were processed and acquired by Illumina Inc. (San Diego, CA) on the basis of a service contract. Targets were prepared using the Illumina RNA amplification kit (Ambion, Austin, TX). cRNA targets were hybridized to Sentrix HumanRef8 BeadChips (>25,000 probes), which were scanned on an Illumina BeadStation 500. Illumina's Beadstudio software was used to assess fluorescent hybridization signals.

Literature profiling. The literature profiling algorithm employed in this study has been previously described in detail (Chaussabel, D. & Sher, A. Mining microarray expression data by literature profiling. Genome Biol 3, RESEARCH0055 (2002), relevant portions incorporated herein by reference). This approach links genes sharing similar keywords. It uses hierarchical clustering, a popular unsupervised pattern discovery algorithm, to analyze patterns of term occurrence in literature abstracts. Step 1: A gene:literature index identifying pertinent publications for each gene is created. Step 2: Term occurrence frequencies were computed by a text processor. Step 3: Stringent filter criteria are used to select relevant keywords (i.e., eliminate terms with either high or low frequency across all genes and retain the few discerning terms characterized by a pattern of high occurrence for only a few genes). Step 4: Two-way hierarchical clustering groups of genes and relevant keywords based on occurrence patterns, providing a visual representation of functional relationships existing among a group of genes.

Modular data mining algorithm. First, one or more transcriptional components are identified that permit the characterization of biological systems beyond the level of single genes. Sets of coordinately regulated genes, or transcriptional modules, were extracted using a novel mining algorithm, which was applied to a large set of blood leukocyte microarray profiles (Figure 1). Gene expression profiles from a total of 239 peripheral blood mononuclear cells (PBMCs) samples were generated using Affymetrix U133A&B GeneChips (>44,000 probe sets). Transcriptional data were obtained for eight experimental groups (systemic juvenile idiopathic arthritis, systemic lupus erythematosus, type I diabetes, liver transplant recipients, melanoma patients, and patients with acute infections: *Escherichia coli, Staphylococcus aureus* and influenza A). For each group, transcripts with an absent flag call across all conditions were filtered out. The remaining genes were distributed among thirty sets by hierarchical clustering (clusters C1 through C30). The cluster assignment for each gene was recorded in a table and distribution patterns were compared among all the genes. Modules were selected using an iterative process, starting with the largest set of genes that belonged to the same cluster in all study groups (i.e. genes that were found in the same cluster in eight of the eight experimental groups). The selection was then expanded from this core reference pattern to include genes with 7/8, 6/8 and 5/8 matches. The resulting set of genes formed a transcriptional module and was withdrawn from the selection pool. The process was then repeated starting with the second largest group of genes, progressively reducing the level of stringency. This analysis led to the identification of 5348 transcripts that were distributed among

twenty-eight modules (a complete list is provided as supplementary material). Each module is assigned a unique identifier indicating the round and order of selection (i.e. M3.1 was the first module identified in the third round of selection).

Analysis of "significance patterns" was performed on gene expression data generated from PBMCs obtained from patients and healthy volunteers using Affymetrix HG-U133A GeneChips that were run on the same Affymetrix system, using standard operating procedures. P values were obtained by comparing 7 groups of patients to their respective healthy control groups (Mann-Whitney rank test). The groups were composed of pediatric patients with: 1) Systemic Lupus Erythomatosus (SLE, 16 samples), 2) Influenza A (16 samples), 3) *Staphylococcus aureus* (16 samples), 4) *Escherichia coli* (16 samples) and 5) *Streptococcus pneumoniae* (14 samples); as well as adult transplant recipients: 6) Liver transplant patients that have accepted the graft under immunosuppressive therapy (16 samples) and 7) bone marrow transplant recipients undergoing graft versus host disease (GVHD, 12 samples). Control groups were also formed taking into account age, sex and project (10 samples in each group). Genes significantly changed ($p<0.01$) in the "study group" (Influenza A and/or SLE) were divided in two sets: over-expressed versus control and under-expressed versus control. P-values of the genes forming the over-expressed set were obtained for the "reference groups" (infections with *E. coli, S. aureus, S. pneumoniae*, Liver transplant recipients and graft versus host disease). P-values of the reference groups were set to 1 when genes were under-expressed. The same procedure was used in the set of genes under-expressed in study group, only this time P-values of the reference group were set to 1 when genes were over-expressed. P-value data were processed with a gene expression analysis software program, GeneSpring, Version 7.1 (Agilent), that was used to perform hierarchical clustering and group genes based on significance patterns.

Modules display distinct "transcriptional behavior". It is widely assumed that co-expressed genes are functionally linked. This concept of "guilt by association" is particularly compelling in cases where genes follow complex expression patterns across many samples. The present inventors discovered that transcriptional modules form coherent biological units and, therefore, predicted that the co-expression properties identified in the initial dataset would be conserved in an independent set of samples. Data were obtained for PBMCs isolated from the blood of twenty-one healthy volunteers. These samples were not used in the module selection process described above.

FIGURE 2 shows gene expression profiles of four different modules are shown (Figure 2: M1.2, M1.7, M2.11 and M2.1). In the graphs of Figure 2, each line represents the expression level (y-axis) of a single gene across multiple samples (21 samples on the x-axis). Differences in gene expression in this example represent inter-individual variation between "healthy" individuals. It was found that within each module genes display a coherent "transcriptional behavior". Indeed, the variation in gene expression appeared to be consistent across all the samples (for some samples the expression of all the genes was elevated and formed a peak, while in others levels were low for all the genes which formed a dip). Importantly, inter-individual variations appeared to be module-specific as peaks and dips formed for different samples in M1.2, M2.11 and M2.1. Furthermore, the amplitude of variation was also characteristic of each module,

with levels of expression being more variable for M1.2 and M2.11 than M2.1 and especially M1.7. Thus, we find that transcriptional modules constitute independent biological variables.

Functional characterization of transcriptional modules. Next, the modules were characterized at a functional level. A text mining approach was employed to extract keywords from the biomedical literature collected for each gene (described in [18]). The distribution of keywords associated to the four modules that were analyzed is clearly distinct (Figure 3). The following is a list of keywords that may be associated with certain modules.

- Keywords highly specific for **M1.2** included *Platelet*, *Aggregation* or *Thrombosis*, and were associated with genes such as **ITGA2B** (Integrin alpha 2b, platelet glycoprotein IIb), **PF4** (platelet factor 4), **SELP** (Selectin P) and **GP6** (platelet glycoprotein 6).
- Keywords highly specific for **M1.3** included *B-cell*, *Immunoglobulin* or *IgG* and were associated with genes such as **CD19**, **CD22**, **CD72A**, **BLNK** (B cell linker protein), **BLK** (B lymphoid tyrosine kinase) and **PAX5** (paired box gene 5, a B-cell lineage specific activator).
- Keywords highly specific for **M1.5** included *Monocyte*, *Dendritic*, *CD14* or *Toll-like* and were associated with genes such as **MYD88** (myeloid differentiation primary response gene 88), **CD86**, **TLR2** (Toll-like receptor 2), **LILRB2** (leukocyte immunoglobulin-like receptor B2) and **CD163.**
- Keywords highly specific for **M3.1** included *Interferon*, *IFN-alpha*, *Antiviral*, or *ISRE* and were associated with genes such as **STAT1** (signal transducer and activator of transcription 1), **CXCL10** (CXC chemokine ligand 10, IP-10), **OAS2** (oligoadenylate synthetase 2) and **MX2** (myxovirus resistance 2).

This contrasted pattern of term occurrence denotes the remarkable functional coherence of each module. Information extracted from the literature for all the modules that have been identified permit a comprehensive functional characterization of the PBMC system at a transcriptional level. Table 2 provides an example of genes that may be used to distinguish between immune responses to, e.g., melanoma and liver transplant.

Table 2: Genes in Module 1.4 Used to Distinguish Immune Responses

moduleID	Entrez ID	Gene Symbol	Gene Title
1.4	55544	RNPC1	RNA-binding region (RNP1, RRM) containing 1
1.4	5930	RBBP6	retinoblastoma binding protein 6
1.4	80273	GRPEL1	GrpE-like 1, mitochondrial (*E. coli*)
1.4	57162	PELI1	pellino homolog 1 (Drosophila)
1.4	9921	RNF10	ring finger protein 10 /// ring finger protein 10
1.4	90637	LOC90637	hypothetical protein LOC90637
1.4	80314	EPC1	Enhancer of polycomb homolog 1 (Drosophila)
1.4	---	---	Full length insert cDNA clone ZB81B12
1.4	5756	PTK9	PTK9 protein tyrosine kinase 9
1.4	55038	CDCA4	cell division cycle associated 4
1.4	5187	PER1	period homolog 1 (Drosophila)
1.4	9205	ZNF237	zinc finger protein 237
1.4	25976	TIPARP	TCDD-inducible poly(ADP-ribose) polymerase
1.4	57018	CCNL1	cyclin L1
1.4	64061	TSPYL2	TSPY-like 2
1.4	81488	GRINL1A	glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A
1.4	22850	KIAA0863	KIAA0863 protein
1.4	23764	MAFF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian)

moduleID	Entrez ID	Gene Symbol	Gene Title
1.4	29035	PRO0149	PRO0149 protein
1.4	7803	PTP4A1	protein tyrosine phosphatase type IVA, member 1
1.4	11171	STRAP	serine/threonine kinase receptor associated protein
1.4	5814	PURB	purine-rich element binding protein B
1.4	5142	PDE4B	phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)
1.4	30836	ERBP	estrogen receptor binding protein
1.4	6782	STCH	stress 70 protein chaperone, microsome-associated, 60kDa
1.4	10950	BTG3	BTG family, member 3
1.4	7037	TFRC	transferrin receptor (p90, CD71)
1.4	54934	FLJ20436	hypothetical protein FLJ20436
1.4	5144	PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
1.4	9929	KIAA0063	KIAA0063 gene product
1.4	143187	VTI1A	Vesicle transport through interaction with t-SNAREs homolog 1A (yeast)
1.4	440309	---	LOC440309
1.4	150094	SNF1LK	SNF1-like kinase /// SNF1-like kinase
1.4	1850	DUSP8	dual specificity phosphatase 8
1.4	9584	RNPC2	RNA-binding region (RNP1, RRM) containing 2
1.4	140735	Dlc2	dynein light chain 2
1.4	54542	MNAB	membrane associated DNA binding protein
1.4	9262	STK17B	serine/threonine kinase 17b (apoptosis-inducing)
1.4	7128	TNFAIP3	tumor necrosis factor, alpha-induced protein 3
1.4	3183	HNRPC	heterogeneous nuclear ribonucleoprotein C (C1/C2)
1.4	5144	PDE4D	Phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
1.4	80311	KLHL15	kelch-like 15 (Drosophila)
1.4	22850	KIAA0863	KIAA0863 protein
1.4	5996	RGS1	---
1.4	468	ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)
1.4	---	---	---
1.4	7430	VIL2	villin 2 (ezrin)
1.4	6627	SNRPA1	small nuclear ribonucleoprotein polypeptide A'
1.4	7750	ZNF198	zinc finger protein 198
1.4	1390	CREM	cAMP responsive element modulator
1.4	10291	SF3A1	splicing factor 3a, subunit 1, 120kDa
1.4	9308	CD83	CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)
1.4	63935	C20orf67	---
1.4	10049	DNAJB6	DnaJ (Hsp40) homolog, subfamily B, member 6
1.4	51526	C20orf111	chromosome 20 open reading frame 111
1.4	55500	ETNK1	ethanolamine kinase 1 /// ethanolamine kinase 1
1.4	79441	C4orf15	chromosome 4 open reading frame 15
1.4	11236	RNF139	ring finger protein 139
1.4	246243	RNASEH1	ribonuclease H1
1.4	3727	JUND	jun D proto-oncogene
1.4	6500	SKP1A	S-phase kinase-associated protein 1A (p19A)
1.4	4204	MECP2	Methyl CpG binding protein 2 (Rett syndrome)
1.4	3189	HNRPH3	heterogeneous nuclear ribonucleoprotein H3 (2H9)
1.4	222161	DKFZp586I1420	hypothetical protein DKFZp586I1420
1.4	266812	NAP1L5	nucleosome assembly protein 1-like 5
1.4	9908	G3BP2	Ras-GTPase activating protein SH3 domain-binding protein 2
1.4	10425	ARIH2	---

moduleID	Entrez ID	Gene Symbol	Gene Title
1.4	55422	ZNF331	Zinc finger protein 331
1.4	8454	CUL1	Cullin 1
1.4	51119	SBDS	Shwachman-Bodian-Diamond syndrome
1.4	6309	SC5DL	sterol-C5-desaturase (ERG3 delta-5-desaturase homolog, fungal)-like
1.4	5277	PIGA	phosphatidylinositol glycan, class A (paroxysmal nocturnal hemoglobinuria) /// phosphatidylinositol glycan, class A (paroxysmal nocturnal hemoglobinuria)
1.4	3422	IDI1	isopentenyl-diphosphate delta isomerase
1.4	63935	C20orf67	chromosome 20 open reading frame 67
1.4	7975	MAFK	v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian)
1.4	7456	WASPIP	Wiskott-Aldrich syndrome protein interacting protein
1.4	55975	KLHL7	kelch-like 7 (Drosophila)
1.4	7128	TNFAIP3	tumor necrosis factor, alpha-induced protein 3
1.4	388796	LOC388796	hypothetical LOC388796
1.4	25852	ARMC8	armadillo repeat containing 8
1.4	54542	MNAB	Membrane associated DNA binding protein
1.4	55422	ZNF331	zinc finger protein 331
1.4	1390	CREM	cAMP responsive element modulator
1.4	10209	SUI1	putative translation initiation factor
1.4	10049	DNAJB6	DnaJ (Hsp40) homolog, subfamily B, member 6
1.4	4929	NR4A2	nuclear receptor subfamily 4, group A, member 2
1.4	1540	CYLD	cylindromatosis (turban tumor syndrome)
1.4	4929	NR4A2	nuclear receptor subfamily 4, group A, member 2
1.4	5805	PTS	6-pyruvoyltetrahydropterin synthase
1.4	10926	ASK	activator of S phase kinase
1.4	10923	PC4	activated RNA polymerase II transcription cofactor 4 /// similar to Activated RNA polymerase II transcriptional coactivator p15 (Positive cofactor 4) (PC4) (p14)
1.4	388796	RNU71A	Hypothetical LOC388796
1.4	133746	JMY	junction-mediating and regulatory protein
1.4	90634	CG018	Hypothetical gene CG018
1.4	10209	SUI1	putative translation initiation factor
1.4	1847	DUSP5	dual specificity phosphatase 5
1.4	7088	TLE1	Transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)
1.4	84275	MGC4399	mitochondrial carrier protein
1.4	---	---	---
1.4	7803	PTP4A1	protein tyrosine phosphatase type IVA, member 1
1.4	55422	ZNF331	zinc finger protein 331
1.4	---	---	CDNA clone IMAGE:30332316, partial cds
1.4	3609	ILF3	interleukin enhancer binding factor 3, 90kDa
1.4	---	---	Homo sapiens, clone IMAGE:4753714, mRNA
1.4	6651	SON	SON DNA binding protein
1.4	11276	AP1GBP1	AP1 gamma subunit binding protein 1
1.4	84124	ZNF394	zinc finger protein 394
1.4	63935	C20orf67	---
1.4	1983	EIF5	eukaryotic translation initiation factor 5
1.4	80063	ATF7IP2	Activating transcription factor 7 interacting protein 2
1.4	285831	LOC285831	hypothetical protein LOC285831
1.4	81873	ARPC5L	actin related protein 2/3 complex, subunit 5-like
1.4	144438	LOC144438	hypothetical protein LOC144438
1.4	10209	SUI1	putative translation initiation factor
1.4	3021	H3F3B	H3 histone, family 3B (H3.3B)
1.4	25948	KBTBD2	kelch repeat and BTB (POZ) domain containing 2

moduleID	Entrez ID	Gene Symbol	Gene Title
1.4	---	---	CDNA FLJ40725 fis, clone TKIDN1000001, highly similar to Translocase of inner mitochondrial membrane 23
1.4	1540	CYLD	cylindromatosis (turban tumor syndrome)
1.4	5144	PDE4D	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
1.4	51182	HSPA14	heat shock 70kDa protein 14
1.4	29080	HSPC128	HSPC128 protein
1.4	8731	RNMT	RNA (guanine-7-) methyltransferase
1.4	3423	IDS	iduronate 2-sulfatase (Hunter syndrome)
1.4	283991	MGC29814	hypothetical protein MGC29814
1.4	1454	CSNK1E	Casein kinase 1, epsilon
1.4	26051	PPP1R16B	protein phosphatase 1, regulatory (inhibitor) subunit 16B
1.4	3422	IDI1	isopentenyl-diphosphate delta isomerase
1.4	5887	RAD23B	RAD23 homolog B (S. cerevisiae)
1.4	5144	PDE4D	Phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
1.4	49854	ZNF295	zinc finger protein 295
1.4	60493	FLJ13149	hypothetical protein FLJ13149
1.4	10950	BTG3	BTG family, member 3

Yet another group that may be used alone or in combination with the genes listed in supplementary table that includes the data shown in Figure 17, denoted P2, and which may include one or more the following genes that are overexpressed, e.g., WARS; IFI53; IFP53; GAMMA-2; FAM46C; FLJ20202; H3F3B; H3.3B; FOXK2; ILF; ILF1; ILF-1; DUSP5; HVH3; ARF6; DKFZp762C186; BRD2; NAT; RNF3; FSRG1; RING3; D6S113E; KIAA9001; RORA; ROR1; ROR2; ROR3; RZRA; NR1F1; DKFZp762C186; DNAJB1; SUI1; CXCR4; HM89; LAP3; NPYR; WHIM; LESTR; NPY3R; HSY3RR; NPYY3R; D2S201E; GRINL1A; CTSB; TRIP-Br2; PDE4B; DPDE4; PDEIVB; PMAIP1; APR; NOXA; BTG2; PC3; TIS21; ASAHL; SON; SUI1; A121; ISO1; HERPUD1; SUP; Mif1; KIAA0025; DUSP2; PAC1; PAC-1; RNF139; RCA1; TRC8; HRCA1; MGC31961; TNFAIP3; A20; TNFA1P2; ARS2; HNRPL; hnRNP-L; P/OKcl.14; C20orf67; C20orf111; HSPC207; dJ1183I21.1; ZNF331; RITA; ZNF361; ZNF463; C20orf67; IER5; SBBI48; ; SUI1; JUN; AP1; CD69; TOB1; H3F3B; H3.3B; FOLR1; TNFAIP3; TCF8; BZP; ZEB; ZEB1; AREB6; ZFHEP; NIL-2A; ZFHX1A; NIL-2-A; DUSP10; MKP5; MKP-5; GGTLA4; MGC50550; dJ831C21.2; PMAIP1; ZC3HAV1; ZAP; FLB6421; ZC3HDC2; FLJ13288; MGC48898; DSIPI; DIP; GILZ; hDIP; TSC-22R; MCL1; TM; EAT; MCL1L; MCL1S; MGC1839; SH3TC1; FLJ20356; CIAS1; FCU; MWS; FCAS; NALP3; C1orf7; PYPAF1; AII/AVP; AGTAVPRL; SLC15A3; PHT2; PTR3; hPTR3; PTDSR; PSR; PTDSR1; KIAA0585; BHLHB2; DEC1; STRA13; Stra14; HMGE; KIAA0063; NR4A2; NOT; RNR1; HZF-3; NURR1; TINUR; NR4A2; NOT; RNR1; HZF-3; NURR1; TINUR; PTS; PTPS; HEAB; CLP1; hClp1; AREG; SDGF; CRDGF; MGC13647; EDG4; LPA2; EDG-4; LPAR2; CREM; ICER; MGC17881; MGC41893; CD83; BL11; HB15; ZNF394; FLJ12298, and combinations thereof.

Module-based microarray data mining strategy. Results from "traditional" microarray analyses are notoriously noisy and difficult to interpret. A widely accepted approach for microarray data analyses includes three basic steps: 1) Use of a statistical test to select genes differentially expressed between

study groups; 2) Apply pattern discovery algorithms to identify signatures among the resulting gene lists; and 3) Interpret the data using knowledge derived from the literature or ontology databases.

The present invention uses a novel microarray data mining strategy emphasizing the selection of biologically relevant transcripts at an early stage of the analysis. This first step can be carried out using for instance the modular mining algorithm described above in combination with a functional mining tool used for in-depth characterization of each transcriptional module (FIGURE 4: top panel, Step 1). The analysis does not take into consideration differences in gene expression levels between groups. Rather, the present invention focuses instead on complex gene expression patterns that arise due to biological variations (e.g., inter-individual variations among a patient population). After defining the transcriptional components associated to a given biological system the second step of the analysis includes the analysis of changes in gene expression through the comparison of different study groups (FIGURE 4: bottom panel, Step 2). Group comparison analyses are carried out independently for each module. Changes at the module level are expressed as the proportion of genes that meet the significance criteria (represented by a pie chart in FIGURE 5 or a spot in FIGURE 6). Notably, carrying out comparisons at the modular level permits to avoid the noise generated when thousands of tests are performed on "random" collections of genes.

Perturbation of modular PBMC transcriptional profiles in human diseases. To illustrate the second step of the microarray data mining strategy described above (FIGURE 4), gene expression data for PBMC samples obtained from two pediatric patient populations composed of eighteen children with systemic lupus erythematosus (SLE) and sixteen children with acute influenza A infection was obtained, compared and analyzed. Each patient cohort was matched to its respective control group (healthy volunteers: eleven and ten donors were matched to the SLE and influenza groups, respectively). Following the analytical scheme depicted in FIGURE 4, a statistical group comparisons between patient and healthy groups for each individual module and measured the proportion of genes significantly changed in each module (FIGURE 5) was performed. The statistical group comparison approach allows the user to focus the analysis on well defined groups of genes that contain minimal amounts of noise and carry identifiable biological meaning. A key to the graphical representation of these results is provided in Figure 4.

The following findings were made: (1) that a large proportion of genes in M3.1 ("interferon-associated") met the significance level in both Flu and SLE groups (84% and 94%, respectively). This observation confirms earlier work with SLE patients [19] and identifies the presence of an interferon signature in patients with acute influenza infection. (2) Equivalent proportions of genes in M1.3 ("B-cell-associated") were significantly changed in both groups (53%), with over 50% overlap between the two lists. This time, genes were consistently under-expressed in patient compared to healthy groups. (3) Modules were also found that differentiate the two diseases. The proportion of genes significantly changed in Module 1.1 reaches 39% in SLE patients and is only 7% in Flu patients, which at a significance level of 0.05 is very close to the proportion of genes that would be expected to be differentially expressed only by chance. Interestingly, this module is almost exclusively composed of genes encoding immunoglobulin

chains and has been associated with plasma cells. However, this module is clearly distinct from the B-cell associated module (M1.3), both in terms of gene expression level and pattern (not shown). (4) As illustrated by module M1.5, gene-level analysis of individual modules can be used to further discriminate the two diseases. It is also the case for M1.3, where, despite the absence of differences at the module-level (FIGURE 4: 53% under-expressed transcripts), differences between Flu and SLE groups could be identified at the gene-level (only 51% of the under-expressed transcripts in M1.3 were common to the two disease groups). These examples illustrate the use of a modular framework to streamline the analysis and interpretation of microarray results.

