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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 November 2012 (29.11.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/162660 A2

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 17/10* (2006.01)
C12N 15/11 (2006.01)

(21) International Application Number:
PCT/US2012/039699

(22) International Filing Date:
25 May 2012 (25.05.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/489,883	25 May 2011 (25.05.2011)	US
61/509,644	20 July 2011 (20.07.2011)	US
61/585,892	12 January 2012 (12.01.2012)	US
61/619,663	3 April 2012 (03.04.2012)	US

(71) Applicant *(for all designated States except US)*: **BROWN UNIVERSITY** [US/US]; Horace Mann Building, 47 George Street, 3rd Floor, Providence, Rhode Island 02912 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KELSEY, Karl** [US/US]; 57 Toxteth St, Brookline, Massachusetts 02446 (US). **HOUSEMAN, Eugene, Andres** [US/US]; 2930 NW Gibson Hill Road, Albany, Oregon 97321 (US). **WIENCKE, John** [US/US]; 42 Cragmont Ave, San Francisco, California 94116 (US). **ACCOMANDO, William, P., Jr.** [US/US]; 111 Medway Street, Apt 14, Providence, Rhode Island 02906 (US). **MARSIT, Carmen** [US/US]; 301 Olde Farms Road, Grantham, New Hampshire 03753 (US).

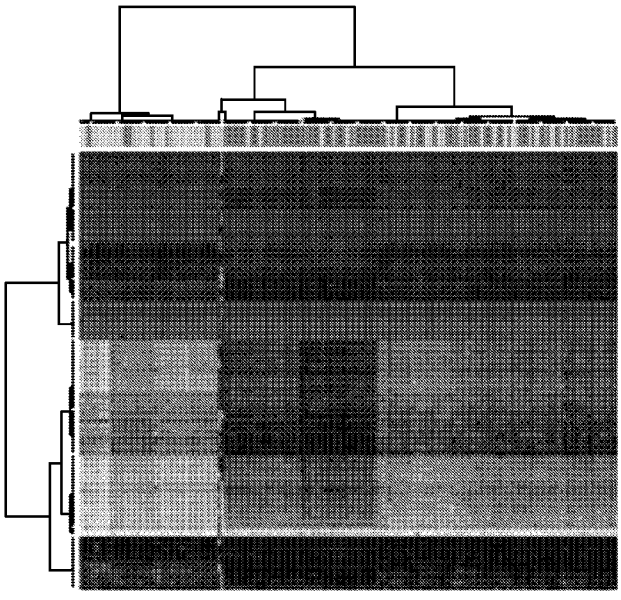
(74) Agents: **GUTERMAN, Sonia, K.** et al.; 88 Black Falcon Avenue, Suite 345, Boston, Massachusetts 00210 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

[Continued on next page]

(54) Title: METHODS USING DNA METHYLATION FOR IDENTIFYING A CELL OR A MIXTURE OF CELLS FOR PROGNOSIS AND DIAGNOSIS OF DISEASES, AND FOR CELL REMEDIATION THERAPIES



Heatmap of HNSCC data (S_1)

Figure 3

(57) Abstract: Methods using DNA Methylation arrays are provided for identifying a cell or mixture of cells and for quantification of alterations in distribution of cells in blood or in tissues, and for diagnosing, prognosing and treating disease conditions, particularly cancer. The methods use fresh and archival samples.

WO 2012/162660 A2

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, 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 (Rule 48.2(g))*

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

Methods using DNA Methylation for identifying a cell or a mixture of cells for prognosis and diagnosis of diseases, and for cell remediation therapies

Related applications

This application claims the benefit of provisional applications having serial numbers 61/489,883 filed May 25, 2011 entitled, "Methods of Immunodiagnostics using DNA Methylation arrays as surrogate measures of the identity of a cell or a mixture of cells"; 61/509,644, filed July 20, 2011 entitled "Methods of Immunodiagnostics using DNA Methylation arrays as surrogate measures of the identity of a cell or a mixture of cells for prognosis and diagnosis of diseases,"; 61/585,892 filed January 12, 2012 entitled, "Methods of Immunodiagnostics using DNA Methylation arrays as surrogate measures of the identity of a cell or a mixture of cells for prognosis and diagnosis of diseases,"; and 61/619,663, filed April 3, 2012 entitled "Methods using DNA Methylation arrays for identifying a cell or a mixture of cells for prognosis and diagnosis of diseases, and for cell remediation therapies" inventors Kelsey K, Houseman EA, Wiencke J, Accomando W and Marsit C, which applications are hereby incorporated herein by reference in their entireties.

Technical field

Methods of determining altered immune cell distribution to diagnose or prognose a disease condition based on determining DNA methylation signatures of specific immune cell type of or mixture of immune cells types are provided.

Background

Leukocytes, commonly called white blood cells, are cells that are primarily responsible for mounting an immune response by a host to pathogens and to foreign antigens. Leukocyte distribution is currently determined by simple histologic or flow cytometric assessments. These methods have significant limitations. In particular, flow cytometry is limited by the following: availability of fluorescent antibody tags, laborious nature of the antibody tagging process, and needs for separation of cells requiring large volumes of fresh cells, expensive technology as well as equipment for detection of cells, and maintaining the integrity of the outer membrane of the cells to preserve labile protein epitopes. Further limitation of methods requiring fresh cells is that the methods are not useful in situations in which prospective studies are impractical, such as in the case of rare diseases, in which large numbers of disease subjects are not available. In these

cases retrospective studies are needed to correlate disease outcome with disease parameters. However, retrospective studies can be performed only if archival samples derived from archived cohort populations could be used to analyze the disease parameters. Currently there are no known methods in which archived samples from patients and normal subjects could be used to provide a quantitative estimate of leukocyte distributions in disease conditions.

Thus there is a need for methods that provide quantification of alterations in distribution of leukocytes in blood or tissues in disease conditions that do not rely upon fresh samples, that are not labor intensive and that do not use expensive technology or equipment.

Summary

In diverse medical conditions such as in disease or in instances of immune-toxic exposure, the leukocyte distribution in blood or tissues contains information about the underlying immune-biology of the medical condition which is useful for diagnosis, prognosis or treatment of the medical condition, or for monitoring response to therapy. Accordingly, an embodiment of the invention provides a method a method for assessing a disease condition in a subject, including: measuring a CD3Z positive T lymphocyte cell number in a sample from the subject by analyzing methylation in the sample of at least one CpG dinucleotide (CpG) in gene CD3Z or in an orthologous or a paralogous gene thereof, such that an amount of a demethylated C of the at least one CpG in the sample is a measure of CD3+ T lymphocyte cell number; and comparing the amount of the demethylated C in the sample from the subject with that in positive control samples from patients with the disease condition, and with that in negative control samples from healthy subjects, such that the disease condition is selected from: an autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, immunosuppression and a cancer. As used herein "subject" refers to any animal, for example, a mammal that is healthy or that has a disease condition for example a human, or a high value agricultural animal or a zoo animal. A "patient" is a subject that either has a disease condition or is in need of obtaining a diagnosis of a disease condition.

A related embodiment of the method includes at least one of: monitoring, diagnosing, prognosing, and measuring response to therapy by comparing the measured CD3+ T lymphocyte cell numbers in the subject after therapy to that in the patients with the disease condition and in the healthy subjects.

An embodiment of the method provides that the inherited disease is an aneuploidy. For example, aneuploidy is selected from trisomy 21, Turner's syndrome, and Klinefelter's syndrome.

The sample used in the method is a fresh sample. For example, the fresh sample is freshly drawn blood, a tumor infiltrate or cells obtained from a lymph node puncture. Alternatively, the sample is an archival sample. For example, the archival sample is archival blood collected and stored on filter paper cards such as a Guthrie card, frozen blood specimens or frozen tissue. Demethylation of DNA is a stable chemical modification of DNA, and archival samples are used to measure cell numbers. Flow cytometry in contrast, requires fresh cells, for detection of cells depends on the availability of protein epitopes, which are labile and not well preserved in archival samples.

In a related embodiment of the method the amount of the demethylated C of the at least one CpG in the CD3Z gene in the sample is at least about 80%, at least about 90%, or at least about 95% of the total amount of the CpG in CD3Z genes in the sample.

An embodiment of the method further involves analyzing the methylation of the CD3Z gene further by amplifying by Polymerase Chain Reaction (PCR) using primer pairs specific for amplification of specific demethylated CpG loci. For example, amplification by PCR involves monitoring quantitative PCR in real time using a MethyLight assay or using digital PCR.

An embodiment of the method further involves analyzing the methylation of the CD3Z gene by a method selected from the group of: Pyrosequencing, Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE), Methylation-sensitive single stranded conformation analysis (MS-SSCA), and High resolution melting analysis (HRM) and digital PCR methods comprising emulsion and nanofluidic partitioning. According to a related embodiment, Methylation-sensitive single-nucleotide primer extension further includes: chemically converting the lymphocyte derived whole genomic DNA with bisulfite; amplifying chemically converted whole genomic DNA; enzymatically fragmenting resulting amplified DNA; hybridizing fragmented DNA to methylation sensitive CpG locus specific DNA oligomers; and labeling by single-base extension using fluorescently labeled nucleotides.

Another embodiment of the method further provides steps for analyzing methylation of differentially methylated regions (DMRs) of gene FOXP3, using primer pairs for amplification of specific loci of demethylated CpG in the FOXP3 gene. Within a gene "loci" as used herein refers to locations of all CpG dinucleotide containing sequences present in that gene, and only one or a few may be differentially demethylated in a specific cell.

A related embodiment of the method further includes: determining a ratio of CpG demethylation of FOXP3 gene DMR to the CpG demethylation of CD3Z gene DMR, in a sample of tumor infiltrate, such that the ratio involves an index of T regulatory cell number to the total T cell number in the infiltrate; and the method further involves diagnosing of a pathological grade of the cancer, so that the index of T regulatory cell number to the total T cell

number in the tumor infiltrate correlates with the grade of the cancer. In a related embodiment, the cancer is selected from: a glioma; an ovarian cancer; a head and neck squamous cell cancer (HNSCC), breast cancer, lung cancer, prostate cancer, colon cancer, pancreatic cancer, bladder cancer, cervical cancer and liver cancer.

In a related embodiment the method further includes prognosing survival of a patient having or needing a diagnosis of glioma or HNSCC, in which amount of demethylation of CD3Z gene DMR in the patient as a percent of total DNA greater than a median value in a sample population of subjects correlates with a prognosis of poor survival.

An embodiment of the invention provides a kit for measuring CD3+ T lymphocyte and FOXP3+ T regulatory cell numbers by analyzing methylation of CpG positions in CD3Z and FOXP3 genes, the kit having sequencing and PCR primers specific for the CD3Z and the FOXP3 gene DMRs and instructions for analyzing and comparing the CpG methylation between healthy subjects and a patient.

An embodiment provides a method for assessing a disease condition by estimating an alteration in proportions of types of leukocytes in a sample from a subject, the method including the steps of: measuring a DNA methylation profile for each type of leukocyte and for unfractionated cells, such that DNA methylation profiles are obtained for a plurality of CpG loci, and obtaining the status of an individual CpG locus by amplifying DNA from each of the types of leukocyte and from the unfractionated cells, such that amplifying comprises hybridizing methylation sensitive locus-specific DNA oligomers corresponding to each CpG locus; ordering CpG loci by ability to distinguish types of leukocytes, such that the ordering of the CpG loci determines differentially methylated DNA regions (DMRs), such that obtaining DMRs comprises statistically minimizing introduction of bias in amount of total methylation status of a large number of CpG loci obtained from the unfractionated cells by employing a Bayesian treatment of prior probabilities of the methylation status at each individual locus, thereby identifying a plurality of CpG loci to include in the measurement, such that an amount of of CpG loci distinguishes DMR signatures among the types of leukocytes and minimizes bias; obtaining DNA methylation profiles comprising DMRs from the types of leukocytes, such that the DNA methylation profiles comprise validating measures of relative amounts of the types of leukocytes, and obtaining DNA methylation profiles of the unfractionated cells as surrogate measures of relative amounts of each leukocyte type in the unfractionated cells; employing an analog of a measurement error model wherein a DNA methylation surrogate y is reverse formulated with respect to the disease outcome z, as

$$y=f(z),$$

such that y denotes a multivariate random variable representing a methylation profile, z denotes a disease outcome or state, and *f* denotes a probability distribution; y, z, and leukocyte distribution, ω are related by the estimator equations,

$$E(y|\omega)=g(\omega), \text{ and}$$

under an assumption $E(z|\omega,y) = E(z|\omega)$, such that, E denotes an expectation of a random variable and ω denotes a subject specific distribution of leukocytes; and, comparing relative amounts of each type of leukocyte in the sample from the subject with those in a control sample, thereby providing an assessment of the disease condition. In related embodiments, the locus-specific DNA oligomers are linked to an array selected from the group of: a glass slide array; a quartz slide array; a fiber optic bundle array, a planar slide array, a micro-well array; a multi-well dish array; a digital PCR array; and a bead array having beads located at known addressable locations on the array. A related embodiment of the method further provides at least one of steps of: monitoring, diagnosing, prognosing and measuring response to therapy of the disease condition.

The method in a related embodiment further includes analyzing sensitivity for correcting bias, such that correcting bias is unrelated to measurement error and is related to errors arising from unprofiled cell types and non-cell mediated profile differences. In related embodiments of the method, fractionated leukocyte types include at least one selected from: CD19+ B lymphocytes, CD15+ granulocytes, CD14+ monocytes, CD56+ Natural Killer cells, and CD3+ T lymphocytes.

In an embodiment of the method the disease condition is Head and Neck Squamous Cell Carcinoma (HNSCC).

According to another embodiment of the method the control sample is taken from the subject at a different point in time for prognosis of the course of the disease condition in the subject. In another related embodiment, the method of assessing disease condition further includes after employing the measurement model, comparing the distribution of leukocytes to the relative amounts in the control sample as a normal standard, such that the normal standard is a statistical measure obtained from a plurality of disease-free subjects.

In a related embodiment the method provides a diagnosis of immunosuppression due to smoking in a currently smoking subject by: determining a ratio of CpG demethylation of FOXP3 gene DMR to the CpG demethylation of CD3Z gene DMR in blood in the currently smoking subject, such that the ratio is an index of T regulatory cell number to the total T cell number; and providing a diagnosis of immunosuppression in the currently smoking subject, such that the value of the index of T regulatory cell number to the total T cell number in the currently smoking subject, greater than the average value in a sample population of currently non-

smoking subjects correlates with immunosuppression due to smoking. In a related embodiment of the method the subject with the currently-smoking or currently non-smoking status is a patient having a cancer, an infection or in need of a transplant.

An embodiment provides a method of predicting a methylation class membership in a bodily fluid sample of a subject for assessing disease status of the subject, in which the methylation class membership corresponds to an epigenetic signature of a plurality of leukocyte types, the method including: measuring amounts of DNA methylation in each of a plurality of leukocyte type populations to determine differentially methylated regions (DMRs); ranking leukocyte DMRs for each leukocyte type according to statistical strength of association of the DMR with each leukocyte type; randomly dividing a data set of control subjects and subjects with a disease into groups having substantially the same numbers of control subjects and subjects with the disease to obtain a training set and a testing set; clustering samples in the training set using a defined number of highest ranked leukocyte DMRs to determine clustering solutions, in which a clustering solution corresponds to the methylation class membership; and predicting methylation class membership for subjects within the testing set by applying the clustering solutions obtained from the training set to the highest ranked leukocyte DMRs in the testing set, such that clinical utility of the predicted methylation class membership is determined by testing association of the predicted methylation class membership with the disease status of the subject.

According to an embodiment of the method, the highest ranked leukocyte DMRs are as shown in Table 21, in which each DMR is identified by chromosomal location and gene name , and the defined number of highest ranked leukocyte DMRs is selected from: least 10, at least 20, at least 30, at least 40 and is 50.

The methylation class membership of the subject in the testing set is predicted for example using a naïve Bayes classifier. Testing the association of the predicted methylation class with disease status includes for example using receiver operating characteristic curves (ROC) and the corresponding area under each curve.

The bodily fluid sample in some embodiments is a fresh sample, for example freshly collected blood or a blood derivative. Alternatively, the bodily fluid is an archival sample, for example stored frozen blood or archival blood collected and stored on a filter paper card such as a Guthrie card.

The method in a related embodiment includes at least one of: diagnosing, monitoring, prognosing and measuring response to therapy of the disease status.

In related embodiments the leukocyte types are selected from the group of: natural killer cells, B Cells, CD4+ T cells, CD8+ T cells, granulocytes and monocytes. The disease according

to an embodiment of the method is exemplified by one of: head and neck squamous cell carcinoma (HNSCC), ovarian cancer, and bladder cancer.

An array is provided as another embodiment for estimating proportions of leukocyte types in a sample from a mammal for assessing a disease condition of the mammal by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the array including: a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, such that the surface at each location is attached to a DNA probe having a specific nucleotide sequence, such that the DNA probe having the specific nucleotide sequence hybridizes to a nucleotide sequence of a methylated form or an ummethylated form of a CpG dinucleotide in a sequence of a gene of the plurality of genes in the sample, such that the array is selected from having: at least 16 probes, at least 64 probes, at least 96 probes, and at least 384 probes.

The plurality of probes, in a related embodiment of the array, has nucleotide sequences that hybridize with a respective plurality of 96 different nucleotide sequences which are found in nature occurring in the plurality of genes. In another related embodiment, the 96 nucleotide sequences have SEQ ID NO: 1 to SEQ ID NO: 96.

In a related embodiment of the array, the addressable locations are wells of a substrate, such that the substrate is selected from: glass slide; quartz slide; fiber optic bundle and planar silica slides. In another related embodiment the surfaces included in the array are particles added to the wells.

In alternative embodiments the addressable locations of the array are defined spots on a glass slide or are microbeads or particles labeled with a code. For example, the particles are microbeads in the form of glass cylinders identifiable with inscribed holographic code.

In various embodiments the disease condition is selected from: an autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, immunosuppression and a cancer.

Another embodiment provides a method for estimating proportions of types of leukocytes in a sample from a subject for assessing a disease condition of the subject by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the method including: providing an array having a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, such that the surface at each location is attached to a DNA probe having a specific nucleotide sequence; reacting genomic DNA in the sample with a bisulfite reagent to convert unmethylated cytosine residues to uracil; hybridizing resulting bisulfite treated genomic DNA with the array to obtain resulting hybridized probes on the array, such that the DNA probes hybridize to a DNA sequence of each of a methylated form and an ummethylated form of a sequence having a CpG dinucleotide in a gene for each of the

plurality of genes; and detecting the methylation status of each of the CpG dinucleotides in each sequence, thereby estimating proportions of types of leukocyte in the sample from the subject for assessing the disease condition of the subject.

In a related embodiment, detecting the methylation status of the CpG dinucleotide sequence includes: extending each hybridized probe of the resulting hybridized probes on the array by primer extension to obtain a resulting primer extension product; ligating the resulting primer extension product to an oligonucleotide complementary to the DNA sequence of a 3' region of the gene to obtain a resulting template for PCR on the array; and amplifying by PCR and measuring amount of resulting PCR product, thereby detecting the methylation status of the CpG dinucleotide containing nucleotide sequence.

In another related embodiment amplifying by PCR further includes: amplifying the resulting template on the array using primers pairs including a 5' primer specific to each of the methylated or the unmethylated form of the CpG dinucleotide containing gene, and a 3'primer specific to the gene containing the CpG dinucleotide, thereby resulting in a first PCR product; amplifying the resulting first PCR product with differentially labeled 5' primers that specifically amplify either the methylated or the unmethylated form of the CpG dinucleotide containing nucleotide sequence containing gene, and a common 3' primer, resulting in a differentially labeled second PCR product, and hybridizing the second PCR product to the CpG dinucleotide containing gene for measuring amount of the second PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

Detecting the methylation status of the CpG dinucleotide sequence, in another related embodiment of the method, includes extending the resulting hybridized probes on the array by single base primer extension with a labeled nucleotide.

The array used in the method, in a related embodiment, includes at least 16 probes, at least 64, at least 96 probes or at least 384 probes. In another related embodiment of the method the plurality of probes on the array hybridizes with a plurality of 96 different nucleotide sequences occurring in the plurality of genes. In yet another related embodiment of the method each probe on the array is complementary to nucleotide sequences having SEQ ID NO: 1 to SEQ ID NO: 96.

In various embodiments of the method, the disease condition assessed is selected from: an autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, and a cancer. Assessing the disease condition using the array, in related embodiments of the method, includes at least one of: monitoring, diagnosing, prognosing, and measuring response to therapy by comparing estimated proportions of types of leukocytes of the subject after therapy to proportions of leukocytes from a healthy subject.

In a related embodiment of the method the sample containing the genomic DNA used to hybridize with the probes on the array is fresh i.e., obtained in real time prior to performing the method. In another related embodiment of the method the sample is archival.

In various embodiments of the method for estimating proportions of leukocytes using the array, the leukocyte types include at least one selected from: CD19+ B lymphocytes, CD15+ granulocytes, CD14+ monocytes, CD56+ natural Killer cells, and CD3+ T lymphocytes.

Another related embodiment provides a kit for estimating proportions of leukocyte types in a sample by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the kit including: an array having: a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, such that the surface at each location is attached to a DNA probe having a specific nucleotide sequence, such that the DNA probe having the specific nucleotide sequence hybridizes to a DNA sequence of a methylated form or an ummethylated form of a CpG dinucleotide in a sequence of a gene of the plurality of genes in the sample, such that the array is selected from having: at least 16 probes, at least 64 probes, at least 96 probes, and at least 384 probes; primers and reagents for detecting the hybridized probes and for detecting the reaction products derived from the hybridized probes; and instructions for using the array with a bisulfite reagent, thereby providing an estimation of proportions of leukocyte types in the sample.

In a related embodiment of the kit, the probes hybridize with a respective plurality of 96 different DNA sequences occurring in the plurality of genes. In yet another related embodiment of the kit the probes have nucleotide sequences complementary to 96 nucleotide sequences having SEQ ID NO: 1 to SEQ ID NO: 96.

The instructions in a related embodiment of the kit include methods for: reacting genomic DNA in the sample with the bisulfite reagent to convert unmethylated cytosine residues to uracil; hybridizing resulting bisulfite treated genomic DNA with probes immobilized to the surfaces to obtain resulting hybridized probes on the array, such that the DNA probes hybridize to a DNA sequence of each of a methylated form and an ummethylated form of a CpG dinucleotide sequence in a gene of the plurality of genes; and detecting the methylation status of the CpG dinucleotide sequence, thereby estimating proportions of leukocyte types in the sample from the subject for assessing the disease condition of the subject.

In a related embodiment of the kit the instructions for detecting the methylation status of the CpG dinucleotide sequence include methods for: extending each hybridized probe of the resulting hybridized probes on the array by primer extension to obtain a resulting primer extension product; ligating the resulting primer extension product to an oligonucleotide complementary to the DNA sequence of a 3' region of the gene to obtain a resulting template for

PCR on the array; and amplifying by PCR and measuring amount of resulting PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

In another related embodiment of the instructions for kit amplifying by PCR include methods for: amplifying the resulting template on the array using primers pairs having a 5' primer specific to each of the methylated or the unmethylated form of the CpG dinucleotide containing gene, and a 3'primer specific to the gene containing the CpG dinucleotide, thereby resulting in a first PCR product; amplifying the resulting first PCR product with differentially labeled 5' primers that specifically amplify each of the methylated and unmethylated form of the CpG dinucleotide sequence containing gene, and a common 3' primer, resulting in a differentially labeled second PCR product, and hybridizing the second PCR product to the CpG dinucleotide containing gene for measuring amount of the second PCR product, to detect the methylation status of the CpG dinucleotide sequence.

Instructions for detecting the methylation status of the CpG dinucleotide sequence, in another related embodiment of the kit, include methods for extending the resulting hybridized probes on the array by single base primer extension with a labeled nucleotide.

Another embodiment of the invention is a method of treating a subject for a disease condition, such that the subject is a human patient and, such that the disease condition is a cancer, the method comprising: obtaining signatures comprising differentially methylated regions (DMRs) from types of leukocytes in a blood sample of the patient, the types of leukocytes comprising at least one selected from: CD19+ B lymphocyte, CD15+ granulocyte, CD14+ monocyte, $CD56^{dim}$ Natural Killer cell, $CD56^{bright}$ Natural Killer cell, and CD3+ T lymphocyte, and from a healthy control human subject not having the cancer; comparing a signature specific for the type of leukocyte in the patient with that in the healthy subject, such that the type of leukocyte specific signature is an indication of amount of cells of the type of leukocyte circulating in blood, and such that a decreased amount of the cells of the type of leukocyte circulating in the blood of the patient compared to the healthy subject is an indicium of the cancer; and, administering a composition comprising the cells of the type of leukocyte to the patient, thereby increasing the amount of the cells of the type of leukocyte in the patient and treating the cancer.

In various embodiments of the method the leukocyte type cell is the $CD56^{dim}$ Natural Killer cell.

The cancer in related embodiments of the method is head and neck squamous cell carcinoma (HNSCC). In embodiments of the method the DMR signature specific for $CD56^{dim}$ Natural Killer cells includes at least one CpG dinucleotide in a region near the promoter of gene *NKp46*. In other embodiments of the method the DMR signature specific for $CD56^{dim}$ Natural

Killer cells is a CpG dinucleotide in a region near the promoter of the gene *NKp46*, such the methylation status of the CpG dinucleotide is quantified by methylation specific quantitative polymerase chain reaction (MS-qPCR) using primers and probes having SEQ ID NOs: 116-118 and 97-99. According to other embodiments of the method, the DMR signature specific for $CD56^{dim}$ Natural Killer cells is a CpG dinucleotide in a region near the promoter of the gene *NKp46*, such that the methylation status of the CpG dinucleotide is quantified by digital PCR involving emulsion and nanofluidic partitioning using primers and probes having SEQ ID NOs: 116-118 and 97-99.

In related embodiments of the method the blood sample is archival. Alternatively the blood sample is fresh.

Brief description of the drawings

Figure 1 is a photograph showing a clustering heatmap for External Validation White Blood Cell Data (S_0). The data were obtained by applying the measurement error formulation described in Examples 1-3. The method delineates effects resulting from immune cell distribution as compared to those resulting from other "non-cell type" alterations in DNA methylation. Methylation array procedure was carried out using Infinium HumanMethylation27 Beadchip Microarrays from Illumina, Inc. (San Diego, CA). The White Blood Cell data were gathered from a set of 46 samples of purified white blood leukocyte subtypes obtained commercially. Light = unmethylated ($Y_{hj} = 0$), black = partially methylated ($Y_{hj} = 0{:}5$), dark = methylated ($Y_{hj} = 1$).

Figure 2 is a chart showing the results of cell mixture reconstruction experiments validating prediction of individual sample profiles. The reconstruction experiments involved six known mixtures of monocytes and B cells and six known mixtures of granulocytes and T cells. Known fractions (Expected) and resulting predictions from Infinium 27K profiles (Observed) percentages of each cell type are shown by shade (dark =100, white=0).

Figure 3 is a photograph showing a clustering heatmap for Target HNSCC data (S_1). The target data set S_1 consisted of arrays applied to whole blood specimens collected in a random subset of individuals involved in an ongoing population-based case-control study (Peters et al., 2005) of head and neck cancer (HNSCC): 92 cases and 92 age and sex matched controls. Blood was drawn at enrollment (prior to treatment in 85% of the cases). Yellow = unmethylated ($Y_{hj} = 0$), black = partially methylated ($Y_{hj} = 0{:}5$), dark = methylated ($Y_{hj} = 1$). The annotation track above the heatmap indicates case-control status.

Figure 4 is a graphical representation of bias sensitivity analysis for HNSCC Data. Bias was assessed by resampling the case coefficients of $\mathbf{B}_1$, a procedure that assumes maximum bias. The abscissa shows the number of assumed non-zero alterations. The dark filled diamond shapes (red in color) indicate median, the thick vertical lines (blue in color) indicates interquartile range, the thin lines (blue in color) represent 95% probability ranges, and the outer dots (black in color) represent 99% probability ranges.

Figure 5 panels A-B are graphs showing Rate-of-Convergence of the Hessian matrix $\mathbf{H}_m$ which allows the determination of the optimal number of CpG sites whose combined methylation status measurements most accurately reflect the exact distribution of different cells in a mixture. The x-axis represents increasing *m*, the number of CpG sites (ordered by F-statistic) included in the model space, on a logarithmic scale.

Figure 5 panel A shows convergence by correlating the Hessian Matrix with the number of CpG sites included in the measurement. The dotted line in (A) shows the tangent at low values of *m*.

Figure 5 panel B shows the Rate of convergence which was calculated by smoothing the first differences of $\log_{10}(\mathrm{tr}\mathbf{H}_m)$. The dotted line (red in color) in (B) corresponds to linear convergence.

Figure 6 is a photograph showing a clustering heatmap for Target Ovarian Cancer data (S_1) (Teschendorff et al., 2009, *PLoS ONE* 4, e8274). Only those cases were included in which blood was collected pre-treatment. After removing four arrays with a preponderance of missing values, the data set consisted of 272 controls and 129 cases having blood drawn prior to treatment. Light = unmethylated ($Y_{hj} = 0$), black = partially methylated ($Y_{hj} = 0{:}5$), dark = methylated ($Y_{hj} = 1$). The annotation track above the heatmap indicates case-control status (cancer case or control).

Figure 7 is a photograph showing a clustering heatmap for Target Down Syndrome Data. The method herein was applied to a trisomy 21 (Down syndrome) data set (Kerkel et al.,. *PLoS Genet* 2010, **6**(11):e1001212) consisting of 29 total peripheral blood leukocyte samples from Down syndrome cases and 21 controls, as well as six T cell samples from cases and four T cell samples from controls (GEO Accession number GSE25395). Light = unmethylated ($Y_{hj} = 0$), black = partially methylated ($Y_{hj} = 0{:}5$), dark = methylated ($Y_{hj} = 1$). The annotation track

above the heatmap indicates case-control and cell type status [Down syndrome case (whole blood), control (whole blood), T cell (pooled cases and controls)].

Figure 8 is a photograph showing a clustering heatmap for Target Obesity Data obtained from applying the method herein to an obesity data set (Wang et al., BMC Med 2010, 8:87) consisting of 7 lean African-Americans and 7 Obese African-Americans (GEO Accession number GSE25301). Yellow = unmethylated (Y_{hj} = 0), black = partially methylated (Y_{hj} = 0:5), grey = methylated (Y_{hj} = 1). The annotation track above the heatmap indicates case-control status (obese and lean).

Figure 9 is a photograph (heatmap) of the methylation profiles of white blood cells obtained from a DNA methylation array analysis described in Example 9. Methylation array procedure was carried out using Infinium HumanMethylation27 Beadchip Microarrays from Illumina, Inc. (San Diego, CA). The number of individual leukocyte samples in each methylation class is shown in the table to the right. The DNA methylation profile distinguishes Lymphocytes from Myeloid Derived Leukocytes. The highest 5000 most variable CpG loci are plotted on the left. Less methylated loci are grey and more methylated loci are black. Recursively partitioned mixture model (RPMM) of autosomal gene Infinium beta values from sorted, human, peripheral blood leukocytes was performed in R version 2.11.1 of Illumina's software which provides convenient mechanisms for loading and analyzing the results of methylation status, and for quality control and basic visualization tasks.

Figure 10 panels A-B are graphical representations of the DNA methylation status of regions in *CD3E* and *CD3Z* genes.

Figure 10 panel A shows DNA methylation status of a region in *CD3E* that was identified from the DNA methylation array analysis (the results of which are shown in Figure 9) as one of the two candidate DMRs with specificity towards CD3+ T cells. The DNA methylation status was measured by pyrosequencing bisulfite converted DNA from different sorted, human, peripheral blood leukocytes.

Figure 10 panel B shows DNA methylation status of a region in *CD3Z* gene that was identified from the DNA methylation array analysis (the results of which are shown in Figure 9) as one of the two candidate DMRs with specificity towards CD3+ T cells. The DNA methylation status of the region in *CD3Z* gene in different sorted, human, peripheral blood leukocytes was measured by MethyLight® qPCR.

Figure 11 is a drawing showing the genomic region containing CD3Z gene, based on information available from the public databases UniProt, RefSeq and GenBank. UniProt is a freely accessible universal protein resource of protein sequence and functional information. RefSeq is a collection that provides integrated and annotated set of sequences including genomic DNA, transcripts and protein. GenBank® is the genetic sequence database of the National Institutes of Health which contains an annotated collection of all publicly available DNA sequences.

Figure 12 is a list of genomic regions used for measuring methylation of CD3Z and FOXP3 gene, for quantitating genome copy numbers, and a list of the corresponding primer and probe sequences. Underlined letters are "C" in CpG motifs.

Figure 13 panels A-C are graphical representations of standard calibration curves which show the relationship between copy numbers of genomic DNA and the signal obtained from quantitaive real time methylation specific PCR. The calibration curves are used for quantifying CD3+ T cells, Tregs (FOXP3 demethylated) and ratios of Tregs/CD3+ T cells. DNA isolated from purified cell types was bisulfite converted and serially diluted into a background of fully methylated commercial DNA standard (Qiagen). The total genomic copy numbers of each sample within a dilution series remained constant. Log dilutions were performed in the appropriate range of Ct values corresponding to test samples (whole blood, tumor specimens). Using cytosine-less: C-less primers genome copy numbers for each test standard were measured to ensure adequate input DNA and to normalize the CD3+ and Treg assay values.

Figure 13 panel A shows the calibration curve for C-less total input. (N= eight replicates); errors denote standard error of the mean Ct value.

Figure 13 panel B shows dilution of isolated normal PanT cells (N= seven replicates).

Figure 13 panel C shows dilution and calibration curve for isolated CD3+CD25+ T cells (N=8 eight replicates).

Calibration curves (Figure 13 panels A,B,C) were used to estimate total input copies, CD3+ T cell and Tregs copies, respectively.

Figure 14 is a drawing and a set of graphical representations showing detection of CD3+ T cell numbers by measuring differential demethylation using MS-qPCR.

Figure 14 panel A is a schematic diagram showing methylation specific primers and probe targeting six CpGs (lollipops) in a region of the *CD3Z* gene identified herein as demethylated in CD3+ T cells.

Figure 14 panel B shows results of real time PCR. The real time PCR Ct values decreased linearly with a ten-fold increase in bisulfite converted CD3+ T cell DNA concentration. Bisulfite converted universal methylated DNA was used to keep total amount of DNA in all samples constant. At least five replicates of each sample were plotted.

Figure 14 panel C shows correlation between T cell levels determined by flow cytometry and *CD3Z* MS-qPCR. Evaluation of CD3+ T cell level by flow cytometry was observed to be highly correlated with T cell quantification by *CD3Z* MS-qPCR in whole blood specimens from glioma patients and healthy donors.

Figure 14 panel D shows correlation between T cell counts obtained using by imunohistochemical staining and *CD3Z* MS-qPCR. CD3+ T cell count by imunohistochemical staining correlates with T cell quantification by *CD3Z* MS-qPCR in excised tumors across histological subtypes. Pearson correlations and F-test p-values are shown in panels B-D.

Figure 15 panels A-C are graphical representations showing T cells and Tregs in the peripheral blood of glioblastoma multiform (GBM) patients and healthy donors determined by MS-qPCR for demethylation of specific CpG loci.

Figure 15 panel A shows comparison of T cell numbers in blood between GBM patients and control subjects measured using *CD3Z* demethylation assay.

Figure 15 panel B shows comparison of Tregs between GBM patients and control subjects measured using *FOXP3* demethylation assay.

Figure 15 panel C is a graph showing comparison of Treg percent of T cells between GBM patients and control subjects determined by the ratio of *FOXP3/CD3Z* demethylation. Wilcoxon rank sum p-values are shown.

Figure 16 panels A-C are graphical representations showing association between cigarette smoking and peripheral blood T cells and Tregs in glioma patients and healthy donors determined by MS-qPCR for demethylation of specific CpG loci.

Figure 16 panel A shows a comparison of peripheral blood T cell levels, determined by *CD3Z* demethylation, among never, former and current cigarette smokers stratified by glioma case status (indicated "cases" on the abscissa).

Figure 16 panel B shows a comparison of peripheral blood Treg levels, determined by *FOXP3* demethylation, among never, former and current cigarette smokers stratified by glioma case status.

Figure 16 panel C shows a comparison of peripheral blood Treg percent of T cells, determined by ratio of *FOXP3* to *CD3Z* demethylation, among never, former and current cigarette smokers stratified by glioma case status. Wilcoxon rank sum p-values are shown.

Figure 17 panels A-C are graphical representations showing levels of T cell and Treg infiltrates in excised glioma tumors determined by MS-qPCR for demethylation of specific CpG loci.

Figure 17 panel A shows T cell levels, determined by *CD3Z* demethyation, in solid glioma samples stratified by tumor grade.

Figure 17 panel B shows Treg levels, determined by *FOXP3* demethyation, in solid glioma samples stratified by tumor grade.

Figure 17 panel C shows Treg percent of T cells, determined by ratio of *FOXP3* to *CD3Z* demethylation, in solid glioma samples stratified by tumor grade. Wilcoxon rank sum p-values are shown.

Figure 18 panels A-C are graphical representations of flow cytometry analysis of CD3+ T cells and total leukocytes in whole blood from glioma cases and controls.

Figure 18 panel A shows a forward and side scatter plot of a representative blood sample showing gating for lymphocytes and counting beads.

Figure 18 panel B shows lymphocyte subpopulation observed using gating for CD3 expression.

Figure 18 panel C shows CD45 gating on all non-bead events. CD45+ low and high cells were added in order to count total CD45+ cells.

Figure 19 panels A-C are photomicrographs and a lie graph that show immunohistochemical (IHC) staining of a representative GBM specimen.

Figure 19 panel A shows CD3 staining. Average number of cells positive for staining was 418.

Figure 19 panel B shows CD8 staining. Average number of cells positive for staining was 296.

Figure 19 panel C shows correlation of CD3 and CD8 staining, Pearson r =.992

Figure 20 is a set of two heatmaps showing results of MS-qPCR and bisulfite pyrosequencing of Magnetic activated cell sorting (MACS) sorted human leukocyte subsets. Abbreviations: B = B lymphocytes, Gran = Granulocytes, Neut = Neutrophils, Mono =

Monocytes, NK = CD56+ Natural killer cells, Nkdim = CD16+CD56dim natural killer cells, NKbr = CD16-CD56bright natural killer cells, NK8+ = CD8+CD56+ natural killer cells, NK8- = CD8-CD56+ natural killer cells, NKT = CD3+CD56+ natural killer T cells, T = CD3+ T lymphocytes, CD8 = CD3+CD8+ T lymphocytes (cytotoxic T cells), CD4 = CD3+CD4+ T lymphocytes (helper T cells), Treg = CD3+CD4+CD25+FOXP3+ regulatory T cells.

Figure 20 panel A is a heatmap of DNA methylation in FOXP3 and CD3Z gene regions assessed by MS-qPCR.

Figure 20 panel B is a heatmap of DNA methylation at three CpG loci in the CD3Z gene assessed by bisulfite pyrosequencing.

Figure 21 panels A-C are graphical representations showing levels of T cell and Treg infiltrates in glioma tissues stratified by histological subtype determined by MS-qPCR for demethylation of specific CpG loci. Abbreviations: PA = Pilocytic Astrocytoma, EP = Ependymoma, OD = Oligodendroglioma, OA = Oligoastrocytoma, AS = Astrocytoma, GBM = Glioblastoma multiforme. Kruskal-Wallis one-way analysis of variance by rank test p-values shown.

Figure 21 panel A shows T cell levels determined by CD3Z demethylation in solid glioma samples stratified by tumor histology.

Figure 21 panel B shows Treg levels determined by FOXP3 demethylation in solid glioma samples stratified by tumor histology.

Figure 21 panel C shows Treg percent of T cells, determined by ratio of FOXP3 to CD3Z demethylation in solid glioma samples stratified by histology.

Figure 22 panels A-C are graphical representations showing Kaplan Meier analysis of time of survival of glioma patients stratified according to whether the level of T cells or Tregs in the tumor infiltrates of the patients are above or below the median level of T cells or Tregs, respectively. Log Rank p-values shown.

Figure 22 panel A shows survival (ordinate) of glioma patients as a function of time (abscissa) in relation to T cell levels as determined by *CD3Z* demethylation.

Figure 22 panel B shows survival of glioma patients in relation to Treg levels as determined by *FOXP3* demethylation.

Figure 22 panel C shows survival of glioma patients in relation to Treg percent of T cells as determined by ratio of *FOXP3* to *CD3Z* demethylation.

Figure 23 panels A-B are representations of results obtained from analysis of DMRs of leukocyte subtypes.

Figure 23 panel A shows a heat map of the methylation status for the highest ranked 50 leukocyte DMRs by leukocyte subtype.

Figure 23 panel B shows a Plot depicting the -log10(P-values) for the highest ranked 50 leukocyte DMRs across three cancer data sets (HNSCC; Ovarian; Bladder). P-values (ordinate) show methylation differences between cancer cases and non-cancer controls and were obtained from individual unconditional logistic regression models fit to each of the 50 leukocyte DMRs. For the HNSCC data set, logistic regression models were adjusted for patient age, gender, smoking status (never, former, current), smoking pack years, weekly alcohol consumption, and HPV serology status. The bladder cancer data set was adjusted for patient age, gender, smoking status, smoking pack years, and family history of bladder cancer. The ovarian cancer data set was adjusted for patient age group (55-60, 60-65, 65-70, 70-75 and >75 years). The horizontal dashed line represents -log10($p = 0.05$).

Figure 24 panels A-B shows results obtained from the DMR profile analysis of the HNSCC data set determining methylation class membership.

Figure 24 panel A left column shows a heat map of the HNSCC testing data set. Rows represent subjects, which are grouped by predicted methylation class membership. Columns represent the highest ranked 50 leukocyte DMRs that were used to generate the methylation classes for the HNSCC testing set. Panel A right column is a bar-plot depicting the percent cancer case/control across the predicted methylation classes in the HNSCC testing set.

Figure 24 panel B shows receiver operating characteristic (ROC) curves based on the predicted methylation classes only in the HNSCC testing set and methylation classes including patient age, gender, smoking status (never, former, current), smoking pack years, weekly alcohol consumption, and HPV serostatus.

Figure 25 shows results obtained from the DMR profile analysis of the Ovarian data set for determining methylation class membership.

Figure 25 panel A is a heat map of the ovarian testing data set. Rows represent subjects which are grouped by predicted methylation class membership. Columns represent the highest ranked ten leukocyte DMRs that were used to generate the methylation classes for the ovarian testing set. Panel A right column is a bar-plot depicting the percent cancer case/control across the predicted methylation classes in the ovarian testing set.

Figure 25 panel B shows ROC curves based on the predicted methylation classes alone in the ovarian testing set and methylation classes plus patient age group (55-60, 60-65, 65-70, 70-75 and >75 years).

Figure 26 shows results obtained from the DMR profile analysis of the bladder data set for determining methylation class membership.

Figure 26 panel A is a heat map of the bladder testing data set. Rows represent subjects, which are grouped by predicted methylation class membership. Columns represent the highest ranked 56 leukocyte DMRs that were used to generate the methylation classes for the bladder testing set. Panel A right column represents a bar-plot depicting the percent cancer case/control across the predicted methylation classes in the bladder testing set.

Figure 26 panel B shows ROC curves based on the predicted methylation classes alone in the bladder testing set and methylation classes plus patient age, gender, smoking status (never, former, current), smoking pack years, and family history of bladder cancer.

Figure 27 panels A-C are graphical representations showing image plots representing the pairwise spearman correlation coefficients.

Figure 27 panel A shows the six CpG loci identified by HNSCC analysis (Langevin SM et al., Epigenetics. 2012 Mar; 7(3):291-9) and the highest ranked 50 leukocyte DMRs used in the present analysis.

Figure 27 panel B shows the seven CpG loci identified by the alternative ovarian analysis and the highest ranked ten leukocyte DMRs used in the present analysis, and (c) the nine CpG loci identified by the bladder analysis reported in (Laird PW, 2003 Nat Rev Cancer 3:253–266) and the highest ranked 56 leukocyte DMRs used in the present analysis.

Figure 27 panel C shows the nine CpG loci identified by the bladder analysis reported in (Shen L et al., 2007 PLoS genetics 3:2023-2036) and the highest ranked 56 leukocyte DMRs used in the present analysis.

Figure 28 is a schematic diagram showing hierarchy of leukocyte subtypes and sample sizes for each of the leukocyte subtypes used in the analysis for determination of methylation class membership.

Figure 29 is a diagram representing the analytic workflow the HNSCC data set (n = 184; 92 HNSCC cases and 92 cancer-free controls). The full HNSCC data set was first divided into equally sized training and testing sets. The training sets were used in development of a classifier

based on leukocyte DMRs. The resulting classifiers were then used to predict methylation class membership for the observations in the respective independent testing sets. The phenotypic importance of the predicted methylation classes in the testing data was examined subsequently.

Figure 30 is a diagram representing the analytic workflow the ovarian cancer data set (n = 401; 128 ovarian cancer cases and 273 cancer-free controls). The full ovarian cancer data set was divided into equally sized training and testing sets. The training sets were used in the development of a classifier based on leukocyte DMRs. The resulting classifiers were then used to predict methylation class membership for the observations in the respective independent testing sets. The phenotypic importance of the predicted methylation classes in the testing data was then examined.

Figure 31 is a diagram representing the analytic workflow of the bladder cancer data set (n = 460; 23 Bladder cancer cases and 237 cancer-free controls). The full bladder cancer data set was divided into equally sized training and testing sets. The training sets were used in the development of a classifier based on leukocyte DMRs. The resulting classifiers were then used to predict methylation class membership for the observations in the respective independent testing sets. The phenotypic importance of the predicted methylation classes in the testing data was then examined.

Figure 32 is a diagram illustrating Semi-Supervised Recursively Partitioned Mixture Models (SS-RPMM) for predicting methylation class membership. The full methylation dataset was randomly divided into training and testing sets. Using the training data only, univariate models (adjusted for potential confounders) were used to identify CpG loci whose methylation is most strongly associated with the clinical variable of interest (i.e., case/control status). RPMM is then fit to the training data using the M CpGs that are most associated with the clinical variable of interest (M is determined using a nested cross-validation procedure) CpGs. The resulting solution is then used in conjunction with an empirical Bayes classifier to predict methylation class membership for the observations in the testing data.

Figure 33 panels A-D show results obtained from SS-RPMM analysis (see Figure 30) of the ovarian cancer data set for determination of methylation class membership.

Figure 33 panel A is a heatmap of the testing set obtained by predicted methylation class using the SS-RPMM procedure. Rows represent subjects and columns represent the seven CpG loci identified by this analysis.

Figure 33 panel B represents percentage of cases/controls obtained by predicted methylation class membership in the testing set.

Figure 33 panel C sows information regarding the seven CpG loci identified by the SS-RPMM analysis.

Figure 33 panel D shows a ROC/AUC (area under the curve) analysis based on the predicted methylation class memberships in the testing set. Dark represents the ROC/AUC based on the predicted methylation classes along and light represents the ROC/AUC using the predicted methylation classes and patient age group.

Figure 34 is a graphical representation showing loci in the gene *NKp46* chosen from candidate NK cell-specific differential DNA methylation markers, selected by DNA methylation and mRNA expression criteria.

Linear mixed effects modeling of DNA methylation microarray data from MACS isolated human leukocytes generated a coefficient estimating differential methylation in NK cells relative to other cell subtypes, shown on the avscissa. Linear modeling of mRNA microarray data from the same isolated cells determined log-fold change in expression between NK cells and each of the following subtypes: T cells, B cells, granulocytes and monocytes. The average of these four log-fold change values is shown on the ordinate. Significance for a particular gene region was achieved when $q < 0.1$ for four mRNA expression linear models as well as the DNA methylation mixed effects model. Candidates for NK cell-specific DNA methylation biomarkers were limited to significant gene loci exhibiting decreased methylation in NK cells (methylation estimate < 0) and within genes that exhibited increased RNA expression (log fold change > 1). The candidate loci are marked with asterisks in the top left quadrant, and *NKp46* loci are marked with grey asterisks.

Figure 35 is a heatmap showing demethylation status of *NKp46* determined by methylation specific quantitative PCR (MS-qPCR) of isolated human leukocyte populations. Individual samples of (MACS) purified white blood cell subtypes were subjected to a MS-qPCR assay that detects demethylated copies of *NKp46* DNA. Extent of *NKp46* methylation is illustrated in this heatmap in which light indicates that all copies of DNA in particular sample were demethylated in the targeted region of *NKp46*, and dark indicates that all copies were methylated.

Figure 36 is a line graph showing linearity of *NKp46* MS-qPCR calibration. Bisulfite converted universal methylated DNA was used to standardize total amount of DNA in all

samples at a constant amount. At least three replicates of each standard are plotted. Real time PCR Ct values decrease linearly with ten-fold increase in bisulfite converted NK cell DNA concentration.

Figure 37 is a bar graph showing prevalence of HNSCC by normal *NKp46* demethylation tertile. Normal *NKp46* demethylation tertile cutoffs were determined from control blood samples only. Higher tertiles indicate higher NK cell levels. HNSCC prevalence (ordinate) refers to the percent of total cases in this example whose *NKp46* demethylation measurements fell within the control derived tertile range. Displayed p-value is from a chi-squared test for trend in proportions.

Figure 38 is a heatmap showing methylation status of selected *NKp46* CpG loci measured by bisulfite pyrosequencing of isolated human leukocytes. The methylation status of eight individual CpG loci near the promoter region of *NKp46* were interrogated by pyrosequencing of bisulfite converted DNA extracted from Magnetic activated cell sorting (MACS) isolated human leukocyte populations. CpG numbers 2 through 7 represent the six loci targeted in the MS-qPCR assay. This heatmap displays methylation levels at each locus ranging from unmethylated (light) to methylated (dark).

Figure 39 is a graph showing percent demethylation (ordinate) of a DNA region in *NKp46* in control and HNSCC patient blood samples (abscissa) assessed by MS-qPCR. The *NKp46* MS-qPCR assay measures the extent of DNA demethylation. A higher level of demethylation indicates a higher level of NK cells within a sample. Wilcoxon rank sum p-value is displayed.

Figure 40 is a listing of DNA sequences of regions in 96 different genes, each sequence having one CpG dinucleotide shown within square brackets and used to determine methylation status of the gene. The DNA sequence surrounding the CpG dinucleotides was used to design probes for the array and for primers for performing the methods for analyzing differential methylation. Also included are the names of the genes, chromosome number indicating the chromosome in which each genes is located, the source of the DNA sequences, Genebank accession numbers, and the coordinate of the CpG dinucleotide in each respective gene.

Figure 41 is a schematic diagram showing different ways of representing effects on measured DNA methylation due to an exposure or a specific phenotype.

Figure 41 panel A depicts the marginal effects (β) on measured DNA methylation. The marginal effects are effects which are not adjusted for white blood cell (WBC) distribution.

Figure 41 panel B depicts the effects on measured DNA methylation adjusted for WBC distribution resulting from exposure or a specific phenotype.

Figure 42 is a set of graphical representations showing the relationship between $\hat{\alpha}$ and $\hat{\beta}$, the effect on measured DNA methylation not adjusted or adjusted for WBC distribution, for the covariate (e.g. age, current smoker status, toe Arsenic concentration and Dye use) of interest over all autosomal CpGs. Dots represents overall methylation as indicated by the first component of the coefficient vector $\hat{\boldsymbol{\beta}}$, corresponding to the intercept (Example 38), light=low, black=moderate, dark=high. The diagonal straight line represents identity ($\hat{\alpha} = \hat{\beta}$). The curve depicts a loess fit to the scatter plot.

Figure 43 is a graphical representation showing fluorescence intensities of CD3Z gene amplified by digital droplet PCR, and a graphical representation showing concentration of CD3Z gene in PCR samples.

Figure 43 panel A shows a fluorescence intensity dot plot for amplification of CD3Z gene by detection of intensities of 6 FAM (6-Carboxyfluorescein). Positive and negative droplets are distinguished by a horizontal line.

Figure 43 panel B shows a correlation of the concentration of copy numbers of CD3Z gene obtained by measuring 6 FAM fluorescence intensities and the expected copy numbers of CD3Z gene obtained by dilution of a known amount of DNA from CD3+ T cells.

Figure 44 is a graphical representation showing fluorescence intensities of FoxP3 gene amplified by digital droplet PCR, and a graphical representation showing concentration of FoxP3 gene in PCR samples.

Figure 44 panel A shows a fluorescence intensity dot plot for amplification of FoxP3 gene by detection of intensities of 6 FAM (6-Carboxyfluorescein). Positive and negative droplets are distinguished by a horizontal line.

Figure 44 panel B shows a correlation of the concentration of copy numbers of FoxP3 gene obtained by measuring 6 FAM fluorescence intensities and the expected copy numbers of FoxP3 gene obtained by dilution of a known amount of DNA from CD3+ T cells.

Figure 45 is a graphical representation showing fluorescence intensities of NKp46 gene amplified by digital droplet PCR, and a table showing concentration of NKp46 gene in the PCR samples amplified under different conditions.

Figure 45 panel A shows a fluorescence intensity dot plot for amplification of NKp46 gene under different conditions by detection of intensities of 6 FAM (6-Carboxyfluorescein). Positive and negative droplets are distinguished by a horizontal line.

Figure 45 panel B is a table showing concentration of NKp46 gene in copies/µl determined under different PCR conditions as fractions of methylated control DNA.

Figure 46 is a graphical representation showing fluorescence intensities of NKp46 gene amplified by digital droplet PCR, and a table showing concentration of NKp46 gene in the PCR samples amplified under different conditions.

Figure 46 panel A shows a fluorescence intensity dot plot for amplification of NKp46 gene by detection of intensities of 6 FAM (6-Carboxyfluorescein). The amplification of demethylated NKp46 locus was performed using C-less and NKp46 DMR specific primers and probes, and results compared. Positive and negative droplets are distinguished by a horizontal line.

Figure 46 panel B is a table showing concentration of NKp46 gene in copies/µl determined with whole blood DNA, Neutrophil DNA, CD16+CD56^{dim} NK cell DNA and CD16+CD56^{bright} NK cell DNA.

Detailed description of the invention

A model of hematopoiesis includes an early restriction point at which multipotent progenitor cells become committed to either lymphoid or myeloid lineages. The standard methods of distinguishing immune cell lineages are inadequate for fully distinguishing lineage commitment and the process of hematopoiesis.

Epigenetics refers to heritable control of gene expression that occurs without changing the sequence of DNA. Chromatin packaging is a mechanism of epigenetic gene regulation which has been implicated in cell lineage commitment and lineage-specific gene expression. Transcriptionally inactive, or silenced, heterochromatin is more tightly packaged around histone proteins than transcriptionally active euchromatin due to differences in DNA methylation patterns and post-translational histone modifications. Due to its accessibility for measurement, DNA methylation is a marker of chromatin packaging. DNA methylation is largely confined to cytosine residues in CpG dinucleotides which, though underrepresented in the genome, are

frequently found in high concentrations called CpG islands. Less methylated CpG islands are highly associated with transcriptional activity and subsequent gene expression, and more methylated CpG islands are highly associated with transcriptional inactivity and gene silencing. Methylation of CpG dinucleotides causes chromatin to become more compact and inaccessible to transcription machinery by moving histones and altering the organization of chromatin and nucleosomes.(Christensen, B.C., *et al.* 2009, *PLoS Genet* **5**, e1000602; Schmidl, C., *et al.* 2009, *Genome Res* **19**, 1165-1174).

In some instances, the overall balance of leukocyte subclasses in circulation or in tissue most prominently influences pathogenesis. For example, incipient cancer cells are recognized and eliminated by cytotoxic T cells (CTLs) and natural killer (NK) cells, and tumorigenesis is also promoted by certain other inflammatory cells, including B-lymphocytes, mast cells, neutrophils, regulatory T cells (Tregs), and others. These cells have been shown to promote angiogenesis, tumor cell proliferation, tissue invasion and metastasis (Hanahan and Weinberg 2011, *Cell*, **144**, 646-74; Ostrand-Rosenberg, 2008, *Curr Opin Genet Dev*, **18**, 11-18). Likewise, higher levels of NK cells and CTLs circulating in the blood and residing in adipose tissues are associated with lower incidence of metabolic diseases such as type II diabetes (Lynch et al., 2009, *Obesity*, **17**, 601-5), and higher levels of M1 macrophages in adipose tissue can induce inflammation and insulin resistance (Anderson et al., 2011, *Curr Opin Lipidol.* **21**, 172-177). Methods of quantifying the composition of lymphocyte populations can be informative regarding the underlying immuno-biology of disease states as well as the immune response to almost all chronic medical conditions. (Chua et al., 2011, *Brit J Cancer* **104**, 1288-1295).

The methods described herein provide a measurement of individual human or animal immune cell numbers or immune cell ratios and in diverse biologic media without the requirement for viable cells or cell sorting or the use of any antibodies or protein markers. The methods are applicable to blood including samples of unsorted blood that is fresh, or is frozen or unfrozen anticoagulant treated peripheral whole blood, finger stick blood, non-anticoagulant treated whole blood, blood clots, isolated mononuclear cells, buffy coat, archival Guthrie card neonatal blood, and to a sample that is a spot, fresh, frozen or is from a tumor such as a formalin-fixed tumor biopsy, and to urine sediment, CNS fluid, fat or other tissue biopsy.

In one embodiment the methods described herein are provided as diagnostic kits for testing laboratories in the form of immune cell specific detection reagents, premixed and optimized plate formatted multiplex assays for immune profiling compatible with specific instrument platforms, applications for in vitro diagnostics of blood, CNS, urine or bronchoalveolar lavage and point of care blood sampling kits for mail-in immune testing and immune monitoring.

The simplified DNA based immuno-diagnostic approach provided herein uses samples that are much smaller volumes of blood than required for earlier methods and that require no processing. These samples can be simply 'spotted' onto a solid phase carrier and transported through the mail or delivered using courier.

In another embodiment, the methods described include development of software that can process the output data of immune specific methylation assays to create immune parameter reports by comparison to different reference and control values.

In an alternate embodiment the methods herein describe a discovery platform which is a bioinformatic integration of empirically derived genome wide methylation analyses with publically available differential gene expression analyses. The merged datasets are then sorted to produce candidates for further examination. The discovery platform is useful to discover clinically useful gene biomarkers.

The methods described herein include a proof-of-principal test of the discovery platform. For the test the goal set was to discover a gene or gene set that provides a marker of CD3+ T cells. The method is applicable to finding a biomarker for any cell. Specifically, the platform identifies gene regions that are 'demethylated' within the target cell population (CD3+ T cell) and completely methylated in non-target cells.

To accomplish this discovery phase for the set goal, normal immune cells from the peripheral blood of different individuals was isolated using flow cytometry antibody based cell sorting. Following purification each of the immune cell subtypes was subjected to methylation discovery analysis using the Infinium genome-wide methylation platform. (Infinium® HumanMethylation27 Beadchip Microarray, developed by Illumina®, Inc., San Diego, CA). The DNA methylation data was then merged with existing gene expression data. Candidates that have high potential to discriminate CD3+ T cells from non-T cells were then further analyzed with two different methylation validation methods (pyrosequencing and quantitative methylation specific PCR i.e. MethylLight). Finally, a quantitative calibration curve was developed by diluting known and measured numbers of CD3+ T cells into a background matrix of fully methylated lymphocyte DNA. The latter procedure reconstructs the conditions of detection that are present in differentiating CD3+ T cells from a mixture of cells in a complex biological sample.

The methods described herein use individual samples of sorted, normal, human, peripheral blood leukocytes shown in Table 15, Example 13, purchased from AllCells®, LLC (Emeryville, CA). These leukocytes were sorted in a column containing antibody-conjugated magnetic beads through a combination of positive and negative selection. DNA from the leukocytes was extracted according to manufacturer's protocol using the DNeasy Blood &

Tissue kit (Qiagen), and subjected to Bisulfite conversion by treatment with sodium bisulfite using the EZ DNA Methylation Kit (Zymo) following the manufacturer's protocol, thereby converting unmethylated cytosine residues to uracil and leaving methylated cytosine residues intact. DNA methylation is measured using a DNA methylation microarray as described in Example 13.