Mapping changes in gene expression at the modular level. Data visualization is paramount for the interpretation of complex datasets and the present invention includes a comprehensive graphical illustration of changes that occur at the modular level. Changes in gene expression levels caused by different diseases were represented for the twenty-eight PBMC transcriptional modules (FIGURE 6). Each disease group is compared to its respective control group composed of healthy donors who were matched for age and sex (eighteen patients with SLE, sixteen with acute influenza infection, sixteen with metastatic melanoma and sixteen liver transplant recipients receiving immunosuppressive drug treatment were compared to control groups composed of ten to eleven healthy subjects). Module-level data were represented graphically by spots aligned on a grid, with each position corresponding to a different module (See Table 1 for functional annotations on each of the modules).

The spot intensity indicates the proportion of genes significantly changed for each module. The spot color indicates the polarity of the change (red: proportion of over-expressed genes, blue: proportion of under-expressed genes; modules containing a significant proportion of both over- and under-expressed genes would be purple-though none were observed). This representation permits a rapid assessment of perturbations of the PBMC transcriptional system. Such "module maps" were generated for each disease. When comparing the four maps, we found that diseases were characterized by a unique modular combination. Indeed, results for M1.1 and M1.2 alone sufficed to distinguish all four diseases (M1.1/M1.2: SLE = +/+; FLU=0/0; Melanoma=-/+; transplant=-/-). A number of genes in M3.2 ("inflammation") were over-expressed in all diseases (particularly so in the transplant group), while genes in M3.1 (interferon) were over-expressed in patients with SLE, influenza infection and, to some extent, transplant recipients. "Ribosomal protein" module genes (M1.7 and M2.4) were under-expressed in both SLE and Flu groups. The level of expression of these genes was recently found to be inversely correlated to disease activity in SLE patients (Bennett *et al.*, submitted). M2.8 includes T-cell transcripts which are under-expressed in lymphopenic SLE patients and transplant recipients treated with immunosuppressive drugs targeting T-cells.

Interestingly, differentially expressed genes in each module were predominantly either under-expressed or over-expressed (FIGURE 5 and FIGURE 6). Yet, modules were purely selected on the basis of similarities in gene expression profiles, not changes in expression levels between groups. The fact that changes in gene expression appear highly polarized within each module denotes the functional relevance

of modular data. Thus, the present invention enables disease fingerprinting by a modular analysis of patient blood leukocyte transcriptional profiles.

Validation of PBMC modules in a published dataset. Next, the validity of the PBMC transcriptional modules described above in a "third-party" dataset was tested. The study from Connolly, et al., who investigated the effects of exercise on gene expression in human PBMCs[20] was tested.

Blood samples were obtained from 35 patients with metastatic melanoma enrolled in three phase I/II clinical trials designed to test the efficacy of a dendritic cell therapeutic vaccine as seen in the table below. Gene expression signatures were generated from blood samples collected prior to the initiation of vaccine therapy, and at least 4 weeks after the last systemic therapy if a patient had undergone such.

Table 3: Clinical and demographic characteristics of 35 patients with metastatic melanoma

ID	Sex	Age	Stage	Diagnosis	Time from Dx to Blood Draw (Months)	Blood Draw	Status	Blood Draw to Time of Death (Months)
MEL 23	F	61	M1a	02/14/00	16	06/28/01	Deceased	28
MEL 24	M	53	M1c	09/01/99	22	07/05/01	Deceased	5
MEL 26	F	52	M1c	10/01/97	45	07/18/01	Deceased	2
MEL 27	F	54	M1a	07/10/93	98	09/14/01	Deceased	5
MEL 29	M	41	M1c	10/26/94	83	09/26/01	Deceased	14
MEL 30	F	58	M1c	10/04/99	23	09/25/01	Deceased	11
MEL 32	M	56	M1b	01/17/94	95	12/17/01	Deceased	24
MEL 34	F	28	M1b	04/01/01	10	02/05/02	Deceased	42
MEL 35	M	29	M1a	08/25/98	43	03/12/02	Deceased	12
MEL 36	F	69	M1b	01/01/85	205	02/19/02	Deceased	7
MEL 40	F	43	M1c	07/02/91	132	07/19/02	Deceased	19
MEL 43	M	60	M1a	08/14/94	100	12/04/02	Alive 05/16/05	*
MEL 44	F	68	M1c	05/01/99	44	01/28/03	Deceased	12
MEL 45	F	53	M1a	02/01/02	10	12/17/02	Alive 5/5/05	*
MEL 46	F	47	M1c	11/18/97	61	12/27/02	Deceased	22
MEL 47	F	35	M1c	03/02/02	10	01/09/03	Deceased	2
MEL 48	M	68	M1b	*1992	>120	03/12/03	Alive 05/10/05	*
MEL 49	M	71	M1b	12/05/97	64	04/10/03	Alive 07/05/05	*
MEL 50	M	52	M1c	07/08/97	69	04/11/03	Deceased	18
MEL 51	M	56	M1c	10/01/01	18	04/16/03	Alive 05/17/05	*
MEL 52	M	42	M1c	03/01/02	13	04/17/03	Deceased	9
MEL 54	M	50	M1b	07/20/90	153	04/25/03	Alive 04/08/05	*
MEL 56	M	71	M1c	03/01/01	26	05/29/03	Deceased	9
MEL 57	F	36	M1b	07/01/02	11	06/05/03	Deceased	20
MEL 58	M	67	M1c	10/01/99	45	07/18/03	Deceased	10
MEL 59	M	61	M1c	unknown	*	07/25/03	Alive 06/22/05	*
MEL 60	M	41	M1c	11/01/02	9	08/14/03	Deceased	7
MEL 61	F	54	M1a	03/03/99	54	09/10/03	Alive 05/18/05	*
MEL 62	M	46	M1b	12/01/01	22	10/09/03	Deceased	5
MEL 63	M	75	M1b	12/01/00	34	10/29/03	Alive 03/16/05	*
MEL 64	F	53	M1b	04/01/00	42	10/30/03	Alive 03/05/04	*
MEL 65	M	62	M1b	08/14/94	111	11/14/03	Alive 05/16/05	*
MEL 68	M	74	M1b	06/09/04	1	07/29/04	Deceased	9
MEL 70	M	67	M1b	04/06/04	5	09/23/04	Alive 8/18/2005	*
MEL 72	M	50	M1c	09/23/04	2	11/10/04	Deceased	*

The Table provides both clinical and demographic characteristics of 35 patients with metastatic melanoma.

A second group of patients included 39 liver transplant recipients maintaining their graft under pharmacological immunosuppressive therapy, time from transplant had a median of 729 days and a range of between 338 and 1905 days. Outpatients coming for routine exams were recruited for this study. All patients received standard treatment regimens with calcineurin inhibitors (e.g., Tacrolimus: n=25; Cyclosporin A: n=13). The main indications for liver transplant were hepatitis C (n = 19) and Laennec's cirrhosis (n = 7). The table below provides both clinical and demographic characteristics of 39 liver transplant recipients.

Table 4: Clinical and demographic characteristics of 39 liver transplant recipients

Patient ID	Age	Sex	Transpl. to Blood Draw (Days)	TAC	CsA	Primary Dx
R1292	47	M	1854	yes		Hepatitis C
R1297	48	M	1868		yes	Hepatitis C
R1308	66	F	1905		yes	Primary Biliary Cirrhosis
R1322	32	F	1821		yes	Hepatitis C
R1323	45	M	1828		yes	Hepatitis C
R1325	50	M	1801		yes	Hepatitis C
R1329	51	F	1781	yes		Laennec's Cirrhosis
R1340	50	M	1829		yes	Laennec's Cirrhosis
R1348	64	F	1802	yes		Fulminant Hepatic Failure
R1355	61	M	1780	yes		Cryptogenic
R1364	42	M	1756	yes		Hepatitis C
R1413	52	M	1856		yes	Laennec's Cirrhosis
R1673	60	F	756	yes		Hepatitis B
R1674	45	M	729	yes		Hepatitis C
R1684	65	F	721	yes		Mx. Carcinoid
R1686	59	M	704	yes		Hepatitis B
R1689	43	M	692	yes		Hepatitis C
R1700	53	M	732		yes	Hepatitis C
R1701	45	M	737	yes		Hepatitis C
R1702	48	M	726	yes		Hepatitis C
R1706	57	M	721	yes		Laennec's Cirrhosis
R1710	50	F	736	yes		Cryptogenic
R1714	63	F	718	yes		Nonalcoholic Steatohepatitis
R1718	53	F	707		yes	Hepatitis C
R1754	51	F	589	yes		Primary Sclerosing Cholangitis
R1771	42	F	812			Hepatitis C
R1787	60	F	794	yes		Laennec's Cirrhosis
R1805	42	M	735	yes		Laennec's Cirrhosis & Postnecrotic Cirrhosis Type C
R1814	45	M	427	yes		Hepatitis C
R1838	66	F	360	yes		Autoimmune Hepatitis
R1839	56	M	354	yes		Cryptogenic
R1841	52	M	363	yes		PSC-UC
R1843	48	M	361		yes	Hepatitis C
R1845	44	F	338	yes		Primary Biliary Cirrhosis
R1846	41	F	338		yes	Hepatitis C
R1847	53	M	358	yes		Laennec's Cirrhosis
R1854	50	M	350		yes	Hepatitis C

R1971	46	M	367		yes	Hepatitis C; Hepatocellular Carcinoma with Cirrhosis
R1974	54	M	341	yes		Hepatitis C

Blood samples were also obtained from 25 healthy donors that constituted the control groups. The table below provides the demographic characteristics of the 25 healthy donors.

Table 5: Demographic characteristics of 25 healthy donors

Healthy Volunteers	Sex	Age
D-001	M	41
D-002	F	53
D-005	F	40
D-007	F	44
D-008	M	40
D-010	M	46
D-011	F	43
D-013	M	58
D-014	F	47
D-015	M	42
D-016	F	40
D-017	M	25
D-018	M	46
D-019	F	40
D-020	M	39
D-021	M	45
D-022	M	50
D-024	F	44
D-025	F	48
D-027	F	43
D-028	F	43
D-029	M	43
D-031	M	35
D-032	F	43
D-033	F	43

Identification of disease-associated blood leukocyte transcriptional signatures. Blood leukocyte gene expression signatures were identified in patients with metastatic melanoma and liver transplant recipients. Each set of patients was compared to a control group of healthy volunteers. Patient samples were divided into a training set, used to identify disease-associated and predictive expression signatures, and an independent test set. This step-wise analysis allowed validation of results in samples that were not used to establish the disease signature. Stringent criteria were employed to select the samples forming training sets in order to avoid confounding the analysis with biological and/or technical factors. The table below illustrates the composition of sample sets taking into account age, gender and sample processing method used for the identification (training) and validation (testing) of expression signatures associated with metastatic melanoma.

Table 6: Composition of sample sets associated with metastatic melanoma.

Training Set					Test				Total n
HV		Fresh	Frozen			Fresh	Frozen		
	Male	7	7	14	Male	5	0	5	
	Female	0	9	9	Female	8	0	8	
		7	16	23		13	0	13	36
Melanoma		Fresh	Frozen			Fresh	Frozen		
	Male	3	9	12	Male	0	11	11	
	Female	0	10	10	Female	0	5	5	
		3	19	22		0	16	16	38
			Total	45			Total	29	

Similarly, the table below provides the composition of sample sets used taking into account age, gender and sample processing method for the identification (training) and validation (testing) of expression signatures associated with liver transplant recipients undergoing immunosuppressive drug therapy.

Table 7: Composition of sample sets associated with liver transplant recipients undergoing immunosuppressive drug therapy.

Training Set					Test				Total n
HV		Fresh	Frozen			Fresh	Frozen		
	Male	12	5	17	Male	0	2	2	
	Female	8	2	10	Female	0	7	7	
		20	7	27		0	9	9	36
LTx		Fresh	Frozen			Fresh	Frozen		
	Male	9	3	12	Male	15	0	15	
	Female	9	1	10	Female	6	0	6	
		18	4	22		21	0	21	43
			Total	49			Total	30	

Table 8 lists the genes differentially expressed in patients with metastatic melanoma in comparison to healthy volunteers. Statistical comparison of a group of twenty-two patients with metastatic melanoma versus twenty-three healthy volunteers, training set, identified 899 differentially expressed genes ($p<0.01$, non parametric Mann-Whitney rank test and >1.25 fold change; 218 overexpressed and 681 underexpressed genes). Tables 8 through 14 are provided in Computer Readable Form (CRF) as part of the Lengthy Table and are incorporated herein by reference. The Tables provide a correlation to the modules from which there were identified, their expression level, various nomenclatures for their individual identification.

FIGURES 7A-7D are images of the hierarchical clustering of genes. FIGURE 7A illustrates the hierarchical clustering of genes that produce a reciprocal expression pattern; which was confirmed in an independent test set shown in FIGURE 7B. FIGURE 7B displays results from 13 healthy volunteers versus 16 patients. Next, class prediction algorithms were applied to the initial training set. These

algorithms yielded 81 genes with the best ability to classify healthy volunteers and patients based on their differential expression as shown in FIGURE 7C and Table 9). Table 9 illustrates the expression levels of a set of transcripts discriminating patients with melanoma from healthy volunteers. Using these 81 genes, an independent test set was classified with 90% accuracy; the class of only three was indeterminate as illustrated in FIGURE 7D.

FIGURES 8A and 8B are plots of the microarray results in an independent set of samples to confirm the reliability of the results by correlating expression levels obtained for the training and test sets. Genes with the best ability to discriminate between patients and healthy volunteers were identified in a training set using a class prediction algorithm (k-Nearest Neighbors). Fold change expression levels between healthy controls and patients were measured for discriminative genes in the training set and an independent test set. Fold change values obtained in training and test sets were correlated: FIGURE 8A illustrates results for metastatic melanoma and produced 81 genes with a Pearson correlation of r^2=0.83 and a $p<0.0001$ and FIGURE 8B illustrates results for liver transplant recipients and produced 65 genes with a Pearson correlation of r^2=0.94 and a $p<0.0001$.

FIGURES 9A-9D are images of the hierarchical clustering of genes for the identification of a blood leukocyte transcriptional signature in transplant recipients under immunosuppressive drug therapy. Samples were divided into a training set (27 healthy, 22 patients) used to identify differentially expressed genes in liver transplant recipients versus healthy volunteers as seen in FIGURE 9A and FIGURE 9C respectively. A test set of nine healthy and 21 patients were used to independently validate this signature as seen in FIGURE 9B and FIGURE 9D. Class comparison identified 2,589 differentially expressed genes (Mann-Whitney test $p < 0.01$, fold change > 1.25). FIGURE 9A illustrates a similar signature in the test set and FIGURE 9B illustrates the class prediction that identified 81 genes. FIGURE 9C shows that the discrimination of the independent test set with 90% accuracy. FIGURE 9D illustrates the class could not be identified for two samples out of 30; one sample was incorrectly predicted (i.e. transplant recipient classified as healthy).

The same analysis strategy was applied to Liver transplant patients. Statistical comparison of a group of 22 transplant recipients versus 27 healthy volunteers identified 2,589 differentially expressed genes ($p<0.01$, non-parametric Mann-Whitney rank test and >1.25 fold change; 938 overexpressed and 1651 underexpressed genes; Table 10). Table 10 illustrates genes that are expressed differentially in liver transplant recipients under treatment with immunosuppressive drugs in comparison to healthy volunteers. Hierarchical clustering of genes produced a reciprocal expression pattern that was observed in both training as seen in FIGURE 9A and independent test sets as seen in FIGURE 9B. Sixty-five classifier genes were established in the training set and are illustrates in FIGURE 9C and Table 11. Table 11 illustrates the expression level of a set of transcripts discriminating liver transplant recipients under treatment with immunosuppressive drugs from healthy volunteers. The sixty-five classifier genes were applied to an independent test set of 9 healthy donors and 21 patients. Samples were correctly classified 90% of the time: class could not be determined in two cases and one sample was misclassified as seen in

FIGURE 9D. The results obtained for both of these sets were highly correlated, e.g., pearson correlation, r^2=0.94, p<0.0001, as seen in FIGURE 8B. Thus, the blood leukocyte transcriptional signatures associated with patients with metastatic melanoma and in liver transplant recipients have been identified and validated.

Module-level analysis of patients PBMC transcriptional profiles was performed. A custom microarray data mining strategy was used to further characterize disease-associated gene expression patterns. The analysis of a set of 239 blood leukocytes transcriptional signatures identified 28 transcriptional modules regrouping 4,742 probe sets. These "transcriptional modules" are sets of genes that follow similar expression patterns across a large number of samples in multiple studies, identified through co-expression meta-analysis. Each module is associated with a unique identifier that indicates the round of selection and order (e.g., M2.8 designates the eighth module of the second round of selection). Upon extraction each transcriptional module is functionally characterized with the help of a literature profiling algorithm (Chaussabel and Sher, 2002).

FIGURES 10 to 13 illustrate detailed statistical comparisons between healthy and disease groups of the module-level analysis. For example, twenty-eight sets of coordinately expressed genes, or transcriptional modules, were identified through the analysis of 239 PBMC microarray profiles. For each of these modules changes in expression levels between groups of healthy volunteers and either patients with metastatic melanoma or transplant recipients were tested. A pie chart indicates the proportion of genes that were significantly changed in each module, where red indicates overexpressed genes and blue illustrates underexpressed genes, with a p<0.05 for the Mann-Whitney test. For each module, keywords extracted from the literature are listed in green along with a functional assessment of relationships existing among the genes.

FIGURE 14 is a plot of the changes observed in a few representative modules represented by a transcriptional profile. Each differentially expressed gene is represented by a line that indicates relative levels of expression across healthy volunteer and patient samples. Peaks and dips respectively indicate relatively higher and lower gene expression in a given patient. Genes that were not significantly different are not represented. The levels of expression of genes associated with a platelet signature changed in opposite directions: 28% of the genes forming this signature (M1.2) were overexpressed in patients with melanoma and 27% were underexpressed in transplant recipients. Furthermore, half of the genes belonging to module M2.1 (cytotoxic cell signature) were underexpressed in transplant recipients. This trend was not observed in patients with melanoma (7% overexpressed with p<0.05 where 5% of changes are expected by chance only). Similarly, a massive down-regulation of genes associated to T cells was observed in transplant recipients (74% of genes in M2.8). This finding most likely reflects pharmacological immunosuppression. In patients with melanoma 29% of these T-cell related genes were down-regulated. In addition, 44% of interferon-inducible genes that form module M3.1 were overexpressed in transplant recipients, while 26% were underexpressed in patients with melanoma. Lists of differentially expressed genes in each module are available in Tables 12 and 13.

Patients with metastatic melanoma and transplant recipients display common transcriptional profiles at the modular level. This analysis identified similarities as well as differences in blood leukocyte transcriptional signatures of patients with metastatic melanoma and liver transplant recipients.

FIGURE 15 is an image of the modular changes observed in both groups of patients vs. their respective healthy control group. The proportion of differentially expressed genes for each module is indicated by a spot of variable intensity. For example, in the overlay a change in transplant is represented by a yellow, while a change in melanoma is indicated by a blue color and a change in both is indicated by a green color. Proportions of underexpressed and overexpressed transcripts are represented on separate grids. Modules that were common between the two groups of patients include M1.4 (regulator of cAMP and NF-kB signaling pathways), M2.6 (including genes expressed in myeloid lineage cells), M3.2 and M3.3 (both M3.2 and 3.3 include factors involved in inflammation; as seen in FIGURES 10-13).

FIGURE 16 is an image illustrating the module-level analysis of the present invention. Common transcriptional signatures in blood from patients with metastatic melanoma and from liver transplant recipients. Expression profiles of genes belonging to blood leukocyte transcriptional modules M1.1, M1.3, M1.4 and M3.2. The total number of probes is indicated for each module (U133A), along with a brief functional interpretation. Keywords extracted by literature profiling are indicated in green. From the total number in each module, the proportion of genes that were significantly changed (Mann-Whitney test, $p<0.05$) in patients compared to the appropriate healthy control group is indicated in a pie chart with overexpressed genes are expressed in red and underexpressed genes are expressed in blue. Graphs represent transcriptional profiles of the genes that were significantly changed, with each line showing levels of expression on the y-axis of a single transcript across multiple conditions (samples, x-axis).

The association between metastatic melanoma and liver transplant phenotype was strongest for M1.4 and M3.2 as seen in FIGURE 16. Interestingly, a majority of underexpressed modules were common to melanoma and transplant groups, with the most striking similarities in the case of M1.1 (including plasma cell associated genes), M1.3 (including B-cell associated genes) and M1.8 (including genes coding for metabolic enzymes and factors involved in DNA replication).

Identification of a common transcriptional signature that is unique to metastatic melanoma and liver transplant patients. The extent of the similarities between patients with metastatic melanoma and liver transplant recipients were specific to these two groups of patients were examined. Statistical group comparison was carried out between patients and healthy controls across all samples (e.g., thirty-eight melanoma, forty-three transplant, thirty-six healthy). Briefly, 323 transcripts were identified that were significantly overexpressed and 918 that were significantly underexpressed in both liver transplant recipients and patients with metastatic melanoma (Mann-Whitney test, $p<0.01$, filtered >1.25 fold change). Next, group comparisons for these transcripts were carried using samples from patients with Systemic Lupus Erythematosus ("SLE"), acute infections (*S. pneumoniae*, *S. aureus*, *E. coli*, and Influenza A) and Graft versus Host Disease ("GVHD") compared to relevant healthy controls. This

analysis yielded p-values whose hierarchical clustering identified distinct significance patterns among transcripts common to the melanoma and transplant groups. This analysis identified sets of genes that changed across all diseases as seen in FIGURE 17 with P1 being ubiquitously overexpressed; P3 being ubiquitously underexpressed, while others were associated more specifically with the melanoma and transplant groups as seen in FIGURE 18 with P2 being overexpressed; and P4 being underexpressed. Table 14 illustrates the significance levels across 8 diseases of the genes forming patterns P1, P2, P3 and P4.

FIGURE 17 is an image of the analysis of significance patterns. Genes expressed at higher levels in both stage IV melanoma or liver transplant patients compared to healthy volunteers were selected. P-values were similarly obtained from gene expression profiles generated in other disease models: in PBMCs obtained from patients suffering from systemic lupus erythematosus (SLE), Graft versus Host Disease (GVHD), or acute infections with influenza virus (Influenza A), *Escherichia coli* (*E. coli*), *Streptococcus pneumoniae* (*Strep. Pneumo.*) or *Staphylococcus aureus* (*Staph. aureus*). Each of these cohorts was compared to the appropriate control group of healthy volunteers accrued in the context of these studies. The genes expressed at significantly higher or lower levels in PBMCs obtained from both patients with melanoma and liver transplant recipients (OVER-XP and UNDER-XP, respectively) were ranked by hierarchical clustering of p-values generated for all the conditions listed above. P-values are represented according to a color scale: Green represents low p-value/significant, while white represents high p-value/not significant. Distinct significant patterns are identified, where P1 and P3 are ubiquitous and P2 and P4 are most specific to melanoma and liver transplant groups.

FIGURE 18 is a chart of the modular distribution of ubiquitous and specific gene signatures common to melanoma and transplant groups. Distribution among 28 PBMC transcriptional modules was determined for genes that form ubiquitous (P1) and specific (P2) transcriptional signatures common to the melanoma and transplant groups. Gene lists of each of the modules were compared in turn to the 109 and 69 transcripts that form P1 and P2. For each module, the proportion of genes shared with either P1 or P2 was recorded. These results are represented by a bar graph of FIGURE 18.