Huehn et al. (U.S. patent publication number 2007/0269823 A1) describes a method for identifying FoxP3-positive regulatory T cells by analyzing the methylation status of CpG positions in the FOXP3 gene, and further describes a method for diagnosing immune status of a mammal by measuring amounts of regulatory T cells thus identified. CpG methylation analysis of FoxP3 gene is also used to determine the quality of *in vitro* generated T regulatory cells and for identifying chemical or biological substances that modulate the expression of the FOXP3 gene in T cells. Specific CpG positions in the mouse FoxP3 gene are identified for analyzing methylation status and primers for amplifying mouse and human CpG dense regions in FOXP3 gene are described.

Olek (U.S. patent publication number 2007/0243161 A1) describes a method for pan-cancer diagnostics involving identification of an amount and/or proportion of stable regulatory T cells in a patient suspected of having cancer by analyzing methylation status of CpG positions in the FOXP3 and/or camta1 genes. Increased amount/proportion of stable regulatory T cells in the patient is indicative of an unspecified cancerous disease. A method of treating cancer by reducing the amount or proportion of stable regulatory T cells and a method for diagnosing survival of a cancer patient by measuring T regulatory cell amounts and/or proportions in patients suspected of having cancer using CpG methylation analysis of FoxP3 and/or camta1genes are described. Increased amounts and/or proportions of stable regulatory T cells in the cancer patient is indicative of a shorter survival.

Olek et al. (International publication number WO 2010/069499 A2) describes a method of identifying T-lymphocytes, in particular CD3+CD4+ and/or CD3+CD8+ cells by analyzing the methylation status of CpG positions in one or more of genes for CD3 multi-protein complex CD3 γ, -δ and –ε, or in other genes. Demethylation is indicative of a CD3+ cell. Olek further describes methods for methylation analysis of CpG positions in CD4+ and/or CD8+ genes, in particular CD8 beta gene, or in other genes, and for determining immune status based on T-lymphocytes identified by methylation analyses, and for monitoring amounts of T-lymphocytes in response to chemical and/or biological substance exposure, in particular CD4+ or CD8+ T lymphocytes.

Shen-Orr et al. 2010, Nature Methods Vol. 7:4, 287-289 describes a cell-type specific significance analysis of microarrays for analyzing differential gene expression for each cell type

in a biological sample from microarray data and relative cell type frequencies. In Shen-Orr's method relative abundance of each cell type in a mix tissue sample is first quantified, and this information is used in combination with microarray gene expression data to deconvolve and compare cell type-specific average expression profiles for groups of mixed tissue samples.

Abbas et al. 2009, PLoS One Vol. 4:7 e6098 describes deconvolution of microarray gene expression data to characterize proportions of cells in a tissue, and further identifies cellular activation patterns in Systematic Lupus Erythematosus.

A method similar to regression calibration is provided herein for determining changes in the distribution of white blood cells between different subpopulations (e.g. cases and controls) using DNA methylation signatures ro DNA methylation profiles, in combination with an external validation set having methylation signatures from purified leukocyte samples. The method is demonstrated with Head and Neck Squamous Cell Carcinoma (HNSCC) cases and matched controls, showing that DNA methylation signatures register known changes in CD4+ and granulocyte populations.

Use of DMRs as markers of immune cell identity is employed herein with a high density methylation platform, and a set of analytical tools for estimating the proportions of immune cells in unfractionated whole blood to determine the DNA methylation signature of each of the principal immune components of whole blood (B cells, granulocytes, monocytes, NK cells, and T cells subsets). A form of regression calibration was determined that considers a methylation signature as a high-dimensional multivariate surrogate for the distribution of white blood cells. This distribution was used to predict or model disease states. As a surrogate, the DNA methylation signature was assumed to be a highly correlated measure of leukocyte distribution, and thus fits into the framework of measurement error models, in which the use of a noisy surrogate marker to investigate an association with a disease outcome of interest results in biased estimates, unless internal or external validation data are obtained to "calibrate" the model and correct the bias (Carroll et al., 2006, *Measurement error in nonlinear models.* Chapman & Hall, Boca Raton, Florida, 2nd edition).

In this case, the problem was complicated by the extremely high dimension of the surrogate. Measurement error problems are formulated as a set of relationships between z, the disease outcome (e.g. case/control status), ω, the gold standard (e.g. leukocyte distribution), and **y**, the surrogate (e.g. DNA methylation). The concept $E(z|\omega)$, was difficult to estimate due to the cost or logistical complications involved in obtaining ω in a large number of samples. Sufficient data for modeling $E(z|\mathbf{y}) = f(\mathbf{y})$ were collected, which provides information about $E(z|\omega)$ through the (often imperfect) association $E(\mathbf{y}|\omega) = g(\omega)$, which is inferred from an external

validation sample (Thurston et al., 2003, *J Stat Plan Inf*, 113, 527-34; Carroll et al., 2006, *Measurement error in nonlinear models*. Chapman & Hall, Boca Raton, Florida, 2nd edition). An additional assumption was that $E(z|\omega,\mathbf{y}) = E(z|\omega)$, i.e. the surrogate provides no information about disease above and beyond the standard for which it serves as a surrogate. The high-dimensional nature of $\mathbf{y}$ renders $f(\mathbf{y})$ difficult to formulate. Although multivariate methods of measurement error correction exist, even in a high-dimensional context (e.g. Li and Yin, 2007, *Ann Stat*, 35, 2143-72) an explicit specification of $f(\mathbf{y})$ is important, which becomes unwieldy as each component of $\mathbf{y}$ contributes a small amount of information about z, and both dimension-reduction strategies and constrained regression strategies entail substantial loss of information. In the present context, specification of $\mathbf{y} = f(z)$ is natural and straightforward. Consequently, a reversal of the modeling equation is here provided, formulating $\mathbf{y} = f(z)$ as part of the modeling strategy, and linking the linear functions f and g in a manner that admits the estimation of ω. In methods herein several major sources of possible bias were identified and methods provided for control and subjection to sensitivity analysis of the sources of the bias.

Examples herein include methods for an estimation technique, theoretical treatment of bias, and a demonstration of the approach through an application to whole blood specimens collected in an example of head and neck squamous cell carcinoma (HNSCC). See Figure 3. Also provided are methods for a sensitivity analysis, demonstrating the impact of possible biases. Simulation study results are shown in examples herein based on the biology in the samples used.

Examples 1-3 herein show a method for determining changes in distribution of white blood cells between different subpopulations (e.g. cases and controls) from DNA methylation signatures, assuming an external validation set consisting of methylation signatures from purified white blood cell (WBC) samples exists. Examples 4, 10 and11 herein demonstrate the methodology using a data set of HNSCC cases and matched controls, inferring from DNA methylation assays alone known changes in CD4+ and granulocyte populations between cases and controls and change in CD4+ populations due to aging. Using previous methods flow cytometry would have been necessary to obtain the same results. A method for assessing the sensitivity of the magnitude estimates to possible biases is also provided. Example 12 validates the method through simulation.

Methods are provide herein for determining changes in the distribution of white blood cell types between different human populations (e.g. cases and controls) using DNA methylation signatures; by using an external validation set having methylation profiles from purified white blood cell components. DNA methylation in peripheral blood was accordingly shown to be a

biomarker for clinical and epidemiological investigation. Studies have attempted to distinguish cancer cases from controls using whole peripheral blood assayed with DNA methylation arrays, including ovarian (Teschendorff et al., 2009, *PLoS ONE* 4, e8274), bladder (Marsit et al., 2011, *J Clin Oncol* 29, 1133-1139), and pancreatic (Pedersen et al., 2011, *PLoS ONE* 6, e18223) cancers. Although these studies have demonstrated discrimination of cases from controls, sound evidence for a biological mechanism has been elusive. Presumably, disease associated alterations in blood methylation have several etiological components driven by endogenous genetic, environmental and disease specific factors. From known developmental associated differences in DNA methylation among specific blood cell types, changes in the distributions of blood cell types alone could account for disease associated DNA methylation. The many diverse types of immune cells in blood make this issue highly complex and problematic to tackle using single cell type assays. Therefore, it is important for the development of this new avenue of biomarker research to delineate effects due to the immune cell distribution itself from other "non cell type" alterations in DNA methylation. The differences among human populations attributed to cell distributions are termed "immunologically mediated".

Immunological explanations for differences in mRNA profiles between cases and controls have been proposed, e.g. Showe et al., 2009, Cancer Res 69: 9202-10 and Kossenkov et al., 2011, Clin Cancer Res 17: 5867-77. The statistical principles described in the method herein apply to mRNA expression profiles and an appropriate validation set S_0 based on mRNA expression arrays. Little to no modification of mathematical expressions and computer code is necessary to apply the statistical principles described in the method herein to analysis of mRNA expression profiles. Under the assumption that the upstream epigenetic control mechanisms are more biologically stable, less variability in measurement of DNA methylation is expected compared with measurement of mRNA expression.

In the methods herein, a solution to partition this component of variation in methylation from other determinants employs multivariate analytic tools including regression coefficients, associated inference, and coefficients of determination measures. These tools were used to evaluate whether the observed DNA methylation differences were due to an immunologically mediated response. Prior measurement error formulations (Thurston et al., 2003, *J Stat Plan Inf*, 113, 527-34; Li and Yin, 2007, *Ann Stat*, 35, 2143-2172) require specification of a logistic regression model for case/control status, conditional on DNA methylation signature, a computationally difficult task that is vulnerabe to model mis-specifications. A reverse formulation was used herein that naturally models the relationship of DNA methylation conditional on known phenotypes. The formulation respects the protocol (DNA methylation assay data collected after sampling from phenotype groups). Other strategies to formulate errors

were found to be unsuccessful. For example, the strategy utilizing Expectation-Maxinlization (EM) algorithm to integrate over the missing data ω (Little and Rubin, 2002, *Statistical Analysis with Missing Data.* Wiley, Hoboken, NJ, 2nd edition) is outside the measurement error literature and within the larger missing-data literature. However, by design, the distribution of ω varied substantially between the data sets S_0 and S_1, severely complicating the approach, with side-effect of introducing feedback from S_1 to S_0, contaminating the gold-standard status of S_0. Another alternative that was found to be unsuccessful was the simpler approach of an empirical Bayes procedure, similar to existing mixture-model approaches (Koestler et al., 2010, *Bioinformatics*, 26, 2578-2585). However, difficulty in specifying the distribution of ξ rendered this approach untenable, and in a separate simulation, attempts to impute ω among S_1 samples using parameters obtained from S_0 samples resulted in extremely biased estimates of ω.

Examples herein show that group level comparisons of blood cell DNA methylation revealed significant immune alterations. Methods for individual level immune cell profiling are applicable also, since methods herein are useful also to clinical and detailed analytical epidemiologic applications that examine individual risk factor information. When $\mathbf{z}_{1i}$ involves an orthogonal (e.g. one-way ANOVA) parameterization and ordinary least squares (OLS) is used to obtain $\mathbf{B}_1$, then equation 5 (Example 3) herein reduces to simple expressions involving the projected quantities $\omega_i = \mathbf{y}_{1i}\mathbf{B}_0(\mathbf{B}_0\ \mathbf{B}_0)^{-1}$. For exploratory purposes, projections ω_i serve as estimates of individual profiles. There is interest in minor immune cell fractions and their role in disease, though the signal strength of cell types comprising < 5% of the total white cell compartment is difficult to quantitate. Examples of such cell types include the regulatory T cell or NK cell fractions, which are implicated in autoimmune and malignant diseases. Optimization of platforms for technical sensitivity to minor subtypes combined with statistical optimization of signature recognition are needed to enhance the approach for testing highly targeted immune hypotheses.

In addition to group level comparisons of blood cell DNA methylation, immune cell profiling at the individual level is important for examining individual risk factors in clinical and detailed analytical epidemiologic applications. As shown in Examples herein, individual immune profiles are theoretically achievable and require extensive validation with a wide array of mixture combinations.

The methods herein have potentially far reaching implications for rapid, simple and complete assessment of the composition of human white blood cell populations, i.e. the immune profile. Currently, assessment of the cellular composition of peripheral blood cannot be accomplished without the use of freshly drawn venous blood that is immediately prepared in a

specially equipped laboratory. A complete assessment of the entire immune profile requires extensive flow cytometric measurements based on protein epitopes on leukocyte membranes that distinguishes subtypes of immune cells that are either too rare or too similar in appearance to be distinguished using simple microscopic approaches. In particular, flow cytometry is limited by the following: cells must be separated, requiring large volumes of fresh cells; detection can be accomplished only by the fluorescent antibody tags available, which require expensive technology to read; the outer cell membrane must be intact, mandating limited utility in many instances.

In contrast, using the methods herein, the application of labor-intensive or expensive steps is required only in the construction of the validation set S_0, which need only be developed once. Once S_0 is available, subsequent interrogation is based on the chemically stable CpG methylation of DNA. Thus the methods herein obviate the need for fresh blood and the preservation of labile protein epitopes. The methods herein are able to also simultaneously assess all of the individual components of the peripheral blood using a highly multiplexed molecular platform and therefore logistically straightforward. Furthermore, the statistical methodology used here is implemented easily with the instrumental output of the methylation arrays, which simplifies the interpretation of the immune profile data from the operator's point of view. The methods herein are immediately deployed in a research framework to cost effectively assess human immune profiles (in fresh or archival samples), to explore the potential of the immune profiles to function as biomarkers, and to address key questions regarding disease pathogenesis. Furthermore, the approach used in the methods herein is readily suited for rapid translation to a broad base of clinical applications such as disease monitoring, diagnosis, prognosis, and response to therapy.

The methods herein are applied to tumor biopsies for immune characterization of cancer patients. Other notable applications exist including the application of the test to urine sediments in patients with autoimmune and diabetic kidney disease or in patients undergoing kidney transplantation. Positive detection of T cells in urine sediment is indicative of immune activation and potential kidney disease progression or acute rejection in the context of kidney transplantation.

Populations of blood lymphocytes can be distinguished morphologically on the basis of size and the presence of a granular cytoplasm.

Small lymphocytes, including all subsets of T- and B cells, are responsible for adaptive immune responses. Sublineages of small lymphocytes are morphologically indistinguishable and are distinguished by cell surface receptors and cellular function. B cells are typically distinguished by expression of the surface molecule CD19. They express immunoglobulins,

which are surface receptors for pathogens. In addition, B cells are capable of further differentiating into effector cells called plasma cells. (Parham, P. *The Immune System*, Garland Science, New York, NY, 2005). Differentiated T cells exhibit a complex of surface molecules which function as antigen receptors, referred to as the T cell receptor (TCR) complex. This complex includes the TCR α plus β, or γ plus δ antigen recognition chains, which are associated with invariant chain subunits CD3γ, δ, ε, and ζ. (Zhang, Z., *et al.* 2007, *Blood* **109**, 4328-4335). In general, T cells are distinguished from other cell lineages by expression of CD3 molecules on the cell surface. The genes that encode CD3 γ, δ, ε, and ζ subunits are *CD3G, CD3D, CD3E* and *CD3Z* respectively. The former three genes are tightly clustered on chromosome 11, whereas *CD3Z* is located on chromosome 1. Differentiated T cells are further divided into two lineages depending on their expression of either CD4 or CD8. The main function of CD8+ T cells, also known as cytotoxic T cells, is to kill infected and transformed cells. The main function of CD4+ T cells is to help other immune cells respond appropriately to sources of infection or malignancy. There are several subsets of CD4+ T cells, including Th1, Th2, Th17 and regulatory T cells. (Parham, P. *The Immune System*, Garland Science, New York, NY, 2005). Regulatory T cells suppress an immune response by influencing the activity of other cell types. They act primarily in the periphery on mature lymphocytes that have exited the main lymphoid tissues and serve as a means of preventing autoimmunity during protective immune responses. Exemplary regulatory T cells are thymus-derived CD4+CD25+Foxp3+ T cells, commonly referred to as Tregs. (Zou, W. 2006, *Nat Rev Immunol* **6**, 295-307). These cells primarily function to maintain peripheral self-tolerance. (Cesana, G.C., *et al.*, 2006, *J Clin Oncol* 24, 1169-1177). Forkhead Box P3 (FOXP3), a transcription factor expressed by Tregs, is an important developmental and functional factor that regulates Treg immunosuppressive functions. (Janson, P.C., Winerdal, M.E. & Winqvist, O. 2009, *Biochim Biophys Acta* 1790, 906-919; Zou, W. 2006, *Nat Rev Immunol* 6, 295-307).

Natural killer (NK) cells are large CD56+ lymphocytes with a granular cytoplasm. They enter infected or malignant tissue to kill damaged cells and secrete cytokines aimed at preventing the spread of disease to other cells or tissues. Thus, NK cells act as effector cells of innate immunity. A subset of CD56+ NK cells that express CD3 surface molecules are NKT cells.

To determine if distinct methylation profiles are indeed associated with leukocyte lineages, statistical clustering of methylation patterns was performed using a modified model-based form of unsupervised clustering known as recursively partitioned mixture modeling (RPMM). (Houseman, E.A., *et al.* 2008, *BMC Bioinformatics*, 2008, **9**, 365).

A locus by locus comparison was performed in which putative leukocyte DMRs were identified from Infinium data in SAS version 9.1 using a macro for locus-by-locus linear modeling that adjusts for control probe and beadchip plate. Infinium beta values for Group 1 leukocyte samples were compared to Infinium beta values for Group 2 leukocyte samples, in which group membership for each phase of the comparison is shown in Table 1.

Table 1. Locus by locus comparison groups

	Group 1 Leukocytes	Group 2 Leukocytes
Phase I	CD3+, Pan-T, CD4, Treg, CD8	NK, B, Mono, Gran, Neut
Phase II	NK	Pan-T, CD4, Treg, CD8, B, Mono, Gran, Neut
Phase III	CD8	CD4, Treg, NK, B, Mono, Gran, Neut

Resultant t-values from each comparison were converted to p-values in R version 2.11.1 of Illumina's software which provides convenient mechanisms for loading and analyzing the results of methylation status, and for quality control and basic visualization tasks.

False discovery rate estimation and Q-values were computed by the Q-value package in R to adjust for multiple comparisons. (Significance was characterized as $Q \leq 0.05$.)

For significant CpG loci ($Q \leq 0.05$), a negative t-value indicates the locus putatively represents a DMR that is unmethylated in group 1 leukocyte lineage(s) and methylated in group 2 leukocyte lineage(s). Conversely, a positive t-value indicates that the locus putatively represents a DMR that is methylated in group 1 leukocyte lineages and unmethylated in group 2 leukocyte lineages. A DMR that is unmethylated in the leukocyte lineage(s) of interest and methylated in other leukocyte lineages would make the best epigenetic biomarker, since unmethylation is associated with transcriptional activity whereas methylation is associated with transcriptional silencing. Therefore, significant CpG loci exhibiting negative t-values are preferred.

In the methods herein, results of locus by locus comparisons were merged with cell type specific gene expression data. (Palmer *et al.*, 2006, *BMC Genomics* 7, 115; Du *et al.*, 2006, *Genomics* 87, 693-703; and Hashimoto *et al.*, 2003, *Blood* 101, 3509-3513) to identify putative DMRs that are in genes associated with altered expression by Group 1 leukocyte lineages compared to Group 2 leukocyte lineages. An exemplary candidate epigenetic biomarker of a specific leukocyte lineage is an unmethylated region of a gene that is highly expressed by the leukocyte lineage, and not expressed by other cell types such as lineage-specific surface molecules,obligate differentiation proteins, and secreted factors. A further candidate is a

methylated region of a gene that is not expressed by the leukocyte lineage and is expressed by all other cell types. Without being limited by any theory or mechanism of action scenarios correlate with chromatin packaging, so that differential DNA methylation plays a large role in regulating leukocyte lineage specific expression of the gene. If no leukocyte lineage specific difference in expression of the gene containing a putative DMR were observed, other modes of gene regulation such as activators, repressors, and enhancers overshadow the role of chromatin packaging in regulating expression of the gene. Alternatively, such a gene is expressed in a temporally or environmentally specific manner that was not elucidated by the gene expression candidate data. Such a putative DMR would not be an ideal target to explore as an epigenetic biomarker of that leukocyte lineage.

In the methods described herein DMR validation is performed for each putative DMR identified from array data using bisulfite pyrosequencing and/or MethyLight quantitative real time PCR assays that measure DNA methylation of the gene region in all sorted human leukocyte samples shown in Table 15, Example 13. Bisulfite pyrosequencing assays were designed using Pyromark Assay Design 2.0 (Qiagen), and carried out on a Pyromark MD pyrosequencer running Pyromark qCpG software (Qiagen). Oligonucleotide primers were obtained from Invitrogen™ by Life Technologies™. The gene region of interest were PCR amplified from bisulfite converted DNA using a biotinylated reverse primer and an unlabelled forward primer. The biotinylated PCR product was complexed with sequencing primers that anneal upstream from the target region, and was then incubated with enzymes and substrates. Then, dNTPs were dispensed in a specific order and light emitted with the incorporation of each nucleotide is measured with a CCD camera. Methylation was quantified by calculating the ratio of cytosine (methylated) to thymine (unmethylated) at each CpG locus.

In the methods described herein methylation status of specific gene regions was calculated using *MethyLight* according to the protocol described by Campan et al. 2009, *Methods Mol Biol* 507, 325-337, with the following modifications: C-less primers and probe were used to determine total DNA input for each sample and control reference rather than ALU-C4 primers and probe. To measure unmethylation, control unmethylated DNA was used as a reference, generating a percent unmethyated reference value which is subsequently converted into percent methylation. Real time PCR primers and flourescent (major groove binding)MGB probes were obtained from Applied Biosystems (Foster City, CA). TaqMan® Universal PCR Mastermix, no AmpErase® UNG was obtained from Applied Biosystems, manufactured by Roche (Branchburg, NJ). Quantitative, real time PCR reactions were performed with Applied Biosystems 7300 Real Time PCR System using Applied Biosystems 7300 system sequence detection software version 1.4.0.25 ©2001-2006.

In the methods herein, a putative DMR identified as being unmethylated in group 1 leukocytes based on Infinium methylation data was shown using bisulfite pyrosequencing or MethyLight® qPCR to be unmethylated in group 1 leukocytes and methylated in group 2 leukocytes and the DMR was confirmed as an unmethylated epigenetic biomarker specific to the group 1 leukocyte lineage(s). A putative DMR shown using bisulfite pyrosequencing or MethyLight® qPCR to be unmethylated in group 1 leukocytes and in some group 2 leukocytes, was not confirmed as an epigenetic biomarker specific to the group 1 leukocyte lineage(s). Instead that DMR represents an epigenetic biomarker of several different human leukocyte lineages including the group 1 lineage(s). A DMR that is partially unmethylated by bisulfite pyrosequencing or MethyLight® qPCR in group 1 leukocytes and methylated in group 2 leukocytes, is a weak epigenetic biomarker of the group 1 leukocyte lineage(s). That DMR is heterogeneously unmethylated in group 1 leukocytes and is homogeneously methylated in group 2 leukocytes and is therefore not useful for distinguishing group 1 from group 2 leukocyte lineages.

If Infinium data suggested that a CpG locus represents a DMR specific to group 1 leukocytes, and bisulfite pyrosequencing or MethyLight qPCR did not find a difference in DNA methylation in that region between group 1 and group 2 leukocyte samples, the region was not considered a DMR that would serve as an epigenetic biomarker of the group 1 leukocyte lineage(s).

These discovery platform criteria successfully identified a unique heretofore unknown sequence of genomic DNA that is specifically marked by CpG demethylation in CD3 positive T cells, not in other hematopoietic peripheral blood cells (Figure 10 panel B). In examples herein it is further shown the DNA methylation status of this region in the promoter of *CD3Z* gene in sorted human peripheral blood leukocytes measured by MethyLight® qPCR confirms that the identified genomic sequence is an immune cell type specific differentially methylated region that is a useful marker to quantify CD3+ T cells in biological specimens such as whole or separated blood and other tissues.

Gliomas are a histologically diverse cancer with few established risk factors and poor prognoses (Kleihues et al. 1993, Brain Pathol 3(3): 255-68; Ohgaki and Kleihues 2005, Acta Neuropathol 109(1): 93-108; Louis et al. 2007, Acta Neuropathol 114(2): 97-109; Ohgaki, and Kleihues 2007, Am J Pathol 170(5): 1445-53). However, immune factors are associated with increased glioma risk and are also thought to play a role in patient outcomes (Wiemels et al. 2009, Int J Cancer. 2009 Aug 1; 125(3):680-7; Yang et al. 2010, J Clin Neurosci 17(11): 1381-5). Patients with glioblastoma multiforme (GBM) exhibit abnormalities (McVicar et al., 1992, J Neurosurg 76(2): 251-60; Ashkenazi et al. 1997, Neuroimmunomodulation 4(1): 49-56) of T

cell response associated with pronounced reductions in T cell numbers in peripheral blood including the suppressive regulatory T cells (Tregs) (Fecci, et al., 2006, Cancer Res 66(6): 3294-302). Despite low T cell and Treg counts, the ratio of Tregs to T cells is clinically relevant in immunosuppression. Currently there is no validated method to quantify this ratio. The quantification of immunosuppression is envisioned herein to help also in characterizing patient tumors. An immunosuppressive environment in glioma is also suggested by the accumulation of tumor infiltrating lymphocytes (TILs) displaying markers of Tregs, (i.e. cell membrane CD4 and CD25 and intracellular staining of the FOXP3 protein).

Epigenetic markers involving the demethylation of the *FOXP3* gene have been determined to be the most specific marker of stable Tregs. (Baron et al., 2007, Eur J Immunol 37(9): 2378-89; Floess et al., 2007 PLoS Biol 5(2): e38; Polansky et al., 2008, Eur J Immunol 38(6): 1654-63). As described in examples herein, by combining information about the *FOXP3* differentially methylated region (DMR) with methylation specific quantitative PCR (MS-qPCR) highly sensitive and accurate counts of Tregs in blood and tissues were obtained. Such DNA-based methods to interrogate specific populations of T cell subsets are far less expensive than flow-cytometry and can be applied to archival specimens. Examples herein show that the DMR marker for CD3+ T cells identified herein is used alone or in conjunction with the previously described Treg DMR marker.

A quantitative assay for CD3+ T cells based on the demethylation of the promoter of a component of the T cell receptor complex: *CD3Z (CD247)* is also described herein. Examples herein show the validity of *CD3Z* demethylation as a CD3+ T cell marker and illustrate its application in patients with glioma that demonstrate the high discriminating value of *CD3Z* demethylation in glioma case-control subject comparisons, histopathological characterization of tumors and patient prognosis.

An understanding of the role played by an altered immune response in etiology facilitates development of more effective therapies and prognostic indicators. Epidemiological studies implicate atopic immune alterations in glioma risk (Wrensch et al., 2005, Am J Epidemiol 161(10): 929-38; Schwartzbaum et al., 2010, Carcinogenesis 31(10): 1770-7). Immune suppression and abnormalities in T cells in glioma patients may prevent antitumor immunity and poses barriers to effective immunotherapeutic strategies (Grauer et al., 2007, Int J Cancer 121(1): 95-105; Sonabend et al., 2008, Anticancer Res 28(2B): 1143-50). Data obtained using novel T cell epigenetic assays described in examples herein demonstrate dramatic decreases in CD3+ T cells and Tregs in peripheral blood from GBM patients. The copy numbers of demethylated *CD3Z* and *FOXP3*, as a percent of total leukocyte copies, were observed to be reduced about two-fold in GBM patients, which was highly statistically significant.

Validation studies herein support the notion that the *CD3Z* MS-qPCR assay using unprocessed archival whole blood is an accurate reflection of T cells as measured by conventional flow cytometry. Previous studies have validated the *FOXP3* demethylation assay as a measure of Tregs in blood and tissues (Baron et al., 2007, Eur J Immunol 37(9): 2378-89). Current steroid use (dexamethasone), temozolomide and radiation exposures as possible factors in these effects among cases were investigated but no significant associations of any factor with these T cell alterations was found. The methods described in examples herein that delineate T cell subsets from DNA facilitate immune cell analyses using blood specimens that have been archived in cohort populations with long-term glioma follow-up data. Nested case control studies within large epidemiologic cohorts are now feasible as a result, allowing for the first time, to test whether T cell and Treg abnormalities precede the diagnosis of glioma.

The balance of suppressive Tregs to total T cells in peripheral blood has been reported to be shifted towards greater suppression in GBM patients and other types of cancer (Beyer and Schultze, 2006, Blood 108(3): 804-11). Ratio of Tregs/T cells in association with cigarette smoking was examined herein. An association of current smoking with higher Treg/T cell ratios was observed. There is strong evidence that cigarette smoke exposure leads to the accumulation of Tregs in respiratory airways in mice (Brandsma et al., 2008, Respir Res 9: 17) and humans (Smyth et al., 2007, Chest 132(1): 156-63) as well as in the gut epithelium of exposed mice (Verschuere et al., 2011 Lab Invest. 91(7):1056-67). Treg/T cell ratios were herein observed to be higher in current smokers versus former smokers (Figure 16). It was subsequently confirmed in an independent population that current but not former cigarette smoking exhibit higher Treg/T cell ratios. Results herein illustrate the need for examination of patient characteristics to include cigarette smoking in diseases that affect Treg levels. New epigenetic methods described herein are useful in promoting these types of studies.