Thus, genes forming transcriptional signatures common to the melanoma and transplant groups can be partitioned into distinct sets based on two properties: (1) coordinated expression as seen in the transcriptional modules of FIGURE 13; and (2) change in expression across diseases as seen in the significance patterns of FIGURE 17. The results from these two different mining strategies were recouped by examining the modular distribution of ubiquitous (P1) and specific (P2) PBMC transcriptional signatures. FIGURE 18 clearly shows that the distribution of P1 and P2 across the 28 PBMC transcriptional modules that have been identified to date is not random. Indeed, P1 transcripts are preferentially found among M3.2 (characterized by transcripts related to inflammation), whereas M1.4 transcripts almost exclusively belonged to P2, which includes genes that are more specifically overexpressed in patients with melanoma and liver transplant recipients.

FIGURE 19 is an illustration of the transcriptional signature of immunosuppression. Transcripts overexpressed most specifically in patients with melanoma and transplant recipients (P1) include repressors of immune responses that inhibit: 1) NF-kB translocation; 2) Interleukin 2 production and signaling; 3) MAPK pathways and 4) cell proliferation. Some of these factors are well characterized anti-inflammatory molecules and others are expressed in anergic T-cells.

Molecular signature of immunosuppression. The genes that were most specifically overexpressed in melanoma and transplant groups (P1) were examined. From the 69 probe sets, 55 unique gene identifiers were identified. A query against a literature database indexed by gene, have developed to aid the interpretation of microarray gene expression data, identified 6527 publications associated with 47 genes, 30 of which were associated with more than ten publications. FIGURE 19 illustrates a remarkable functional convergence among the genes forming this signature and includes genes encoding molecules that possess immunoregulatory functions (e.g., anti-proliferative genes: BTG2, TOB1, AREG, SUI1 or RNF139; anti-inflammatory genes: TNFAIP3); inhibitors of transcription: (SON, ZC3HAV1, ZNF394); stress-induced molecules (HERPUD1); while others possess well established immunosuppressive properties. For example, dual specificity phosphatases 2, 5 and 10 (DUSP2, 5, 10) interfere with the MAP kinases ERK1/2, which are known targets of calcineurin inhibitors such as Tacrolimus/FK506. DUSP10 selectively dephosphorylates stress activated kinases(Theodosiou et al., 1999). Interestingly, DUSP5 was found to have a negative feedback role in IL2 signaling in T-cells(Kovanen et al., 2003). CREM, FOXK2 and TCF8 directly bind the IL2 promoter and can contribute to the repression of IL-2 production in T cell anergy(Powell et al., 1999). BHLHB2 (Stra13) negatively regulates lymphocyte development and function in vivo(Seimiya et al., 2004). CIAS1 codes for the protein Cryopyrin, which regulates NF-kappa B activation and production of proinflammatory cytokines. Mutations of this gene have been identified in several inflammatory disorders(Agostini et al., 2004). DSIPI, a leucine zipper protein, is known to mediate the immunosuppressive effects of glucocorticoids and IL10 by interfering with a broad range of signaling pathways (NF-kappa B, NFAT/AP-1, MEK, ERK 1/2), leading to the general inhibition of inflammatory responses in macrophages and down-regulation of the IL2 receptor in T cells. Noatably, the expression of DSIPI in immune cells was found to be augmented after drug treatment (dexamethasone)(D'Adamio et al., 1997) or long term exposure to tumor cells (Burkitt Lymphoma)(Berrebi et al., 2003).

Other immunosuppressive molecules, which did not belong to P1, were also found overexpressed in melanoma and transplant groups. Notably, DDIT4, another dexamethasone-induced gene which was recently found to inhibit mTOR, the mammalian target for rapamycin (Corradetti et al., 2005). Thus, this endogenous factor appears capable of reproducing the action of potent immunosuppressive drugs. HMOX1, a cytoprotective molecule that also demonstrates anti-inflammatory properties. Most recently, HMOX1 expression was found to be induced by FOXP3 and to mediate immunosuppressive effects of CD4+ CD25+ regulatory T cells (Choi et al., 2005). Accordingly, an increase in transcriptional activity of HMOX1 has been correlated with favorable outcomes in experimental transplant models (Soares et al.,

1998). Both DDIT4 and HMOX1 genes were also overexpressed in patients with acute *E. coli* or *S. aureus* infections. The immunophilin FKBP1A (FKBP12), a member of the FK506-binding protein family, is a key mediator of T-cell immunosuppression by the drugs FK506 (tacrolimus) and rapamycin (Xu et al., 2002). Expression of this gene was elevated in comparison to healthy donors in all patient groups.

Blood is an accessible tissue and lends itself to comparative analyses across multiple diseases. Parmacological and tumor-mediated immunosuppression would produce a common transcriptional signature in blood leukocytes. Metastatic melanoma and transplant recipients disease-associated transcriptional signatures were identified in the blood of patients. These signatures were identified and confirmed through several analytical approaches. Analysis of transcriptional modules identified alterations in blood leukocytes transcriptional components associated to cell types (e.g., Plasma cells, B-cells, T-cells, Cytotoxic cells) and to immune reactions (e.g., Inflammation, Interferon). Furthermore, using both transcriptional modules and gene expression levels similarities between blood transcriptional signatures in patients with metastatic melanoma and liver transplant recipients were identified. However, this common transcriptional signature could not be entirely attributed to immunosuppression. For instance, expression levels of B-cell associated genes (M1.3) were not only decreased in the melanoma and transplant groups, but also in patients with acute influenza infection and systemic lupus erythematosus (SLE) (53% of genes were underexpressed, in comparison to healthy controls; Chaussabel et al.). Conversely, nearly 40% of the genes associated with plasma cells (M1.1) were overexpressed in SLE patients and there was no change in patients with acute influenza infection (7% of the genes were overexpressed at $p<0.05$), whereas expression levels were significantly decreased in both patients with melanoma and transplant recipients (61% and 62% of the genes in M1.1, respectively). In order to select the most specific transcripts common between melanoma and transplant signatures a gene-level analysis was carried out across a total of eight groups of patients. This led to the identification of a set of transcripts that was most specifically overexpressed in immunosuppressed patients. The identified set of genes showed marked functional convergence and included genes coding for repressors of Interleukin-2 transcription, inhibitors of NF-kB or MAPK pathways, and anti-proliferative molecules. Interestingly, these signatures are consistent with the mechanism of action of drugs used for pharmacological immunosuppression, which inhibit the activity of calcineurin, a calcium-dependent serine threonine protein phosphatase responsible for the nuclear translocation of NF-AT and NF-kappaB upon T-cell activation. Indicating a functional convergence between immunosuppressive mechanisms operating in patients with advanced melanoma and pharmacologically treated transplant recipients. The fact that the transcripts more specifically induced in immunosuppressed patients include glucocorticoids-inducible genes (e.g., DSIPI, CXCR4, JUN) and hormone nuclear receptors (NR4A2 and RORA)(Winoto and Littman, 2002) suggest a possible role for steroid hormones in tumor-mediated immunosuppression.

Patients with metastatic melanoma display an endogenous transcriptional signature of immunosuppression similar to that induced by pharmacological treatments in patients who underwent

liver transplant. The present invention provides a method and apparatus to identify patients at high risk of melanoma progression. In addition the present invention also provides a method and apparatus for monitoring indicators of immunosuppression could help adjusting the dosage of immunosuppressive drugs and balance risks of rejection and side effects for liver transplant recipients.

Examples of patient information and processing of blood samples include the following. Blood was obtained after informed consent as approved by the institutional IRB (Liver transplant recipients: 002-1570199-017; patients with melanoma: 000-048, 002-094; 003-187). Blood samples were obtained in acid citrate dextrose yellow-top tubes (BD Vaccutainer) at the Baylor University Medical Center in Dallas, TX. Samples were immediately delivered at room temperature to the Baylor Institute for Immunology Research, Dallas, TX, for processing. Fresh PBMCs isolated via Ficoll gradient were either stored in liquid nitrogen (e.g., viable freezing) or immediately lysed in RLT buffer, containing ß-mercaptoethanol (Qiagen, Valencia, CA). Total RNA was extracted from cells previously frozen in liquid nitrogen ("frozen") or from cells that were lysed immediately after isolation ("fresh"), using the RNEASY® Mini Kit according to the manufacturer's recommended protocol (Qiagen, Valencia, CA). This parameter was taken into account in the experimental design taking into account the age, gender and sample processing method used for the identification (training) and validation (testing) of expression signatures associated with metastatic melanoma and liver transplant recipients undergoing immunosuppressive drug therapy.

Microarray assays. Total RNA was isolated using the RNEASY® kit (Qiagen, Valencia, CA) according to the manufacturer's instructions and the RNA integrity was assessed using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA). Although, the skilled artisan will recognize that other methods of isolation may be used. From 2-5 micrograms of total RNA, double-stranded cDNA containing the T7-dT (24) promoter sequence (Operon Biotechnologies, Huntsville, AL) was generated. This cDNA was then used as a template for in vitro transcription single round amplification with biotin labels (Enzo BioArray HighYield RNA Transcript Labeling Kit from Affymetrix Inc, Santa Clara, CA). Biotinylated cRNA targets were purified using the Sample Cleanup Module and subsequently hybridized to human U133A GeneChips (Affymetrix Inc, Santa Clara, CA) according to the manufacturer's standard protocols. Affymetrix U133A GeneChips that contain 22,283 probe sets, represented by ten to twenty unique probe pairs (perfect match and its corresponding mismatch), which allow detection of 14,500 different genes and expressed sequence tags (ESTs). Arrays were scanned using a laser confocal scanner (Agilent). The samples were processed by the same team, at the same core facility, and were randomized between each array run. Raw data are deposited with GEO (www.ncbi.nlm.nih.gov/geo/).

Data analysis. For each Affymetrix U133A GENE CHIP® raw intensity data were normalized to the mean intensity of all measurements on that array and scaled to a target intensity value of 500 (TGT) in Affymetrix Microarray Suite 5.0. With the aid of GeneSpring software, version 7.2, the measurement for each gene per patient sample array was divided by the median of that gene's measurement from the cohort of healthy volunteers. A filter was applied based on Affymetrix flag calls: probe sets were

selected if "Present" in at least 75% of samples in either group (healthy controls or patients). This step insured a more reliable intensity measurement of the genes used in downstream analyses. Class comparison was performed using a non-parametric ranking statistical analysis test (Mann-Whitney) applied to the selected set of genes. In the vertical direction, hierarchical clusters of genes were generated using the Pearson correlation around zero, Genespring's standard correlation measure. Normalized gene expression data were examined with a nonparametric univariate analysis (Fisher's exact test) to identify genes potentially discriminating two different groups. A supervised learning algorithm, the K-Nearest Neighbors Method, was applied that assigned a sample to pre-defined classes in three steps: 1) identification of genes (observations) that have strong correlations to classes to be distinguished; 2) confirmation that identified genes distinguish pre-defined classes; and 3) validation with "unknown samples".

Identification of transcriptional modules. A total of 239 blood leukocyte gene expression profiles were generated using Affymetrix U133A&B GENECHIPS (>44K probe sets). Transcriptional data were obtained for 8 groups including Systemic Juvenile Idiopathic Arthritis, SLE, liver transplant recipients, melanoma patients, and patients with acute infections: *Escherichia coli*, *Staphylococcus aureus* and Influenza A. For each group, transcripts that were present in at least 50% of all conditions were segregated into 30 clusters (k-means clustering: clusters C1 through C30). The cluster assignment for each gene was recorded in a table and distribution patterns were compared among all the genes. Modules were selected using an iterative process, starting with the largest set of genes that belonged to the same cluster in all study groups (i.e. genes that were found in the same cluster in 8 of the 8 groups). The selection was then expanded from this core reference pattern to include genes with 7/8, 6/8 and 5/8 matches. The resulting set of genes formed a transcriptional module and was withdrawn from the selection pool. The process was then repeated starting with the second largest group of genes, progressively reducing the level of stringency. This analysis led to the identification of 4742 transcripts that were distributed among 28 modules. Each module is attributed a unique identifier indicating the round and order of selection (e.g., M3.1 was the first module identified in the third round of selection).

Analysis of significance patterns. Gene expression data were generated for PBMCs obtained from patients and healthy volunteers using Affymetrix HG-U133A GENECHIPS. P values were obtained for six reference datasets by comparing groups of patients to their respective healthy control groups (Mann-Whitney rank test). The groups were composed of patients with: 1) Systemic Lupus Erythematosus (SLE, 16 samples), 2) Influenza A (16 samples), 3) *Escherichia coli* (16 samples), 4) *Staphylococcus aureus* (16 samples), and 5) *Streptococcus pneumoniae* (14 samples); and 7) bone marrow transplant recipients undergoing graft versus host disease (GVHD, 12 samples). Control groups were also formed taking into account age, sex and project (10 samples in each group). Genes significantly changed ($p<0.01$) in the "study group" (Melanoma and Transplant) were divided in two sets: overexpressed versus control and underexpressed versus control. P-values of the genes forming the overexpressed set were obtained for the "reference groups" (SLE, GVHD and infections with influenza virus, *E. coli*, *S. aureus*,

S. pneumoniae). P-value data were processed with a gene expression analysis software program, GeneSpring, Version 7.2 (Agilent), which was used to perform hierarchical clustering and group genes based on significance patterns.

Example 2. Determination and Analysis of Patterns of Significance are used to identify ubiquitous and disease-specific gene expression signatures in patient peripheral blood leukocytes.

The use of gene expression microarrays in patient-based research creates new prospects for the discovery of diagnostic biomarkers and the identification of genes or pathways linked to pathogenesis. Gene expression signatures were generated from peripheral blood mononuclear cells isolated from over one hundred patients with conditions presenting a strong immunological component (patient with autoimmune, graft versus host and infectious diseases, as well as immunosuppressed transplant recipients). This dataset permitted the opportunity to carry out comparative analyses and define disease signatures in a broader context. It was found that nearly 20% of overlap between lists of genes significantly changed versus healthy controls in patients with Systemic Lupus Erythematosus (SLE) and acute influenza infection. Transcriptional changes of 22,283 probe sets were evaluated through statistical group comparison performed systematically for 7 diseases versus their respective healthy control groups. Patterns of significance were generated by hierarchical clustering of p-values. This "Patterns of Significance" approach led to the identification of a SLE-specific "diagnostic signature", formed by genes that did not change compared to healthy in the other 6 diseases. Conversely, "sentinel signatures" were characterized that were common to all 7 diseases. These findings allow for the use of blood leukocyte expression signatures for diagnostic and early disease detection.

Briefly, blood is a reservoir and migration compartment for immune cells exposed to infectious agents, allergens, tumors, transplants or autoimmune reactions. Leukocytes isolated from the peripheral blood of patients constitute an accessible source of clinically-relevant information and a comprehensive molecular phenotype of these cells can be obtained by microarray analysis. Gene expression microarrays have been extensively used in cancer research, and proof of principle studies analyzing Peripheral Blood Mononuclear Cell (PBMC) samples isolated from patients with Systemic Lupus Erythematosus (SLE) lead to a better understanding of mechanisms of disease onset and responses to treatment. Two main applications have been found for gene expression microarrays in the context of patient-based research: (1) the discovery of biomarkers and establishment of diagnosis / prognosis signatures (e.g. prediction of survival of breast cancer patients) (2) the identification of genes/pathways involved in pathogenesis, leading for instance to the discovery of the role of interleukin-1 in the pathogenesis of systemic onset juvenile idiopathic arthritis. However, the analysis of microarray data still constitutes a considerable challenge. The ability to simultaneously acquire data for tens of thousands of features in a single test is one of the most appealing characteristic of microarrays, but it can also be a major shortcoming7. This 'curse of dimensionality' is compounded by the fact that the numbers of samples analyzed is usually small. The imbalance between the numbers of genes and conditions analyzed considerably weakens data interpretation capabilities. A microarray gene expression database was created that constitutes samples

obtained from patients with diseases that possess a strong immune component. The meta-analysis strategy of the present invention allows for the identification of ubiquitous as well as disease-specific signatures.

Processing of Blood Samples. Blood samples were collected by venipuncture and immediately delivered at room temperature to the Baylor Institute for Immunology Research, Dallas, TX, for processing. Peripheral blood mononuclear cells (PBMCs) from 3-4 ml of blood were isolated via Ficoll gradient and immediately lysed in RLT reagent (Qiagen, Valencia, CA) with beta-mercaptoethanol (BME) and stored at –80°C prior to the RNA extraction step.

Microarray analysis. Total RNA was isolated using the RNeasy kit (Qiagen, Valencia, CA) according to the manufacturer's instructions and RNA integrity was assessed by using an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA). Target labeling was performed according to the manufacturer's standard protocol (Affymetrix Inc, Santa Clara, CA). Biotinylated cRNA targets were purified and subsequently hybridized to Affymetrix HG-U133A GeneChips (22,283 probe sets). Arrays were scanned using an Affymetrix confocal laser scanner. Microarray Suite, Version 5.0 (MAS 5.0; Affymetrix) software was used to assess fluorescent hybridization signals, to normalize signals, and to evaluate signal detection calls. Normalization of signal values per chip was achieved using the MAS 5.0 global method of scaling to the target intensity value of 500 per GeneChip. A gene expression analysis software program, GeneSpring, Version 7.1 (Agilent), was used to perform statistical analysis, hierarchical clustering and classification of samples.

Development and Analysis of Patterns of Significance. Gene expression data were generated for PBMCs obtained from patients and healthy volunteers using Affymetrix HG-U133A GeneChips that were run on the same Affymetrix system, using standard operating procedures. P values were obtained by comparing 7 groups of patients to their respective healthy control groups (Mann-Whitney rank test). The groups were composed of pediatric patients with: 1) Systemic Lupus Erythomatosus (SLE, 16 samples), 2) Influenza A (16 samples), 3) *Staphylococcus aureus* (16 samples), 4) *Escherichia coli* (16 samples) and 5) *Streptococcus pneumoniae* (14 samples); as well as adult transplant recipients: 6) Liver transplant patients that have accepted the graft under immunosuppressive therapy (16 samples) and 7) bone marrow transplant recipients undergoing graft versus host disease (GVHD, 12 samples). Control groups were also formed taking into account age, sex and project (10 samples in each group). Genes significantly changed ($p<0.01$) in the "study group" (Influenza A and/or SLE) were divided in two sets: over-expressed versus control and under-expressed versus control. P-values of the genes forming the over-expressed set were obtained for the "reference groups" (infections with *E .coli, S. aureus, S. pneumoniae*, Liver transplant recipients and graft versus host disease). P-values of the reference groups were set to 1 when genes were under-expressed. The same procedure was used in the set of genes under-expressed in study group, only this time P-values of the reference group were set to 1 when genes were over-expressed. P-value data were processed with a gene expression analysis software program, GeneSpring, Version 7.1 (Agilent), that was used to perform hierarchical clustering and group genes based on significance patterns.

Identification of blood leukocytes transcriptional signatures associated with acute Influenza A infection and SLE. Microarray gene expression data obtained from pediatric patients with either SLE or acute influenza A infections were used to identify transcriptional signatures characteristic of these two diseases (Figure 20). Statistical comparison of a similar number of patients (18 samples) to their respective control group (10 samples) identified: (1) 1826 differentially expressed genes that formed the Influenza signature (of those 703 were over-expressed (red) relative to controls and 1123 were under-expressed (blue), see Figure (20A); 2) 3382 differentially expressed genes formed the SLE signature (of those 1019 were over-expressed relative to controls and 2363 were under-expressed, see Figure 20B).

Figure 20 shows a statistical group comparison between patients and their respective controls. Figure 20A. Microarray expression obtained for PBMC isolated from 16 children with acute Influenza A infection (FLU) and 10 healthy volunteers (HV) were compared (Mann-Whitney rank test, $p<0.01$). Out of 1826 differentially expressed genes, 703 were over-expressed and 1123 under-expressed in patients. Figure 20B. An equivalent number of children with Systemic Lupus Erythomatosus (SLE) were compared to their respective set of 10 healthy volunteers (HV) (Mann-Whitney rank test, $p<0.01$). Out of 3382 differentially expressed genes, 1019 were over-expressed and 2363 under-expressed in patients. Figure 20C. Comparison of over-expressed and under-expressed gene lists obtained for SLE and FLU samples relative to their respective control groups (healthy volunteers).

Transformed expression levels are indicated by color scale, with red representing relatively high expression and blue indicating relatively low expression compared to the median expression for each gene across all donors.

Analysis of significance patterns. Next, the specificity of these signatures for each disease was determined. A substantial overlap was found between the sets of genes that were differentially expressed in FLU and SLE (Figure 20C), with 279 over-expressed and 490 under-expressed genes common to both diseases (19% and 16% of similarities, respectively). This observation was used to determine whether a specific disease signature could be obtained in the context of a broader set of diseases.

In order to address this question the analysis was extended to PBMC transcriptional datasets obtained for patients with acute infections caused by bacteria (*E. coli, S. aureus and S. pneumoniae*) as well as transplant recipients (liver recipients who have accepted the allograft under pharmacological immunosuppressive therapy and bone marrow recipients with graft versus host disease). Patterns of significance were analyzed for genes that were specifically over-expressed in SLE compared to Influenza (Figure 1, 740 genes). This approach allowed the visualization of the significance of changes in levels of gene expression for each disease compared to its respective control group (age and sex matched healthy volunteers). Genes were arranged according to significance patterns by hierarchical clustering.

Of the 4 patterns identified, 2 were found to be largely specific to SLE (Figure 21: P1 – 98 genes, and P3 – 193 genes). In conclusion, the method was used to identify sets of genes, particularly among P3, that displayed a high degree of specificity for SLE when compared to 6 other diseases.

Figure 21 is an analysis of patterns of significance for genes over-expressed in SLE patients but not in patients with acute Influenza A infection. The genes used for this analysis were significantly over-expressed in patients with SLE compared to their respective control group (Mann-Whitney $P<0.05$) and not in patients with acute influenza A infection were selected for this analysis (740 genes). P values were obtained for five additional groups of patients: *E. coli, S. aureus, S. pneumoniae*, Liver transplant recipients and patients with graft vs host disease. The values were imported into a microarray data analysis software package (see methods for details). Four patterns were identified: SLE-1 to 4. Significance levels are indicated by color scale, with darker green representing lower P-values and white indicating a P-value of 1.

Identification of a common disease signature. An important proportion of genes from Figure 21 were induced ubiquitously (P2 – 222 genes and P4 – 225 genes). This finding suggests that these different diseases may share common transcriptional components in the blood constituting a "sickness" signature. In order to investigate this possibility sets of genes were analyzed that were shared between Influenza and SLE signatures (Figure 20C: 279 genes over-expressed, and 490 under-expressed).

Figure 22 shows Patterns of Significance for genes common to Influenza A and SLE. Genes overexpressed (left panel, OVER) and underexpressed (right panel, UNDER) in both patients with Influenza A (FLU) and SLE were examined in the context of other diseases: acute infections with *E. coli, S. aureus, S. pneumoniae*, liver transplant recipients (transplant) and bone marrow recipients with graft versus host disease (GVHD). Significance levels are indicated by color scale, with dark green representing lower P-values and white indicating a P-value of 1.

Patterns of significance were generated for these genes across all 7 diseases as described above. Three subsets were identified among the genes that were over-expressed in patients with Influenza A infection and SLE: one changing in most diseases, another presenting significant differences in all diseases, while the third was more specific to Influenza and SLE (Figure 22A, respectively P1, P2 and P3). Equivalent patterns can be found upon analysis of a set of under-expressed genes common to Influenza and SLE (Figure 22B, P4-7). Interestingly, the group of patient with significance patterns that were the most similar to Influenza and SLE had Graft Versus Host Disease. The parallelism was particularly striking for the set of under-expressed genes (Figure 22B).

Functional analysis of significance patterns. Finally, functional annotations associated to the patterns identified on Figure 22 were extracted. Genes associated with "defense response" were preferentially found in two patterns (P2-3 on Figures 3 & 4; Fisher's test for over-representation of this functional category: $p<0.0005$). These genes were expressed at higher levels compared to healthy. The list includes Defensin alpha 3, Azurocidin 1, Stabilin 1 (P2); the tumor necrosis factor family member TRAIL, and Galectin 3 binding protein (P3). Conversely under-expressed genes belonging to patterns P4-6 were preferentially associated with "structural constituent of ribosome" (Fisher's test for over-representation of this functional category in P4-6: $p<0.0001$). These genes include multiple ribosomal protein family

members (e.g. RPS10, RPL37, and RPL13). Genes belonging to the set of over-expressed genes the most specific to Influenza and SLE (P3) were preferentially associated to "interferon response" ($p<0.0001$, e.g. myxovirus resistance 1, interferon alpha-inducible protein 16, double stranded RNA inducible protein kinase), while genes in P1 were uniquely associated to "heavy metal binding" ($p<0.0001$, reflecting an overabundance of members of the metallothionein family).

Figure 23 is a functional analysis of genes shared by patients with Influenza infection and Lupus grouped according to significance patterns. Sets of genes forming the different patterns indicated on Figure 22 (P1-7) were subjected to functional analyses. The histograms indicate the percentage of genes associated to a given annotation for each of the sets. Over-expressed genes = red, under-expressed = blue. P1 n= 71 genes; P2 n=118; P3 n=85; P4 n=117; P5 n=184; P6 n=120; P7 n=46.

The comparative analysis of PBMC transcriptional patterns identified disease-specific as well as ubiquitous expression signatures. Different degrees of disease specificity were observed among the genes found to be common between the transcriptional profiles of PBMCs obtained from patients with Influenza infection and SLE. Differences in significance patterns were translated into distinct functional associations. Indeed, the genes that were most specific to Influenza and SLE relative to 5 other diseases were the most strongly associated to biological themes such as: "Interferon induction" (over-expressed genes; Figures 22 and 23: P3) or "structural constituent of ribosome" (under-expressed genes; Figures 22 and 24: P4). These observations permit to validate the relevance of this approach. This analysis facilitates the interpretation of microarray data by placing disease signatures in a much broader context.

In addition to contributing to a better understanding of disease processes the meta-analysis of PBMC transcriptional datasets has important implications for clinical diagnostic with: (1) the identification of discriminatory disease-specific signatures; as the screening of tens of thousands of potential markers will in most cases permit to pinpoint a limited number of transcripts that uniquely characterize a disease; and (2) the identification of a sentinel signature; as sets of genes for which expression changes in a wide range of health disorders could potentially be used in a screening assay for early disease detection.

Patients with metastatic melanoma display an endogenous transcriptional signature of immunosuppression similar to that induced by pharmacological treatments in patients who underwent liver transplant. The present invention provides a method and apparatus to identify patients at high risk of melanoma progression. In addition, the present invention also provides a method and apparatus for monitoring indicators of immunosuppression could help adjusting the dosage of immunosuppressive drugs and balance risks of rejection and side effects for liver transplant recipients.

It will be understood that particular embodiments described herein are shown by way of illustration and not as limitations of the invention. The principal features of this invention can be employed in various embodiments without departing from the scope of the invention. Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, numerous equivalents to the specific

procedures described herein. Such equivalents are considered to be within the scope of this invention and are covered by the claims.

All publications and patent applications mentioned in the specification are indicative of the level of skill of those skilled in the art to which this invention pertains. All publications and patent applications 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.

All of the compositions and/or methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this invention have been described in terms of preferred embodiments, it will be apparent to those of skill in the art that variations may be applied to the compositions and/or methods and in the steps or in the sequence of steps of the method described herein without departing from the concept, spirit and scope of the invention. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the invention as defined by the appended claims.

REFERENCES

Agostini, L., Martinon, F., Burns, K., McDermott, M. F., Hawkins, P. N., and Tschopp, J. (2004). NALP3 forms an IL-1beta-processing inflammasome with increased activity in Muckle-Wells autoinflammatory disorder. Immunity 20, 319-325.

Barrett, W. L., First, M. R., Aron, B. S., and Penn, I. (1993). Clinical course of malignancies in renal transplant recipients. Cancer 72, 2186-2189.

Berrebi, D., Bruscoli, S., Cohen, N., Foussat, A., Migliorati, G., Bouchet-Delbos, L., Maillot, M. C., Portier, A., Couderc, J., Galanaud, P., et al. (2003). Synthesis of glucocorticoid-induced leucine zipper (GILZ) by macrophages: an anti-inflammatory and immunosuppressive mechanism shared by glucocorticoids and IL-10. Blood 101, 729-738.

Bordea, C., Wojnarowska, F., Millard, P. R., Doll, H., Welsh, K., and Morris, P. J. (2004). Skin cancers in renal-transplant recipients occur more frequently than previously recognized in a temperate climate. Transplantation 77, 574-579.

Carroll, R. P., Ramsay, H. M., Fryer, A. A., Hawley, C. M., Nicol, D. L., and Harden, P. N. (2003). Incidence and prediction of nonmelanoma skin cancer post-renal transplantation: a prospective study in Queensland, Australia. Am J Kidney Dis 41, 676-683.

Chaussabel, D., and Sher, A. (2002). Mining microarray expression data by literature profiling. Genome Biol 3, RESEARCH0055.

Choi, B. M., Pae, H. O., Jeong, Y. R., Kim, Y. M., and Chung, H. T. (2005). Critical role of heme oxygenase-1 in Foxp3-mediated immune suppression. Biochem Biophys Res Commun 327, 1066-1071.

Corradetti, M. N., Inoki, K., and Guan, K. L. (2005). The stress-inducted proteins RTP801 and RTP801L are negative regulators of the mammalian target of rapamycin pathway. J Biol Chem 280, 9769-9772.

D'Adamio, F., Zollo, O., Moraca, R., Ayroldi, E., Bruscoli, S., Bartoli, A., Cannarile, L., Migliorati, G., and Riccardi, C. (1997). A new dexamethasone-induced gene of the leucine zipper family protects T lymphocytes from TCR/CD3-activated cell death. Immunity 7, 803-812.

Gabrilovich, D. (2004). Mechanisms and functional significance of tumour-induced dendritic-cell defects. Nat Rev Immunol 4, 941-952.

Gerlini, G., Romagnoli, P., and Pimpinelli, N. (2005). Skin cancer and immunosuppression. Crit Rev Oncol Hematol 56, 127-136.

Jachimczak, P., Apfel, R., Bosserhoff, A. K., Fabel, K., Hau, P., Tschertner, I., Wise, P., Schlingensiepen, K. H., Schuler-Thurner, B., and Bogdahn, U. (2005). Inhibition of immunosuppressive effects of melanoma-inhibiting activity (MIA) by antisense techniques. Int J Cancer 113, 88-92.

Kovanen, P. E., Rosenwald, A., Fu, J., Hurt, E. M., Lam, L. T., Giltnane, J. M., Wright, G., Staudt, L. M., and Leonard, W. J. (2003). Analysis of gamma c-family cytokine target genes. Identification of dual-specificity phosphatase 5 (DUSP5) as a regulator of mitogen-activated protein kinase activity in interleukin-2 signaling. J Biol Chem 278, 5205-5213.

Lee, J. H., Torisu-Itakara, H., Cochran, A. J., Kadison, A., Huynh, Y., Morton, D. L., and Essner, R. (2005a). Quantitative analysis of melanoma-induced cytokine-mediated immunosuppression in melanoma sentinel nodes. Clin Cancer Res 11, 107-112.

Lee, Y. R., Yang, I. H., Lee, Y. H., Im, S. A., Song, S., Li, H., Han, K., Kim, K., Eo, S. K., and Lee, C. K. (2005b). Cyclosporin A and tacrolimus, but not rapamycin, inhibit MHC-restricted antigen presentation pathways in dendritic cells. Blood.

Liyanage, U. K., Moore, T. T., Joo, H. G., Tanaka, Y., Herrmann, V., Doherty, G., Drebin, J. A., Strasberg, S. M., Eberlein, T. J., Goedegebuure, P. S., and Linehan, D. C. (2002). Prevalence of regulatory T cells is increased in peripheral blood and tumor microenvironment of patients with pancreas or breast adenocarcinoma. J Immunol 169, 2756-2761.

Monti, P., Leone, B. E., Zerbi, A., Balzano, G., Cainarca, S., Sordi, V., Pontillo, M., Mercalli, A., Di Carlo, V., Allavena, P., and Piemonti, L. (2004). Tumor-derived MUC1 mucins interact with differentiating monocytes and induce IL-10highIL-12low regulatory dendritic cell. J Immunol 172, 7341-7349.

Powell, J. D., Lerner, C. G., Ewoldt, G. R., and Schwartz, R. H. (1999). The -180 site of the IL-2 promoter is the target of CREB/CREM binding in T cell anergy. J Immunol 163, 6631-6639.

Puente Navazo, M. D., Valmori, D., and Ruegg, C. (2001). The alternatively spliced domain TnFnIII A1A2 of the extracellular matrix protein tenascin-C suppresses activation-induced T lymphocyte proliferation and cytokine production. J Immunol 167, 6431-6440.

Seimiya, M., Wada, A., Kawamura, K., Sakamoto, A., Ohkubo, Y., Okada, S., Hatano, M., Tokuhisa, T., Watanabe, T., Saisho, H., et al. (2004). Impaired lymphocyte development and function in Clast5/Stra13/DEC1-transgenic mice. Eur J Immunol 34, 1322-1332.

Soares, M. P., Lin, Y., Anrather, J., Csizmadia, E., Takigami, K., Sato, K., Grey, S. T., Colvin, R. B., Choi, A. M., Poss, K. D., and Bach, F. H. (1998). Expression of heme oxygenase-1 can determine cardiac xenograft survival. Nat Med 4, 1073-1077.

Theodosiou, A., Smith, A., Gillieron, C., Arkinstall, S., and Ashworth, A. (1999). MKP5, a new member of the MAP kinase phosphatase family, which selectively dephosphorylates stress-activated kinases. Oncogene 18, 6981-6988.

Viguier, M., Lemaitre, F., Verola, O., Cho, M. S., Gorochov, G., Dubertret, L., Bachelez, H., Kourilsky, P., and Ferradini, L. (2004). Foxp3 expressing CD4+CD25(high) regulatory T cells are overrepresented in human metastatic melanoma lymph nodes and inhibit the function of infiltrating T cells. J Immunol 173, 1444-1453.

Winoto, A., and Littman, D. R. (2002). Nuclear hormone receptors in T lymphocytes. Cell 109 Suppl, S57-66.

Woltman, A. M., van der Kooij, S. W., Coffer, P. J., Offringa, R., Daha, M. R., and van Kooten, C. (2003). Rapamycin specifically interferes with GM-CSF signaling in human dendritic cells, leading to apoptosis via increased p27KIP1 expression. Blood 101, 1439-1445.

Xu, X., Su, B., Barndt, R. J., Chen, H., Xin, H., Yan, G., Chen, L., Cheng, D., Heitman, J., Zhuang, Y., et al. (2002). FKBP12 is the only FK506 binding protein mediating T-cell inhibition by the immunosuppressant FK506. Transplantation 73, 1835-1838.

What is claimed is:

1. A method of identifying a subject with melanoma comprising:

determining a dataset that comprises the level of expression of one or more melanoma expression vectors; and

displaying each of the melanoma expression vectors with a separate identifier.

2. The method of claim 1, wherein the one or more melanoma expression vectors comprise three or more genes selected from Table 2, Table 8, Table 9, Table 12 or a combination thereof.

3. The method of claim 1, wherein the six or more genes disposed on a microarray, the genes selected from: RNA-binding region (RNP1, RRM) containing 1; retinoblastoma binding protein 6; GrpE-like 1, mitochondrial (*E. coli*); pellino homolog 1 (Drosophila); ring finger protein 10; hypothetical protein LOC90637; Enhancer of polycomb homolog 1 (Drosophila); Full length insert cDNA clone ZB81B12; PTK9 protein tyrosine kinase 9; cell division cycle associated 4; period homolog 1 (Drosophila); zinc finger protein 237; TCDD-inducible poly(ADP-ribose) polymerase; cyclin L1; TSPY-like 2; glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A; KIAA0863 protein; v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian); PRO0149 protein; protein tyrosine phosphatase type IVA, member 1; serine/threonine kinase receptor associated protein; purine-rich element binding protein B; phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila); estrogen receptor binding protein; stress 70 protein chaperone, microsome-associated, 60kDa; BTG family, member 3; transferrin receptor (p90, CD71); hypothetical protein FLJ20436; phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila); KIAA0063 gene product; Vesicle transport through interaction with t-SNAREs homolog 1A (yeast); LOC440309; SNF1-like kinase; dual specificity phosphatase 8; RNA-binding region (RNP1, RRM) containing 2; dynein light chain 2; membrane associated DNA binding protein; serine/threonine kinase 17b (apoptosis-inducing); tumor necrosis factor, alpha-induced protein 3; heterogeneous nuclear ribonucleoprotein C (C1/C2); Phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila); kelch-like 15 (Drosophila); KIAA0863 protein; activating transcription factor 4 (tax-responsive enhancer element B67); villin 2 (ezrin); small nuclear ribonucleoprotein polypeptide A'; zinc finger protein 198; cAMP responsive element modulator; splicing factor 3a, subunit 1, 120kDa; CD83 antigen (activated B lymphocytes, immunoglobulin superfamily); DnaJ (Hsp40) homolog, subfamily B, member 6; chromosome 20 open reading frame 111; ethanolamine kinase 1; chromosome 4 open reading frame 15; ring finger protein 139; ribonuclease H1; jun D proto-oncogene; S-phase kinase-associated protein 1A (p19A); Methyl CpG binding protein 2 (Rett syndrome); heterogeneous nuclear ribonucleoprotein H3 (2H9); hypothetical protein DKFZp586I1420; nucleosome assembly protein 1-like

5; Ras-GTPase activating protein SH3 domain-binding protein 2; Zinc finger protein 331; Cullin 1; Shwachman-Bodian-Diamond syndrome; sterol-C5-desaturase (ERG3 delta-5-desaturase homolog, fungal)-like; phosphatidylinositol glycan, class A (paroxysmal nocturnal hemoglobinuria); isopentenyl-diphosphate delta isomerase; chromosome 20 open reading frame 67; v-maf musculoaponeurotic fibrosarcoma oncogene homolog K (avian); Wiskott-Aldrich syndrome protein interacting protein; kelch-like 7 (Drosophila); tumor necrosis factor, alpha-induced protein 3; hypothetical LOC388796; armadillo repeat containing 8; Membrane associated DNA binding protein; zinc finger protein 331; cAMP responsive element modulator; putative translation initiation factor; DnaJ (Hsp40) homolog, subfamily B, member 6; nuclear receptor subfamily 4, group A, member 2; cylindromatosis (turban tumor syndrome); nuclear receptor subfamily 4, group A, member 2; 6-pyruvoyltetrahydropterin synthase; activator of S phase kinase; activated RNA polymerase II transcription cofactor 4 (related to Activated RNA polymerase II transcriptional coactivator p15 (Positive cofactor 4) (PC4) (p14)); Hypothetical LOC388796; junction-mediating and regulatory protein; Hypothetical gene CG018; putative translation initiation factor; dual specificity phosphatase 5; Transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila); mitochondrial carrier protein; protein tyrosine phosphatase type IVA, member 1; zinc finger protein 331; CDNA clone IMAGE:30332316, partial cds; interleukin enhancer binding factor 3, 90kDa; Homo sapiens, clone IMAGE:4753714, mRNA; SON DNA binding protein; AP1 gamma subunit binding protein 1; zinc finger protein 394; eukaryotic translation initiation factor 5; Activating transcription factor 7 interacting protein 2; hypothetical protein LOC285831; actin related protein 2/3 complex, subunit 5-like; hypothetical protein LOC144438; putative translation initiation factor; H3 histone, family 3B (H3.3B); kelch repeat and BTB (POZ) domain containing 2; CDNA FLJ40725 fis, clone TKIDN1000001, highly similar to Translocase of inner mitochondrial membrane 23; cylindromatosis (turban tumor syndrome); phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila); heat shock 70kDa protein 14; HSPC128 protein; RNA (guanine-7-) methyltransferase; iduronate 2-sulfatase (Hunter syndrome); hypothetical protein MGC29814; Casein kinase 1, epsilon; protein phosphatase 1, regulatory (inhibitor) subunit 16B; isopentenyl-diphosphate delta isomerase; RAD23 homolog B (S. cerevisiae); Phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila); zinc finger protein 295; hypothetical protein FLJ13149; and BTG family, member 3 and combinations thereof.

4. The method of claim 1, wherein the six or more genes disposed on a microarray, the genes selected from: WARS; IFI53; IFP53; GAMMA-2; FAM46C; FLJ20202; H3F3B; H3.3B; FOXK2; ILF; ILF1; ILF-1; DUSP5; HVH3; ARF6; DKFZp762C186; BRD2; NAT; RNF3; FSRG1; RING3; D6S113E; KIAA9001; RORA; ROR1; ROR2; ROR3; RZRA; NR1F1; DKFZp762C186; DNAJB1; SUI1; CXCR4; HM89; LAP3; NPYR; WHIM; LESTR; NPY3R; HSY3RR; NPYY3R; D2S201E;

GRINL1A; CTSB; TRIP-Br2; PDE4B; DPDE4; PDEIVB; PMAIP1; APR; NOXA; BTG2; PC3; TIS21; ASAHL; SON; SUI1; A121; ISO1; HERPUD1; SUP; Mif1; KIAA0025; DUSP2; PAC1; PAC-1; RNF139; RCA1; TRC8; HRCA1; MGC31961; TNFAIP3; A20; TNFA1P2; ARS2; HNRPL; hnRNP-L; P/OKcl.14; C20orf67; C20orf111; HSPC207; dJ1183I21.1; ZNF331; RITA; ZNF361; ZNF463; C20orf67; IER5; SBBI48; ; SUI1; JUN; AP1; CD69; TOB1; H3F3B; H3.3B; FOLR1; TNFAIP3; TCF8; BZP; ZEB; ZEB1; AREB6; ZFHEP; NIL-2A; ZFHX1A; NIL-2-A; DUSP10; MKP5; MKP-5; GGTLA4; MGC50550; dJ831C21.2; PMAIP1; ZC3HAV1; ZAP; FLB6421; ZC3HDC2; FLJ13288; MGC48898; DSIPI; DIP; GILZ; hDIP; TSC-22R; MCL1; TM; EAT; MCL1L; MCL1S; MGC1839; SH3TC1; FLJ20356; CIAS1; FCU; MWS; FCAS; NALP3; C1orf7; PYPAF1; AII/AVP; AGTAVPRL; SLC15A3; PHT2; PTR3; hPTR3; PTDSR; PSR; PTDSR1; KIAA0585; BHLHB2; DEC1; STRA13; Stra14; HMGE; KIAA0063; NR4A2; NOT; RNR1; HZF-3; NURR1; TINUR; NR4A2; NOT; RNR1; HZF-3; NURR1; TINUR; PTS; PTPS; HEAB; CLP1; hClp1; AREG; SDGF; CRDGF; MGC13647; EDG4; LPA2; EDG-4; LPAR2; CREM; ICER; MGC17881; MGC41893; CD83; BL11; HB15; ZNF394; FLJ12298 and combinations thereof.

5. The method of claim 1, wherein the one or more melanoma expression vectors comprise six or more genes over-expressed, under-expressed or combinations thereof selected from Table 12.

6. The method of claim 1, wherein the melanoma expression vectors comprises genes related to platelets, platelet glycoproteins, platelet-derived immune mediators, MHC/ribosomal proteins, MHC class I molecules, Beta 2-microglobulin, ribosomal proteins, hemoglobin genes or combinations thereof.

7. The method of claim 1, wherein the melanoma expression vectors comprises genes related to interferon-inducible genes, signaling molecules, kinases, RAS family members or combinations thereof.

8. The method of claim 1, wherein the level of expression of one or more melanoma expression vectors comprise the mRNA expression level, protein expression level or both mRNA expression level and protein expression level.

9. The method of claim 1, wherein the level of expression comprises a mRNA expression level and is quantitated by a method selected from the group consisting of polymerase chain reaction, real time polymerase chain reaction, reverse transcriptase polymerase chain reaction, hybridization, probe hybridization and gene expression array.

10. The method of claim 1 further comprising the step of detecting one or more polymorphisms in the one or more melanoma expression vectors.

11. The method of claim 1, wherein the level of expression is determined using at least one technique selected from the group consisting of polymerase chain reaction, heteroduplex analysis, single stand conformational polymorphism analysis, ligase chain reaction, comparative genome hybridization, Southern blotting, Northern blotting, Western blotting, enzyme-linked immunosorbent assay, fluorescent resonance energy-transfer and sequencing.

12. A computer implemented method for determining the phenotype of a sample comprising:

obtaining a plurality of sample probe intensities;

diagnosing melanoma based upon the plurality of sample probe intensities; and

calculating a linear correlation coefficient between the plurality of sample probe intensities and a reference probe intensities; and

accepting the tentative phenotype as the phenotype of the sample if the linear correlation coefficient is greater than a threshold value.

13. A computer readable medium comprising computer-executable instructions for performing the method for determining the phenotype of a sample comprising:

obtaining a plurality of sample probe intensities;

diagnosing melanoma based upon the sample probe intensities for two or more metastatic melanoma expression vectors selected from one or more the genes listed in Tables 9 to 14; and

calculating a linear correlation coefficient between the sample probe intensities and a reference probe intensity; and

accepting the tentative phenotype as the phenotype of the sample if the linear correlation coefficient is greater than a threshold value.

14. A method of identifying a subject with immunosuppression associated with transplants comprising:

determining the level of expression of one or more immunosuppression expression vectors; and

displaying each of the melanoma expression vectors with a separate identifier.

15. The method of claim 14, wherein the one or more immunosuppression expression vectors comprise three or more genes selected from Table 10, Table 11, Table 13 or a combination thereof.

16. The method of claim 14, wherein the level of expression of one or more immunosuppression associated expression vectors comprise the mRNA expression level, protein expression level or both mRNA expression level and protein expression level.

17. The method of claim 14, wherein the level of expression comprises a mRNA expression level and is quantitated by a method selected from the group consisting of polymerase chain reaction, real time polymerase chain reaction, reverse transcriptase polymerase chain reaction, hybridization, probe hybridization, and gene expression array.

18. The method of claim 14, further comprising the step of detecting one or more polymorphisms in the one or more immunosuppression associated expression vectors.