Similar to many types of cancer CD4+ T helper cells and Tregs have been shown to infiltrate the human glioma tumor microenvironment (Nishikawa and Sakaguchi, 2010, Int J Cancer 127(4): 759-67). In glioma studies using IHC to quantify T cells in FFPE preparations CD4+ T cell numbers were reported to increase with tumor grade, whereas CD8+ T cells appear in equal frequencies across glioma grades (Heimberger et al., 2008, Clin Cancer Res 14(16): 5166-72). Results herein indicate increased CD3Z demethylated cells according to grade (Figure 17). Immunohistochemical IHC analysis herein showed that mostly these cells were CD8+ cells with very few CD4+ cells. Examples herein also show that ependymal tumor cells and some significant fraction of grade II Oligodendrogliomas (OD) and Astrocytomas (AS) tumors contain significant numbers of T cells and Tregs (Figure 21). As progression of lower grade to higher grade brain tumors is a common and serious clinical problem results herein show that epigenetic

analyses are useful for characterizing low grade OD and AS tumors as well as Ependymomas (EP). Compared to previous reports (El Andaloussi and Lesniak, 2006, Neuro Oncol 8(3): 234-43; El Andaloussi and Lesniak, 2007, J Neurooncol 83(2): 145-52; Heimberger et al., 2008, Clin Cancer Res 14(16): 5166-72; Heimberger et al., 2008, Neuro Oncol 10(1): 98-103) analysis herein using the MS-qPCR showed significantly increased ratio of Treg/CD3+ Tcells within glioma tumor tissues of different pathological grade (Figure 17). Results herein showed also how the ratio of Tregs/CD3+ Tcells increases with tumor grade in comparison to blood. Thus, until the present results, there was no evidence of a specific accumulation of Tregs in human brain tumors. The survival data in examples herein show significant associations of immune parameters with patient survival (Figure 22).

Without being limited by any theory or mechanism of action, observations herein of a close linear relationship between flow cytometry of CD3+ T cells and *CD3Z* demethylation that was identical among glioma cases and controls argues against a cancer related effect on CD3Z demethylation such as downregulation of CD3Z through a posttranslational effect on CD3Z proteins mediated by up regulation of lysosomal or proteasomal degradation pathways. Another issue concerning the validity of CD3Z demethylation as a CD3+ T cell marker in cancer tissues is that DNA demethylation may take place in transformed cells and thus 'mimic' a lymphocyte signal. To ascertain that the observed CD3Z demethylation was taking place in CD3+ T cells and not due to DNA demethylation taking place in transformed cells *CD3Z* and *FOXP3* demethylation in brain tumor cells lines and in human GBM xenografts which cannot contain human T cells was assessed. These samples contained non-detectable levels of *CD3Z* or *FOXP3* demethylation. Normal brain tissue was also uniformly devoid of T cell signals, consistent with the specificity of the MS-qPCR in tumor as reflecting infiltration of immune cells. Some subtypes of NK cells ($CD56^{dim}CD16^{bright}$) utilize CD3Z in NK receptor signaling (Lanier, 2006, Trends Cell Biol 16(8): 388-90). The contribution of CD3Z expressing and demethylated NK cells to the overall CD3Z demethylated signal in peripheral white blood cells is estimated to be very small. Furthermore, NK cells have not been observed in glioma tissues.

The fundamental innovation in the epigenetic analyses described herein is a shift in immunodiagnostics away from proteomic-based approaches to one that is based on quantifying cell type specific DNA methylation events. This new approach produces gains in versatility, sensitivity, feasibility and throughput compared with conventional flow cytometry or IHC and does so at a lower cost. The high chemical stability of cytosine methylation marks within genomic DNA and the fact that differentiation within the immune system is tightly linked with gene specific DNA methylation events makes quantification of immune cells through epigenetic analyses a unique approach. The method combines the intrinsic chemical stability of DNA with

the high sensitivity of qPCR methods. Automation and liquid robotic handling in processing and analysis add further to the power of the methodology and open avenues for investigations in the immunoepidemiology of glioma and many other diseases.

Methods herein show that blood-based DNA methylation signatures across a complex cellular mixture of WBCs are useful for distinguishing solid tumor cancer cases in which there are well-defined immune-mediated responses and controls. As tumorigenesis elicits a distinct immune response (Camilleri-Brot S et al., 2004, Ann Oncol 15:104–112; Wang Y et al., 2005, Am J Clin Pathol 124:392–401; Rui L et al., 2011 Nat Immunol 12:933–940), the result is a hematopoietic shift in WBC populations, which can be precisely discerned by applying the unique epigenetic signature of differing lineages. The aggregate methylation signature in blood that distinguishes cancer cases from controls corresponds to the epigenetic signatures that define leukocyte subtypes.

To understand the role of immune-mediated responses to tumorigenesis in defining distinct signatures of blood-based DNA methylation between cancer cases and cancer-free controls in examples herein, the epigenetic landscape of WBCs was obtained by identifying DMRs among leukocyte subtypes. This analysis revealed that nearly all of the highest ranking 50 leukocyte DMRs (Example 25) were differentially methylated between disease cases and normal controls for HNSCC and ovarian cancers, with a smaller fraction differentially methylation between bladder cancer cases and controls. Among the eight overlapping CpG loci that were found to be significantly differentially methylated between cancer cases and controls across the three data sets, the direction of the relationships was similar for HNSCC and ovarian cancer cases compared to controls. These findings show that HNSCC and ovarian cancer elicit similar shifts in leukocyte compositions in the hematopoietic system.

Of the seven overlapping DMRs (CD72, PACAP, FGD2, SLC22A18, GSTP1, NFE2, ASGR2) several are located within genes with either established or alleged involvement in immune differentiation or function, *viz.*, CD72, PACAP and FGD2 (Kumanogoh and Kikutani, 2001, Trends Immunol 22:670–676; Parnes and Pan, 2000, Immunol Rev 176:75–85; Tan et al., 2009, Proc Natl Acad Sci 106:2012– 2017; Huber C et al., 2008, J Biol Chem 283:34002–34012). CD72, a member of the C-type lectin superfamily, negatively regulates B cell coreceptor signaling (Kumanogoh and Kikutani, 2001) and has been shown to act as a unique inhibitory receptor on NK cells regulating cytokine production (Alcon VL et al., 2009, Eur J Immunol 39:826–832). Moreover, PACAP has been implicated as an intrinsic regulator of regulatory T cell abundance after inflammation36 and FGD2 has been shown to play a role in leukocyte signaling and vesicle trafficking in cells specialized to present antigen in the immune system (Huber C et al., 2008, J Biol Chem 283:34002–34012).

In the model described herein containing the DNA methylation profile for the highest ranking 50 leukocyte DMRs, patient age, gender, smoking status, smoking pack years, weekly alcohol consumption, and HPV serological status (Table 19, Example 13), HNSCC cancer was predicted with high degree of sensitivity and specificity. Similarly high prediction performance was obtained for ovarian cancer using the DNA methylation profile for the highest ranking ten leukocyte DMRs and patient age group. Prediction performance for bladder cancer, based on the methylation profile of the highest ranking 56 DMRs, patient age, gender, smoking status, smoking pack years, and family history of bladder cancer, was lower than that observed for HNSCC and ovarian cancer. One explanation for the differences in magnitude for discriminating cancer cases and controls among cancer types is underlying differences in the magnitude of shift in leukocyte subtypes. Cancers characterized by a pronounced immunologic response such as HNSCC and ovarian cancer (Alhamarneh O et al., 2008, Head Neck 30:251–261; Zhang L et al., 2003, N Engl J Med 348:203-213; Tomsova M et al., 2008, Gynecol Oncol 108:415-420; Sato E et al., 2005, Proc Natl Acad Sci 102:18538-18543; Curiel TJ et al., 2004, Nat Med 10:942-949), correspond to more discernable shifts in leukocyte sub-population, thus resulting in greater discrimination of blood-derived DNA methylation using leukocyte DMRs for these cancers compared to bladder cancer.

Substantial correlation was also obtained in methylation of the loci identified via the semi-supervised recursively partitioned mixture model (SS-RPMM) analyses and the leukocyte DMRs that defined the methylation classes discovered for the HNSCC and ovarian data sets. A diagram illustrating the analytic framework for SS-RPMM is provided in Figure 32. The SS-RPMM25 procedure is specifically designed to construct methylation classes that are based on an optimal number of informative features (loci whose methylation is most strongly associated with cancer case/control status). The results demonstrate that the methylation classes identified through SS-RPMM for the HNSCC and ovarian data sets are in large part due to systematic hematopoietic changes in WBC populations in response to tumorigenesis. The 56 leukocyte DMRs used in the bladder profile analysis were less correlated with the nine CpG loci identified via the previously reported SSRPMM analysis of this data set (Marsit CJ et al., 2011, J Clin Oncol 29:1133-1139). Alternative biological epigenetic mechanisms may be operative in bladder cancer in addition to the epigenetic signatures characteristic of leukocyte subtypes, and contribute independently to the blood-derived differences in DNA methylation between bladder cancer cases and controls.

Examples herein provide evidence that observed differences in blood-derived DNA methylation in cancer cases are largely explained by systematic differences in the methylation signatures of leukocyte sub-populations. These findings signify that different cancers elicit a

discernible, unique immune response evident in peripheral blood. These results have important implications for research into the immunology of cancer. Further, the approach of observing differences in blood derived DNA methylation provides a completely novel tool for the study of the immune profiles of diseases where only DNA can be accessed; that is, this approach has utility not only in cancer diagnostics and risk-prediction, but can also be applied to future research (including stored specimens) for any disease where the immune profile holds medical information. The approach represents an extremely simple, yet truly powerful and important new tool for medical research and may serve as a catalyst for future non-invasive disease diagnostics.

Natural killer (NK) cells are a key element of the innate immune system implicated in human cancer. To examine NK cell levels in archived blood samples from a study of human head and neck squamous cell carcinoma (HNSCC), a DNA-based quantification method described in methods herein was developed (Examples 27-36).

Head and neck squamous cell carcinoma (HNSCC) is strongly associated with alterations in the immune system and it is postulated that progression of HNSCC tumors is linked to immune evasion or failure of the immune system to fight the cancer (Duray A, et al., 2010, Clinical & developmental immunology, 2010:701657; Pries R, and Wollenberg B, 2006, Cytokine Growth Factor Rev, 17:141-6; Wulff S et al., 2009, Anticancer research, 29:3053-7; Kuss I et al., 2004. Clin Cancer Res, 10:3755-62; Kuss I et al., 2005, Adv Otorhinolaryngol, 62:161-72). Natural killer (NK) cells are of particular interest in the context of HNSCC and other cancers, since they are able to recognize and destroy pre-cancerous and malignant cells (Kim R et al., 2007, Immunology, 121:1-14; Ostrand-Rosenberg S. 2008, Curr Opin Genet Dev, 18:11-8; Whiteside TL, 2006, Cancer Treat Res, 130:103-24; Parham P. The Immune System. 2nd ed. New York, NY: Garland Science; 2005). Natural killer cell infiltration into solid tumor tissue has been associated with improved survival in studies of many different types of cancer (Ishigami S et al., 2000 Cancer, 88:577-83; Kondo E et al., 2003, Dig Surg, 20:445-51; Villegas FR et al., 2002, Lung Cancer 2002;35:23-8). Immune suppression is frequently seen in patients with head and neck cancer (Duray A, et al., 2010, Clinical & developmental immunology, 2010:701657; Pries R, and Wollenberg B, 2006, Cytokine Growth Factor Rev, 17:141-6; Wulff S et al., 2009, Anticancer research, 29:3053-7; Kuss I et al., 2004. Clin Cancer Res, 10:3755-62; Kuss I et al., 2005, Adv Otorhinolaryngol, 62:161-72). Diminished NK cell and natural killer T (NKT) cell activity and number have been observed in the peripheral blood of patients with HNSCC (Wulff S et al., 2009, Anticancer research, 29:3053-7; Molling JW et al., 2007, J Clin Oncol, 25:862-8).

A novel DMR is identified herein that distinguishes NK cells from other leukocytes to facilitate the quantification of NK cells in archived blood samples from a case control study of HNSCC. Many chemical exposures, such as tobacco and alcohol, as well as viral factors, such as human papilloma virus (HPV), are known or suspected to be causal factors in HNSCC (Furniss CS et al., 2009 Annals of oncology : official journal of the European Society for Medical Oncology / ESMO, 20:534-41; Applebaum KM et al., 2007, Journal of the National Cancer Institute, 99:1801-10) and may independently affect immune profiles (Mehta H et al., 2008, Inflammation research, 57:497-503; Wansom D et al., 2010, Archives of otolaryngology--head & neck surgery 2010;136:1267-73; Gao B et al., 2011 American journal of physiology Gastrointestinal and liver physiology 300:G516-25). Unlike previous studies, data shown herein evaluates the effects of these factors on the depression in NK immune profile. Patient risk factors and disease characteristics (e.g. tumor location) are evaluated herein in relationship to NK cells to determine the independent associations of HNSCC with innate immune parameters.

NK cell-specific DNA methylation was identified by analyzing DNA methylation and mRNA array data from purified blood leukocyte subtypes (NK, T, B, monocytes, granulocytes), and confirmed via pyrosequencing and methylation specific quantitative PCR (MS-qPCR). NK cell levels in archived whole blood DNA from 122 HNSCC patients and 122 controls from a study population were assessed by MS-qPCR. Details of this study population have been previously described (Applebaum KM et al., 2007, Journal of the National Cancer Institute, 99:1801-10). Briefly, peripheral blood from 122 control donors and 122 HNSCC patients was collected between December 1999 and December 2003 in the greater Boston area. Population based control subjects with no prior history of cancer were from the same region as cases, and were frequency matched on age and gender. Study approval was obtained from the Brown University Institutional Review Board. All subjects provided written informed consent for participation in this study. Venous anticoagulated whole blood was drawn into sodium citrate and stored at -20 °C prior to DNA isolation.

Pyrosequencing and MS-qPCR (Figure 39) confirmed that a demethylated DNA region in NKp46 distinguishes NK cells from other leukocytes, and serves as a quantitative NK cell marker. Demethylation of NKp46 was significantly lower in HNSCC patient blood samples compared with controls ($p < 0.001$). Individuals in the lowest NK tertile had over 5-fold risk of being a HNSCC case, controlling for age, gender, HPV16 status, cigarette smoking, alcohol consumption, and BMI (OR = 5.6, 95% CI: 2.0, 17.4) (Figure 37). Cases did not show differences in NKp46 demethylation based on disease treatment or tumor site.

The results of this study indicate a significant depression in NK cells in HNSCC patients that is unrelated to exposures associated with the disease. DNA methylation biomarkers

of NK cells represent an alternative to conventional flow cytometry that can be applied in a wide variety of clinical and epidemiologic settings including archival blood specimens.

Understanding of immune cell level alterations associated with cancer and other diseases has, until now, been restricted by the limitations of immunodiagnostic methods. Described herein is a new method for measuring NK cell levels in human blood and tissue based on cell-lineage specific DNA methylation that can be applied to samples regardless of handling and storage procedures. This is a step forward in immune cell detection and quantification that is applicable to many types of clinical samples. Applying the method to a case-control study of HNSCC (Examples 27-36) revealed a case-associated decrease in circulating NK cells that is independent of known risk factors and treatments. This shows that it is important to monitor NK cell levels in patients with HNSCC, and that it may be worthwhile to pursue future immune therapies may be designed aimed at restoring circulating NK cells in patients with HNSCC.

A variety of methods are available as bases for methodology used to analyze CpG methylation states. These methods can be divided roughly into two types: gene-specific and global methylation analysis. A large number of techniques have been developed for gene-specific CpG methylation analysis. Early studies used methylation sensitive restriction enzymes to digest DNA followed by Southern detection or PCR amplification. Bisulfite reaction based methods such as methylation specific PCR (MSP) and bisulfite genomic sequencing PCR are commonly used currently. Global methylation analysis measures the overall level of methyl cytosines in genome by methods such as chromatography or methyl accepting capacity assay. Further, methylation hot-spots or methylated CpG islands in the genome may also be identified by several of the recently developed genome-wide screen methods such as Restriction Landmark Genomic Scanning for Methylation (RLGS-M), and CpG island microarray.

The gene-specific method MethyLight is a highly sensitive high-throughput quantitative methylation assay, capable of detecting methylated alleles in the presence of a 10000-fold excess of unmethylated alleles using fluorescence-based real-time PCR technology that requires few or minor further manipulations after the PCR step. Eads CA et al., Nucl. Acids Res. (2000) 28 (8): e32-00. For example, a MethylLight assay is commercially available from QIAGEN, Inc. Valencia, CA.

In another embodiment of the method, analyzing the methylation of any gene, e.g., the CD3Z gene through amplification by Polymerase Chain Reaction (PCR) is performed using digital PCR. Digital PCR is an improved method of PCR useful to overcome difficulties associated with conventional PCR. Conventional PCR assumes that amplification of nucleic acid is exponential and nucleic acids are quantified by comparing the number of amplification cycles and amount of PCR end-product to those of a reference sample. In practice however, several

factors interfere with this calculation, making measurements uncertainties and inaccurate and hence unsuitable for highly sensitive measurements.

In digital PCR, a sample is partitioned so that individual nucleic acid molecules within the sample are localized and concentrated within many separate regions. Molecules can be counted by estimating by using a Poisson distribution. Each partition contains "0" or "1" molecules, or a negative or positive reaction, respectively. After PCR amplification, nucleic acids are quantified by counting the regions that contain PCR end-product, which is a count of positive reactions. A system for digital PCR based on integrated fluidic circuits (chips) having integrated chambers and valves for partitioning samples is commercially available. For example a digital PCR system is available from Life Technologies (Grand Island, NY 14072USA) and QuantaLife QuantaLife Pleasanton, CA USA).

A skilled person will recognize that many suitable variations of the methods may be substituted for or used in addition to those described above and in the claims. It should be understood that the implementation of other variations and modifications of the embodiment of the invention and its various aspects will be apparent to one skilled in the art, and that the invention is not limited by the specific embodiments described herein and in the claims. The present application mentions various patents, scientific articles, and other publications, each of which is hereby incorporated herein in its entirety by reference.

The invention having now been fully described, it is exemplified by the following examples and claims which are for illustrative purposes only and are not meant to be further limiting.

Examples

Example1: Statistical methods for using DNA methylation arrays as surrogate measures of cell mixture distribution

In the framework for measurement of methylation status of CpG sites in cell mixtures $\mathbf{Y}_{0h}$ represents an $m \times 1$ vector of methylation assay values, e.g. average beta values from an Infinium bead-array product corresponding to a purified blood sample consisting of a homogenous cellular population (e.g. monocytes or granulocytes), with the qualitative characterization of the cell type indicated by a $d_0 \times 1$ covariate vector $\mathbf{w}_h$. Here, $h \in \{1,...,n_0\}$, and the m individual values correspond to CpG sites on a DNA methylation microarray, possibly pre-selected to correspond to putative DMRs for distinguishing different cellular types. Correspondingly, $\mathbf{Y}_{1i}$ represents an $m \times 1$ vector of methylation assay values for the same CpG sites (in the same order) as $\mathbf{Y}_{0h}$, but corresponding to a heterogeneous mixture of cells (e.g.

peripheral whole blood) from a human subject. Here, $i \in \{1,...,n_1\}$, n_1 is the number of target specimens, and $\mathbf{z}_{1i}$ is a $d_1 \times 1$ covariate vector representing an intercept as well as phenotypes or exposures corresponding to the subject, e.g. $d_1 = 2$ for a simple case/control study without confounders. Here the goal is to understand the associations between $\mathbf{Y}_{1i}$ and $\mathbf{z}_{1i}$ in terms of associations between $\mathbf{Y}_{0h}$ and $\mathbf{w}_{0h}$, i.e. to infer changes in mixtures of cell types associated with phenotypes or exposures, using DNA methylation as a surrogate measure of cell mixture. Thus, there are two data sets, $S_0 = \{(\mathbf{Y}_{01}, \mathbf{w}_1),...,(\mathbf{Y}_{0n_0}, \mathbf{w}_{n_0})\}$, the set of data from "purified" cell samples effectively representing external validation or gold-standard data and $S_1 = \{(\mathbf{Y}_{11}, \mathbf{z}_1),...,(\mathbf{Y}_{1n_1}, \mathbf{z}_{n_1})\}$, representing surrogate data collected from a target population. To this end following linear models are provided:

$$\begin{aligned} \mathbf{Y}_{0h} &= \mathbf{B}_0\mathbf{w}_{0h} + \mathbf{e}_{0h} \\ \mathbf{Y}_{1i} &= \mu_1 + \mathbf{B}_1\mathbf{z}_{1i} + \mathbf{e}_{1i}, \end{aligned} \tag{1}$$

where $\mathbf{B}_0$ and $\mathbf{B}_1$ are, respectively, $m \times d_0$ and , $m \times d_1$ matrices and $\mathbf{e}_0$ and $\mathbf{e}_1$ are error vectors. For simplicity a one-way ANOVA parameterization for $\mathbf{w}$ is assumed. Slight generalizations to account for design complications met in practice is described in Example 2.

A reasonable regression parameterization for $\mathbf{z}$ is also assumed, including an intercept, and for convenience, the first column of $\mathbf{B}_0$ is denoted as μ_1, the $m \times 1$ intercept. The error vectors $\mathbf{e}_0$ and $\mathbf{e}_1$ may reflect independence among arrays h and i, or else may have more complex random effects structure accounting for technical effects or biological replication; however, their substructures are incidental to this analysis, with the exception of the fine details of the bootstrap procedure proposed below.

To implement a surrogacy relation, the following linking regression model is proposed:

$$\mathbf{B}_1 = \mathbf{1}_m\gamma_0^{\mathrm{T}} + \mathbf{B}_0\Gamma + \mathbf{U}, \tag{2}$$

where Γ is a $d_0 \times d_1$ matrix that summarizes associations between the rows of $\mathbf{B}_{0j}$ and $\mathbf{B}_{1i}$ and $\mathbf{U}$ is a matrix of errors. Substituting equation (2) into (1), writing $\mathbf{B}_0 = (\mathbf{b}_{01},...,\mathbf{b}_{0d_0})$ explicitly in terms of its columns and writing $\Gamma^{\mathrm{T}} = (\gamma_1,...,\gamma_{d_0})$, it follows that

$$\mathbf{Y}_{1i} = \sum_{l=0}^{d_0} \mathbf{b}_{0l}(\gamma_l^{\mathrm{T}}\mathbf{z}_{1i}) + (\mathbf{1}_m\gamma_0^{\mathrm{T}} + \mathbf{U})\mathbf{z}_{1i} + \mathbf{e}_{1i}. \tag{3}$$

To impart a biological interpretation, it is assumed assume that the DNA assayed in S_1 arises as a mixture of DNA from cell types profiled in S_0, with mixture coefficients whose population average, conditional on $\mathbf{z}$, are $\{\omega_1^{(\mathbf{z})},...,\omega_{d_0}^{(\mathbf{z})}\}$, so that

$$\mathrm{E}(\mathbf{Y}_{1i} \mid \mathbf{z}_{1i} = \mathbf{z}) = \xi^{(\mathbf{z})} + \sum_{l=1}^{d_0} \mathbf{b}_{0l} \omega_l^{(\mathbf{z})}, \tag{4}$$

where the $m \times 1$ vector $\xi^{(\mathbf{z})}$ represents cell types excluded from consideration among the purified samples in S_0, or else non-cell specific methylation, including alterations at the molecular level in the maintanence of DNA methylation patterns themselves (possibly exposure related, age, or disease related). It follows from (3) and (4) that the mixture coefficients are recoverable from Γ, $\omega_l^{(\mathbf{z})} = \gamma_l^{\mathrm{T}} \mathbf{z}_{1i}$, provided $\xi^{(\mathbf{z})}$ is orthogonal to the column space of $\mathbf{B}_0$. As discussed in detail in the Example 3 bias can arise if differences in $\xi^{(\mathbf{z})}$ between distinct values of $\mathbf{z}$ have nonzero projection onto the column space of $\mathbf{B}_0$, although the magnitude of anticipated biases can be assessed through sensitivity analysis as sown in Example 11.

It is possible to assign interpretations to the components of variation in (3). SS_o represents overall variability in $\mathbf{Y}_{1i}$, i.e. $SS_o = \sum_{i=1}^{n_1} \| \mathbf{Y}_{1i} - \mu_1 \|^2$, where $\mu_1 = \mathrm{E}(\mathbf{Y}_{1i})$. From multivariate probability theory it is straightforward to show that $SS_o = SS_e + SS_v + SS_u$, where $SS_e = \sum_{i=1}^{n_1} \| \mathbf{e}_{1i} \|^2$, $SS_v = \sum_{i=1}^{n_1} (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)^{\mathrm{T}} \Gamma^{\mathrm{T}} \mathbf{B}_0^{\mathrm{T}} \mathbf{B}_0 \Gamma (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)$, and $SS_u = \sum_{i=1}^{n_1} \{ (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)^{\mathrm{T}} \mathbf{U}^{\mathrm{T}} \mathbf{U} (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1) + m (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)^{\mathrm{T}} \gamma_0 \gamma_0^{\mathrm{T}} (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1) \}$. SS_e measures variation unexplained by the covariates $\mathbf{z}_{1i}$, presumed to represent a combination of technical noise and unsystematic biological heterogeneity. SS_v measures variability explained by mixtures of profiles in the set S_0, and SS_u measures variability in systematic biological heterogeneity that nevertheless remains unexplained by mixtures of profiles in S_0, presumably due to some process other than differences in mixtures of cell types. Thus two partial coefficient of determination measures are proposed: $R_{1,0}^2 = SS_v / SS_o$, which represents the proportion of total variation in S_1 explained by S_0, and $R_{1,1}^2 = SS_v / (SS_o - SS_e)$, which represents the proportion of systematic variation in S_1 explained by S_0. It is noted that $R_{1,1}^2$ is poorly defined when $SS_o \approx SS_e$.

Estimation proceeds by applying an appropriate linear model, e.g. ordinary least squares, linear mixed effects models (Wang and Petronis, 2008, DNA Methylation Microarrays: Experimental Design and Statistical Analysis. Chapman & Hall, Boca Raton, Florida), limma (Smyth, 2004, Stat Appl Genet and Mol Biol, 3(1), 3), or surrogate variable analysis (Teschendorff et al., 2011, Bioinformatics, 27(11), 1496–505), to obtain estimates $\hat{\mathbf{B}}_0$ and $\hat{\mathbf{B}}_1$. Estimates of γ_0 and Γ are then obtained by projecting $\hat{\mathbf{B}}_1$ onto the column space of $\tilde{\mathbf{B}}_0 = (\mathbf{1}_m, \mathbf{B}_0)$, as described in detail in the Example 3. Standard errors can be obtained in one of three ways. The simplest estimator, SE_0, is the "naive" estimator from simple least squares theory, ignoring the fact that $\hat{\mathbf{B}}_0$ and $\hat{\mathbf{B}}_1$ are estimates, i.e. potentially variable. To account for variation in estimating $\hat{\mathbf{B}}_1$, a simple alternative is to use a nonparametric bootstrap procedure. For each bootstrap iteration t, sampling is performed with replacement from S_1 (or sample errors in a manner consistent with a hierarchical experimental design) to obtain $S_1^{(t)}$, producing bootstrap estimates $\hat{\mathbf{B}}_1^{(t)}$ from which "single-bootstrap" standard errors SE_1 are computed. Finally, it is possible to account for variation in estimating $\mathbf{B}_0$ by also bootstrapping S_0; because of potentially small sample sizes n_0, using a parametric bootstrap is proposed herein. A "double-bootstrap" standard error estimator, SE_2, is computed from these two sets of bootstraps. The double-bootstrap has the additional benefit over the single-bootstrap, in that it can be used to assess bias due to measurement error (variability) in $\hat{\mathbf{B}}_0$. Estimation details are provided in Example 3.

Beyond bias due to measurement error, which is easily corrected using the double-bootstrap procedure, there are additional sources of potential bias. For example, a univariate z_{1i} representing case/control status is considered, where $\delta \equiv \xi^{(1)} - \xi^{(0)} = \mathbf{B}_0 \alpha$ for some $d_0 \times 1$ vector $\alpha \neq \mathbf{0}$. In such a situation, there will be a bias equal to $\boldsymbol{\alpha}$ in estimating the mixture differences. Example 2 provides a detailed analysis of such biases, and proposes a sensitivity analysis procedure for assessing the magnitude of possible bias in a given data set.

In the examples herein the method for inferring changes in the distribution of white blood cells between different subpopulations is used for analysis of population data. It is possible to use S_0 to predict distribution of leukocytes in a single sample having DNA methylation profile $\mathbf{Y}^*$. Equating the intercept term of $\mathbf{B}_1$ in (1) with $\mathbf{Y}^*$ and applying (2), mixing proportion estimates $\Gamma^* = (\tilde{\mathbf{B}}_0^{\mathrm{T}} \tilde{\mathbf{B}}_0)^{-1} \tilde{\mathbf{B}}_0^{\mathrm{T}} \mathbf{Y}^*$ is obtained. Estimates can be further refined with the use of quadratic programming techniques (Goldfarb and Idnani, 1983, Math Prog, 27,

1–33), restricting the components of Γ^*, $\gamma_l^* \geq 0$ in minimizing $\| \mathbf{Y}^* - \mathbf{B}_0 \Gamma^* \|^2$ with respect to Γ^*. Such individual projections of methylation profiles on the column space spanned by S_0 facilitate the application of the fundamental ideas proposed above to individual, clinically-based diagnostic procedures.

It is noted that DNA methylation arrays are typically focused on the comparison of methylated to unmethylated CpG dinucleotides, not quantifying actual amounts of DNA. Therefore, information on cell mixtures from DNA methylation is limited to distributions, not actual counts, as one might obtain from flow cytometry. In addition, it is possible to model $\mathbf{z}_{1i}$ directly as a function of mixture coefficients Γ^* obtained individually via the constraint $\gamma_l^* \geq 0$.