19. The method of claim 14, wherein the level of expression is determined using at least one technique selected from the group consisting of polymerase chain reaction, heteroduplex analysis, single stand conformational polymorphism analysis, ligase chain reaction, comparative genome hybridization, southern blotting, northern blotting, western blotting, enzyme-linked immunosorbent assay, fluorescent resonance energy-transfer and sequencing.

20. The method of claim 14, wherein the one or more immunosuppression associated expression vectors are derived from a leukocyte.

21. A computer implemented method for determining the propensity for immunosuppression in a sample comprising:

obtaining a plurality of sample probe intensities;

diagnosing immunosuppression based upon the sample probe intensities; and

calculating linear correlation coefficient between the sample probe intensities and reference probe intensities; and

accepting the tentative phenotype as the phenotype of the sample if the linear correlation coefficient is greater than a threshold value.

22. A microarray for identifying a human subject with melanoma comprising:

disposing four or more genes on a substrate selected from the group consisting of six or more genes selected from Table 2, Table 8, Table 9, Table 12 or a combination thereof.

23. A microarray for identifying a human subject predisposed to immunosuppression comprising:

disposing four or more genes on a substrate selected from the group consisting of six or more genes selected from Table 10, Table 11, Table 13 or a combination thereof.

24. A method for displaying transcriptome vector data comprising:

separating one or more genes into one or more modules to visually display an aggregate gene expression vector value for each of the modules; and

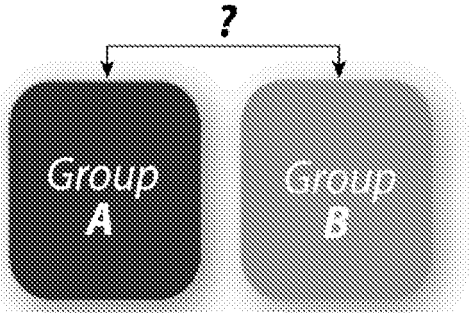
displaying the aggregate gene expression vector value for overexpression, underexpression or equal expression of the aggregate gene expression vector value in each module.

25. The method of claim 23, wherein overexpression is identified with a first identifier and underexpression is identified with a second identifier.

26. The method of claim 23, wherein overexpression is identified with a first identifier and underexpression is identified with a second identifier, wherein the first identifier is a first color and the second identifier is a second color, wherein first and second identifiers are superimposed to provide a combined color.

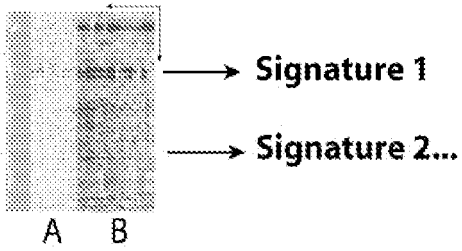
a. Gene-level analysis

I. Statistical testing:



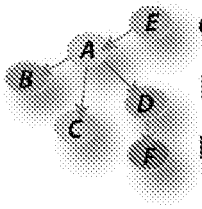
Class comparison/prediction

II. Pattern discovery



III. Functional annotation/analysis

Signature 1
Signature 2...

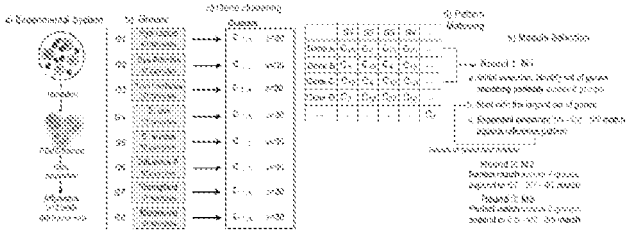


Gene Ontologies
Literature profiling
Pathway analysis

Figure 1a

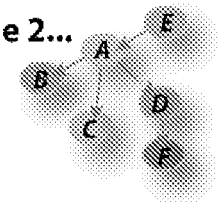
b. Module-level analysis

I. Module extraction algorithm



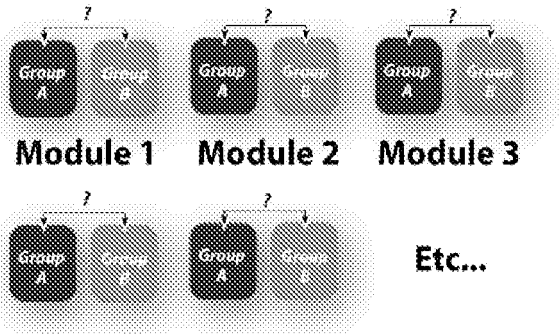
II. Functional annotation/analysis

Module 1
Module 2...



Gene Ontologies
Literature profiling
Pathway analysis

III. Statistical testing:



Etc...

Class comparison/prediction, module-by-module

IV. Visualization / Interpretation:

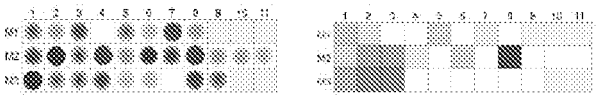


Figure 1b

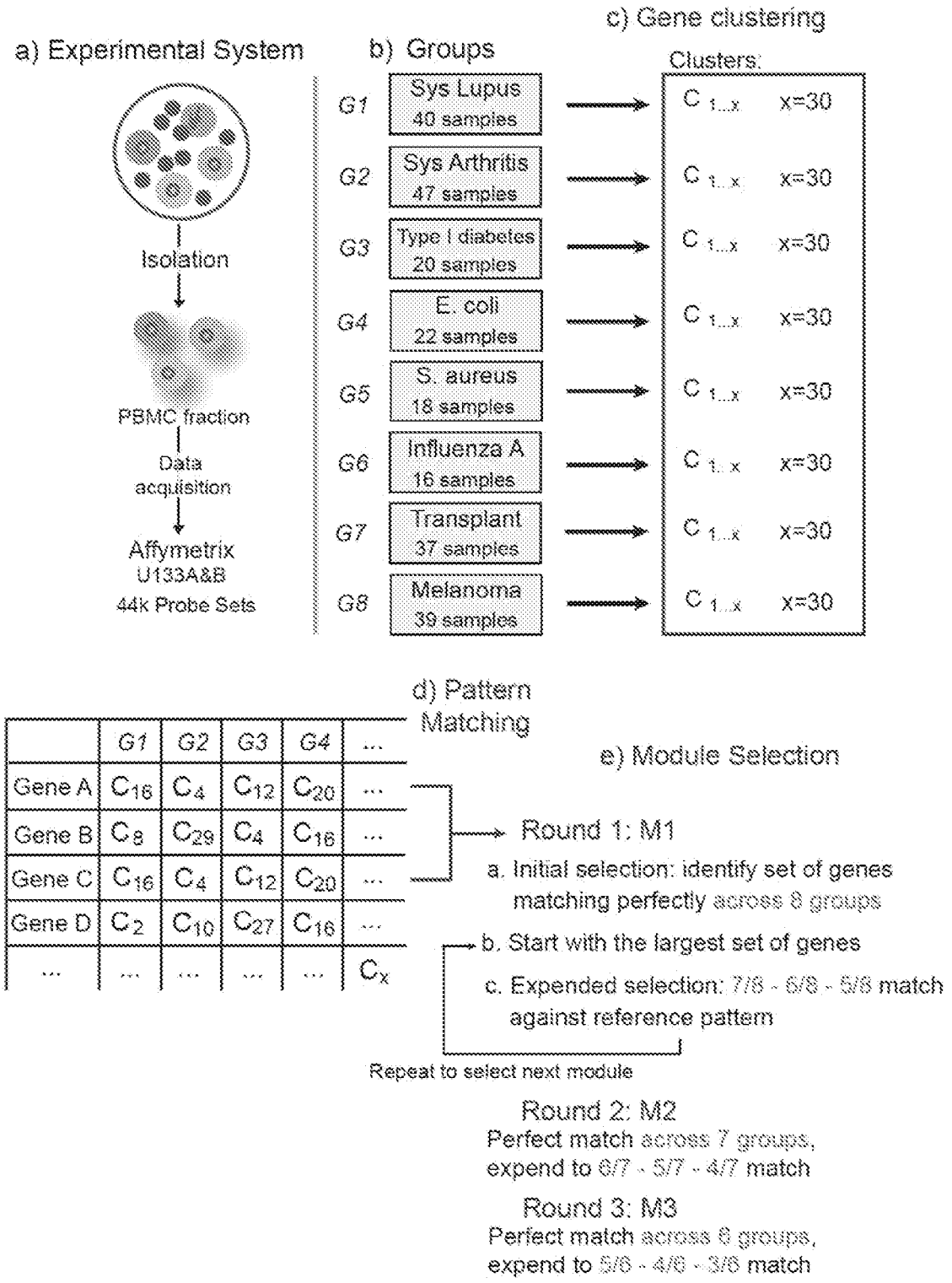

a) Experimental System
Isolation
PBMC fraction
Data acquisition
Affymetrix U133A&B 44k Probe Sets
b) Groups
G1 Sys Lupus 40 samples
G2 Sys Arthritis 47 samples
G3 Type I diabetes 20 samples
G4 E. coli 22 samples
G5 S. aureus 18 samples
G6 Influenza A 16 samples
G7 Transplant 37 samples
G8 Melanoma 39 samples
c) Gene clustering
Clusters:
C 1...x x=30
d) Pattern Matching
G1 G2 G3 G4 ...
Gene A C16 C4 C12 C20 ...
Gene B C8 C29 C4 C16 ...
Gene C C16 C4 C12 C20 ...
Gene D C2 C10 C27 C16 ...
... ... Cx
e) Module Selection
Round 1: M1
a. Initial selection: identify set of genes matching perfectly across 8 groups
b. Start with the largest set of genes
c. Expended selection: 7/8 - 6/8 - 5/8 match against reference pattern
Repeat to select next module
Round 2: M2
Perfect match across 7 groups, expend to 6/7 - 5/7 - 4/7 match
Round 3: M3
Perfect match across 6 groups, expend to 5/6 - 4/6 - 3/6 match

Figure 1c

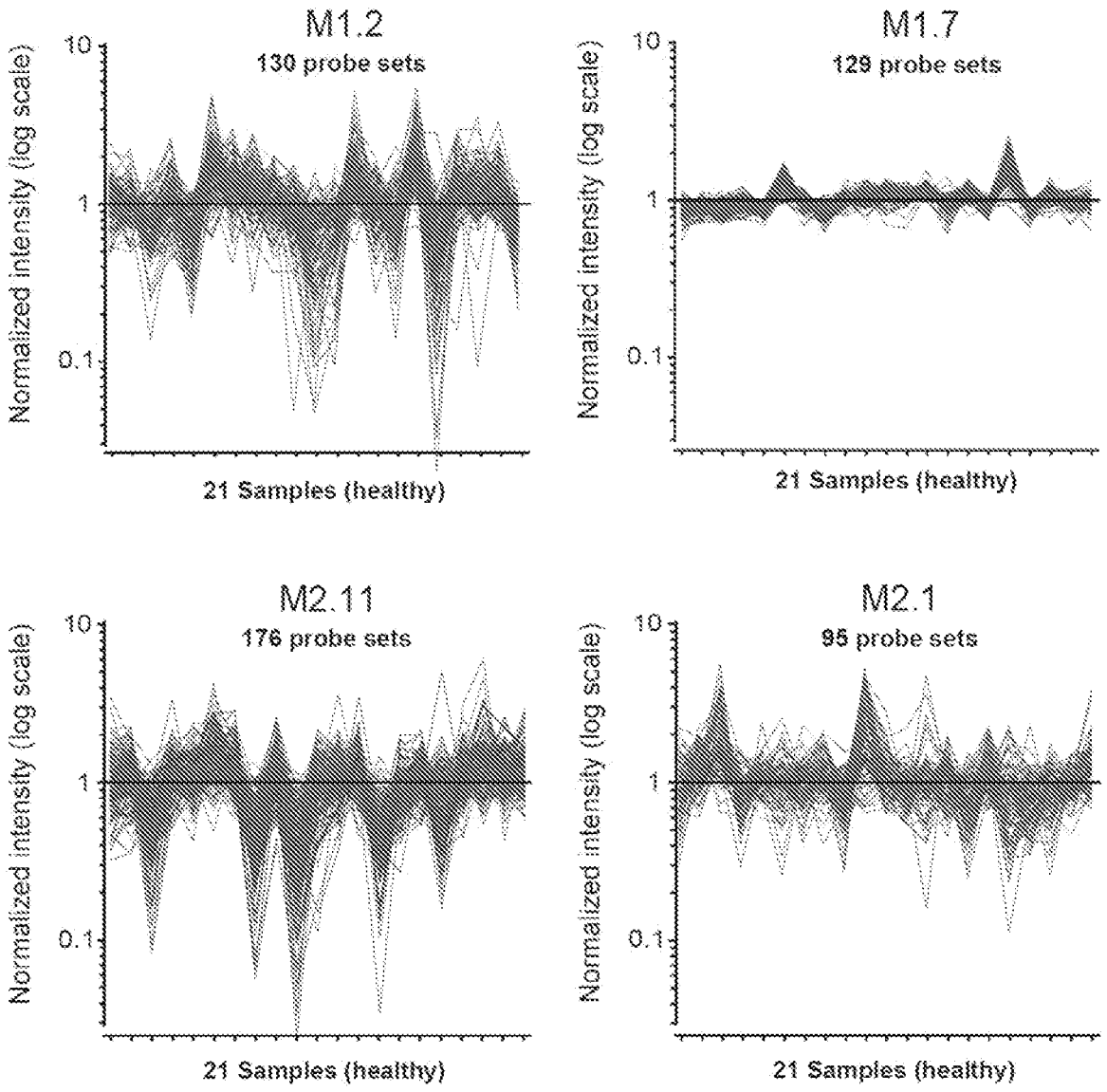

M1.2
130 probe sets
M1.7
129 probe sets
M2.11
176 probe sets
M2.1
95 probe sets
Normalized intensity (log scale)
10
1
0.1
21 Samples (healthy)

Figure 2

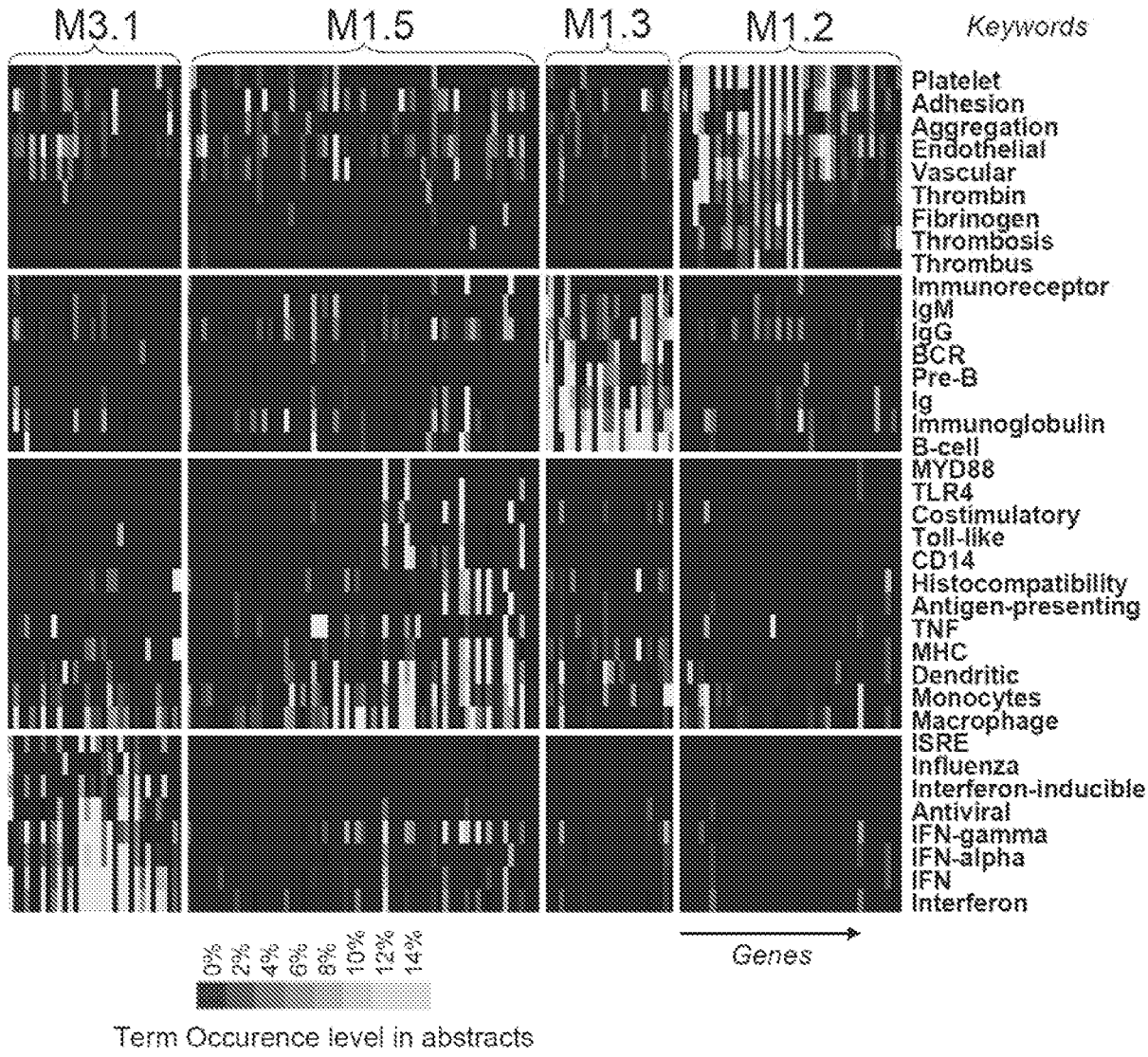

M3.1
M1.5
M1.3
M1.2
Keywords
Platelet
Adhesion
Aggregation
Endothelial
Vascular
Thrombin
Fibrinogen
Thrombosis
Thrombus
Immunoreceptor
IgM
IgG
BCR
Pre-B
Ig
Immunoglobulin
B-cell
MYD88
TLR4
Costimulatory
Toll-like
CD14
Histocompatibility
Antigen-presenting
TNF
MHC
Dendritic
Monocytes
Macrophage
ISRE
Influenza
Interferon-inducible
Antiviral
IFN-gamma
IFN-alpha
IFN
Interferon
0%
2%
4%
6%
8%
10%
12%
14%
Genes
Term Occurence level in abstracts

Figure 3

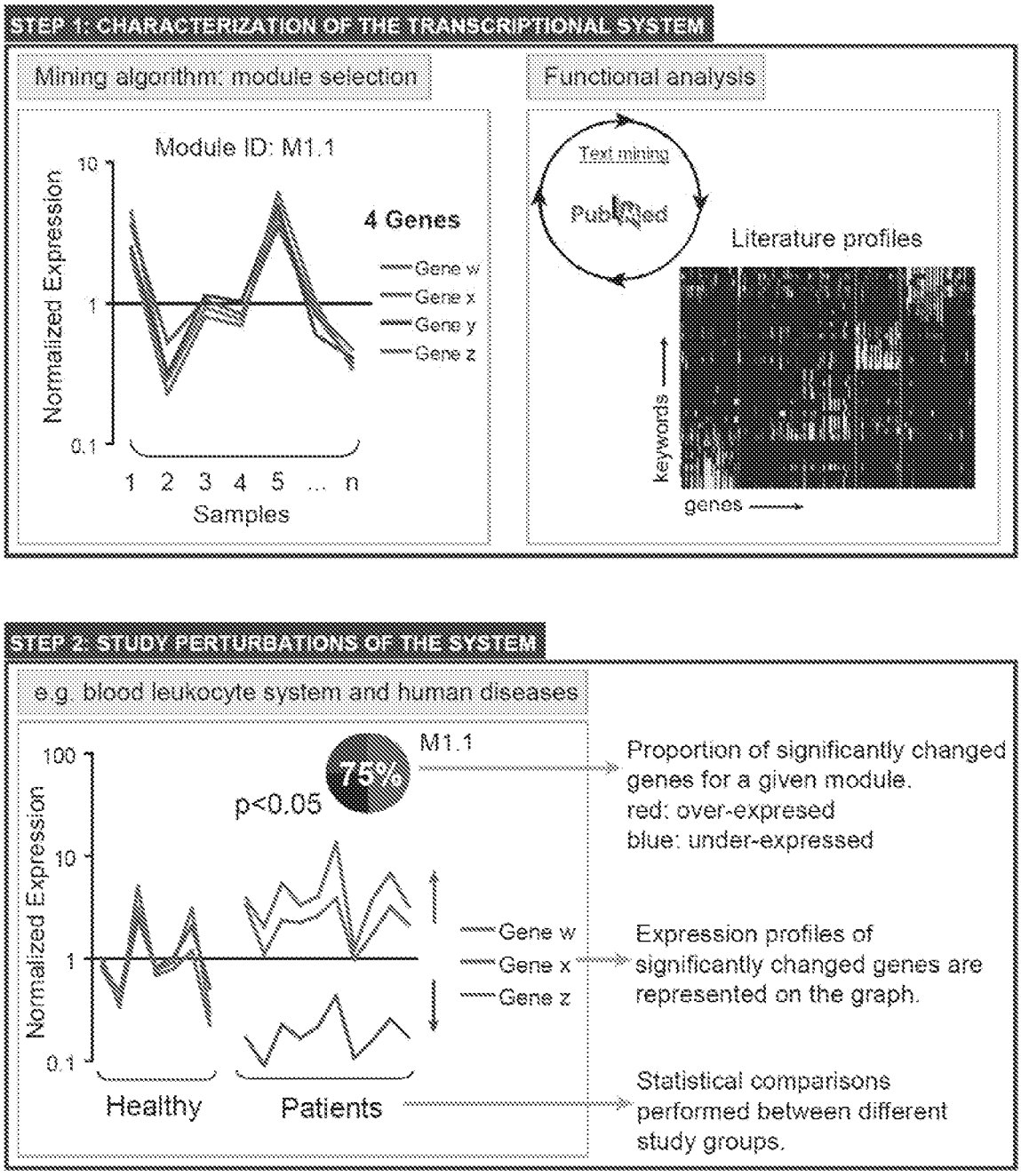

STEP 1: CHARACTERIZATION OF THE TRANSCRIPTIONAL SYSTEM
Mining algorithm: module selection
Module ID: M1.1
Normalized Expression
10
1
0.1
4 Genes
Gene w
Gene x
Gene y
Gene z
1 2 3 4 5 ... n
Samples
Functional analysis
Text mining
PubMed
Literature profiles
keywords
genes
STEP 2: STUDY PERTURBATIONS OF THE SYSTEM
e.g. blood leukocyte system and human diseases
100
10
1
0.1
Normalized Expression
M1.1
75%
p<0.05
Gene w
Gene x
Gene z
Healthy
Patients
Proportion of significantly changed genes for a given module.
red: over-expresed
blue: under-expressed
Expression profiles of significantly changed genes are represented on the graph.
Statistical comparisons performed between different study groups.