Example 2: General designs for the treatment of methylation assay data obtained from purified cells S_0

Because the cell types assembled in S_0 potentially involve hierarchical relationships corresponding to cell lineage, designs that are more general than a one-way ANOVA parameterization may be necessary for $\mathbf{w}$. If cell-type interpretations can be extracted from S_0 via a $d_0 \times d_0^*$ contrast matrix $\mathbf{L}$ (i.e. $\mathbf{B}_0\mathbf{L}$ identifies the mean methylation for d_0^* cell types), then interpretations can be obtained by simply replacing $\hat{\mathbf{B}}_0$ with $\hat{\mathbf{B}}\mathbf{L}$ in the projection used to estimate γ_0 and Γ and their standard errors. The case of CD4+ and CD8+ T cells, both of which are the primary components of the T-lymphocyte group is considered as an example. In this example one sample is purified CD4+ T cells, another sample is purified CD8+ T cells, and yet another sample is T-lymphocyte cells that have not been purified to more specific lineages. Such was the case for S_0 in the examples. The CD4+ sample may be identified as $\mathbf{w}_{0h} = (1,1,0)^T$, the CD8+ sample as $\mathbf{w}_{0h} = (1,0,1)^T$, and the latter, less specific sample as $\mathbf{w}_{0h} = (1,0,0)^T$. Then an appropriate contrast $\mathbf{L}$ for identifying CD4+ and CD8+ samples would be constructed as a 3 x 2 matrix with columns $(1,1,0)^T$ and $(1,0,1)^T$. This approach was used in the examples 6-9 below, and was also employed in the simulations.

Example 3: Estimation details and bias

Estimation : A two-stage estimation procedure is here introduced. The first stage of analysis involves estimation of $\mathbf{B}_0$ and $\mathbf{B}_1$ by appropriate linear models, e.g. ordinary least squares (OLS) regression estimator $\hat{\mathbf{B}}_0^T = \left[\sum_{h=1}^{n_0} \mathbf{z}_{0h}\mathbf{z}_{0h}^T\right]^{-1}\left[\sum_{h=1}^{n_0} \mathbf{z}_{0h}^T\mathbf{Y}_{0h}^T\right]$ and a similar estimator for

$(\hat{\mu}_1, \hat{\mathbf{B}}_1)^{\mathrm{T}}$; a procedure such as *limma*; or else locus-by-locus linear mixed effects models that adjust for technical (e.g. chip) effects. The second stage of analysis, estimation of γ_0 and Γ, proceeds as follows:

$$(\hat{\gamma}_0, \hat{\Gamma}^{\mathrm{T}})^{\mathrm{T}} = \tilde{\mathbf{B}}_1^{\mathrm{T}} \tilde{\mathbf{B}}_0 (\tilde{\mathbf{B}}_0^{\mathrm{T}} \tilde{\mathbf{B}}_0)^{-1}, \quad (5)$$

where $\tilde{\mathbf{B}}_0 = (\mathbf{1}_m, \hat{\mathbf{B}}_0)$. Let $\hat{\mathbf{r}}_\gamma = \hat{\mathbf{B}}_1 - \mathbf{1}_m \hat{\gamma}_0 - \hat{\mathbf{B}}_0 \hat{\Gamma}$, $\hat{\Sigma}_\gamma \equiv (\hat{\sigma}_{rs}^{(\gamma)})_{rs} = (m - d_0 - 1)^{-1} \hat{\mathbf{r}}_\gamma^{\mathrm{T}} \hat{\mathbf{r}}_\gamma$, $\mathbf{V}_0 = m(\tilde{\mathbf{B}}_0^{\mathrm{T}} \tilde{\mathbf{B}}_0)^{-1}$, and $\mathbf{V}_0 = (v_{rs}^{(0)})_{rs}$. Naive standard error estimates for the $(r,s)^{th}$ element of $(\hat{\gamma}_0, \hat{\Gamma}^{\mathrm{T}})$ can be obtained by computing $(m^{-1} v_{ss}^{(0)} \hat{\sigma}_{rr}^{(\gamma)})^{1/2}$. The naive standard error estimates fail to account for the variability in estimating $\hat{\mathbf{B}}_0$ and $\hat{\mathbf{B}}_1$, and are consequently biased, as demonstrated in the simulations, Example 12.

A nonparametric bootstrap procedure is used as an alternative. For each bootstrap iteration t, with replacement from S_1 is sampled, (or sample errors in a manner consistent with a hierarchical experimental design, e.g. taking into account chip effects), to obtain $S_1^{(t)}$. From $S_1^{(t)}$ an estimate of $\hat{\mathbf{B}}_1^{(t)}$ is obtained, and then $\hat{\gamma}_0^{(t)}$ and $\hat{\Gamma}^{(t)}$ are computed by replacing $\hat{\mathbf{B}}_1$ with $\hat{\mathbf{B}}_1^{(t)}$ in (**S1**). After resampling a large number T times, standard errors are obtained empirically from the bootstrap sets $\{\hat{\gamma}_0^{(t)}\}_{t=1,\ldots,T}$ and $\{\hat{\Gamma}^{(t)}\}_{t=1,\ldots,T}$. This method of estimation is called the "single bootstrap" to distinguish it from an alternative that accounts for variability in estimation of $\hat{\mathbf{B}}_0$ as well.

Because S_0 will typically consist of small sample sizes per cell type, a nonparametric bootstrap procedure for estimating variation in $\hat{\mathbf{B}}_0$ may not perform well. Therefore a parametric bootstrap is used. Let Ω_j be the variance-covariance matrix for the j^{th} row of $\hat{\mathbf{B}}_0$. A resampled matrix $\hat{\mathbf{B}}_0^{(t)}$ is formed by adding, to each row j of $\hat{\mathbf{B}}_0$, a zero-mean multivariate normal vector with variance-covariance Ω_j, or a corresponding multivariate t-distribution with $n_0 - d_0$ degrees of freedom. Then $\hat{\gamma}_0^{(t)}$ and $\hat{\Gamma}^{(t)}$ are computed from (**S1**) by replacing $\hat{\mathbf{B}}_0$ with $\hat{\mathbf{B}}_0^{(t)}$ (in addition to the previously mentioned replacement). This method is referred to as the "double bootstrap". The double bootstrap ignores correlation between CpG sites within a single validation sample, and given the relative purity assumed for these samples and adequate correction for technical effects, this is reasonable to first order. As is demonstrated in

Examples 6-9 and simulations (Example 10), there is negligible difference between the single and double bootstrap, so the incorporation of additional complexity to model cross-CpG correlations is unlikely to produce much benefit. However, the double-bootstrap has the additional benefit over the single-bootstrap, in that it can be used to assess bias due to measurement error (variability) in $\hat{\mathbf{B}}_0$.

Bias: There are several potential sources of bias in this analysis. The first arises from measurement error in $\mathbf{B}_0$, and the others arise from biological non-orthogonality.

It can be shown that first form of bias, from measurement error, manifests as a multiple of $\mathbf{\Gamma}$ on the order of $\mathbf{V}_0\overline{\Omega}$, where $\overline{\Omega} = m^{-1}\sum_{j=1}^{m}\Omega_j$. However, it is easily assessed using the double-bootstrap procedure described above, by subtracting $\hat{\gamma}_0$ from $T^{-1}\sum_{t=1}^{T}\hat{\gamma}_0^{(t)}$ and $\hat{\Gamma}$ from $T^{-1}\sum_{t=1}^{T}\hat{\Gamma}^{(t)}$, and bias correction can be implemented by subtracting this term from the estimate.

Biases induced by biological non-orthogonality are more insidious. For example, a univariate z_{1i} is considered representing case/control status, where $\delta \equiv \xi^{(1)} - \xi^{(0)} = \mathbf{B}_0\alpha$ for some $d_0 \times 1$ vector $\alpha \neq \mathbf{0}$. In such a situation, there will be a bias equal to α in estimating the mixture differences. Non-orthogonal δ may arise from two distinct sources. One occurs when some cell types have not been profiled in S_0, so that $\sum_{l=0}^{d_0}\omega_l^{(z)} < 1$. The other may arise when some non-cell-mediated biological process (i.e. distinct from a change in cellular mixtures) nevertheless results in methylation profiles that appear similar to those that distinguish cell types profiled in S_0,. To this end, model represented by equation (4) is elaborated follows:

$$\mathrm{E}(\mathbf{Y}_{1i} \mid \mathbf{z}_{1i1} = z) = \sum_{l=1}^{d_0}(\mathbf{B}_0\varepsilon_l + \lambda_l^{(z)})\omega_l^{(z)} + \sum_{q=1}^{Q}(\tilde{\mu}_q + \tilde{\lambda}_q^{(z)})\tilde{\omega}_q^{(z)}, \qquad (6)$$

where $q \in \{1,...,Q\}$ indexes unprofiled cell types (or free DNA), each with methylation profile $\tilde{\mu}_q$, and in mixture proportions $\omega_l^{(z)}$ and $\tilde{\omega}_q^{(z)}$, $\sum_{l=1}^{d_0}\omega_l^{(z)} + \sum_{q=1}^{Q}\tilde{\omega}_q^{(z)} = 1$. Here $\lambda^{(z)}$ denotes an "abnormal", or at least non-functional, non-cell-mediated process that is specific to disease status (and may affect different cell types in different degrees of intensity).

Let $\mathbf{P} = (\widetilde{\mathbf{B}}_0^{\mathrm{T}}\widetilde{\mathbf{B}}_0)^{-1}\widetilde{\mathbf{B}}_0^{\mathrm{T}}$, and denote difference between case and control parameters using Δ, e.g. $\Delta\omega_l = \omega_l^{(1)} - \omega_l^{(0)}$ and $\Delta\mathrm{E}(\mathbf{Y}_{1i}) = \mathrm{E}(\mathbf{Y}_{1i} \mid \mathbf{z}_{1i1} = 1) - \mathrm{E}(\mathbf{Y}_{1i} \mid \mathbf{z}_{1i1} = 0)$. It follows from equation (6) that

$$\mathbf{P}\Delta\mathrm{E}(\mathbf{Y}_{1i}) = \sum_{l=1}^{d_0}\varepsilon_l\Delta\omega_l + \sum_{q=1}^{Q}\mathbf{P}\mu_q\Delta\widetilde{\omega}_q + \sum_{l=1}^{d_0}\mathbf{P}\Delta(\lambda_l\omega_l) + \sum_{q=1}^{Q}\mathbf{P}\Delta(\lambda_q\widetilde{\omega}_q). \quad (7)$$

The values $\Delta\widetilde{\omega}_q$ may need to shift in order to accommodate any shifts in $\Delta\omega_l$, since the model constrains $\sum_{l=1}^{d_0}\Delta\omega_l + \sum_{q=1}^{Q}\Delta\widetilde{\omega}_q = 0$. The first term on the right hand side of (6) is the target quantity, identifying the desired mixture weights. The second term will be negligible if all profiles $\widetilde{\mu}_q$ are approximately orthogonal to the columns of $\mathbf{B}_0$, or else the differences $\Delta\widetilde{\omega}_q$ are all small. This condition will be satisfied if S_0 is exhaustive in the sense that $1 - \sum_{l=1}^{d_0}\omega_l^{(z)}$ is negligible.

Mathematically, it is difficult to further characterize the latter two terms, without specifying what kinds of non-cell-mediated processes are likely. For example, even if $\Delta\lambda_q = 0$ for a particular value of q, it may nevertheless still produce a bias if $\Delta\widetilde{\omega}_q \neq 0$. Conversely, even if $\Delta\omega_l = 0$, bias can result from a nonzero difference $\Delta\lambda_l$ (e.g. different methylation intensities at island shores due to distinct risk profiles) if $\Delta\lambda_l$ is not annihilated by $\mathbf{P}$. Only processes that are equal in intensity in both cases and controls and across all cell types will be differenced out of equation (7). Thus, a key consideration is whether $\mathbf{P}$ annihilates the methylation signature corresponding to a given non-cell-mediated biological process. In order to examine this issue more carefully, a Bayesian view is adopted to characterize a prior expectation of bias as a function of prior probabilities for individual CpG sites. The goal, in part, is to understand the potential for bias, given the number m of CpG sites chosen to be measured in S_0, with the goal of selecting m in a manner consistent with minimizing bias.

Assuming that the CpGs under consideration are ordered in advance (e.g. randomly or by F-statistic $F_j = d_0^{-1}\hat{\mathbf{B}}_{0j\cdot}\Omega_j^{-1}\hat{\mathbf{B}}_{0j\cdot}^{\mathrm{T}}$, and that the dependence of $\mathrm{tr}\mathbf{H}_m = \widetilde{\mathbf{B}}_0^{\mathrm{T}}\widetilde{\mathbf{B}}_0$ is explicitly written on m. If the CpGs are randomly ordered, then $\mathrm{tr}\mathbf{H}_m = O(m)$, otherwise it is possible that $\mathrm{tr}\mathbf{H}_m = O(m^{1-\zeta})$, $\zeta > 0$ reflecting a diminishing rate of return by adding additional non-informative CpG sites. Then $\delta = \sum_{l=1}^{d_0}\mathbf{P}\Delta(\lambda_l\omega_l) + \sum_{q=1}^{Q}\mathbf{P}\Delta(\widetilde{\lambda}_q\widetilde{\omega}_q)$ is decomposed by the number

k of CpG sites affected by all alterations that distinguish cases from controls. k is fixed, $k \in \mathrm{J}_m = \{1,..., m\}$; each of the $C(m,k) = m!/[k!(m-k)!]$ subsets $\mathrm{J}_{kl} \subset \mathrm{J}_m$ of k indices corresponds to a vector δ_{kl} representing the mean methylation difference between case and control over all systematic biological processes that result in changes at the k specific CpG sites represented by the k indices, and only those k CpG sites. Thus δ_{kl} has at most k nonzero values. The bias resulting from such processes is $\mathbf{H}_m^{-1}\tilde{\mathbf{B}}_0^{\mathrm{T}}\delta_{kl} = O(km^{\zeta-1})$. A prior probability π_{kl} is assumed that the subset J_{kl} could correspond to one or more biological processes that distinguish cases from controls. It follows from this view that the prior expectation of δ is

$$\mathrm{E}[\delta|(\pi_{kl})_{kl}] = \sum_{k=1}^{m}\sum_{l=1}^{C(m,k)}\pi_{kl}\delta_{kl} = O\left(\sum_{k=1}^{m}\sum_{l=1}^{C(m,k)}\pi_{kl}km^{\zeta-1}\right). \quad (8)$$

If a prior probability over all sets of CpG sites in the genome is constructed so that CpG sites are considered independent, and each CpG site is assigned a uniform prior probability of π_0, then $\pi_{kl} \equiv \pi_0^k(1-\pi_0)^{m-k}$ and, from (8),

$$\mathrm{E}(\delta|\pi_0) = O\left(m^{\zeta}\sum_{k=1}^{m}C(m-1,k-1)\pi_0^k(1-\pi_0)^{m-k}\right) = \pi_0(1-\pi_0)O(m^{\zeta}). \quad (9)$$

The bias does not depend on m if $\mathrm{tr}\mathbf{H}_m = O(m)$, i.e. random ordering. Random ordering renders the size of $\mathrm{E}(\delta|\pi_0)$ theoretically independent of m, it does so at the cost of including many potentially noninformative CpGs, early on at low values of m, and these may be possible sources of bias in practice, without offering any modeling benefit in return. If the CpG sites are ordered by level of informativeness, then potentially $\mathbf{H}_m = O(m^{1-\zeta})$, and there will be a small increasing prior expectation of bias, motivating judicious choice of m. The key, then, is to order the CpGs in terms of their ability to distinguish different types profiled in S_0, choosing m large enough to distinguish all signatures from one another, but small enough that the $\mathrm{E}(\delta|\pi_0)$ is reasonably low, in a relative sense. Naturally, different choices of prior π_{kl} in (8) will lead to different conclusions about the magnitude of bias. If the set J_m of CpG sites used in S_0 and S_1 oversample those known to have less modifiable methylation states, e.g. away from so-called shore regions (Doi A et al., 2009, Nat Genet 41: 1350-3), then π_0 is effectively lowered, and so

will be the corresponding expected prior bias. It is worth emphasizing that this analysis concerns only a Bayesian prior, not the actual biological truth. In choosing CpG sites among those assayed in S_0 and S_1, a potentially negative outcome would be to have included a number of sites that also happen to represent systematic, non-cell-mediated biological differences between cases and controls in S_1, in which case biased estimates will be inevitable. In summary, bias in the proposed estimation procedure is controlled by selecting a sufficiently exhaustive list of cell types to profile in S_0, and by choosing m judiciously.

Example 4: Proof of concept of Measurement Error Model for determining changes in distribution of white blood cells between different subpopulations

In this example, general features of the method herein are described that can be used with existing methylation data sets as benchmarks for validating the proposed method to demonstrate its clinical or epidemiological utility. Examples 6-9 that follow show application of the method to specific data sets. The data analyses involve DNA methylation data obtained by the Infinium HumanMethylation27 Beadchip Microarrays from Illumina, Inc. (San Diego, CA). A subset of m = 100 CpG sites on the array was used and the subset was selected as described below. In Examples 6-9, S_0 consisted of 46 white blood cell samples; the sorted, normal, human, peripheral blood leukocyte subtypes were purchased from AllCells®, LLC (Emeryville, CA) and were isolated from whole blood using a combination of negative and positive selection with highly specific cell surface antibodies conjugated to magnetic beads; materials and protocols were obtained from Miltenyi Biotec, Inc. (Auburn, CA). These 46 samples are summarized in Table 2 and depicted by the clustering heatmap in Figure 1. T lymphocytes that express CD4 or CD8 constitute over 95% of the T cell class. The pan-T cell type was further refined to CD4+, CD8+, and "other" Pan-T cells subtypes.

In summary, the covariate vector $\mathbf{w}_h$ consisted of indicators for five cell types and another two indicators for CD4+ and CD8+ T cell subtypes. A generalization of the one-way ANOVA parameterization assumed above for $\mathbf{w}_h$ (Example 2) was necessary to account for the ambiguous status of some Pan-T cells. For each CpG site, a linear mixed effects model with a random intercept for bead chip was used to estimate $\mathbf{B}_0$; 27 additional whole blood control samples (replicates from the same individual) were used to assist in estimating chip effects, since otherwise the data set would have been sufficiently sparse to risk confounding between cell type and chip. These "array controls" were indicated with an additional term in $\mathbf{w}_{0h}$. For

each CpG site, a linear mixed effects model with a random intercept for bead chip was used to estimate the corresponding row of $\mathbf{B}_0$ and $\mathbf{B}_1$.

From S_0, F statistics were computed and used to order each of the 26,486 autosomal CpGs by decreasing level of informativeness with respect to blood cell types. Figure 5A depicts the relationship $\log_{10}$ tr$\mathbf{H}_m$ by $\log_{10}$ (m) for increasing array sizes. Figure 5B depicts the relationship $\partial\log_{10}$ tr($\mathbf{H}_m$)/ ∂log(m) by $\log_{10}(m)$ for increasing array sizes, obtained by smoothing the first differences of the curve depicted in Figure 5 panel A via loess smoother. Figure 5 panel A also shows the tangent (obtained from the loess curve) at low values of m. For $O(m)$ convergence, Figure 5 panel A should show a linear association with slope equal to one, and the curve in Figure 5 panel B should show a curve close to the value of 1.0. Neither is the case, i.e. convergence is sub-linear in m. It is noted that the rate of convergence dropped precipitously after about 6,000 CpG sites, but was notably slower than $O(m)$ even after m = 10. In the range of 1-1000 CpG sites the convergence rate appeared parabolic with a minimum of about 0.85, starting to stabilize in the m = 100 - 300 range. Thus, maximum informativeness was provided by the highest ranking m = 100 - 300 CpG sites, with m > 300 reflecting diminishing returns from adding additional CpGs. Therefore, a moderately low value of m in this range, m = 100, consistent with the size of a small custom microarray chip was chosen.

Table 2. Sorted white blood cells in S_0

Short Name	Description	Number
B cells	CD19+ B-lymphocytes	6
Granulocytes	CD15+ granulocytes	8
Monocytes	CD14+ monocytes	5
NK	CD56+ Natural Killer (NK) cells	11
T cells (CD4+)[1,2]	CD3+CD4+ T-lymphocytes	8
T cells (CD8+)[1,3]	CD3+CD8+ T-lymphocytes	2
T cells (NKT)[1]	CD3+CD56+ natural killer	1
T cells (other)[1]	CD3+ T-lymphocytes	5

[1] Considered as a member of the "pan-T cell" group.
[2] Pan-T cell further refined as also belonging to the "CD4+" group.
[3] Pan-T cell further refined as also belonging to the "CD8+" group.

Example 5: Cell mixture experiment for validating the method for determining changes in distribution of white blood cells between different subpopulations

In this example is described a laboratory reconstruction experiment, which validates the concept on which the method herein is based that DNA methylation retains substantial information about cell mixtures. The results of applying the method herein to several different target data sets S_1 is described in Examples 6-9.

For the HNSCC and ovarian cancer data sets, from which bead chip data were available, a linear mixed effects model with a random intercept for bead chip was used to estimate the corresponding row of B1. For the remaining data sets, no bead chip data were available; consequently, ordinary least squares was used. 250 bootstrap iterations were used for each example and each of the two bootstrap methods of standard error estimation.

An experiment was conducted which involved six known mixtures of monocytes and B cells and six known mixtures of granulocytes and T cells. Figure 2 presents both the known fractions ("Expected") and the resulting predictions ("Observed") from Infinium 27K profiles, as described above. As Figure 2 shows, accuracy of prediction is within 10%, and often less than 5%, with the largest errors occurring for granulocytes, as shown in Table 3. It is noted that the sum of the individual observed predictions for each individual profile ranged from 98.9% to 102.7% even though the constraints of the projection do not explicitly constrain the sum to 100%; this provides additional evidence that the DNA methylation profile captures information about cell mixtures.

Table 3. Summary statistics for errors in cell mixture reconstruction Results*

	B cell	Granulocyte	Monocyte	NK	T cell
minimum	0.0	0.3	0.0	0.0	0.0
median	0.1	6.5	1.1	2.1	0.3
maximum	5.5	10.0	4.1	6.4	5.3

* |Observed% - Expected%|

Example 6: Application of the methods herein to the subpopulations of head and neck cancer patients and controls

This example describes the application of the method herein for determining changes in the distribution of white blood cells between different subpopulations to patients having head and neck squamous cell carcinoma (HNSCC). The target data set S_1 was obtained from arrays applied to whole blood specimens collected in a random subset of individuals involved in an ongoing population-based case-control study (Peters et al., 2005, Cancer Epidemiol Biomarkers Prev, 14(2), 476–82) of head and neck cancer (HNSCC): 92 cases and 92 age and sex matched controls. Blood was drawn at enrollment (prior to treatment in 85% of the cases). Mean age among the subjects arrayed in this study was 60 years, and there were 56 females and 128 males, consistent with the higher incidence of the disease in men. Thus, the covariate vector **z** consisted of an indicator for case/control status, an indicator for male sex, and age (in decades) centered at the mean. The clustering heatmap in Figure 3 depicts the raw DNA methylation data in S_1. Table

4 presents coefficient case status, double-bootstrap bias estimates (estimates of bias arising from measurement error), as well as naive, single-bootstrap, and double-bootstrap standard error estimates. Each of these quantities is measured in percentage points (%). Estimates of bias arising from measurement error (i.e. substituting estimated quantities for known ones in a two-stage statistical procedure) were almost always less than half a percentage point, and for significant coefficient estimates, always towards the null.

The proportion of CD4+ T-lymphocytes decreased in cases compared with controls, with a bias-corrected estimate of -10:4 percentage points and approximate 95% confidence interval (-13:1%;-3:3%); the proportion of NK cells decreased, with a bias-corrected estimate of -1.5 percentage points and 95% confidence interval (-2:2%;-0:75%); and the proportion of granulocytes increased, with a bias-corrected estimate of 7.6 percentage points and 95% confidence interval (4:2%; 10:9%). There was also some evidence of an increase in CD8+ T-lymphocytes, with an estimate of 4.5 percentage points and 95% confidence interval (4:5%; 7:0%). As shown in Table 5 the proportion of CD4+ T-lymphocytes decreased by 3.3 percentage points (-4:4%;-2:2%) per decade of age, and CD8+ T-lymphocytes increased by 2.0 percentage point (1:0%; 3:0%) per decade. All other coefficients were insignificant.

For this analysis, $R^2_{1,0}$ was estimated at 14.2%, and $R^2_{1,1}$ was estimated at 93:9%. Thus, a small but non-negligible proportion of total variation (systematic variation + unexplained biological heterogeneity + technical noise) appeared to have been driven by changes in cell population between cases and controls and as a result of aging. The SS_e comprised 85% of total variation, so a substantial portion of variability in DNA methylation appeared to remain unexplained (presumably due, in large part, to technical noise). However, almost all of the systematic variation was explained by changes in cell population.

These results were consistent with previous studies, as HNSCC patients are known to display an absolute and relative increase in myeloid derived granulocytes (Trellakis et al., 2011, Int J Cancer, Epub ahead of print, DOI: 10.1002/ijc.25892) and also displayed an alteration in lymphoid T cell homeostasis that leads to decreases in CD4+ T cells (Kuss et al., 2004, Clin Cancer Res, 10(11), 3755–62; Kuss et al., 2005, Adv Otorhinolaryngol, 62, 161–72). In addition, the proportion of Treg cells (a subclass of CD4+ T cells) is known to decrease from infancy to adulthood (Mold et al., 2010, Science, 330(6011), 1695–9). The bias estimates obtained from the double-bootstrap procedure allow the correction of bias arising from measurement error. However, there is no statistical procedure for correcting the other possible sources of bias, those arising from changes in distribution among unprofiled cell types as well as non-immune-mediated methylation differences. Example 7 presents a detailed sensitivity

analysis which shows that the magnitude of the resulting bias is likely to be small, less than a percentage point.

Table 4. Estimates for HNSCC analysis (case vs. control)

	Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
(Intercept, γ_0)	-0.62	-0.02	0.41	0.52	0.52	0.23
B Cell	-0.45	0.04	0.30	0.77	0.76	0.55
Granulocyte	7.51	-0.07	0.50	1.73	1.71	<0.0001
Monocyte	0.49	0.10	0.50	0.47	0.48	0.31
NK	-1.43	0.06	0.56	0.37	0.38	0.00017
T Cell (cd4+)	-9.08	1.32	1.95	1.15	1.39	<0.0001
T Cell (cd8+)	3.06	-1.46	1.96	0.98	1.27	0.016

Est = Regression coefficient estimate (x 100%).
$Bias_2$ = Double-bootstrap bias estimate (x 100%).
SE_0 = Naive standard error (x 100%)
SE_1 = Single-bootstrap standard error (x 100%).
SE_2 = Double-bootstrap standard error (x 100%).
P-values were computed using SE_2.

Table 5. Estimated Regression Coefficients for Sex and Age in HNSCC Data Set

		Est	$Bias_2$	SE_0	SE_1	SE_2		P-value
Sex	(Intercept, γ_0)	0.12	0.00	0.24	0.57	0.57		0.83
	B Cell	0.38	0.01	0.17	0.85	0.84		0.65
	Granulocyte	-0.29	-0.08	0.28	1.82	1.81		0.87
	Monocyte	0.13	0.01	0.29	0.47	0.47		0.78
	NK	0.49	0.05	0.32	0.40	0.40		0.22
	T Cell (cd4+)	-1.80	0.45	1.12	1.25	1.20		0.13
	T Cell (cd8+)	0.82	-0.44	1.12	1.03	1.04		0.43
(Age - 60)/10	(Intercept, γ_0)	-0.20	-0.02	0.15	0.24	0.24		0.40
	B Cell	0.24	0.01	0.11	0.34	0.33		0.47
	Granulocyte	1.12	-0.01	0.19	0.67	0.67		0.096
	Monocyte	0.13	0.02	0.19	0.20	0.20		0.54
	NK	-0.22	0.02	0.21	0.15	0.15		0.14
	T Cell (cd4+)	-2.75	0.56	0.73	0.53	0.57	<	0.0001

	T Cell (cd8+)	1.44	-0.56	0.73	0.46	0.50		0.0038

Est = Regression coefficient estimate (× 100%)
$Bias_2$ = Double-bootstrap bias estimate (× 100%).
SE_0 = Naive standard error (× 100%).
SE_1 = Single-bootstrap standard error (× 100%).
SE_2 = Double-bootstrap standard error (× 100%).
P-values were computed using SE_2.

Example 7: Application of the methods herein to subpopulations of ovarian cancer cases and controls

In this example the method herein for inferring changes in the distribution of white blood cells between different subpopulations (e.g. cases and controls) was applied to an ovarian cancer data set (Teschendorff et al., 2009, PLoS ONE, 4(12), e8274). DNA methylation data for blood samples were obtained from Gene Expression Omnibus (Accession number GSE19711). Only those cases in which blood was collected pre-treatment were used ere. After removing four arrays with a preponderance of missing values, the data set consisted of 272 controls and 129 cases in which blood was collected prior to treatment. A clustering heatmap displaying the DNA methylation data is shown in Figure 6. In this analysis, **z** consisted of case-control status, age (categorized in five-year increments), and two bisulfite conversion efficiency measures. Tables 6-8 presents result for case-control status and estimated regression coefficients for age in ovarian cancer data set. $R^2_{1,0}$ was estimated at 17.8%, and $R^2_{1,1}$ was estimated at 86:1%.

Table 6. Estimates for Ovarian Cancer Analysis (Case vs. Control)

	Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
(Intercept, γ_0)	-0.05	-0.05	0.41	0.19	0.20	0.81
B Cell	-1.36	0.02	0.29	0.22	0.23	<0.0001
Granulocyte	8.97	-0.04	0.49	1.02	1.00	<0.0001
Monocyte	0.55	0.06	0.49	0.29	0.30	0.066
NK	-2.09	0.01	0.55	0.31	0.34	<0.0001
T Cell (cd4+)	5.64	0.18	1.93	1.06	1.34	<0.0001
T Cell (cd8+)	-0.35	-0.17	1.93	0.95	1.19	0.77

Est = Regression coefficient estimate (x 100%).
$Bias_2$ = Double-bootstrap bias estimate (x 100%).
SE_0 = Naive standard error (x 100%)
SE_1 = Single-bootstrap standard error (x 100%).
SE_2 = Double-bootstrap standard error (x 100%).
P-values were computed using SE2.