Figure 4

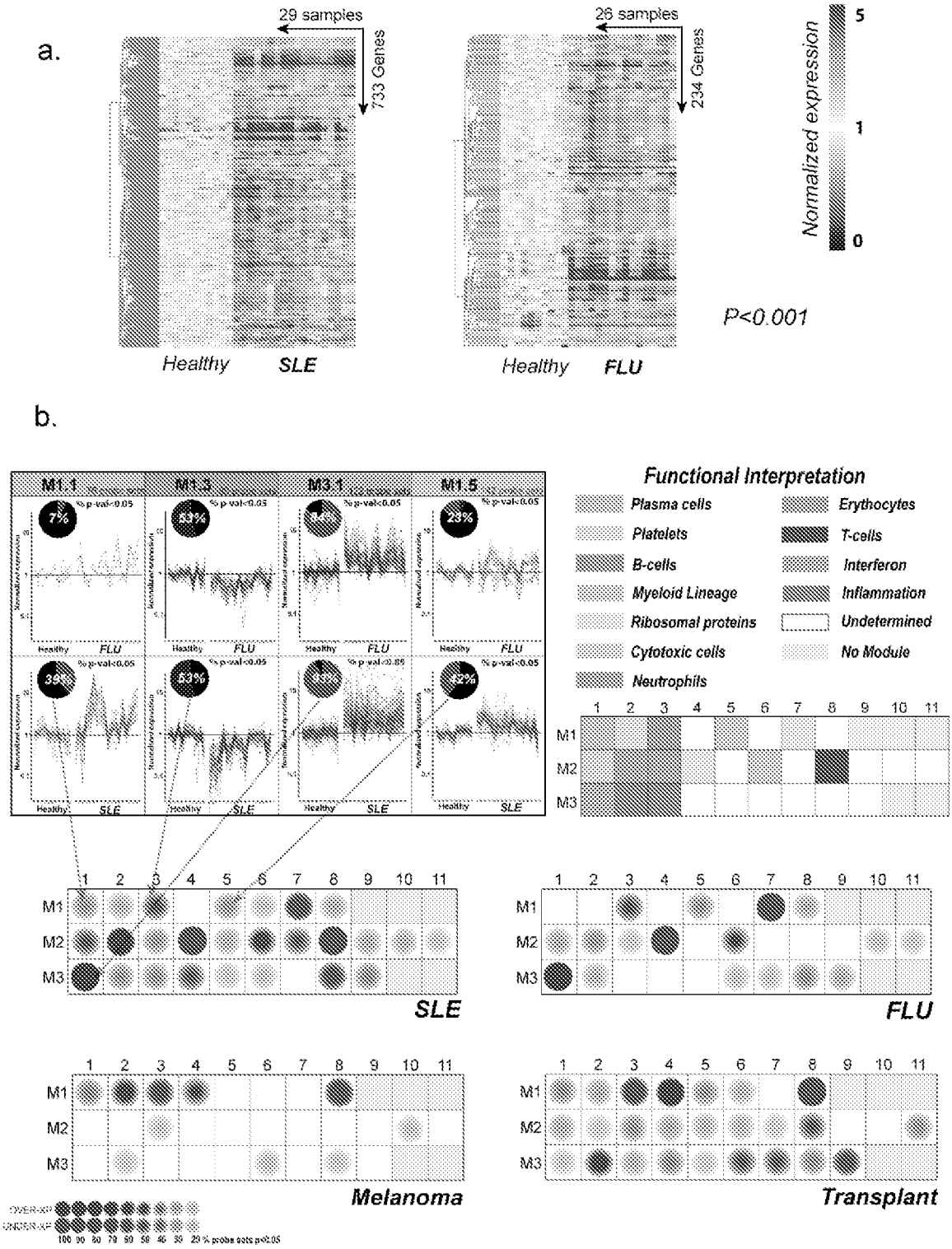

a.
29 samples
733 Genes
26 samples
234 Genes
Normalized expression
5
1
0
P<0.001
Healthy
SLE
Healthy
FLU
b.
M1.1
M1.3
M3.1
M1.5
Functional Interpretation
Plasma cells
Platelets
B-cells
Myeloid Lineage
Ribosomal proteins
Cytotoxic cells
Neutrophils
Erythocytes
T-cells
Interferon
Inflammation
Undetermined
No Module
M1
M2
M3
SLE
FLU
Melanoma
Transplant

Figure 5

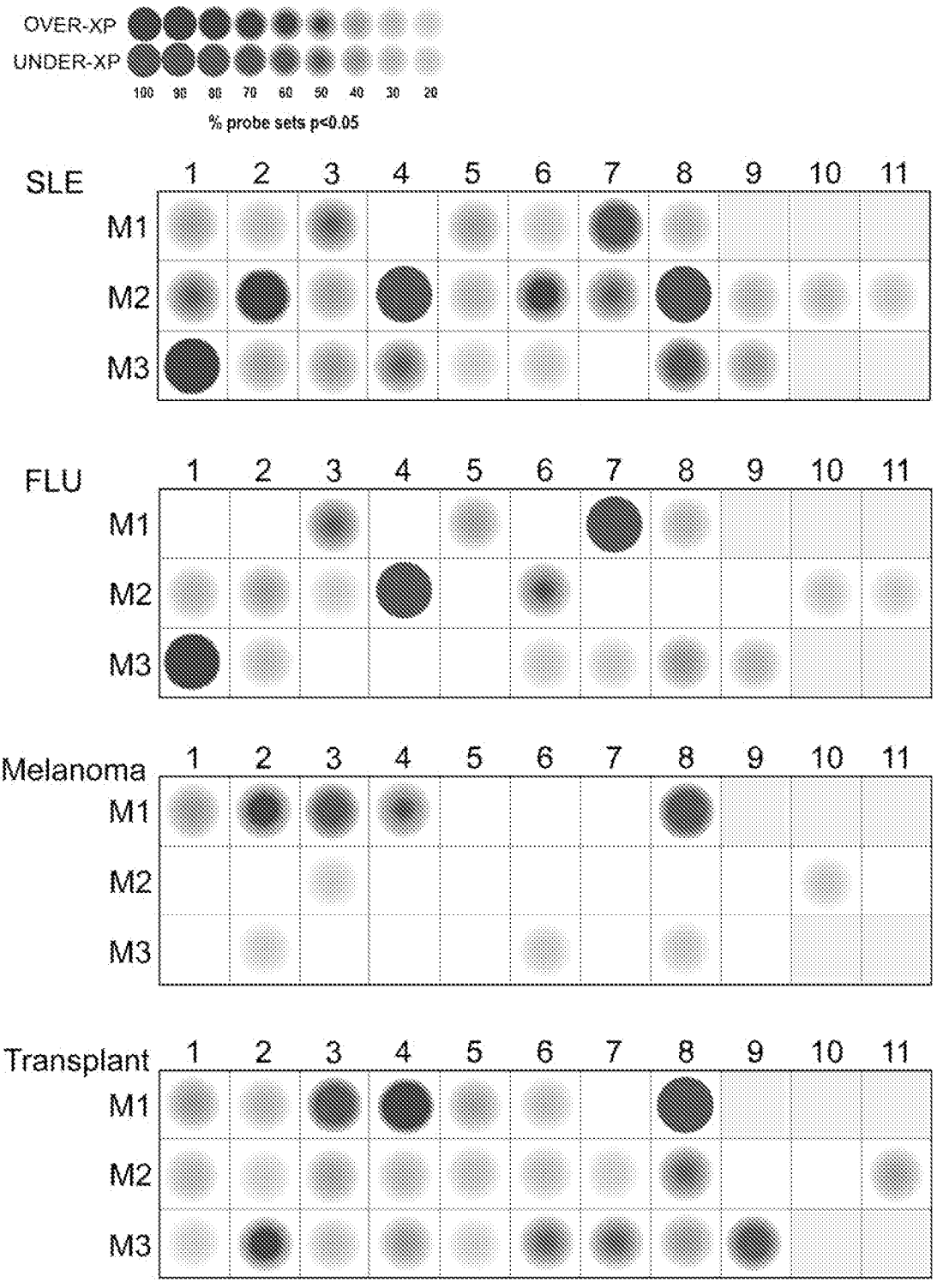

OVER-XP
UNDER-XP
100 90 80 70 60 50 40 30 20
% probe sets p<0.05
SLE
1 2 3 4 5 6 7 8 9 10 11
M1
M2
M3
FLU
1 2 3 4 5 6 7 8 9 10 11
M1
M2
M3
Melanoma
1 2 3 4 5 6 7 8 9 10 11
M1
M2
M3
Transplant
1 2 3 4 5 6 7 8 9 10 11
M1
M2
M3

Figure 6

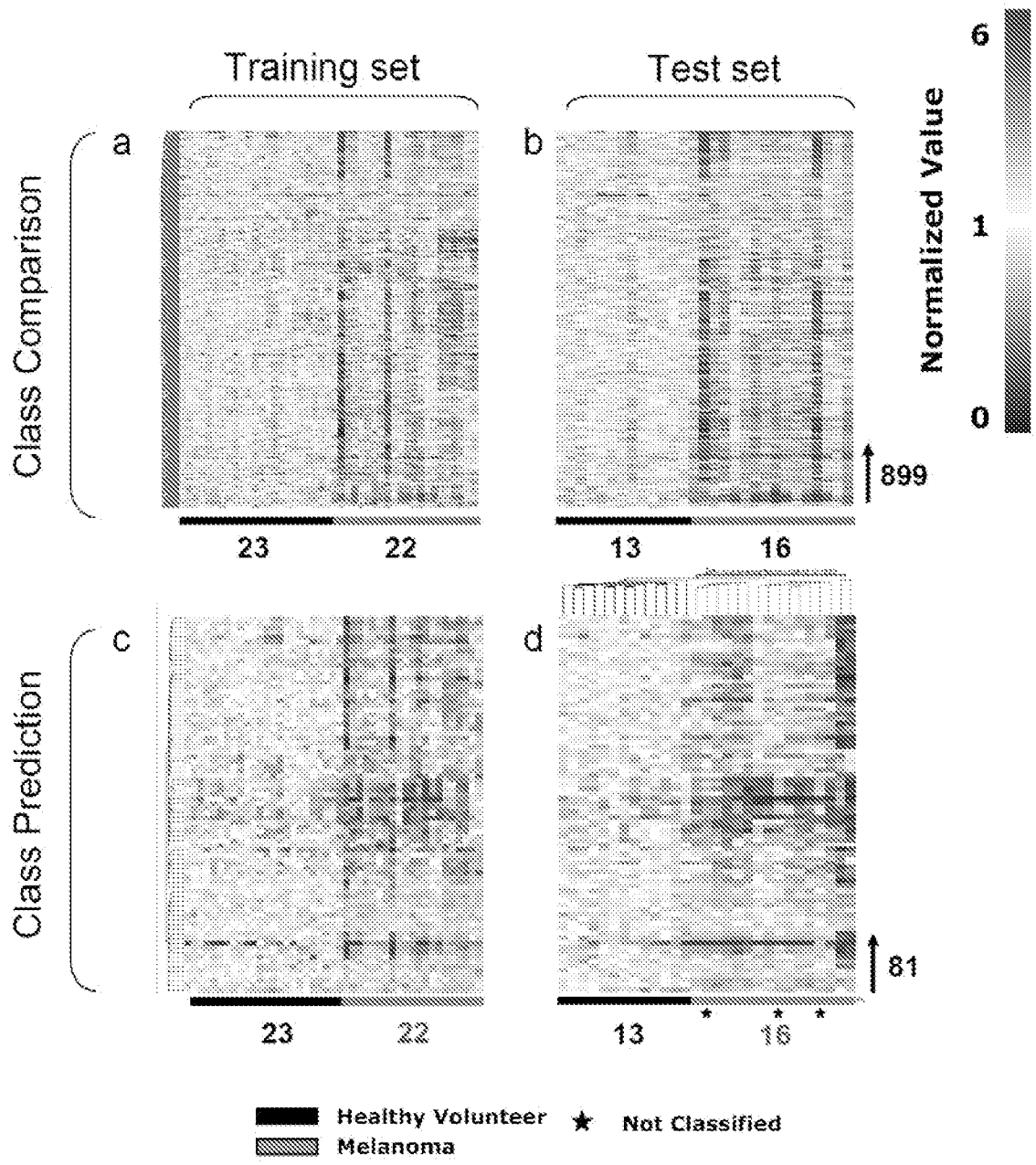

Training set
Test set
Class Comparison
a
b
Normalized Value
6
1
0
899
23
22
13
16
Class Prediction
c
d
81
23
22
13
16
Healthy Volunteer
Melanoma
Not Classified

FIGURE 7

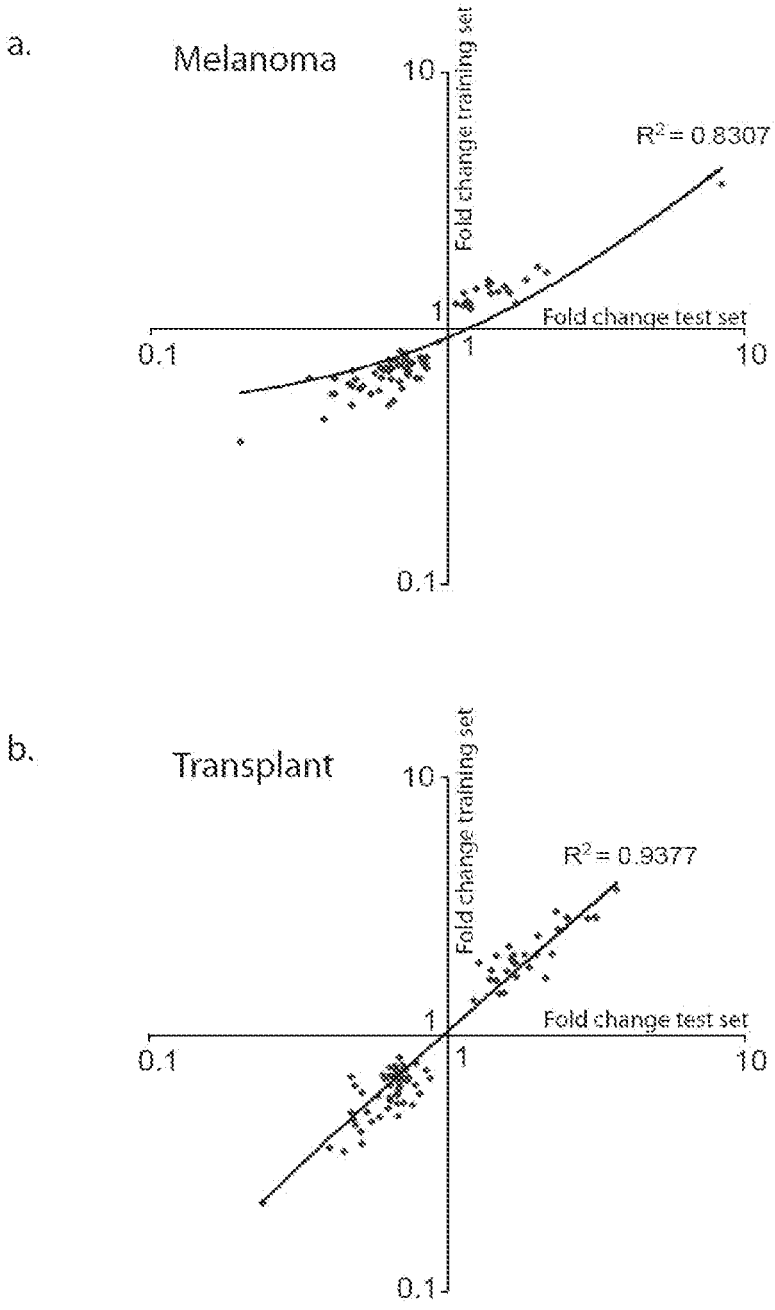

a.
Melanoma
Fold change training set
10
1
0.1
R^2 = 0.8307
Fold change test set
0.1
1
10
b.
Transplant
Fold change training set
10
1
0.1
R^2 = 0.9377
Fold change test set
0.1
1
10

FIGURE 8

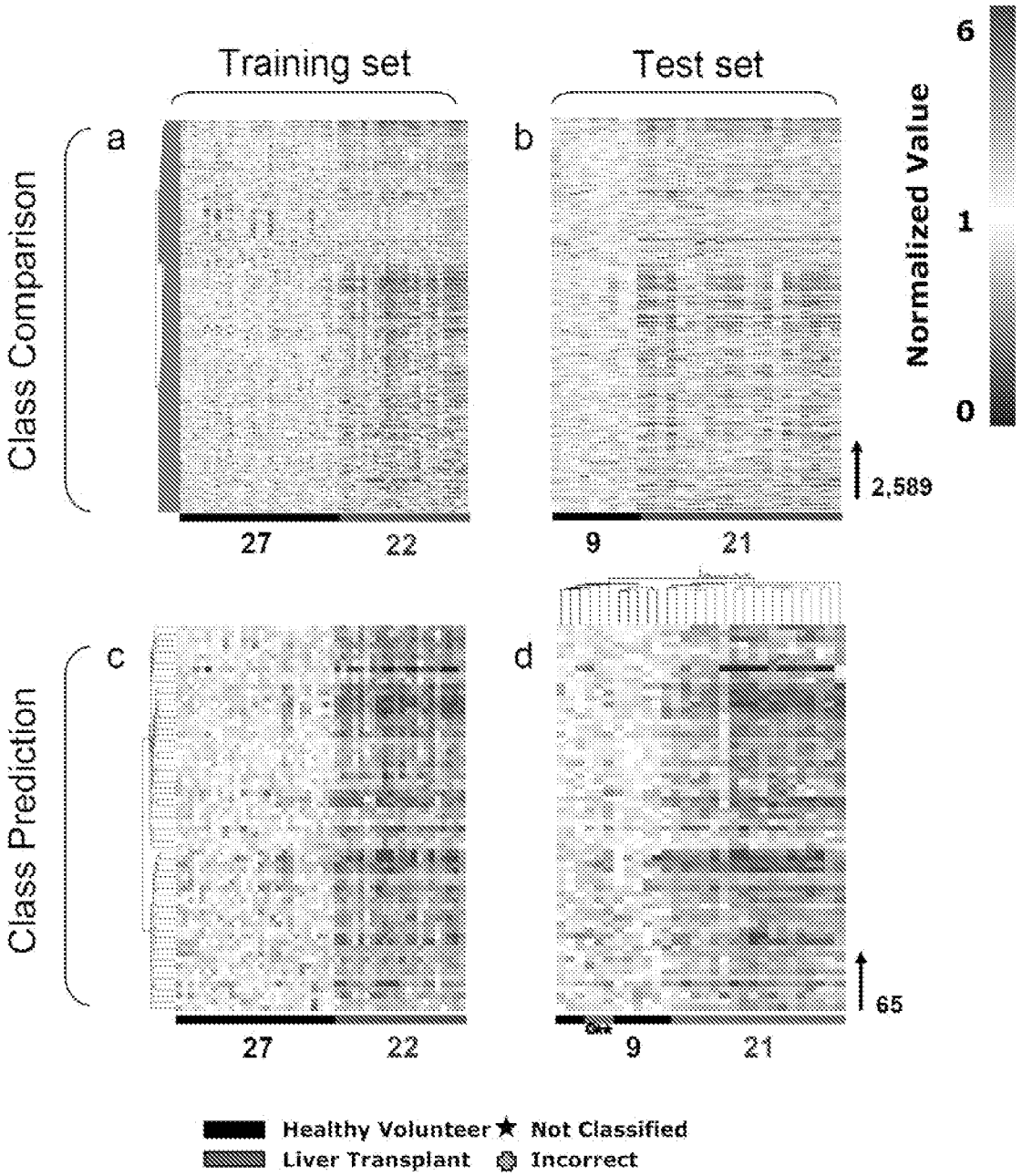

Training set
Test set
Class Comparison
Class Prediction
a
b
c
d
Normalized Value
6
1
0
2,589
65
27
22
9
21
Healthy Volunteer
Liver Transplant
Not Classified
Incorrect

FIGURE 9

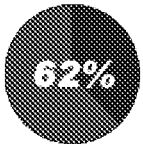
M	p-val <0.05		Functional assessment
	Melanoma	Transplant	
1.1 n=69			**Plasma cells.** Includes genes coding for Immunoglobulin chains (e.g. IGHM, IGJ, IGLL1, IGKC, IGHD) and the plasma cell marker CD38. Ig, Immunoglobulin, Bone, Marrow, PreB, IgM, Mu
1.2 n=96			**Platelets.** Includes genes coding for platelet glycoproteins (ITGA2B, ITGB3, GP6, GP1A/B), and platelet-derived immune mediators such as PPPB (pro-platelet basic protein) and PF4 (platelet factor 4). Platelet, Adhesion, Aggregation, Endothelial, Vascular
1.3 n=47			**B-cells.** Includes genes coding for B-cell surface markers (CD72, CD79A/B, CD19, CD22) and other B-cell associated molecules: Early B-cell factor (EBF), B-cell linker (BLNK) and B lymphoid tyrosine kinase (BLK). Immunoreceptor, BCR, B-cell, IgG
1.4 n=87			**Undetermined.** Includes regulators and targets of the cAMP signaling pathway (JUND, ATF4, CREM, PDE4, NR4A2, VIL2), as well as repressors of TNF-alpha mediated NF-kB activation (CYLD, ASK, TNFAIP3). Replication, Repression, Repair, CREB, Lymphoid, TNF-alpha
1.5 n=130			**Myeloid lineage.** Includes molecules expressed by cells of the myeloid lineage (CD86, CD163, FCGR2A), some of which being involved in pathogen recognition (CD14, TLR2, MYD88). This set also includes TNF family members (TNFR2, BAFF) Monocytes, Dendritic, MHC, Costimulatory, TLR4, MYD88
1.6 n=28			**Undetermined.** This set includes genes coding for signaling molecules, e.g. the zinc finger containing inhibitor of activated STAT (PIAS1 and PIAS2), or the nuclear factor of activated T-cells NFATC3. Zinc, Finger, P53, RAS
1.7 n=127			**MHC/Ribosomal proteins.** Almost exclusively formed by genes encoding MHC class I molecules (HLA-A,B,C,G,E) + Beta 2-microglobulin (B2M) or Ribosomal proteins (RPLs, RPSs). Ribosome, Translational, 40S, 60S, HLA
1.8 n=86			**Undetermined.** Includes genes encoding metabolic enzymes (GLS, NSF1, NAT1), and factors involved in DNA replication (PURA, TERF2, EIF2S1) Metabolism, Biosynthesis, Replication, Helicase

FIGURE 10

M	p-val <0.05 Melanoma	p-val <0.05 Transplant	Functional assessment
2.9 n=122			**Undetermined.** Related to M2.8. Includes genes encoding molecules that associtate to the cytoskeleton (Actin related protein 2/3, MAPK1, MAP3K1, RAB5A). Also present are T-cell expressed genes (FAS, ITGA4/CD49D, ZNF1A1). ERK, Transactivation, Cytoskeletal, MAPK, JNK
2.10 n=44			**Undetermined.** Includes genes coding for Immune-related cell surface molecules (CD36, CD86, LILRB), cytokines (IL15) and molecules involved in signaling pathways (FYB, TICAM2 - Toll-like receptor pathway). Myeloid, Macrophage, Dendritic, Inflammatory, Interleukin
2.11 n=77			**Undetermined.** Includes kinases (UHMK1, CSNK1G1, CDK6, WNK1, TAOK1, CALM2, PRKCI, ITPKB, SRPK2, STK17B, DYRK2, PIK3R1, STK4, CLK4, PKN2) and RAS family mem--bers (G3BP, RAB14, RASA2, RAP2A, KRAS). Replication, Repress, RAS, Autophosphorylation, Oncogenic
3.1 n=80			**Interferon-inducible.** This set includes interferon-inducible genes: antiviral molecules (OAS1/2/3/L, GBP1, G1P2, EIF2AK2/PKR, MX1, PML), chemokines (CXCL10/IP-10), signaling molecules (STAT1, STAT2, IRF7, ISGF3G) ISRE, Influenza, Antiviral, IFN-gamma, IFN-alpha, Interferon
3.2 n=230			**Inflammation I.** Includes genes encoding molecules involved in inflammatory processes (e.g. IL8, ICAM1, C5R1, CD44, PLAUR, IL1A, CXCL16), and regulators of apoptosis (MCL1, FOXO3A, RARA, BCL3/6/2A1, GADD45B). TGF-beta, TNF, Inflammatory, Apoptotic, Lipopolysaccharide
3.3 n=202			**Inflammation II.** Includes molecules inducing or inducible by inflammation (IL18, ALOX5, ANPEP, AOAH, HMOX1, SERPINB1), as well as lysosomal enzymes (PPT1, CTSB/S, NEU1, ASAH1, LAMP2, CAST) Inflammatory, Defense, Lysosomal, Oxidative, LPS
3.4 n=323			**Undetermined.** Includes protein phosphatases (PPP1R12A, PTPRC, PPP1CB, PPM1B), and phosphoinositide 3-kinase (PI3K) family members (PIK3CA, PIK3C2A, PIK3R1, PIP5K3) Ligase, Kinase, KIP1, Ubiquitin, Chaperone
3.5 n=19			**Undetermined.** Composed of only a small number of transcripts. Includes hemoglobin genes (HBA1, HBA2, HBB). No keyword extracted.