Table 7. Estimated Regression Coefficients for Age in Ovarian Cancer Data Set

		Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
Age 55-60	(Intercept, γ_0)	-1.24	-0.05	0.37	0.41	0.40	0.0021
	B Cell	0.40	0.04	0.27	0.50	0.49	0.42
	Granulocyte	0.91	0.04	0.45	2.04	2.02	0.65
	Monocyte	0.85	0.12	0.45	0.59	0.58	0.15
	NK	-0.25	0.10	0.50	0.55	0.55	0.65
	T Cell (cd4+)	-2.79	0.63	1.76	2.13	1.96	0.15
	T Cell (cd8+)	2.22	-0.84	1.77	1.81	1.59	0.16
Age 60-65	(Intercept, γ_0)	-0.72	-0.07	0.35	0.39	0.39	0.070
	B Cell	0.54	0.07	0.25	0.49	0.49	0.27
	Granulocyte	0.71	0.06	0.42	1.99	1.98	0.72
	Monocyte	0.27	0.08	0.42	0.58	0.58	0.64
	NK	-0.24	0.06	0.47	0.55	0.55	0.65
	T Cell (cd4+)	-3.54	0.80	1.66	2.02	1.97	0.072
	T Cell (cd8+)	2.84	-0.97	1.66	1.85	1.64	0.084
Age 65-70	(Intercept, γ_0)	-0.53	-0.08	0.40	0.41	0.41	0.19
	B Cell	-0.03	0.07	0.29	0.51	0.51	0.96
	Granulocyte	2.46	0.02	0.48	2.17	2.17	0.26
	Monocyte	0.85	0.12	0.48	0.64	0.64	0.18
	NK	-0.89	0.07	0.54	0.59	0.60	0.14
	T Cell (cd4+)	-6.12	1.48	1.89	2.18	2.12	0.0038
	T Cell (cd8+)	4.37	-1.64	1.89	1.87	1.71	0.011
Age 70-75	(Intercept, γ_0)	-1.20	-0.07	0.40	0.41	0.41	0.0037
	B Cell	0.29	0.07	0.29	0.48	0.48	0.55
	Granulocyte	2.13	-0.05	0.48	2.05	2.04	0.30
	Monocyte	0.76	0.12	0.48	0.60	0.60	0.21
	NK	-0.51	0.19	0.54	0.56	0.55	0.36
	T Cell (cd4+)	-6.82	1.97	1.89	2.16	2.12	0.0013
	T Cell (cd8+)	5.35	-2.20	1.90	1.89	1.79	0.0028
Age 75+	(Intercept, γ_0)	-0.31	-0.09	0.49	0.46	0.45	0.49
	B Cell	0.13	0.08	0.35	0.54	0.53	0.81
	Granulocyte	1.10	-0.15	0.58	2.12	2.11	0.60
	Monocyte	1.73	0.12	0.59	0.64	0.63	0.0065
	NK	-0.30	0.13	0.66	0.60	0.59	0.61
	T Cell (cd4+)	-6.54	1.31	2.30	2.29	2.18	0.0027
	T Cell (cd8+)	2.73	-1.37	2.31	2.06	1.86	0.14

Est = Regression coefficient estimate (× 100%)

$Bias_2$ = Double-bootstrap bias estimate (× 100%).

SE_0 = Naive standard error (× 100%).

SE_1 = Single-bootstrap standard error (× 100%).

SE_2 = Double-bootstrap standard error (× 100%).

P-values were computed using SE_2.

Table 8. Estimated Regression Coefficients for Bisulfite Conversion in Ovarian Cancer Data Set

		Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
BSC1	(Intercept, γ_0)	-0.08	0.00	0.14	0.09	0.10	0.39
(Green/1000)	B Cell	-0.10	0.00	0.10	0.10	0.10	0.30
	Granulocyte	0.13	0.04	0.17	0.40	0.40	0.74
	Monocyte	0.13	-0.01	0.17	0.12	0.12	0.26
	NK	-0.09	0.00	0.19	0.14	0.14	0.53
	T Cell (cd4+)	0.51	-0.14	0.65	0.48	0.51	0.32
	T Cell (cd8+)	-0.23	0.11	0.66	0.40	0.47	0.62
BSC2	(Intercept, γ_0)	0.25	0.00	0.14	0.08	0.08	0.0027
(Green/1000)	B Cell	0.07	0.00	0.10	0.08	0.08	0.40
	Granulocyte	0.07	0.01	0.17	0.38	0.37	0.84
	Monocyte	-0.18	0.01	0.17	0.10	0.10	0.075
	NK	0.10	0.00	0.19	0.12	0.12	0.41
	T Cell (cd4+)	-0.65	0.20	0.67	0.41	0.50	0.20
	T Cell (cd8+)	0.63	-0.21	0.68	0.34	0.45	0.16

Est = Regression coefficient estimate (× 100%)
$Bias_2$ = Double-bootstrap bias estimate (× 100%).
SE_0 = Naive standard error (× 100%).
SE_1 = Single-bootstrap standard error (× 100%).
SE_2 = Double-bootstrap standard error (× 100%).
P-values were computed using SE_2.
It is noted that coefficients are given as % / 1000 units fluorescence, and that standard deviations for BSC1 and BSC2 were 1950 and 2169, respectively.

Compared with controls, data obtained from cases showed significant increases in granulocytes and significant decreases in B cells, NK cells, and CD4+ T cells. Cases also showed marginally significant increases in monocytes. These results are consistent with previous literature, in which it has been demonstrated that ovarian cancer patients experience decreases in B and T lymphocytes (den Ouden et al., 1997, Eur J Obstet Gynecol Reprod Biol, 72, 73–77; Bishara et al., 2008, Reprod Biol, 138, 7175; Cho et al., 2009, Cancer Immunol Immunother, 58, 1523), increases in monocytes (den Ouden et al., 1997, Eur J Obstet Gynecol Reprod Biol, 72, 73–77; Bishara et al., 2008, Reprod Biol, 138, 7175) and (somewhat equivocally) increases in eosinophil granulocytes (Bishara et al., 2008, Reprod Biol, 138, 7175). Additionally, there were significant systematic decreases in CD4+ T cells with increasing age, with a gradient consistent in direction and somewhat consistent in magnitude with the corresponding effect found in the HNSCC data set. The CD8+ T cell coefficients for were all

positive, with gradient consistent in direction and somewhat consistent in magnitude with the corresponding effect found in the HNSCC data set. No bisulfite conversion coefficient was significant, and all coefficients were of small magnitude (Table 8; generally less than 1 percentage point per standard deviation).

Example 8: Application of the methods herein to subpopulations of Down Syndrome patients and controls

The method herein was applied to trisomy 21 (Down syndrome) data set (Kerkel et al.,. *PLoS Genet* 2010, **6**(11):e1001212) consisting of 29 total peripheral blood leukocyte samples from Down syndrome cases and 21 controls, as well as six T cell samples from cases and four T cell samples from controls (GEO Accession number GSE25395). Because of the potential for bias induced by copy number amplification four CpG sites on Chromosome 21were excluded, resulting in $m = 96$ CpG sites that were used for analysis. A clustering heatmap displaying the DNA methylation data is shown in Figure 7. In one analysis data from cases and controls were compared using the total leukocyte samples only, and in another total leukocytes to T cells were compared, pooling cases and controls. Coefficient estimates are provided in Table 9. The only significant difference between cases and controls was in B cell distribution, with bias-corrected estimated decrease of 4.8%, 95% confidence interval (- 6:2%; - 3:5%). This result is consistent with known immune characteristics of Down Syndrome, including deficiencies in both B and T cells (Verstegen et al., 2010, Pediatr Res, 67, 563–9; Ram and Chinen, 2011, Clin Exp Immunol, 164, 9–16). However, in the comparison between total leukocytes and T cells, all coefficients except B Cell and NK were highly significant, in directions consistent with comparison of a sample of purified T cells to a generic whole blood sample. In fact, an estimate of the cellular composition of the T cell samples can be obtained by a simple linear transformation of Γ estimates (adding intercept terms with the T cell coefficients); this operation produces values that are not significantly distinct from zero for all cell types except CD4+ and CD8+, whose bias-corrected estimates were, respectively, 75.9%, 95% confidence interval (67%; 85%) and 8.6%, 95% confidence interval (0%; 17%), for cases and controls consistent with the known distribution of these T cells. For the analysis of case vs. control within total leukocytes, $R^2_{1,0}$ was estimated at 4.5%, and $R^2_{1,1}$ was estimated at 67:6%. For the analysis of total leukocyte vs. T cell with pooled cases and controls, $R^2_{1,0}$ was estimated at 81.4%, and $R^2_{1,1}$ was estimated at 98:9%. The latter set of coefficients of determination indicates that a substantial portion of variation is explained by composition of leukocytes, which is the expected result for such an analysis.

Table 9. Estimates for Down syndrome analysis (case vs. control, total leukocyte vs. T Cell)

		Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
Case Status	Intercept, γ_0	2.02	-0.10	0.86	1.17	1.17	0.084
(total leukocytes)	B Cell	-4.87	-0.03	0.62	0.70	0.69	< 0.0001
	Granulocyte	3.85	0.15	1.02	3.01	2.98	0.20
	Monocyte	0.12	0.11	1.03	0.97	0.96	0.90
	NK	-0.63	-0.06	1.16	0.83	0.82	0.44
	T Cell (cd4+)	-0.30	-0.37	4.02	2.49	2.66	0.91
	T Cell (cd8+)	-1.89	0.35	4.03	2.47	2.42	0.43
T Cell	Intercept, γ_0	-0.97	0.07	1.7	1.4	1.6	0.54
(cases+controls)	B Cell	-0.51	0.02	1.2	1.2	1.2	0.67
	Granulocyte	-56.21	0.49	2.1	3.4	3.4	< 0.0001
	Monocyte	-5.13	-0.37	2.1	1.1	1.3	< 0.0001
	NK	0.07	0.34	2.3	1.5	1.7	0.97
	T Cell (cd4+)	60.18	-2.89	8.1	3.2	5.2	< 0.0001
	T Cell (cd8+)	3.00	2.34	8.2	3.3	5.4	0.58

Est = Regression coefficient estimate (× 100%).

$Bias_2$ = Double-bootstrap bias estimate (× 100%).

SE_0 = Naive standard error (× 100%).

SE_1 = Single-bootstrap standard error (× 100%).

SE_2 = Double-bootstrap standard error (× 100%).

P-values were computed using SE_2.

Example 9: Application of the methods herein to obesity in an African American population

The method herein was also applied to an obesity data set (Wang et al., 2010) consisting of seven lean African-Americans and seven Obese African-Americans (GEO Accession number GSE25301). Figure 8 shows a clustering heatmap displaying the DNA methylation data. In this analysis, **z** consisted of obesity status. Obese subjects had an estimated increase of 12 percentage points in granulocytes, bias-corrected 95% confidence interval (3:4%; 20%) and an estimated decrease of 4 percentage points in NK cells, bias-corrected 95% confidence interval (-7:7%;-0:9%) (Table 10). No significant differences were found for other blood cell types. The specific immunological differences estimated by the method herein are consistent with known immunological perturbations associated with type II diabetes (Lynch et al., 2009, Obesity, 17(3), 601–5; Anderson et al., 2011, Curr Opin Lipidol, 21(3), 172–7.).

Table 10. Estimated Regression Coefficients for Data Set concerning Obesity in African Americans

		Est	$Bias_2$	SE_0	SE_1	SE_2	P-value
Obese	Intercept, γ_0	0.96	-0.09	1.08	0.85	0.84	0.25
	B Cell	0.70	-0.03	0.78	1.16	1.14	0.54
	Granulocyte	12.25	0.51	1.30	4.27	4.27	0.0041
	Monocyte	-0.70	-0.01	1.31	1.57	1.54	0.65
	NK	-4.42	-0.13	1.46	1.75	1.73	0.011
	T Cell (cd4+)	-6.97	-0.29	5.11	6.27	5.49	0.20
	T Cell (cd8+)	-2.29	0.22	5.13	4.97	4.36	0.60

Est = Regression coefficient estimate (× 100%).
$Bias_2$ = Double-bootstrap bias estimate (× 100%).
SE_0 = Naive standard error (× 100%).
SE_1 = Single-bootstrap standard error (× 100%).
SE_2 = Double-bootstrap standard error (× 100%).
P-values were computed using SE_2.

Example 10: Additional analyses

In this example a special case was considered in which subject population was such that for this population $\mathbf{z} = 0$ and the population was sufficiently homogeneous with respect to blood cell distribution to admit sensible characterization of that distribution. In such case it is possible to recover estimates from $\hat{\Gamma}$. The results of such an analysis applied to the HNSCC case/control data set is shown in Table 11 below.

Table 11: White Blood Cell Distribution in HNSCC Controls

	Est	SE_2	$Bias_2$	BC-Est	95% Conf. Int.
B Cell	7.9	0.5	0.1	7.8	(6.8,8.9)
Granulocyte	42.2	1.2	-0.1	42.3	(39.9,44.6)
Monocyte	9.9	0.7	0.3	9.6	(8.3,10.9)
NK	7.9	0.7	0.2	7.7	(6.3,9.1)
T Cell (cd4+)	15.2	3.0	-0.1	15.3	(9.5,21.2)
T Cell (cd8+)	7.6	3.0	0.4	7.2	(1.4,13.0)

Est = Regression coefficient estimate (× 100%), normalized so that estimates sum to 90%.
SE_2 = Double-bootstrap standard error (× 100%).
$Bias_2$ = Double-bootstrap bias estimate (× 100%).
BC-Est = bias-corrected estimate.

If the coefficients represented a complete profiling of blood cell types, the estimates should sum approximately to one, even though the model does not explicitly constrain them so. In this case, the original bias corrected estimates (of leukocyte distribution in HNSCC controls) summed to 133%. The table shows the values re-normalized to 90%, the anticipated proportion of the cell types. The resulting estimated distribution of leukocytes is consistent with the literature (Alberts B et al., 2008, Molecular Biology of the cell. New York, NY: Taylor and Francis, 5th edition)

An additional analysis was also conducted in which S_0 consisted of only samples with pure CD4+ or CD8+ cells and S_1 to consisted only of samples having the less purified T-lymphocytes. For such S_1, there were no covariates, so **z** consisted only of an intercept. The following unnormalized bias-corrected estimates: 69.0% CD4+, 95% confidence interval (54%; 84%), and 32.5% CD8+, 95% confidence interval (19%; 46%). This is consistent with known proportions of these specific cell types among T lymphocytes.

Example 11: Sensitivity analysis

The bias estimates evident from the double-bootstrap procedure admit the possibility of correcting the bias arising from measurement error. There is no statistical procedure for correcting the other possible sources of bias, those arising from unprofiled cell types and non-cell-mediated profile differences, i.e. methylation difference signatures $\boldsymbol{\delta}$ with nonzero projection onto the space spanned by the WBC signatures. It is possible to conduct a sensitivity analysis using the theory presented under "Bias" (equations 6-9). It is shown that the magnitude of the bias is likely to be small, less than a percentage point.

Detailed analysis

A method of sensitivity analysis to estimate the magnitude of bias arising from unprofiled cell types and non-cell-mediated profile differences is described below for the HNSCC data set presented in Example 6 and Figure 4.

For each value of $k \in \mathbb{J}_m$, k elements are randomly sampled, $\mathbb{J}_k \subset \mathbb{J}_m$ without replacement, then k rows of $\mathbf{B}_1$ are sampled without replacement, $\boldsymbol{\delta}^*$ is set equal to the m x d_1 zero matrix, and the rows indicated by $\mathbb{J}_k$ are substituted by the k rows selected from $\mathbf{B}_1$. The matrix $\boldsymbol{\delta}^*$ served as a representative of the sum of processes having systematic methylation changes at k locations, of total magnitude consistent with the observed data (under the conservative assumption that *no* systematic methylation difference is cell mediated), and $\boldsymbol{\alpha}^* = (\mathbf{B}_0\ \mathbf{B}_0)^{-1}\mathbf{B}_0$ $\boldsymbol{\delta}^*$ represented the corresponding bias in $\boldsymbol{\Gamma}$. If, as in this situation, the goal was to assess the

sensitivity to bias in column of $\mathbf{B}_1$ (i.e. Case Status), the uninteresting columns of $\boldsymbol{\delta}^*$ or $\boldsymbol{\alpha}^*$could be simply deleted. Replicating this resampling procedure 100, 000 times, an approximation to the distribution of possible biases corresponding to processes involving exactly k CpG sites was generated. Figure 4 displays the results of such an analysis, showing the distribution of $(\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*)^{-1/2}$ for various values of k. It is noted that the relationship of median values to m was consistent with the theory presented in Example 12 under the subheading "Additional simulations." The median values of $(\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}^*)$ had an almost perfect linear relationship with m. The magnitude of the bias was small: for the more likely low values of k, the bias was 0.1 to 0.25 of a percentage point. In addition, this analysis was conservative in that it assumed all of the effect in $\mathbf{B}_1$ was due to non-cell-mediated processes, a strongly conservative assumption. In addition, for various choices of π_0 over a range of small magnitudes, the expected bias over the uniform posterior implied by π_0 was computed by iterated expectation, first by computing the mean bias for each choice of k, then forming the expectation over the binomial distribution $Bin(100, \pi_0)$, As noted in details described under "Bias" in Example 3 the result scaled linearly with π_0. The constant of proportionality was estimated to be 2.08 percentage points. In summary, if the prior expectation is of even moderate size (~0.1) that any one CpG among the 100 selected for this application will show systematic differentiation between cases and controls, then the implied bias would be expected to be less than a percentage point.

Example 12: Simulations

To verify the properties of the proposed methodology, extensive simulation studies were conducted. Simulation parameters were obtained from the HNSCC data set, and most simulations assumed no sources of biological bias (DNA methylation changes arising from processes not mediated by the profiled leukocytes, including shifts in distribution within cell types not profiled). In every simulation, S_0 was specified to consist of five B cell samples, ten granulocyte samples, five monocyte samples, 15 NK samples, five general T cell samples, eight specific CD4+ T cell samples, and two specific CD8+ T cell samples. Estimates from the external validation set S_0, described above, were used for mean methylation profiles among WBC types, using the $m = 100$ most informative CpG sites.

$n_1/2$ cases and $n_0/2$ controls, were specified, $no \in \{100, 200, 500\}$. Among the controls, methylation profiles were generated by a white blood cell population of 7% B cells, 62% granulocytes, 6% monocytes, 2% NK cells, and 13% were T cells, of which 65% were CD4+ cells and 35% were CD8+ cells, and the remaining 5% were unspecified (and assumed to have mean equal to the unsorted T-lymphocytes). Among cases, one of the following scenarios was

specified: a 4% reduction in CD4+ cells, a 2% reduction in CD8+ cells, and an 8% increase in granulocytes (alternative with changes in both CD4+ and CD8+, "Strong Alternative I"); a 6% reduction in CD4+ cells, and an 8% increase in granulocytes (alternative with changes in CD4+ and not CD8+, "Strong Alternative II"); a weaker alternative with half the effects of Strong Alternative I ("Mixed Alternative" elaborated upon below); and two null scenarios with no changes in cell population, each with a different assumption about $\boldsymbol{\delta}$. It is noted that these changes reflect absolute changes in percentage points, not relative changes. It is also noted that these values were actually used to generate Dirichlet-distributed mixture weights for each simulated subject, with Dirichlet parameters equal to a precision parameter (10 corresponding to "noisy", and 100 corresponding to "precise") times the mean weight described above.

Residual effects $\xi_i^{(0)}$ for controls were set equal to 0.1 times estimated intercept μ_1 and residual effects $\xi_i^{(1)}$ for cases were set equal to 0.08 or 0.09 times μ_1 plus multiples 10θ of the column of $\mathbf{U}$ corresponding to case. The constants of proportionality 0.1, 0.08, and 0.09 were chosen to correspond to assumed contributions of ξ to an overall methylation signature presumed to be dominated by profiled populations of white blood cells in specified proportions, with 0.08 used for the strong alternatives and 0.09 used for the Mixed Alternative. The constant 10 was used to amplify the scale of $\boldsymbol{\delta}$ so that its effect could be detected in simulation; it is noted that $\mathbf{U}$ was orthogonal to the white blood cell profiles, by construction.

It is noted also that the individual, Dirichlet-generated subject weights did not necessarily sum to one, and the difference from 1 was not applied as a multiplier; thus the resulting ξ corresponded to the situation $\mathbf{P}\boldsymbol{\mu}_q = \mathbf{0}$, where $P = (\mathbf{B}_0\ \mathbf{B}_0)^{-1}\ \mathbf{B}_0$ along with orthogonal contributions from the λ terms of (6). The multiplier $\theta = 0$ was used for strong alternatives, and the "Strong Null" case (i.e. no methylation differences between cases and controls) and $\theta = 0.5$ was used for the Mixed Alternative, and $\theta = 1$ was used for the "Mixed Null" with case/control differences not mediated by cellular population differences.

A simple normal error structure for e_{oh} and e_{oi} was specified, with no chip effects, and with variance equal to the sum of chip and residual variance estimated (individually for each CpG) for the HNSCC data. For each simulation, 50 bootstraps were used to estimate standard errors. 1000 simulations were run for each scenario. Table 12 presents results for $n_1 = 200$ with precise mixture weights (small within-status heterogeneity in distribution), and Table 13 presents results for $n_1 = 200$ with noisy mixture weights (larger within-status heterogeneity). The tables show mean estimate, simulation standard deviation, median estimates for the three types of proposed standard errors, and proportion of p-values (obtained from z-scores constructed using the double-bootstrap standard error) falling below $\alpha = 0.05$ and $\alpha = 0.01$.

In all cases, the bias in estimation was minimal. Both types of bootstrap produced similar standard error estimates, which were close to the simulation standard deviation and often quite different from the naive standard error estimate. Under null scenarios, the rejection probabilities were tolerably close to their nominal values, and for alternatives, power could be quite high, even with this modest design.

Results for Coefficients of Determination

Results for the coefficients of determination are provided in Table 14. $R^2_{1,0}$ decreased with decreasing strength of the alternative, falling to zero under both null scenarios. For strong alternatives, $R^2_{1,1}$ was frequently close to 1.0. For the Mixed Alternative, $R^2_{I,I}$ had a lower, and still high values ranging from about 0.85 to 0.90. For the mixed null result, $R^2_{1,1}$ typically had lower values, from about 0.05 to 0.20. In the Strong Null case, $R^2_{1,1}$ covered a broader range among moderately low values; note, however, that this scenario effectively represents 0/0, i.e. a poorly defined value. Scenarios with $n_1 \in \{100, 500\}$ produced similar results, with simulation standard deviations and power adjusted accordingly, and still having practical utility.

Additional Simulations

Additional simulations, were conducted which assumed bias arising from processes not profiled by the profiled leukocytes. For these scenarios, ξ^0 was set to $\hat{\mu}_1$, and $\xi^1=\xi^0$ except for a set of CpG sites randomly selected among the m dimensions of the array (once and for all before all 1000 simulations); among those dimensions j, ξ^1_j was set to $1-\hat{\mu}_{1j}$, reflecting a \reversal" of methylation state. Estimates were biased towards the null, on the order of about a percentage point.

Table 12. Simulation results (precise mixtures, $n_1 = 200$)

Strong Alternative I ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.07	1.00	0.92	0.97	0.98	0.057	0.018
Granulocyte	8.0	8.02	0.73	0.39	0.73	0.73	1.000	1.000
Monocyte	0.0	0.01	0.48	0.43	0.47	0.47	0.055	0.013
NK	0.0	-0.09	1.08	1.02	1.02	1.05	0.066	0.015
T Cell (cd4+)	-4.0	-4.06	0.81	0.80	0.78	0.81	0.999	0.989
T Cell (cd8+)	-2.0	-1.93	0.83	0.81	0.78	0.81	0.653	0.419

Strong Alternative II ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.00	0.97	0.92	0.97	0.99	0.048	0.016
Granulocyte	8.0	8.00	0.71	0.39	0.72	0.72	1.000	1.000
Monocyte	0.0	0.03	0.48	0.42	0.47	0.47	0.063	0.016
NK	0.0	0.03	1.04	1.02	1.01	1.05	0.052	0.014
T Cell (cd4+)	-6.0	-5.83	0.76	0.80	0.77	0.80	1.000	1.000
T Cell (cd8+)	0.0	-0.22	0.81	0.81	0.80	0.81	0.064	0.014

Mixed Alternative ($\theta = 0.5$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	-0.02	1.02	1.10	0.96	0.98	0.065	0.011
Granulocyte	4.0	3.99	0.75	0.47	0.73	0.73	1.000	0.995
Monocyte	0.0	0.02	0.49	0.51	0.47	0.47	0.060	0.015
NK	0.0	0.04	1.05	1.22	1.01	1.04	0.054	0.009
T Cell (cd4+)	-2.0	-2.07	0.82	0.96	0.79	0.83	0.695	0.471
T Cell (cd8+)	-1.0	-0.95	0.82	0.96	0.78	0.82	0.203	0.082

Mixed Null ($\theta = 1$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.00	1.04	1.58	0.96	1.02	0.066	0.017
Granulocyte	0.0	0.03	0.73	0.67	0.74	0.74	0.055	0.014
Monocyte	0.0	-0.01	0.47	0.73	0.47	0.48	0.054	0.013
NK	0.0	-0.01	1.12	1.76	1.01	1.09	0.063	0.014
T Cell (cd4+)	0.0	0.01	0.87	1.38	0.80	0.90	0.054	0.013
T Cell (cd8+)	0.0	-0.02	0.88	1.39	0.79	0.89	0.057	0.015

Strong Null ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	-0.01	0.99	0.90	0.96	0.96	0.068	0.014
Granulocyte	0.0	0.03	0.72	0.38	0.74	0.73	0.052	0.013
Monocyte	0.0	-0.01	0.47	0.42	0.47	0.47	0.055	0.013
NK	0.0	-0.01	1.06	1.00	1.01	1.02	0.059	0.020
T Cell (cd4+)	0.0	0.00	0.81	0.78	0.80	0.82	0.054	0.013
T Cell (cd8+)	0.0	-0.01	0.81	0.79	0.79	0.80	0.054	0.015

Est = Mean regression coefficient estimate (× 100%); SD = SD regression coefficient estimate (× 100%).

SE_0 = Naive standard error (× 100%); SE_1 = Single-bootstrap standard error (× 100%).

SE_2 = Double-bootstrap standard error (× 100%).

pow(α) = $Pr\{P_2 < \alpha\}$, where P_2 is the p-value computed from SE_2.

Table 13. Simulation Results (Noisy Mixtures, $n_1 = 200$)

Strong Alternative I ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	-0.06	1.39	0.92	1.36	1.34	0.065	0.019
Granulocyte	8.0	7.87	2.02	0.39	2.00	1.99	0.974	0.897
Monocyte	0.0	0.05	1.03	0.42	1.04	1.02	0.049	0.012
NK	0.0	-0.02	1.21	1.02	1.16	1.18	0.061	0.010
T Cell (cd4+)	-4.0	-4.00	1.23	0.79	1.21	1.22	0.903	0.739
T Cell (cd8+)	-2.0	-1.97	1.05	0.80	1.02	0.98	0.517	0.298

Strong Alternative II ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	-0.08	1.38	0.92	1.36	1.34	0.063	0.017
Granulocyte	8.0	7.90	2.03	0.39	1.99	1.98	0.973	0.905
Monocyte	0.0	0.10	1.07	0.42	1.04	1.02	0.054	0.019
NK	0.0	0.02	1.17	1.02	1.14	1.18	0.053	0.009
T Cell (cd4+)	-6.0	-5.70	1.19	0.80	1.13	1.16	0.999	0.986
T Cell (cd8+)	0.0	-0.23	1.08	0.81	1.10	1.04	0.066	0.015

Mixed Alternative ($\theta = 0.5$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.05	1.42	1.10	1.34	1.34	0.066	0.016
Granulocyte	4.0	4.00	2.01	0.47	2.02	2.01	0.500	0.291
Monocyte	0.0	0.01	1.06	0.51	1.03	1.02	0.072	0.020
NK	0.0	-0.02	1.24	1.22	1.13	1.16	0.064	0.013
T Cell (cd4+)	-2.0	-2.11	1.30	0.95	1.26	1.28	0.391	0.191
T Cell (cd8+)	-1.0	-0.94	1.08	0.96	1.05	1.02	0.163	0.052

Mixed Null ($\theta = 1$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.06	1.41	1.59	1.36	1.37	0.062	0.016
Granulocyte	0.0	0.04	2.08	0.67	2.06	2.05	0.056	0.008
Monocyte	0.0	-0.02	1.05	0.73	1.03	1.03	0.058	0.020
NK	0.0	0.01	1.26	1.76	1.14	1.22	0.066	0.011
T Cell (cd4+)	0.0	-0.01	1.42	1.38	1.31	1.36	0.067	0.016
T Cell (cd8+)	0.0	0.00	1.19	1.39	1.08	1.10	0.073	0.011

Strong Null ($\theta = 0$)

	Truth	Est	SD	SE_0	SE_1	SE_2	pow(0.05)	pow(0.01)
B Cell	0.0	0.06	1.37	0.91	1.36	1.32	0.065	0.017
Granulocyte	0.0	0.03	2.07	0.38	2.06	2.05	0.055	0.009
Monocyte	0.0	-0.02	1.04	0.42	1.03	1.02	0.057	0.021
NK	0.0	0.01	1.19	1.01	1.14	1.16	0.053	0.018
T Cell (cd4+)	0.0	-0.04	1.38	0.79	1.31	1.31	0.069	0.015
T Cell (cd8+)	0.0	0.01	1.11	0.79	1.08	1.03	0.065	0.016

Est = Mean regression coefficient estimate (× 100%); SD = SD regression coefficient estimate (× 100%).

SE_0 = Naive standard error (× 100%); SE_1 = Single-bootstrap standard error (× 100%).

SE_2 = Double-bootstrap standard error (× 100%).

$\text{pow}(\alpha) = Pr\{P_2 < \alpha\}$, where P_2 is the p-value computed from SE_2.

Table 14. Results for coefficients of determination

		Median $R^2_{1,0}$	Median $R^2_{1,1}$
		(Interquartile Range)	(Interquartile Range)
Precise Mixtures	Strong Alternative I ($\theta = 0$)	0.13 (0.12-0.15)	0.98 (0.97-0.98)
$n_1 = 200$	Strong Alternative II ($\theta = 0$)	0.13 (0.12-0.15)	0.98 (0.97-0.98)
	Mixed Alternative ($\theta = 0.5$)	0.04 (0.03-0.05)	0.88 (0.85-0.91)
	Mixed Null ($\theta = 1$)	0.00 (0.00-0.00)	0.10 (0.05-0.17)
	Strong Null ($\theta = 0$)	0.00 (0.00-0.00)	0.25 (0.15-0.38)
Noisy Mixtures	Strong Alternative I ($\theta = 0$)	0.05 (0.03-0.06)	0.98 (0.97-0.98)
$n_1 = 200$	Strong Alternative II ($\theta = 0$)	0.05 (0.03-0.06)	0.98 (0.97-0.98)
	Mixed Alternative ($\theta = 0.5$)	0.01 (0.01-0.02)	0.89 (0.81-0.94)
	Mixed Null ($\theta = 1$)	0.00 (0.00-0.01)	0.46 (0.28-0.64)
	Strong Null ($\theta = 0$)	0.00 (0.00-0.01)	0.72 (0.55-0.85)

Example 13: Identification of a unique DMR in CD3Z gene

Individual samples of sorted, normal, human, peripheral blood leukocytes as shown in Table 15, were purchased from AllCells®, LLC (Emeryville, CA). These leukocytes were sorted in a column with antibody-conjugated magnetic beads using a combination of positive and negative selection. Genomic DNA from the leukocytes was extracted according to manufacturer's protocol using the DNeasy Blood & Tissue kit (Qiagen) or the AllPrep DNA/RNA/Protein Mini Kit according to manufacturer's protocol (Cat. No. 8004, QIAGEN, Valencia, CA), then quantified by NanoDrop ND-1000 Spectrophotometer (NanoDrop

Technologies, Inc. , Wilmington, DE) and stored at -20 °C. The extracted genomic DNA was subjected to Bisulfite conversion by treatment with sodium bisulfite using the EZ DNA Methylation Kit (Zymo) following the manufacturer's protocol, thereby converting unmethylated cytosine residues to uracil and leaving methylated cytosine residues intact.

Table 15: Sorted leukocytes from AllCells®, LLC

Cell Lineage	Abbreviation	N
CD3+ T Lymphocytes	Pan-T	5
CD3+CD4+ T Lymphocytes	CD4	2
CD3+CD4+CD25+ Regulatory T Lymphocytes	Treg	6
CD3+CD8+ T Lymphocytes	CD8	2
CD56+ Natural Killer Cells (Large Granular Lymphocytes)	NK	3
CD19+ B Lymphocytes	B	5
CD14+ Monocytes	Mono	4
CD15+ Granulocytes	Gran	5
CD16+ Neutrophils	Neut	4

Analysis of the methylation status of the bisulfate converted DNA was performed using DNA methylation microarray, Infinium® HumanMethylation27 Beadchip Microarray, (Illumina®, Inc. ,San Diego, CA). This microarray quantifies the methylation status of 27,578 CpG loci from 14,495 genes, with a redundancy of 15-18-fold. Bisulfite converted, genomic DNA from sorted human peripheral blood leukocytes was subjected to whole genome amplification. The purified whole genome amplified DNA was hybridized to locus-specific DNA oligomers linked to individual bead types corresponding to each CpG locus, unmethylated or methylated. Allele-specific primer annealing was followed by specific single-base extension using labeled ddNTPs. Extension only occurs if the bead type matches the methylation status of the genomic DNA.

The array was fluorescently stained, scanned, and fluorescent intensities of each of the unmethylated and methylated bead types were measured. The ratio of fluorescent signals is computed from both alleles using the following equation: $\beta=(max(M,0))/(|U| + |M|) + 100$. The β-value is a continuous variable ranging from 0 (unmethylated) to 1 (completely methylated) that represents the methylation at each CpG site and is used in subsequent statistical analyses. Data were assembled with BeadStudio methylation software from Illumina, Inc. (San Diego, CA). Bibikova, M., *et al.*, *Epigenomics* **1**, 177-200 (2009).

A comparison of methylation in sorted normal human immune cells was observed to produce distinct profiles of methylation markers for further consideration. As shown in Figure 9 DNA Methylation profiles distinguished lymphocytes from myeloid derived leukocytes. Recursively partitioned mixture model (RPMM) of autosomal gene Infinium beta values from sorted, human, peripheral blood leukocytes was performed in R version 2.11.1 of Illumina's software which provides convenient mechanisms for loading and analyzing the results of methylation status, and for quality control and basic visualization tasks.

Candidate DNA regions with high potential to discriminate CD3+ T cells from non-T cells were chosen based on the criteria of being differentially demethylated and differentially overexpressed in CD3+ T cells compared with other cell types (monocytes, granulocytes, NK cells, and B cells). Two quantitative methylation methods, bisulfite pyrosequencing and MS-qPCR, were used to confirm array methylation.

The highest ranking 5000 most variable CpG loci were plotted on the left (Figure 9 left panel), such that the less methylated loci appear as grey and more methylated loci appear as black. The number of individual leukocyte samples in each methylation class is shown in Figure 9 in the table to the right. The algorithm for prioritizing these candidates described herein yielded CD3E and CD3Z as specific DMR for identifying CD3+ T cells.

Example 14: Patient characteristics and biological samples for determining CD3+ T cell distribution in glioma cases and controls

Whole blood samples from glioma patients (N=94) and controls (N=71) were obtained from the UCSF San Francisco Adult Glioma Study (AGS) for these examples (Table 16). The patients included in this example were diagnosed between 1997 and 2011. Details of subject ascertainment through the rapid case ascertainment program of San Francisco regional population-based registry or the UCSF Neuro-oncology Clinic have been described (Wrensch M et al., 2007, Clin Cancer Res 13(1): 197-205; Felini MJ et al. 2009, Cancer Causes Control 20(1): 87-96; Wrensch M et al., 2009, Nat Genet 41(8): 905-8; Christensen BC et al., 2011, J Natl Cancer Inst 103(2): 143-53). Pertinent data for this analysis included age at histological diagnosis, gender, vital status, and survival time between diagnosis date and date of death for those deceased or between diagnosis date and date of last contact for those alive, and any of cigarette smoking history and exposure to steroids, chemotherapy and radiation therapy.

A panel of 120 fresh frozen glioma tumors from the UCSF Brain Tumor Research Center tissue bank, obtained under appropriate institutional review board approval, which were previously characterized for molecular features (Christensen BC et al., 2011, J Natl Cancer Inst 103(2): 143-53; Zheng S et al., 2011, Neuro Oncol 13(3): 280-9) was chosen for tumor MS-

qPCR and IHC studies (Table 16). Tumor samples were defined as secondary GBM if the patients had prior histological diagnosis of a low-grade glioma. All ages are given at the time of surgery, which occurred at UCSF between 1990 and 2003. This tumor set contained the following histological subtypes: 2 pilocytic astrocytoma (PA), 15 ependymoma grade II (EPII), 20 oligodendroglioma grade II (ODII), 16 oligoastroglioma grade II (OAII), 3 oligoastroglioma grade III (OAIII), 23 astrocytoma grade II (ASII), 4 astrocytoma grade III (ASIII) and 37 astrocytoma grade IV, also called glioblastoma multiforme grade IV (GBM), ten of which were recurrent and five of which were secondary.

Sorted, normal, human, peripheral blood leukocyte subtypes were isolated from different non-diseased individuals' whole blood by MACS using a combination of negative and positive selection with highly specific cell surface antibodies conjugated to magnetic beads. The purity of separated cells was determined with flow cytometry to be >97%.

Example 15: Bisulfite pyrosequencing and MS-qPCR assays for validating CD3Z, CD3E and FOXP3 specific DMRs

The demographic characteristics of donors for all samples (N = 285) used in MS-qPCR analysis is as shown in Table 16. CpGenome Universal Methylated DNA (Cat. No. S7821, Millipore Corp., Temecula, CA), purified T cell and Treg DNA were bisulfite converted at the same time. All bisulfite pyrosequencing assays were designed using Pyromark Assay Design 2.0 (QIAGEN), and carried out using a Pyromark MD pyrosequencer running Pyromark qCpG software (QIAGEN). Custom oligonucleotide primers used in bisulfite pyrosequencing were obtained from Invitrogen (Life Technologies Co, Carlsbad CA). For MS-qPCR reactions, primers and TaqMan major groove binding (MGB) probes with 5' 6FAM and 3' non-fluorescent quencher (NFQ) as well as TaqMan 1000 RXN Gold with Buffer A Pack were obtained from Applied Biosystems (Part No. 4304971, 4316034 and 4304441, Applied Biosystems, Foster City, CA). The primer and probe sequences are shown in Table 17 and Figure 12. Solutions for MS-qPCR: 10X TaqMan Stabilizer containing 0.1% Tween-20, 0.5% gelatin were prepared weekly. Each reaction of 20 µl contained 5 µl DNA, 11.9 µl PreMix, 3 µl OligoMix, and 0.1 µl Taq DNA polymerase. Cycling was performed using a 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, CA); 50 cycles at 95 °C for 15 sec and 60 °C for 1 min after 10 min at 95 °C preheat. All samples were run in triplicate using the absolute quantification method. Copy number of the target locus in each sample was determined by reference to a four-point standard curve, which was based on known copies of bisulfite converted template.

Table 16. Demographic characteristics of donors for all samples (N = 285) used in MS-qPCR analysis

Characteristic	Control Blood samples (n = 94)	Case Blood samples (n = 71)	Excised Tumors (n = 120)
Age			
Median (range)	57 (22-87)	57 (20-86)	41 (1-78)
Mean (standard deviation)	55 (16.5)	56 (13)	41 (15)
Gender, No (%)			
Female	43 (46%)	26 (36%)	42 (35%)
Male	51 (54%)	45 (64%)	78 (65%)
Race, No (%)			
White, Non-Hispanic	78 (83%)	67 (95%)	102 (85%)
Hispanic	3 (3%)	3 (4%)	7 (6%)
Asian	6 (7%)	0 (0%)	4 (3%)
Black	5 (6%)	0 (0%)	0 (0%)
Other	1 (1%)	1 (1%)	7 (6%)

Quantification of total bisulfite converted DNA copies for all standard and biological samples was determined by reference to the C-less qPCR assay as described previously (Weisenberger DJ et al., 2008, Nucleic Acids Res 36(14): 4689-98.; Campan M et al., 2009, Methods Mol Biol 507: 325-37). In this procedure one determines the relative amounts of a bisulfite converted sample through the use of a TaqMan PCR reaction using primers and probes that recognize a DNA strand that does not contain cytosines, and hence is able to amplify the total amount of DNA (bisulfite-converted or unconverted) in a PCR reaction well. The absolute copy number in DNA Standard Solution (Cambio Ltd. Cambridge, UK) was used to calibrate the C-less reaction and assuming 3.3 pg = 1 genome copy. Universal methylated DNA and purified CD3+ T cell and Treg DNA (all bisulfite converted) were quantified at the same time. Since C-less primers hybridize to both strands of the standard DNA (non-bisulfite converted) and bisulfite converted samples allow for only single strand hybridization during the first cycle, the resultant copy number in bisulfite samples is multiplied by two. After C-less assay, the copy number of the different standards: universal methylated, CD3+ T cell and Treg DNA was used to create standard curves for CD3Z and FOXP3. To create a calibration curve known quantities of CD3+ T cell or Treg DNA were spiked into universal methylated DNA in ratios that maintained a constant total copy number in each reaction across the dilution scheme. The latter procedure mimics the conditions of detection that exist in differentiating different relative numbers of CD3+ T cells and Tregs within a mixture of cells in a complex biological sample. For absolute quantification of CD3Z, the four-point standard curve used 10,000, 1,000, 100, and

10 bisulfite converted CD3+ T cell DNA copies; absolute quantification of FOXP3 used, 5,000, 500, 50 and 5 bisulfite converted Treg cell DNA copies.

Table 17. Primer and probe sequences for MS-qPCR assays

Oligonucleotide Name	Sequence (5' to 3')
C-less Fwd	TTGTATGTATGTGAGTGTGGGAGAGA (SEQ ID NO: 97)
C-less Rev	TTTCTTCCACCCCTTCTCTTCC (SEQ ID NO: 98)
C-less Probe	(6FAM) CTCCCCCTCTAACTCTAT (MGB,NFQ) (SEQ ID NO: 99)
CD3Z Fwd	GGATGGTTGTGGTGAAAAGTG (SEQ ID NO: 100)
CD3Z Rev	CAAAAACTCCTTTTCTCCTAACCA (SEQ ID NO: 101)
CD3Z Probe	(6FAM) CCAACCACCACTACCTCAA (MGB,NFQ) (SEQ ID NO: 102)
FOXP3 Fwd	GGGTTTTGTTGTTATAGTTTTTG (SEQ ID NO: 103)
FOXP3 Rev	TTCTCTTCCTCCATAATATCA (SEQ ID NO: 104)
FOXP3 Probe	(6FAM) CAACACATCCAACCACCAT (MGB,NFQ) (SEQ ID NO: 105)

MGB: major groove binding
FAM: 6-Carboxyfluorescein
NGQ: NFQ
C-less qPCR assay: Campan M et al., 2009, Methods Mol Biol, 507:325-37; Weisenberger DJ et al., 2008, Nucleic Acids Res 2008; 36:4689-98

The CD3E specific DMR DNA methylation status of the DMR in CD3E gene was measured by pyrosequencing bisulfite converted DNA from sorted, human, peripheral blood leukocytes. Figure10 panel A. The CD3Z specific DMR, DNA methylation status of the DMR in CD3Z gene was measured by MethyLight® qPCR. of converted DNA from sorted, human, peripheral blood leukocytes (Figure 10 panel B). The genomic region containing the CD3Z DMR is shown in Figure 11.

Standard calibration curves were used to determine if the newly identified CD3Z DMR was useful to quantify CD3+ T cells, Tregs (FOXP3 demethylated) and ratios of Tregs/CD3+ T cells in biological specimens such as whole or separated blood or other tissues. To obtain these curves quantitative real time methylation specific PCR was performed. DNA isolated from purified cell types was bisulfite converted and serially diluted into a background of fully methylated commercial DNA standard (Qiagen). This method is referred to herein as "CS-DM assay" or assays.

It was observed that the total genomic copy numbers of each sample within a dilution series remained constant. Log dilutions were prepared to include the appropriate range of Ct values corresponding to test samples (whole blood, tumor specimens). Using cytosine less: C-less primers genome copy numbers for each test standard were measured to ensure adequate

input DNA and to normalize the CD3+ and Treg assay values. The calibration curve for C-less total input is shown in Figure 13 panel A (N=8 replicates); errors denote standard error of the mean Ct value. Figure 13 panel B shows dilution of isolated normal PanT cells (N=7 replicates) and Figure 13 panel C shows dilution and calibration curve for isolated CD3+CD25+ T cells (N=8 replicates). For samples to be tested these calibration curves (Figure 13 panels A-C) were used to estimate total input copies, CD3+ T cell, and Tregs copies, respectively.

The results show that the DNA methylation status of this region identified herein in the promoter of *CD3Z* gene in sorted human peripheral blood leukocytes, which was validated as an immune cell type specific differentially methylated region (Figure 10 panel B) was observed to be useful to quantify CD3+ T cells in biological specimens such as whole or separated blood, or other tissues.

Example 16: Flow cytometry of blood lymphocytes in whole blood for quantification of CD3+ T cells

Levels of CD3+ T cells in whole blood were quantified by flow cytometry for comparison with CD3+ T cell levels determined using *CD3Z* Ms-qPCR assay. Venous whole blood samples were collected in citrate EDTA and processed using a lysis no wash protocol (Invitrogen, Carlsbad, CA cat# GAS-010). Cells were labeled by direct staining with the appropriate fluorochrome-conjugated antibodies (eBioscience Inc, San Diego, CA), and were incubated for 20 minutes in the dark at 4 °C; CD3-fluorescein isothiocyanate (FITC, cat #11-0038-41), anti-CD4-allophycocyanin (APC, cat #17-0048-41), anti-CD8-phycoerythrin (PE, cat #12-0086-41), and anti-CD45-PerCP-Cy5.5 (cat #45-0459-41). Isotype control mAbs were used as negative controls. Accucheck counting beads (Invitrogen, Carlsbad CA cat # PCB100) were used for quantifying leukocyte numbers. Acquisition was preformed within 48 hrs of blood draw on a FACScalibur flow cytometer using Cell-Quest Software (Becton Dickinson, Franklin Lakes, NJ). For CD3+ cells a minimum of 10,000 events were collected on the lymphocyte gate that was set on the forward scatter vs. side scatter (FSC vs. SSC) and then gated on CD3+ cells. CD45+ counts were obtained by first gating on all non-bead events using the FSC vs. SSC. A CD45+ histogram plot of the non-bead events was then created, CD45+ cells were gated. Examples are seen in Figure 18. Absolute counts (number cells per µl) were obtained by taking the number of cells counted, divided by total number of beads counted, multiplied by the known concentration of beads. Flowjo software (TreeStar Inc, Ashland, OR) was used for data analysis.

Example 17: Tumor immunohistochemistry (IHC) for measuring levels of tumor infiltrating lymphocytes (TIL) in glioma tumors

Slides were prepared from a 5 micron slice of each FFPE tumor block. Slides were stained using a Benchmark XT instrument per manufacturer's instructions (Ventana, Tucson, AZ). CD3 antibody (Dako, Carpinteria, CA cat # A0452) was added in a 1:600 dilution, and incubated for 30 minutes. CD8 antibody (Dako, Carpinteria, CA cat # M7103) was added in a 1:200 dilution and incubated for 60 minutes. CD4 antibody (Leica Microsystems, Buffalo Grove IL, cat # NCL-L-CD4-368) was added in a 1:50 dilution, and incubated for 2 hours. Slides were counterstained with hematoxylin. Each slide was scanned at a magnification of 10X to identify four suitable fields that were then scored at 25X magnification. Examples are seen in Figure 19 panels A-C. The numbers of positive staining cells were recorded and the average count per four fields calculated. Photomicrographs was taken and scored for specimens with very high cell counts to increase accuracy. Samples were also examined to see if they contained predominantly perivascular and/or parenchymal infiltrates. A blind comparison of observation by two individuals was carried out to ensure uniform interpretation. Data from tumor IHC were analyzed in combination with *CD3Z* MS-qPCR data to determine association between the two data sets. (see Example 19)

Example 18: Statistical analysis of differential methylation in CD3+ T cells for identification of cell-specific DMRs

To identify putative cell specific DMRs, MACS sorted leukocyte DNA methyation data consisting of un-normalized average beta values from the Illumina HumanMethyation27 microrray were calculated from probe intensities using Illumina GenomeStudio. Locus by locus comparisons of DNA methyation between the sorted cell types were performed using a linear mixed effects model (controlling for beadchip) in SAS version 9.2, thereby generating estimates and p-values for differential methyation in CD3+ T cells compared to other cell types. Resultant p-values were adjusted for multiple comparisons using the qValue package in the software program R project for statistical computing, version 2.13 available for downloading from the internet, and q-values of less than 0.05 were considered significant. All correlations, F-tests, Wicoxon rank sum and Kruskal-Wallis one-way analysis of variance by ranks tests were carried out in R version 2.11.1 and survival analysis was performing using the survival pack in R version 2.11.1.

Example 19: Discovery and validation of CD3Z demethylation as a marker of CD3+ T cells

The search for genes containing DMRs specific for CD3+ T cells using methods herein revealed candidate CpG sites within the genes encoding several components of the T cell receptor (TCR) complex; namely, CD3D, CD3E, CD3G, and CD3Z. Myeloid derived blood cells (granulocytes, neutrophils, monocytes) and B-lymphocytes contained methylated CpG sites within CD3D, CD3E, CD3G and CD3Z loci compared with T cells, which were demethylated. CD3Z was also unmethylated in CD16+ NK cells, but was methylated in CD16- NK cells. The promoter regions of the CD3D, CD3E and CD3G genes are CpG sparse compared with CD3Z, which contains a CpG island that is optimally suited for designing MS-qPCR assays (Fig. 1 panel A). For these reasons the CD3Z locus was analyzed for the development of a CD3+ T cell epigenetic marker. CD3Z is significantly overexpressed ($p = 0.0001$; Palmer, Diehn et al. 2006) and demethylated ($q = 0.00026$) in CD3+ T cells compared with non-T cells. Pyrosequencing of CD3Z showed the extent of differences in demethylation among immune cell lineages, which approaches complete demethylation in CD3+ T cells and nearly complete methylation in other cell lineages (Figure 20 panels A-B).

Bisulfite converted universal methylated DNA and DNA from purified CD3+ Tcells were used to prepare a four point calibration curve to estimate CD3+ T cell numbers in mixtures of cells (Figure 14 panel B). Total amount of DNA was held constant at all four points. Log Linear PCR kinetics were demonstrated over a range of CD3+ T cell DNA inputs corresponding to 10 to 100000 genomic copies, indicating that the MS-qPCR assay was able to detect a few demethylated cells within a background of many thousands of methylated cells.

Whole blood samples from 46 healthy controls and 20 patients with glioma were then used to compare flow cytometry quantification of CD3+ T cells with the CD3Z MS-qPCR assay (Figure 14 panel C). The MS-qPCR measurements were observed to correlate highly with conventional flow measurement of T cells as a fraction of total blood leukocytes (Pearson $R = 0.93$; F test $p < 2.2x10\text{-}16$). The uniform regression and close correspondence of the two methods was true for both glioma patients (labeled "cases") and the healthy controls. These data show that the disease process itself and treatment exposures did not influence the demethylation assay.

The correlation of CD3+ T cells detected by IHC and MS-qPCR was assessed in a set of FFPE samples; the results indicated a significant association of IHC score with CD3Z demethylation (Pearson $R = 0.85$; F test $p = 3.4x10^{-11}$; Figure 14 panel D). Most CD3+ TILs were CD8+ and only a few stained positively for CD4+ (Figure 19). Glioma cell lines (A172, T98G) were also studied; both expressed Foxp3 copy numbers < 0.06% of total input. Analysis of two autopsy brain specimens revealed Foxp3 copy numbers <0.04% of total input. These values show limits of detection of the assay which were observed to be much lower than values

observed in patient blood or tumor samples. These results demonstrate the specificity of the CD3Z epigenetic assay for detecting CD3+ immune cells within a background of tumor cells.

Example 20: Determination of T cells and Tregs levels in peripheral blood by *CD3Z* and *FOXP3* MS-qPCR assays in glioma cases and controls

The utility of the epigenetic assays using archived frozen blood specimen samples was tested by performing a case control analysis of CD3Z and FOXP3 demethylation in glioma patients and control subjects to measure CD3+ T cell and Treg levels, respectively, in stored peripheral blood specimens from the University of San Francisco Adult Glioma Study (AGS). Results of MS-qPCR assays are summarized in Table 18. The total inputs of DNA from whole blood from the 94 controls and 71 glioma cases were not significantly different from each other. In patients with grade IV glioblastoma multiforme (GBM), peripheral blood CD3+ T cell levels were observed to be significantly lower (Wilcoxon p = 1.7x10-9; Figure 15 A), peripheral blood Treg levels were observed to be significantly lower (Wilcoxon p=5.2x10-11; Figure 15 B) and peripheral blood Treg/ CD3+ T cell ratios were observed to be moderately lower (Wilcoxon p = 0.024; Figure 15 C) compared to healthy controls. In glioma patients and controls subjects, levels of T cells and Tregs were positively correlated (Pearson R = 0.61, F test p < 2.2x10-16). Use of dexamethasone or chemotherapy was not associated with T cell measures. The GBM case patients received steroid treatments prior to blood sampling. In healthy controls, but not glioma patients, people who had smoked were observed to have higher peripheral blood CD3+ T cell levels than those who had never smoked (Wilcoxon p = 0.08, Figure 16 panel A) and current smokers had significantly higher levels of peripheral blood Tregs than former smokers (Wilcoxon p = 0.01) and never smokers (Wilcoxon p = 0.002; Figure 16 panel B). Furthermore, the ratio of Tregs / CD3+ T cells was significantly elevated in the peripheral blood of current smokers compared to former smokers (Wilcoxon p = 0.01) and never smokers (Wilcoxon p = 0.03) among healthy controls, and trended towards elevated levels in current smokers compared to former smokers (Wilcoxon p = 0.17) and never smokers (Wilcoxon p = 0.14; Figure 16 panel C).

Table 18. Summary of MS-qPCR measurements for all samples (N = 285)

Sample Description	**Percent Demethylation**, Median (Range)		
	CD3Z	*FOXP3*	*FOXP3/CD3Z*
Blood samples (n = 165)	17.6 (2.1-44.4)	0.8 (0.06-3.2)	4.5 (0.9-20.2)
Controls (n = 94)	21.7 (4.7-44.4)	1.0 (0.2-3.2)	4.8 (1.0-20.2)
Never Smokers (n = 44)	19.3 (4.7-32.1)	1.0 (0.2-2.5)	4.8 (1.0-11.7)
Former Smokers (n = 42)	22.4 (8.8-43.4)	1.1 (0.2-2.2)	4.4 (1.8-10.5)
Current Smokers (n = 8)	23.4 (5.7-44.4)	1.6 (0.8-3.2)	7.4 (3.6-20.2)
Glioma Cases (n = 71)	11.2 (2.1-37.7)	0.5 (0.06-2.5)	4.1 (0.9-14.8)

Never Smokers (n = 31)	11.3 (2.7-37.7)	0.5 (0.06-2.5)	3.8 (1.3-11.5)
Former Smokers (n = 29)	12.7 (3.3-32.8)	0.5 (0.06-1.7)	4.1 (0.9-12.8)
Current Smokers (n = 11)	9.6 (2.1-27.8)	0.5 (0.1-1.2)	5.1 (2.3-14.8)
Non-GBM (n = 6)	18.5 (3.5-26.6)	0.9 (0.2-1.6)	6.0 (3.8-7.1)
GBM (n = 65)	10.5 (2.1-37.7)	0.5 (0.06-2.5)	4.1 (0.9-14.8)
Excised Tumors (n = 120)	0.5 (0.03-18.7)	0.03 (0-1.5)	5.1 (0-100)
Grades I, II & III (n = 83)	0.3 (0.03-3.9)	0.02 (0-0.5)	3.4 (0-100)
Pilocytic Astrocytoma (n = 2)	1.4 (1.0-1.9)	0 (0-0)	0 (0-0)
Ependymoma (n = 15)	0.5 (0.09-3.0)	0.03 (0-0.3)	3.4 (0-29.4)
Oligodendroglioma (n = 20)	0.2 (0.04-1.6)	0 (0-0.2)	0 (0-57.3)
Oligoastrocytoma (n = 19)	0.25 (0.04-3.9)	0.05 (0-0.4)	10.5 (0-100)
Astrocytoma (n = 27)	0.3 (0.03-2.0)	0 (0-0.5)	0 (0-100)
Grade IV, GBM (n = 37)	1.1 (0.17-18.7)	0.08 (0-1.5)	7.8 (0-47.4)

Example 21: Determination of T cells and Tregs levels in tumor infiltrates by *CD3Z* and *FOXP3* MS-qPCR assays in excised glioma tumors.

The demethylation assays of CD3Z and FOXP3 were used to measure levels of tumor infiltrating CD3+ T cells and Tregs, respectively, in 120 fresh frozen glioma tumors from the UCSF Brain Tumor Research Center tissue bank. Results of MS-qPCR assays are summarized in Table 18. Increased glioma tumor grade and higher levels of both CD3+T cell (Wilcoxon p = 5.7x10-7; Figure 17 panel A) and Treg (Wilcoxon p = 0.00014; Figure 17 panel B) in tumor infiltrates were observed ro be significantly associated. In grade IV glioma tumor tissues the median level of Treg percentage of T cells was observed to be higher than that of control blood samples (Table 18), and higher than that of lower grade tumors (Figure 17 panel C). Data from MS-qPCR showed significant differences among glioma tumor histologies in levels of CD3+ T cells (Kruskal-Wallis p = 8.6x10-7; Figure 21 panel A), Tregs (Kruskal-Wallis p = 0.00011; Figure 21 panel B) and Treg/CD3+ T cell ratios (Kruskal-Wallis p = 0.018; Figure 21 panel C). Poorer patient survival was associated with and higher levels of tumor infiltrating CD3+ T cells (Log-Rank p-value = 0.014; Figure 22 panel A) and Tregs (Log-Rank p-value = 0.039; Figure 22 panel B) measured by MS-qPCR.

Example 22: Kaplan-Meier survival curves for glioma cases show association of lower Treg with improved survival

Survival of glioma patients were correlated with the incidence of CD3+ T cells and Tregs as measured by *CD3Z* demethylation assays. (Figure 22 panels A-C). Both univariate and multivariate survival analyses were performed. Kaplan-Meier survival curves for glioma cases were stratified by median values of *CD3Z* demethylation assays. For depicting the survival results in panels A-C, patients were divided into two groups. In each panel the top trace

represents survival data of the group of patients for whom the measured variable (methylation status of CD3+ T cells, or of Tregs, or a ratio Tregs/T cells) was below the median observed for that variable, and the bottom trace represents survival data of the group of patients for whom the measured variable was above the median observed for that variable.

The results show that after controlling for age, gender and grade the *CD3Z* demethylation assays for CD3+ and CD3+Tregs in glioma tumor tissue were significantly associated (Figure 22 panels A-C) with poorer patient survival.

A CD3+ T cell *CD3Z* demethylation assay was performed which showed that lower CD3+T cell/total input in glioma tumor tissue was significantly associated (Figure 22 panel A) with improved survival (Log-Rank p-value = 0.0144). A Treg CS-DM *CD3Z* demethylation assays was performed which showed (Figure 22 panel B) that lower Treg/total input in glioma tumor tissue was significantly associated with improved survival (Log-Rank p-value = 0.0385). A measurement of Treg/ CD3+ T cell ratio was performed by *CD3Z* demethylation assay which showed (Figure 22 panel C) that lower Treg percentage of CD3+ T cells in glioma tumor tissue was significantly associated with improved survival (Log-Rank p-value = 0.4558).