FIGURE 11

M	p-val <0.05		Functional assessement
	Melanoma	Transplant	
2.1 n=72			**Cytotoxic cells.** Includes cytotoxic T-cells and NK-cells surface markers (CD8A, CD2, CD160, NKG7, KLRs), cytolytic molecules (granzyme, perforin, granulysin), chemokines (CCL5, XCL1) and CTL/NK-cell associated molecules (CTSW). NK, Killer, Cytolytic, CD8, Cell-mediated, T-cell, CTL, IFN-g
2.2 n=44			**Neutrophils.** This set includes innate defense molecules that are found in neutrophil granules (Lactotransferrin: LTF, defensin: DEAF1, Bacterial Permeability Incresing protein: BPI, Cathelicidin antimicrobial protein: CAMP...). Granulocytes, Neutrophils, Defense, Myeloid, Marrow
2.3 n=94			**Erythrocytes.** Includes hemoglobin genes (HGBs) and other erythocyte-associated genes (erythrocytic ankyrin: ANK1, Glycophorin C: GYPC, hydroxymethylbilane synthase: HMBS, erythroid associated factor: ERAF). Erythrocytes, Red, Anemia, Globin, Hemoglobin
2.4 n=118			**Ribosomal proteins.** Related to M1.7. Includes genes encoding ribosomal proteins (RPLs, RPSs), Eukaryotic Translation Elongation factor family members (EEFs) and Nucleolar proteins (NPM1, NOAL2, NAP1L1). Ribonucleoprotein, 60S, nucleolus, Assembly, Elongation
2.5 n=242			**Undetermined.** This module includes genes encoding immune-related (CD40, CD80, CXCL12, IFNA5, IL4R) as well as cytoskelton-related molecules (Myosins, Dedicator of Cytokinesis, Syndecan 2, Plexin C1, Dystrobrevin). Adenoma, Interstitial, Mesenchyme, Dendrite, Motor
2.6 n=110			**Myeloid lineage.** Related to M1.5. Includes genes expressed in myeloid lineage cells (ITGB2/CD18, Lymphotoxin beta receptor, Myeloid related proteins 8/14 Formyl peptide receptor 1), such as Monocytes and Neutrophils. Granulocytes, Monocytes, Myeloid, ERK, Necrosis
2.7 n=43			**Undetermined.** This module is largely composed of transcripts with no known function. Only 20 genes associated with literature, including a member of te chemokine-like factor superfamily (CKLFSF8). No keywords extracted.
2.8 n=104			**T cells.** Includes T-cell surface markers (CD5, CD6, CD7, CD26, CD28, CD96) and molecules expressed by lymphoid lineage cells (lymphotoxin beta, IL2-inducible T cell kinase, TCF7, T-cell differentiation protein mal, GATA3, STAT5B) Lymphoma, T-cell, CD4, CD8, TCR, Thymus, Lymphoid, IL2

FIGURE 12

M	p-val <0.05 Melanoma	p-val <0.05 Transplant	Functional assessment
3.6 n=65			**Undetermined.** This set includes mitochondrial ribosomal proteins (MRPLs, MRPs), mitochondrial elongation factors (GFM1/2), Sortin Nexins (SN1/8/14), as well as lysosomal ATPases (ATP6V1C/D). Ribosomal, T-cell, Beta-catenin
3.7 n=62			**Undetermined.** Includes genes encoding proteasome subunits (PSMA2/5, PSMB5/8); ubiquitin protein ligases HIP2 and STUB1, as well as components of ubiquitin ligase complexes (SUGT1). Spliceosome, Methylation, Ubiquitin
3.8 n=183			**Undetermined.** Includes genes encoding for several enzyme families: **aminomethyltransferase, arginyltransferase, aspargine synthetase, diacylglycerol kinase, inositol phosphatases, methytransferases, helicases...** CDC, TCR, CREB, Glycosylase
3.9 n=254			**Undetermined.** Includes genes encoding for kinases (IBTK, PRKRIR, PRKDC, PRKCI) and phosphatases (e.g. PTPLB PPP2CB/3CB, PTPRC, MTM1, MTMR2). Chromatin, Checkpoint, Replication, Transactivation

FIGURE 13

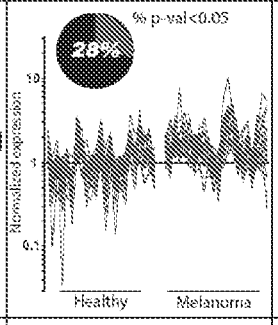
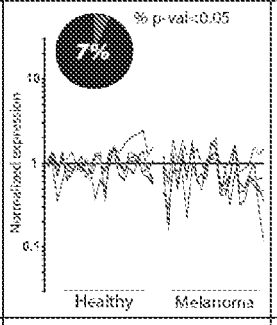
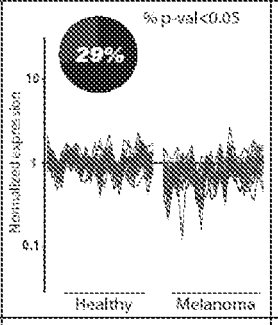
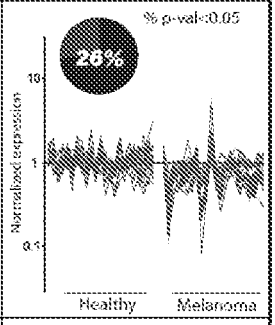
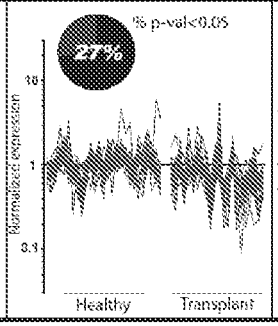
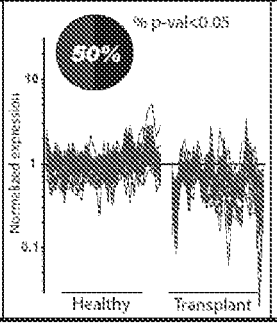
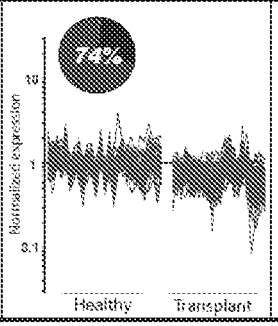
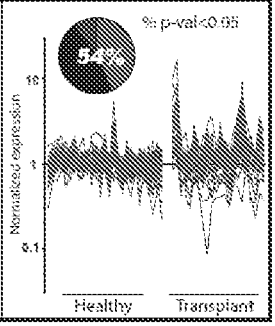
	M1.2 n=96 probes	M2.1 n=72 probes	M2.8 n=104 probes	M3.1 n=80 probes
	Platelets. Includes genes encoding platelet glycoproteins (ITGA2B, ITGB3, GP6, GP1A/B), and platelet-derived immune mediators such as Pro-platlet basic protein and Platelet Factor 4. Platelet, Aggregation, Adhesion, Endothelial, Vascular	**Cytotoxic cells.** Includes T-cells and NK-cells surface markers (CD8A, CD2, CD16, NKG7, KLRs), cytolytic molecules (granzyme, granulysin perforin), and other CTL/NK-cell associated molecules (e.g CTSW). NK, Killer, Cytolytic, CD8, T-cell, CTL, IFN-g	**T cells.** Includes T-cell surface markers (CD5, CD6, CD7, CD26, CD28, CD96) and molecules expressed by lymphoid lineage cells (lymphotoxin-beta, GATA3, ITK, TCF7, STAT5B). Lymphoma, T-cell, CD4, CD6, TCR, Thymus, Lymphoid, IL2	**Interferon-inducible.** This set of IFN-inducible genes includes: antiviral molecules (MX1, GBP1, OAS1/2/3/L, G1P2, PML, PKR), chemokines (IP-10), and signaling molecules (STAT1/2, ISGF3G, IRF7). ISRE, Influenza, Antiviral, IFN-gamma, IFN-alpha, Interferon
Mel				
LT				

FIGURE 14

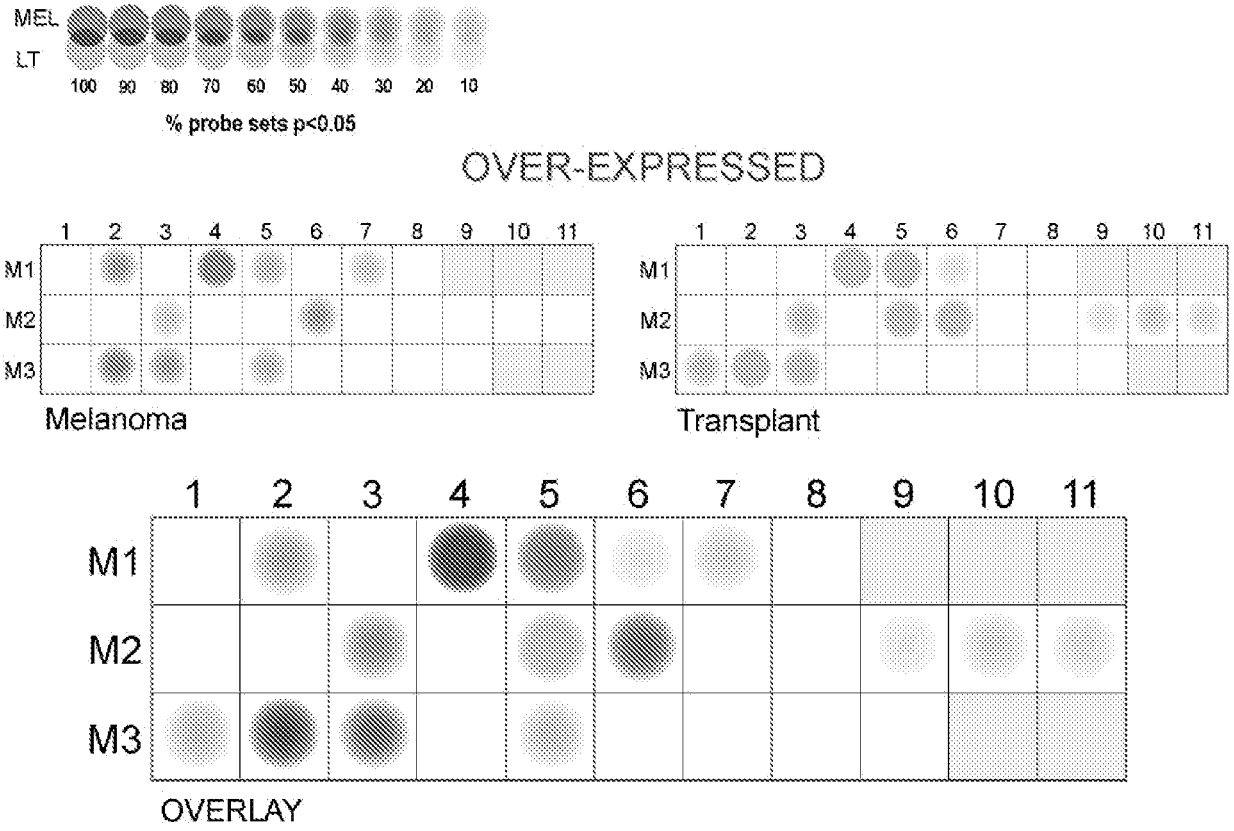

MEL
LT
100 90 80 70 60 50 40 30 20 10
% probe sets p<0.05
OVER-EXPRESSED
1 2 3 4 5 6 7 8 9 10 11
M1
M2
M3
Melanoma
Transplant
OVERLAY
UNDER-EXPRESSED
Melanoma
Transplant
OVERLAY

FIGURE 15

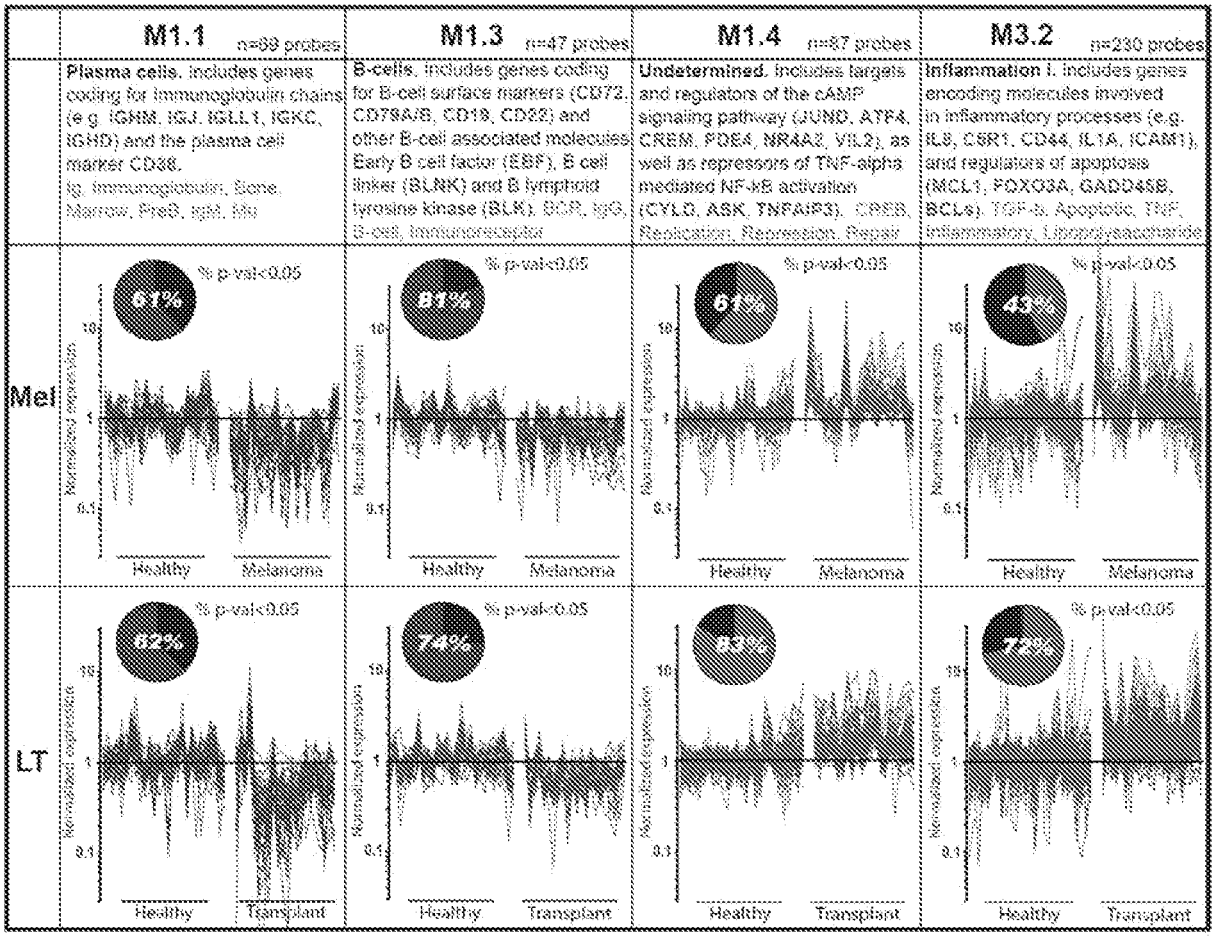

M1.1 n=89 probes
M1.3 n=47 probes
M1.4 n=87 probes
M3.2 n=230 probes
Plasma cells. Includes genes coding for immunoglobulin chains (e.g. IGHM, IGJ, IGLL1, IGKC, IGHD) and the plasma cell marker CD38.
Ig, Immunoglobulin, Bone Marrow, PreB, IgM, Mu
B-cells. Includes genes coding for B-cell surface markers (CD72, CD79A/B, CD19, CD22) and other B-cell associated molecules. Early B cell factor (EBF), B cell linker (BLNK) and B lymphoid tyrosine kinase (BLK). BCR, IgG, B-cell, Immunoreceptor
Undetermined. Includes targets and regulators of the cAMP signaling pathway (JUND, ATF4, CREM, PDE4, NR4A2, VIL2), as well as repressors of TNF-alpha mediated NF-kB activation (CYLD, ASK, TNFAIP3). CREB, Replication, Repression, Repair
Inflammation I. Includes genes encoding molecules involved in inflammatory processes (e.g. IL8, C5R1, CD44, IL1A, ICAM1), and regulators of apoptosis (MCL1, FOXO3A, GADD45B, BCL6). TGF-b, Apoptotic, TNF, Inflammatory, Lipopolysaccharide
Mel
LT
% p-val<0.05
61%
81%
61%
43%
82%
74%
83%
72%
Normalized expression
10
1
0.1
Healthy
Melanoma
Transplant

FIGURE 16

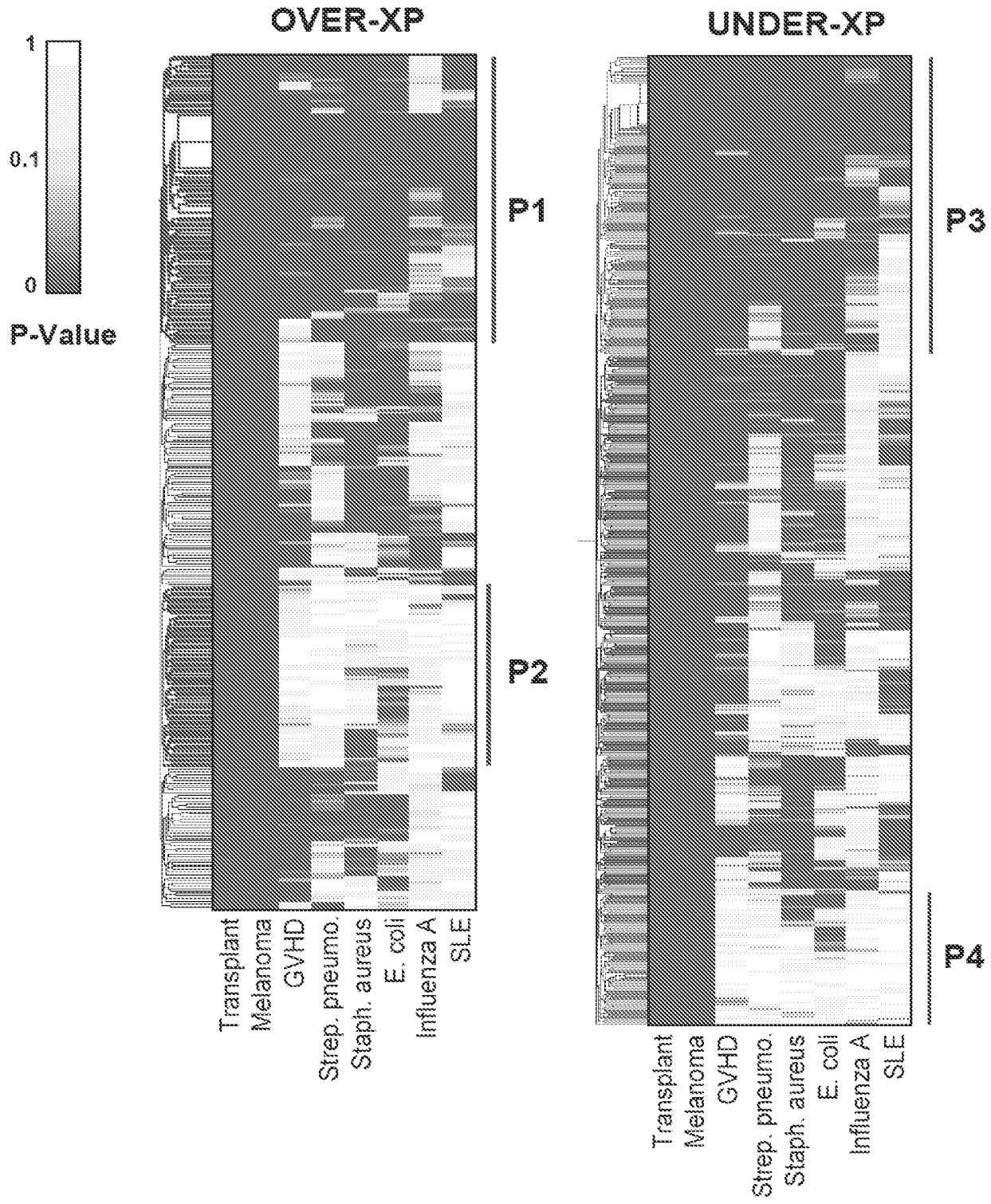

OVER-XP
UNDER-XP
1
0.1
0
P-Value
P1
P2
P3
P4
Transplant
Melanoma
GVHD
Strep. pneumo.
Staph. aureus
E. coli
Influenza A
SLE

FIGURE 17

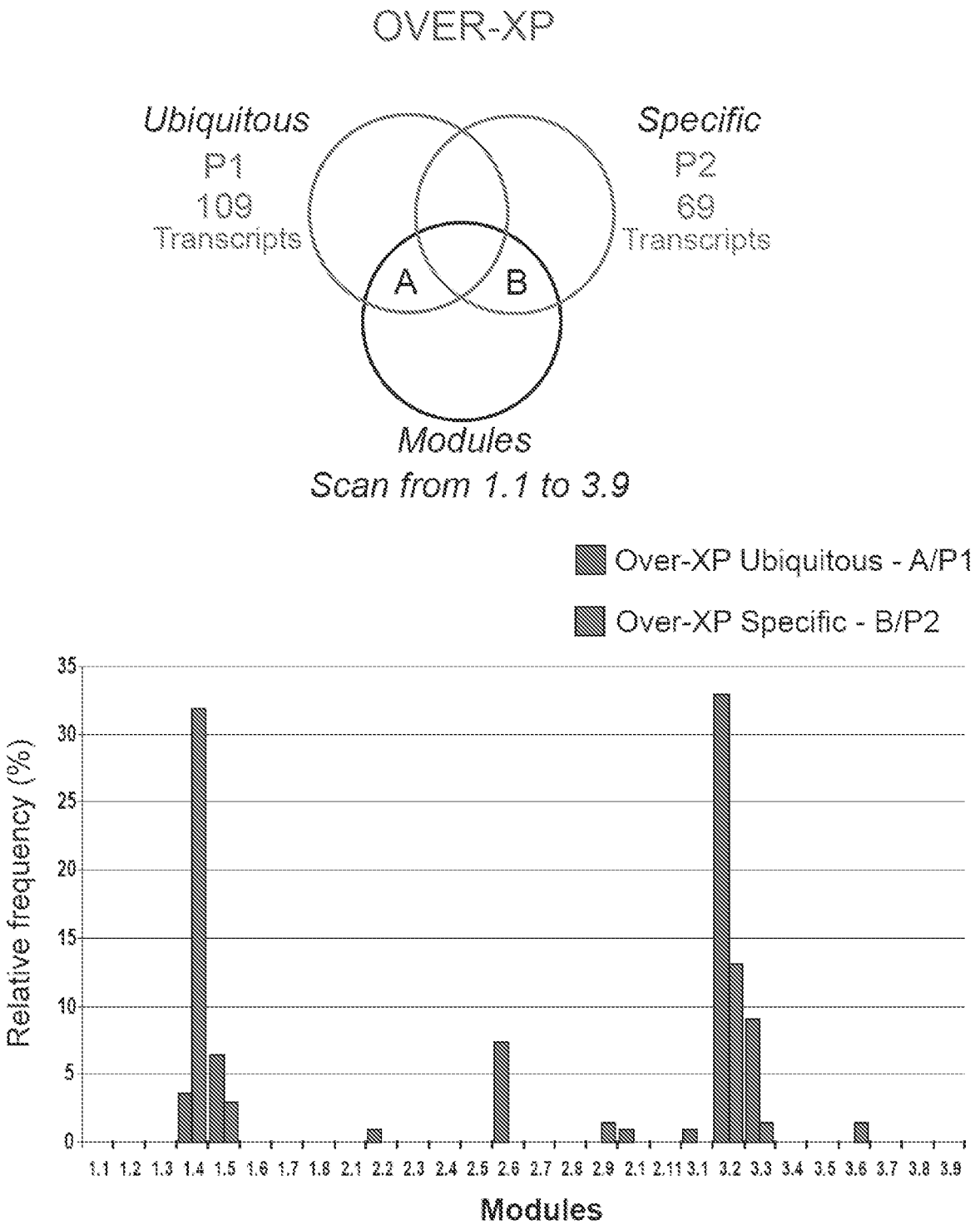

OVER-XP
Ubiquitous
P1
109
Transcripts
Specific
P2
69
Transcripts
A
B
Modules
Scan from 1.1 to 3.9
Over-XP Ubiquitous - A/P1
Over-XP Specific - B/P2
Relative frequency (%)
0
5
10
15
20
25
30
35
1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 2.1 2.11 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9
Modules

FIGURE 18

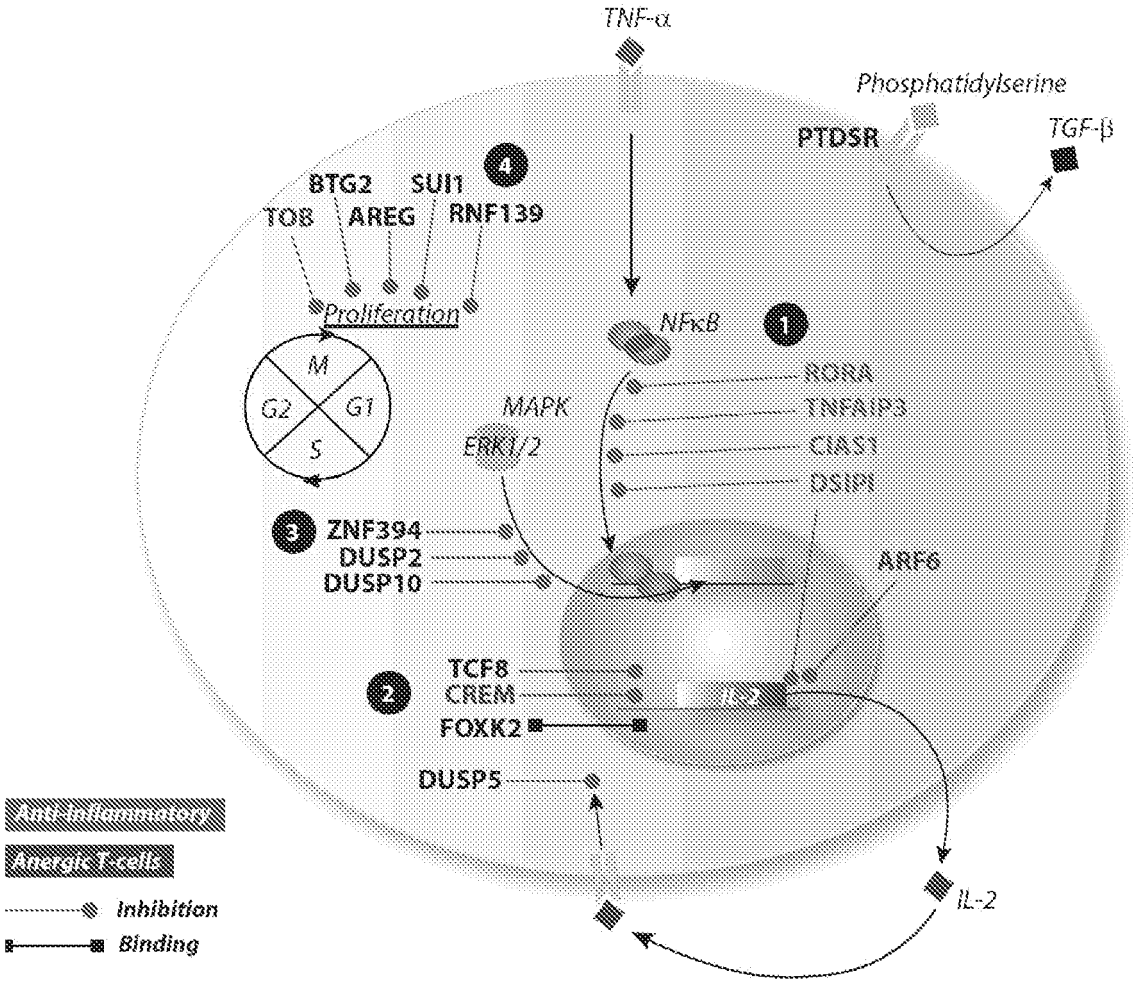

TNF-α
Phosphatidylserine
PTDSR
TGF-β
BTG2
SUI1
TOB
AREG
RNF139
4
Proliferation
M
G2
G1
S
NFκB
1
RORA
TNFAIP3
CIAS1
DSIPI
MAPK
ERK1/2
3
ZNF394
DUSP2
DUSP10
ARF6
TCF8
CREM
2
FOXK2
DUSP5
IL-2
Anti-inflammatory
Anergic T-cells
Inhibition
Binding

FIGURE 19

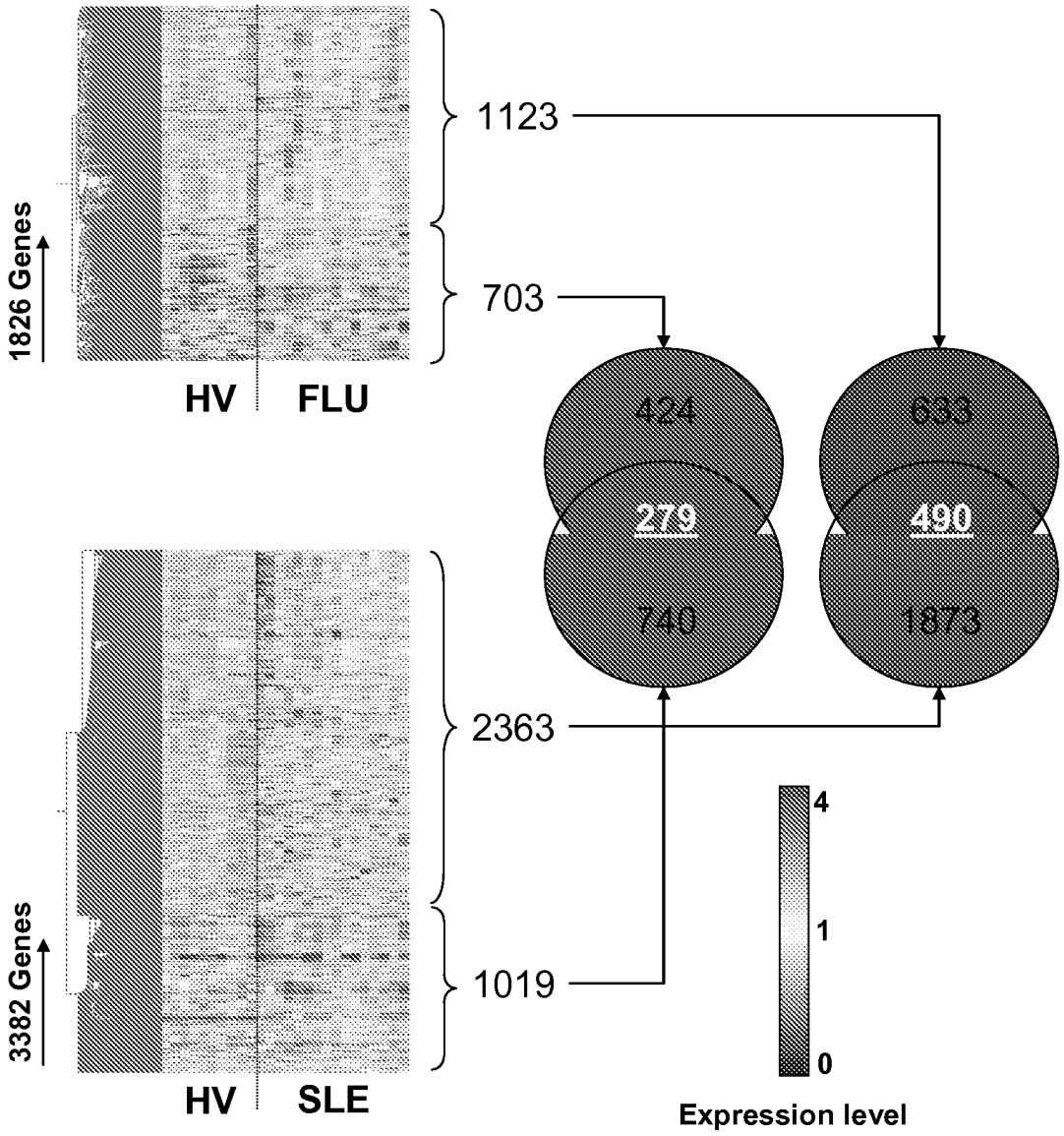

1826 Genes
1123
703
HV
FLU
424
279
740
633
490
1873
2363
1019
3382 Genes
HV
SLE
4
1
0
Expression level

FIGURE 20

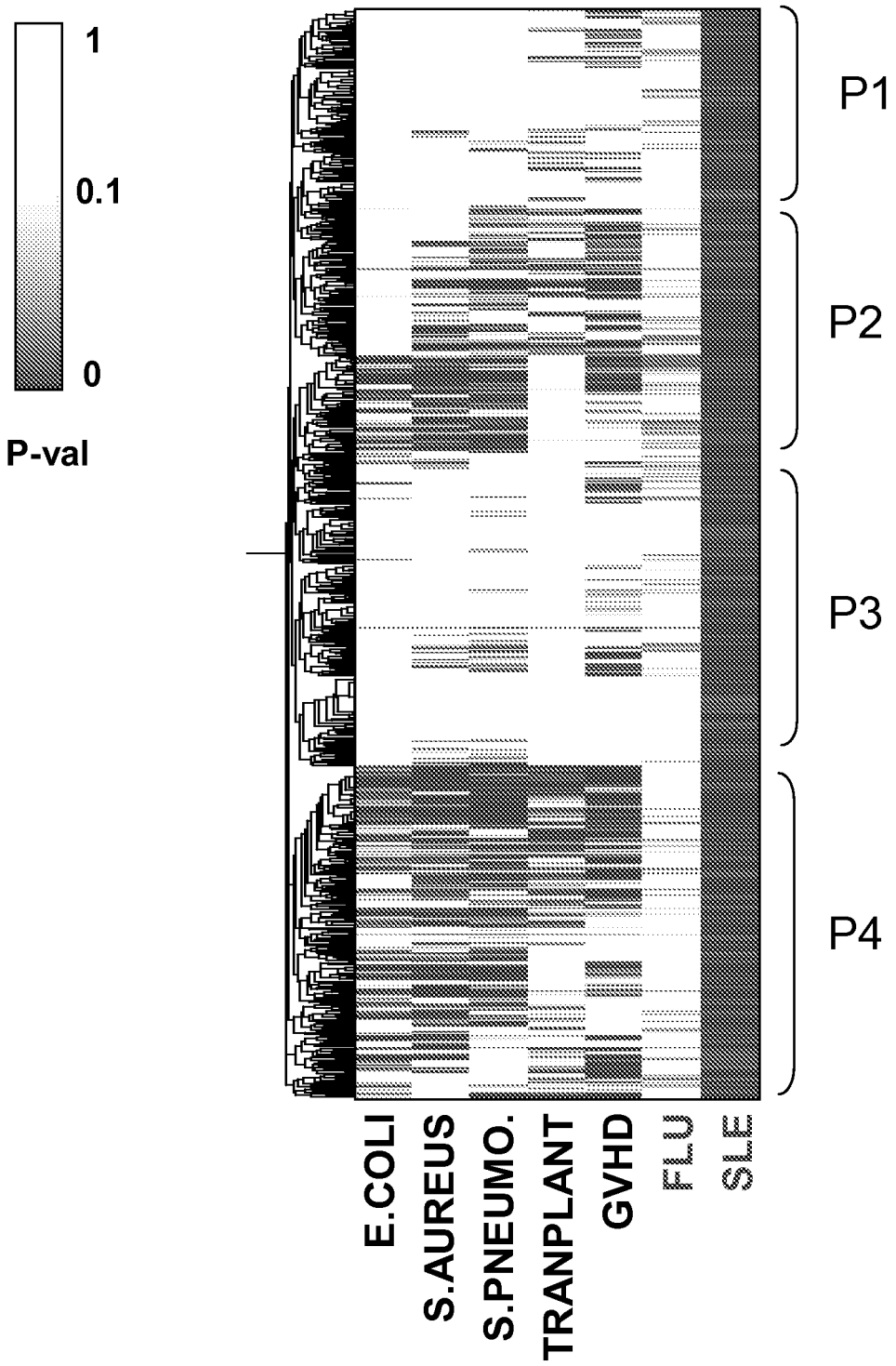

1
0.1
0
P-val
P1
P2
P3
P4
E.COLI
S.AUREUS
S.PNEUMO.
TRANPLANT
GVHD
FLU
SLE

FIGURE 21

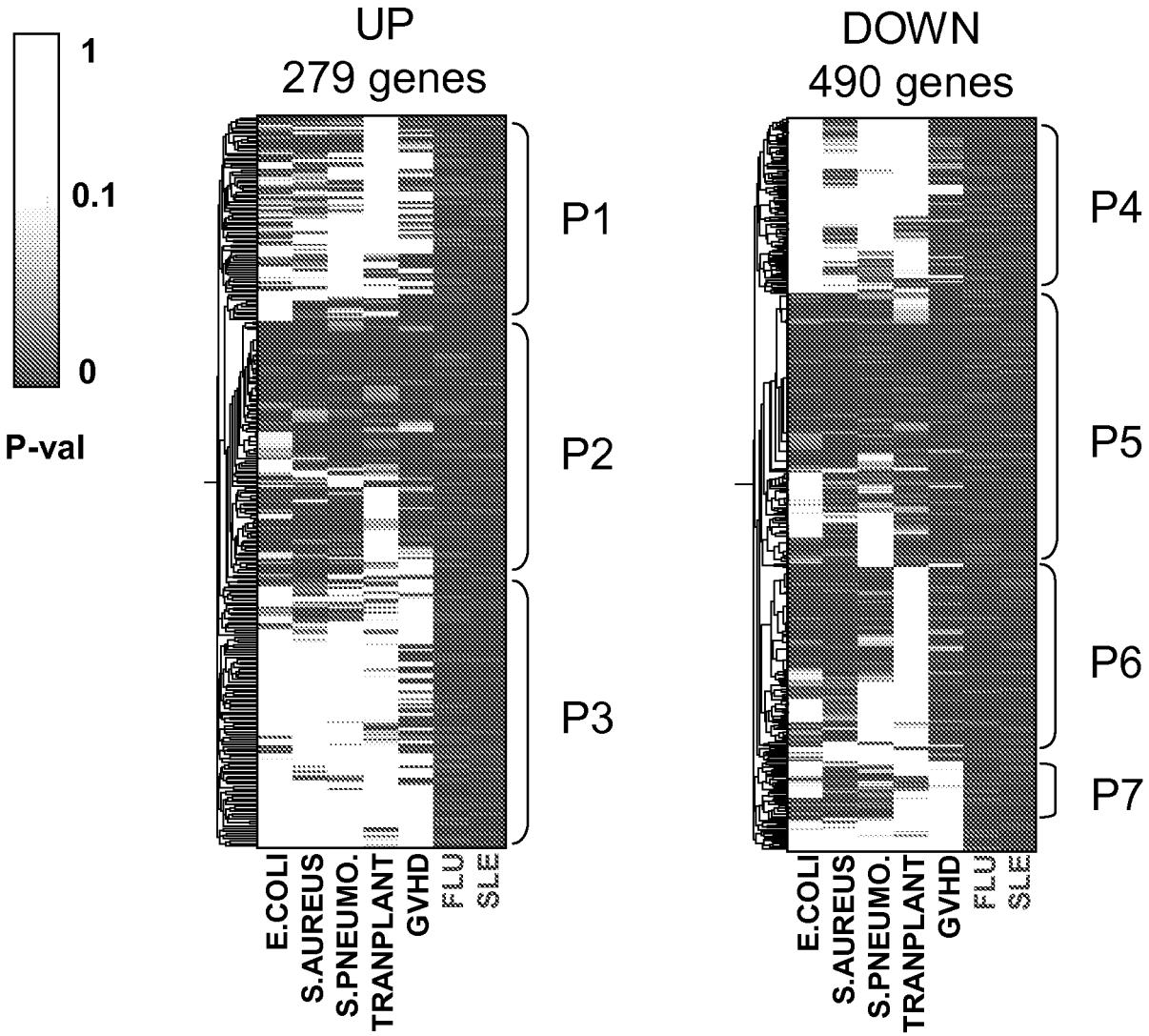

1
0.1
0
P-val
UP
279 genes
DOWN
490 genes
P1
P2
P3
P4
P5
P6
P7
E.COLI
S.AUREUS
S.PNEUMO.
TRANPLANT
GVHD
FLU
SLE

FIGURE 22

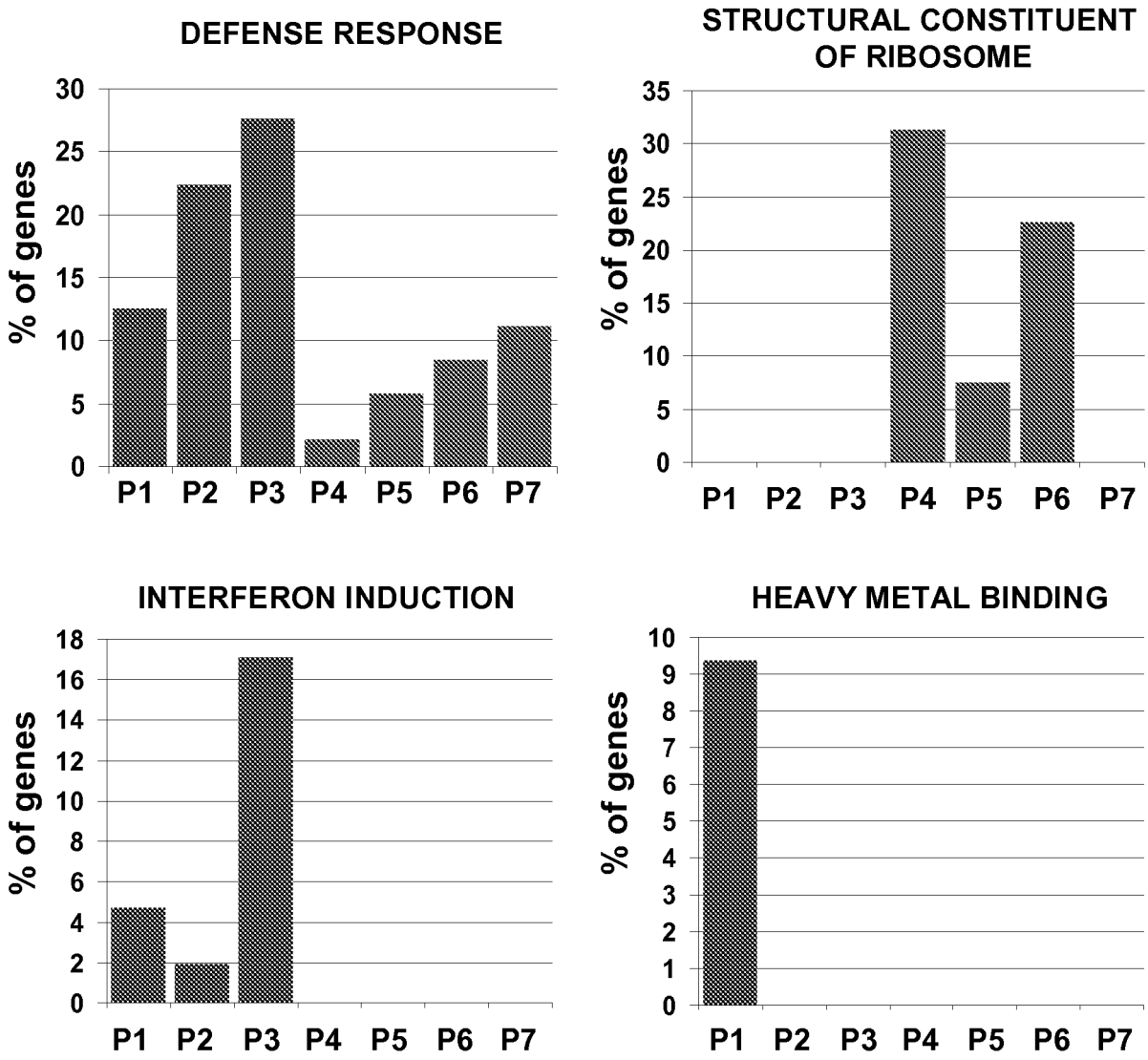

DEFENSE RESPONSE
% of genes
0
5
10
15
20
25
30
P1 P2 P3 P4 P5 P6 P7
STRUCTURAL CONSTITUENT OF RIBOSOME
% of genes
0
5
10
15
20
25
30
35
P1 P2 P3 P4 P5 P6 P7
INTERFERON INDUCTION
% of genes
0
2
4
6
8
10
12
14
16
18
P1 P2 P3 P4 P5 P6 P7
HEAVY METAL BINDING
% of genes
0
1
2
3
4
5
6
7
8
9
10
P1 P2 P3 P4 P5 P6 P7

FIGURE 23