Example 23: Cells, and cancer patient and control datasets for determining DNA methylation based epigenetic signatures for differentiating patients and controls

Sorted, normal, human peripheral blood leukocyte subtypes were isolated from whole blood by magnetic activated cell sorting (MACS) (AllCells LLC, Emeryville, CA). The purity of separated cells was confirmed with flow cytometry to be >97%. Genomic DNA was extracted and purified from cell pellets using a commercially available method (Qiagen, Valencia, CA), treated with sodium bisulfite (Zymo Research, Irvine, CA) and subjected to methylation profiling using the Infinium HumanMethyation27 BeadArray (Illumina, San Diego, CA). This same platform was used for the analysis of samples from the case-control studies described below.

The HNSCC data set consists (Table 19) of 92 incident cases from the greater Boston area and 92 cancer-free population-based control subjects from the same region (Applebaum KM et al., Int J Cancer 124:2690–2696, 2009). The clinical characteristics for this study population are contained in Table 19. The ovarian cancer data set (Teschendorff AE et al., 2009, PLoS One 4:e8274, 2009) is publicly available from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/, Accession number GSE19711), and consists of 266 postmenopausal women diagnosed with primary epithelial ovarian cancer (131 pre-treatment and 135 post-treatment cases) from the UK Ovarian Cancer Population Study (UKOPS). Controls (n = 274) were cancer-free postmenopausal women for which annual serum samples

were available. To avoid potential biases due to therapy, only pre-treatment ovarian cases were included in the analysis. The bladder cancer data set (Marsit CJ et al., 2011, J Clin Oncol 29:1133-1139) consists of 223 incident bladder cancer cases identified from the New Hampshire state cancer registry and 237 population controls from the same region (Karagas MR et al., 1998, Environ Health Perspect 106:1047–1050; Wallace K et al., 2009, Cancer Prev Res 2:70–73). Table 20 provides a summary of the participant characteristics.

Table19. Characteristics of the study population in the HNSCC data set.

Characteristics	**Cases** (n = 92)	**Controls** (n = 92)
Age, median years (range)	58 (31-84)	59 (32-86)
Gender, n (%)		
Male	64 (69.6%)	64 (69.6%)
Female	28 (30.4%)	28 (30.4%)
Smoking history, n (%)		
Never	17 (18.5%)	32 (34.8%)
Former	59 (64.1%)	47 (51.1%)
Current	16 (17.4%)	13 (14.1%)
Pack-years*, median (range)	40.0 (0.8-135.0)	24.5 (0.5-85.0)
Alcohol history, median drinks/week (range)	15.7 (0-307.0)	5.6 (0-140.6)
HPV16 (E6, E7 or L1 seropositivity), n (%)		
Negative	66 (71.7%)	83 (90.2%)
Positive	26 (28.3%)	9 (9.8%)
Tumor Site, n (%)		
Oral cavity	39 (42.4%)	---
Pharynx	35 (38.0%)	---
Larynx	18 (19.6%)	---
Stage, n (%)		
I	9 (12.5%)	---
II	9 (12.5%)	---
III	14 (19.4%)	---
IV	40 (55.6%)	---
*Restricted to ever-smokers (current + former)		

Table 20. Characteristics of the study population in the Bladder cancer data set.

Characteristics	Controls		Cases	
	No.	%	No.	%
Total No.	237		223	
Age, years				
Median	65		66	
Range	28-74		25-74	
Sex				
Male	158	48	171	52
Female	79	60	52	40
Family history of bladder cancer*				
No	224	53	199	47

Yes	7	44	9	56
Smoking history				
Never	72	64	40	36
Former	126	53	111	47
Current	39	35	72	66
Tumor stage/grade designation				
Carcinoma in situ	NA		6	3
Noninvasive low grade (grade 1-2)	NA		140	63
Noninvasive high grade (grade 3)	NA		17	7
Invasive	NA		60	27
* Data on family history were not available for 13 subjects				

Example 24: Statistical analysis of differences in methylation status in leucocyte subsets for determining signatures based on leukocyte DMRs

The analytic strategy was aimed toward examining the extent to which peripheral blood DNA methylation of non-hematopoietic cancers is driven by the epigenetic signatures that define leukocyte subtypes. Linear mixed-effects models were used to assess differences in methylation across the leukocyte subtypes and controlled for the large number of comparisons using false discovery rate (fdr) estimation. Leukocyte DMRs were subsequently ranked based on their strength of association and the highest ranking 50 DMRs were examined across the three cancer data sets between cancer cases and cancer-free controls.

An analysis was performed that capitalized on the aggregate methylation signatures across a collection of leukocyte DMRs. Each one of the full cancer data sets was split into equally sized training and testing sets. Samples in the training sets were then clustered using leukocyte DMRs. Clustering analysis was achieved using the Recursively Partitioned Mixture Model20 (RPMM), a hierarchical model-based method for clustering used for the clustering of array-based methylation data ((Christensen BC et al., 2009, PLoS Genet 5:e1000602; Christensen BC et al., 2011, J Natl Cancer Inst 103:143–153; Hinoue T et al., 2012, Genome Res. 22(2):271-82; Koestler DC et al., 2010, Bioinformatics 26:2578–2585). Based on the RPMM fit to the training sets, methylation class membership for the observations in the respective testing sets was predicted and the association between predicted methylation class and cancer case/control status were assessed.

The detailed statistical methodologies employed in the analysis are shown in Examples 25-26. Analyses were carried out using the R statistical package, R project for statistical computing, version 2.13 R available for downloading from the internet.

Example 25: Prediction of methylation class membership based on epigenetic signatures from leukocyte derived DMRs

Genome-wide DNA methylation was profiled in 46 samples of magnetic antibody sorted, normal human peripheral blood leukocyte subtypes (including B cells, granulocytes, monocytes, NK-cells, CD4+ T cells, CD8+ T cells, and Pan-T cells; Figure 28) using the Infinium HumanMethylation27 BeadArray. To discern leukocyte subtype DMRs, an association between methylation and leukocyte subtype for each of 26,486 autosomal CpG loci was examined. This data revealed 10,370 significantly differentially methylated CpGs among the leukocyte subtypes (fdr q-value < 0.05), which were ranked by q-value (Table 22 and Figure 24 panel A). The highest ranking 50 DMRs (Table 21) from this ranked list were selected for use in the case-control analyses. Since the publically available ovarian cancer data set included both pre- and post-treatment cases, only pre-treatment cases (n = 131) were considered in subsequent analyses to avoid potential biases resulting from therapy. Using unconditional logistic regression models, adjusted for available and relevant confounders (Figure 24 panel A), a substantial proportion of the 50 selected leukocyte DMRs were found to be significantly differentially methylated between cancer cases and cancer-free controls at the $\alpha = 0.05$ threshold (48, 47, and 8 out of 50, permutation p-values = <0.001, <0.001, 0.085, for HNSCC, ovarian cancer, and bladder cancer, respectively; Figure 24 panel B).

Eight of the leukocyte DMRs that were significantly differentially methylated in cancer cases compared to controls were observed to be common to the three cancer types (Figure 24 panel B). In HNSCC and ovarian cancer, seven of these eight leukocyte DMRs were hypomethylated in cases relative to controls, whereas all 8 DMRs were hypermethylated in bladder cancer cases relative to controls (Table 22).

To extend on the aggregate methylation signatures across a collection of leukocyte DMRs, classifiers based on profiles of leukocyte DMRs obtained from the subset analysis were developed and tested and the performance of these classifiers for successfully discriminating cancer cases from cancer-free controls was assessed. The workflow of the DMR methylation profile analysis is shown in Figures 29-31. For each of the three cancer data sets, a cross-validation procedure (Christensen BC et al., 2011, J Natl Cancer Inst 103:143–153) was implemented on the training sets only to determine the number of highest ranking leukocyte DMRs (M) for subsequent clustering analysis of the training sets. The highest ranking 50, 10, and 56 leukocyte DMRs from the respective cross-validation procedures using the 10,370 putative DMRs initially identified were selected to cluster the observations in the HNSCC, ovarian cancer, and bladder cancer training sets respectively. The resultant clustering solutions were used to predict methylation class membership for the subjects within the respective

independent testing sets. Figures 24 panel A, 25 panel A and 26 panel A depict heat maps of the respective testing sets by predicted methylation class for each cancer data set. Methylation classes derived from leukocyte subtype DMRs were significantly associated with cancer case status within each cancer type (permutation χ^2 p-values <0.0001, <0.0001, 0.03, HNSCC, ovarian cancer, and bladder cancer data sets respectively), supporting the phenotypic relevance of predicted methylation classes based on leukocyte DMRs.

For the HNSCC testing set, subjects predicted to be in the right most classes of the dendrogram (classes beginning with R) were six-fold more likely to be HNSCC cases compared to subjects in the left most classes (classes beginning with L) (OR = 5.99; 95% CI [1.96, 18.36]), controlling for age, gender, smoking, alcohol consumption, and HPV serostatus. Assessing the clinical utility of the predicted methylation classes in HNSCC demonstrated that methylation classes derived from the highest ranking 50 leukocyte DMRs were highly predictive of HNSCC case/control status (area under the curve (AUC) = 0.82 95% CI [0.74, 0.91]), which increased to 0.92 (0.87, 0.98 with age, gender, smoking, alcohol consumption, and HPV serostatus included in the model (Figure 24 panel B).

For ovarian cancer, subjects predicted to be in the right most classes were approximately ten-fold more likely to be ovarian cancer cases compared to subjects in the left most classes (OR = 9.87, 95% CI [4.63, 21.10]), controlling for age. Additionally, the predicted methylation classes in the ovarian cancer data demonstrated remarkably high sensitivity and specificity for predicting ovarian cancer case/control status (AUC = 0.83 95% CI [0.77, 0.89]), which increased to AUC = 0.86 95% CI [0.81, 0.92] with age included in the model (Figure 25 panel B).

In the bladder cancer data, subjects in the right most classes were nearly twice as likely to be bladder cancer cases compared to subjects in the left most (OR = 1.94 95% CI [0.95, 3.98], adjusted for age, gender, smoking and family history of bladder cancer). The clinical utility of the predicted methylation classes in the bladder cancer data was lower than that observed for HNSCC and ovarian cancer (bladder AUC = 0.67 95% CI [0.60, 0.73] and adjusted AUC = 0.77 95% CI [0.71, 0.83] with age, gender, smoking, and family history in the model) (Figure 26 panel B).

Utilizing leukocyte-derived DMRs to differentiate cases and controls resulted in methylation profiles that were consistent, and in the case of HNSCC and ovarian tumors, considerably better in terms of their prediction performance compared to previously published results using the same data sets (Teschendorff AE et al., 2009, PLoS One 4:e8274; Marsit CJ et al., 2011, J Clin Oncol 29:1133-1139; Langevin SM et al., Epigenetics. 2012 Mar; 7(3):291-9). For the HNSCC and ovarian data sets there was a high degree of correlation in the methylation

status of leukocyte DMRs and CpG loci identified by previous analytic strategies (Langevin SM et al., Epigenetics. 2012 Mar; 7(3):291-9; mean absolute spearman correlations = 0.68 and 0.75, respectively; Figure 27 panels A and B). In contrast, the highest ranking 56 DMRs in the bladder data set were found to be less correlated with the CpG loci used to form the methylation classes in a previous study using the same data set (mean absolute spearman correlation = 0.11; Figure 27 panel C).

Table 21. The highest ranking 50 differentially methylated regions (DMRs) among the leukocyte subtypes (false discovery rate q-values < 0.001 for all)

CpG Name	Chromosome	Gene Name	F-statistic
cg03801286	21	KCNE1	373.63
cg25634666	11	FOLR3	369.50
cg24777950	14	CTSG	350.66
cg17356733	21	IFNGR2	291.97
cg02497428	16	IGSF6	291.35
cg24211388	6	AIF1	285.92
cg03330678	17	9-Sep	284.79
cg00546897	21	LOC284837	279.64
cg24841244	11	CD3D	271.62
cg11283860	1	SLC45A1	271.09
cg27485921	2	ATP6V1E2	267.19
cg00974864	1	FCGR3B	260.62
cg07730301	11	ALDH3B1	252.52
cg07728874	11	CD3D	250.67
cg17496921	19	TSPAN16	246.58
cg26661623	17	ASGR2	242.83
cg18920397	1	LY9	238.64
cg27461196	19	FXYD1	236.64
cg20720686	7	POR	232.23
cg09303642	12	NFE2	231.34
cg23140706	12	NFE2	224.95
cg08458487	10	SFTPD	217.67
cg20748065	7	POR	217.63
cg18589858	11	SLCO2B1	217.14
cg10287137	11	P2RY2	215.31
cg25587233	9	PPP2R4	207.25
cg08044694	19	BRD4	202.50
cg18084554	19	ARID3A	198.61
cg13650156	7	PILRA	197.87
cg18854666	2	SLC11A1	197.42
cg17173423	11	MS4A3	195.50
cg22242539	17	SERPINF1	194.11
cg02780988	17	KRTHA6	193.25
cg10266490	1	ACOT11	192.62

Table 22. Methylation differences between cancer cases and controls for the eight overlapping differentially methylated leukocyte DMRs. Mean delta-beta refers to the difference in mean methylation between cancer cases and controls (i.e. βcases – βcontrols).

cg27606341	5	FYB	191.23
cg15512851	6	FGD2	185.34
cg20070090	1	S100A8	183.43
cg11058932	7	TSGA13	183.31
cg13500819	5	PACAP	182.82
cg15880738	11	CD3G	182.73
cg07285167	1	CSF3R	182.16
cg09868035	20	C20orf135	179.56
cg01980222	6	TREM2	178.94
cg21019522	11	SLC22A18	176.20
cg16097772	12	LYZ	172.89
cg21969640	12	GPR84	172.51
cg12971694	9	CD72	172.43
cg22224704	11	GSTP1	172.40
cg07239938	19	ELA2	170.70
cg02240622	15	PLCB2	169.99

	Mean delta-beta (95% CI)		
Gene Locus	HNSCC	Ovarian	Bladder
C20orf135	-0.05 (-0.07, -0.03)	-0.06 (-0.08, -0.05)	0.02 (0.0, 0.04)
PACAP	0.02 (0.00, 0.04)	0.04 (0.02, 0.05)	0.02 (0.0, 0.04)
FGD2	-0.05 (-0.07, -0.03)	-0.06 (-0.07, -0.04)	0.02 (0.01, 0.04)
SLC22A18	-0.05 (-0.07, -0.04)	-0.05 (-0.06, -0.04)	0.02 (0.01, 0.04)
GSTP1	-0.05 (-0.07, -0.04)	-0.06 (-0.07, -0.05)	0.02 (0.01, 0.04)
NFE2	-0.04 (-0.05, -0.03)	-0.04 (-0.05, -0.03)	0.02 (0.0, 0.03)
ASGR2	-0.06 (-0.08, -0.04)	-0.05 (-0.07, -0.04)	0.02 (0.01, 0.04)
SLC11A1	-0.05 (-0.07,-0.04)	-0.05 (-0.04,-0.06)	0.02 (0.0, 0.04)

Example 26. Statistical analysis of methylation differences in leukocyte DMRs between cancer cases and cancer–free controls for determining epigenetic signatures specific to each group

Linear mixed-effects models were used to assess differences in methylation across the leukocyte subtypes, modeling arcsine square-root transformed methylation as the response1, leukocyte subtype as a fixed effect covariate, and a random effect term for plate/BeadChip. False discovery rate (fdr) estimation was used to control for the large number of comparisons and putative leukocyte DMRs were defined as those with fdr q-value < 0.05. Leukocyte DMRs

were then ranked based on their strength of association using the F-statistics that resulted from the respective linear mixed-effects models.

Methylation differences among the highest ranking 50 leukocyte DMRs were examined between cancer cases and cancer-free controls using a series of unconditional logistic regression models that were adjusted using available and relevant covariate information. A leukocyte DMR was considered differentially methylated if the nominal p-value from the unconditional logistic regression model was less than 0.05. Permutation tests were then applied to each of the three data sets to determine if the number of differentially methylated leukocyte DMRs was significantly greater than expected by chance. Specifically, samples were randomly permuted (same permutation across the highest ranking 50 DMRs) and an unconditional logistic regression model was fit to the resampled data. For each data set 1000 permutations were considered to generate a null distribution of the number of differentially methylated leukocyte DMRs. Permutation p-values were then obtained by comparing the observed number of differentially methylated leukocyte DMRs to the respective null distribution.

The leukocyte DMR profile analysis involved splitting the full cancer data sets into equally sized training and testing sets (Figures 29-32). Samples in the training set were clustered using the highest ranking M leukocyte DMRs, where M was determined from the total pool of putative DMRs using the previously described cross-validation procedure (Sincic N and Herceg Z, 2011, Curr Opin Oncol 23:69-76). Clustering analysis was achieved using the Recursively Partitioned Mixture Model3 (RPMM), a hierarchical model-based method for clustering that has been extensively used for the clustering of array-based methylation data (Cui HM, 2007, Dis Markers 23:105-112; Wilhelm-Benartzi CS et al., 2010, Carcinogenesis 31:1972-1976; Schwartzman J et al., 2011, Epigenetics 6:1248-1256, 2011). Based on the RPMM fit to the training data, a naive Bayes classifier was used to predict methylation class membership for the observations in the independent testing set. Associations between predicted methylation class and cancer case/control status were assessed using permutation χ^2 tests and unconditional logistic regression models adjusted for available and relevant confounders. The clinical utility of the identified methylation classes were investigated using receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC).

Pairwise spearman correlation coefficients were computed between the highest ranking M leukocyte DMRs and the CpG loci identified from the corresponding semi-supervised RPMM2 (SS-RPMM) analysis of the HNSCC, ovarian, and bladder cancer data sets. A diagram illustrating the analytic framework for SS-RPMM is provided in Figure 32. Briefly SS-RPMM is a statistical methodology for identifying classes of methylation that are associated with a

phenotype of interest and has been successfully applied in several of settings (Christensen BC et al., 2009, Cancer Res 69:227–234; Marsit CJ et al., 2006, Cancer Res 66:10621-10629, 2006).

The same training and testing sets were used for the HNSCC and bladder cancer data sets as were used in the references Langevin SM et al., Epigenetics. 2012 Mar; 7(3):291-9 and Christensen BC et al., 2009, Cancer Res 69:227–234, to compare the results of the present analysis to previously published results, and to provide additional insight with respect to the findings of those studies. The ovarian cancer data set was also analyzed using SS-RPMM strategy described in Langevin SM et al., Epigenetics. 2012 Mar; 7(3):291-9 and Christensen BC et al., 2009, Cancer Res 69:227–234, and the results are shown in Figure 33. Following the logic above, the training sets used for the SS-RPMM analysis were applied to the leukocyte DMR profile analysis of the ovarian data.

Analyses were carried out using the R statistical package, R project for statistical computing, version 2.13 R available for downloading from the internet.

Example 27: Methylation analysis by DNA Methylation Microarray for NK cell specific DMR

Normal human peripheral blood leukocytes were isolated by magnetic activated cell sorting (MACS; Miltenyi Biotec Inc., Auburn, CA) and purity was confirmed by fluorescence activated cell sorting (FACS). The major cell types obtained included NK cells (n=9), B cells (n=5), T cells (n=16), monocytes (n=5), and granulocytes (n = 8). DNA and RNA were co-extracted from MACS sorted leukocytes using AllPrep DNA/RNA mini kit (Qiagen Inc., Valencia, CA). DNA from archived blood was extracted with DNeasy Blood & Tissue kit (Qiagen Inc., Valencia, CA). DNA was treated with sodium bisulfite according to the EZ DNA Methylation Kit (Zymo Research Corporation, Irvine, CA).

Methylation analysis was performed using The Infinium® HumanMethylation27 Beadchip Microarray (Illumina Inc., San Diego, CA), which quantifies the methylation status of 27,578 CpG loci from 14,495 genes, with a redundancy of 15-18 fold. The ratio of fluorescent signals was computed from both alleles using the following equation: $\beta=(\max(M,0))/(|U| + |M|) + 100$. The resultant β-value is a continuous variable ranging from 0 (unmethylated) to 1 (completely methylated) that represents the methylation at each CpG site and is used in subsequent statistical analyses. Data were assembled with the methylation module of GenomeStudio software (Illumina, Inc., San Diego, CA; Bibikova M et al., 2009, Epigenomics 2009;1:177-200)

Example 28: Validation of DNA Methylation Microarray results for identifying NK cell-specific DMRs by pyrosequencing

Pyrosequencing assays to validate microarray results were designed using Pyromark Assay Design 2.0 (Qiagen Inc., Valencia, CA), and carried out on a Pyromark MD pyrosequencer running Pyromark qCpG 1.1.11 software (Qiagen Inc., Valencia, CA). Oligonucleotide primers were obtained from Life Technologies™ (Grand Island, NY).

Example 29: Protein expression analysis by mRNA expression array for identifying NK cell-specific DMRs

The Whole-Genome DASL HT Assay Kit (Illumina Inc., San Diego, CA) was used to obtain simultaneous profiles of more than 29,000 mRNA transcripts. Data were assembled with the expression module of GenomeStudio software (Illumina Inc., San Diego, CA). The mRNA expression array data was used in combination with DNA methylation array data to identify NK cell-specific DNA methylation.

Example 30: Methylation specific quantitative polymerase chain reaction (MS-qPCR) analysis for quantification of *NKp46* demethylation

Primers and TaqMan major groove binding (MGB) probes (Table 23) with 5' 6-FAM (6-Carboxyfluorescein) and 3' non-fluorescent quencher (NFQ) as well as TaqMan® 1000 RXN Gold with Buffer A Pack were obtained from Life Technologies™ (Grand Island, NY). MS-qPCR was performed using solutions and conditions according to Campan M et al., 2009, Methods Mol Biol, 507:325-37 with the following modifications. A solution of 10X TaqMan® Stabilizer containing 0.1% Tween-20, 0.5% gelatin was prepared weekly. Each reaction of 20 µl contained 5 µl DNA, 11.9 µl preMix, 3 µl oligoMix, and 0.1 µl Taq DNA polymerase. Cycling was performed using a 7900HT Fast Real- Time PCR System (Applied Biosystems, Foster City, CA); 50 cycles at 95 °C for 15 sec and 60 °C for 1 min after 10 min at 95 °C preheat. All samples were run in triplicate using the absolute quantification method.

Table 23. MS-qPCR oligonucleotide sequences

Oligonucleotide name	**Sequence**
NKp46 forward primer	ATTAGGTTGGTAGAATTTGAGT (SEQ ID NO: 116)
NKp46 reverse primer	CCCATTCCCCTTCCACA (SEQ ID NO: 117)
NKp46 probe	(6FAM) CTCACCAACACAAAACAA (MGB, NFQ) (SEQ ID NO: 118)
C-less forward primer	TTGTATGTATGTGAGTGTGGGAGAGA (SEQ ID NO: 97)
C-less reverse primer	TTTCTTCCACCCCTTCTCTTCC (SEQ ID NO: 98)
C-less probe	(6FAM) CTCCCCCTCTAACTCTAT (MGB,NFQ) (SEQ ID NO: 99)

MGB: major groove binding

FAM: 6-Carboxyfluorescein
NGQ: NFQ
C-less qPCR assay: Campan M et al., 2009, Methods Mol Biol, 507:325-37; Weisenberger DJ et al., 2008, Nucleic Acids Res 2008; 36:4689-98

Quantification of total bisulfite converted DNA copies was performed by reference to the C-less qPCR assay (Campan M et al., 2009, Methods Mol Biol, 507:325-37; Weisenberger DJ et al., 2008, Nucleic Acids Res 2008;36:4689-98). C-less primers and probes recognize a DNA sequence without cytosines; hence, the assay amplifies the total amount of DNA in a PCR reaction regardless of bisulfite conversion or methylation status. A conversion factor was used for a diploid human cell, which is 6.6 picograms (pg) of DNA (3.3 pg per copy) to calculate copy number.

Normal human blood DNA quantified by UV absorption (Nanodrop, Inc) was used to generate a four point standard curve with 30,000 copies, 3,000 copies, 300 copies and 30 copies of genomic DNA. This standard curve was included on each sample plate to obtain quantification of DNA from Ct values. Since C-less primers hybridize to both strands of the standard DNA (non-bisulfite converted) and since bisulfite converted samples hybridize to a single strand during the first cycle, the resultant copy number obtained from bisulfite treated samples was multiplied by two. Bisulfite converted, universal methylated DNA standard (Zymo Research Corperation, Valencia, CA) and bisulfite converted, isolated NK cell DNA were quantified at the same time using the C-less assay. Resultant copy number measurements were used to prepare a calibration curve for the NKp46 demethylation assay. NK cell DNA in known copy numbers was spiked into universal methylated DNA in ratios that maintained a constant total number of DNA copies (10,000 copies) in each reaction across the dilution scheme. This mimics conditions for detecting different relative numbers of NK cells within a complex mixture of cells in a biological sample. For absolute quantification of NKp46 demethylation, the four-point standard curve used 10,000 copies, 1,000 copies, 100 copies, and 10 copies of bisulfite converted NK cell DNA.

Example 31: Statistical modeling of the DNA methylation microarray data for estimation of differential methylation

A linear mixed effects model was applied to the Illumina Infinium® HumanMethylation27 data using SAS (SAS Institute Inc., Cary, NC). Cell type was designated as the fixed effect and beadchip plate was the random effect. For this example, the fixed effect groups were NK cells and non-NK cells, which included pan T lymphocytes, CD4+ T-lymphocytes, Tregs, CD8+ T-lymphocytes, B-lymphocytes, granulocytes and monocytes. Coefficients were generated that estimated differential methylation were generated such that, for

any particular locus, a negative coefficient indicated less methylation in NK cells than in the other cell types. Resultant p-values were adjusted for multiple comparisons using the "qvalue" package in the software, the R project for statistical computing available for downloading from the internet.

Example 32: Statistical modeling of the RNA expression array for estimation of differential RNA expression

Linear models were applied to the Illumina Whole-Genome DASL HT using the "limma" package in the software, the R project for statistical computing. RNA expression for MACS isolated NK cells was compared to each of the following MACS isolated leukocytes: pan T-lymphocytes, CD4+ T-lymphocytes, Tregs, CD8+ T-lymphocytes, B lymphocytes, ganulocytes and monocytes. Thus, estimates were obtained for log-fold changes in RNA expression between NK cells and each of the aforementioned cell types, in which a positive value indicated higher RNA expression in NK cells compared to a particular cell type. Resultant p values were adjusted for multiple comparisons using the "qvalue" package in R project for statistical computing. NK cell specific differential RNA expression was considered significant only if the seven q-values were each less than 0.1.

Example 33: Statistical analysis of the (MS-qPCR) data

Statistical analyses were carried out in R project for statistical computing. A generalized linear model analysis and F-test were performed to determine log linear PCR kinetics for the NK cell standard curve. To test for univariate associations between continuous NKp46 demethylation measurements and discrete variables, Wilcoxon rank sum tests (for dichotomous variables, such as case status) and Kruskal-Wallis one-way analysis of variance tests were employed. To test for univariate associations between continuous NKp46 demethylation and other continuous variables linear regression analysis, calculation of Pearson product-moment correlations and F-tests were performed. A chi-squared test for trends in proportions was applied to identify trends in HNSCC prevalence by control-determined demethylation tertiles. Multivariate logistic regression analyses were performed using the "glm" function with family set to binary.

Example 34: NKp46 demethylation is a biomarker of NK cells

Analysis of DNA methylation and RNA expression microarray data from MACS isolated (FACS validated) normal human leukocytes were integrated to identify putative, NK cell-specific DMRs that could potentially serve as reliable biomarkers of the cell type. The list of

candidate gene regions was narrowed to CpG loci that were significantly demethylated in NK cells ($q < 0.1$, coefficient < 0) and that were located within genes whose RNA expression was significantly elevated in NK cells ($q < 0.1$, log fold-change > 1). These candidates are marked as darkened asterisks in the top left quadrant of Figure 34. Pyrosequencing and MS-qPCR of bisulfite converted DNA from the MACS isolated leukocytes confirmed that a region near the promoter of *NKp46* is demethylated in NK cells, and is methylated in T cells, B cells, granulocytes, and monocytes (Figures 35 and 38). Furthermore, the $CD56^{dim}$ subset of NK cells showed complete demethylation in the *NKp46* region, whereas $CD56^{bright}$ NK cells exhibited only partial demethylation in the region as measured by MS-qPCR. The *NKp46* MS-qPCR assay was optimized to fit a log-linear relationship between lower Ct values (more demethylated copies of *NKp46*) and increased NK cell DNA content (Pearson $R = -0.996$, $p < 2.2 \times 10^{16}$; Figure 36).

Example 35: Samples from HNSCC patients have diminished circulating NK cells

The calibrated *NKp46* MS-qPCR assay was used to measure the level of circulating NK cells in the peripheral blood of patients with HNSCC and cancer free controls. The demographics of the study population are shown in Table 24.

Table 24. Demographic characteristics

Characteristic	Total (N = 244)	Controls (n = 122)	HNSCC (n = 122)	Oral (n = 43)	Pharyngeal (n = 53)	Laryngeal (n = 26)
Age						
Mean (SD)	61 (12)	62 (12)	61 (12)	60 (15)	60 (10)	64 (9.5)
Median (Range)	60 (29-87)	60 (31-87)	60 (29-86)	59 (29-86)	60 (41-86)	64 (50-83)
Gender						
Male, No.(%)	178 (73%)	89 (73%)	89 (73%)	27 (63%)	41 (77%)	21 (81%)
Female, No.(%)	66 (27%)	33 (27%)	33 (27%)	16 (37%)	12 (23%)	5 (19%)
HPV 16 Serology						
L1+, No.(%)	33 (14%)	4 (3%)	29 (24%)	6 (14%)	22 (42%)	1 (4%)
E6+, No.(%)	41 (17%)	4 (3%)	37 (30%)	2 (5%)	32 (60%)	3 (12%)
E7+, No.(%)	28 (11%)	2 (2%)	26 (21%)	1 (2%)	23 (43%)	2 (8%)
E6+ and E7+, No.(%)	25 (10%)	0 (0%)	25 (20%)	0 (0%)	23 (43%)	2 (8%)
E6+ or E7+, No.(%)	44 (18%)	6 (5%)	38 (31%)	3 (7%)	32 (60%)	3 (12%)
Cigarette Smoking Status						

Never, No.(%)	65 (27%)	41 (34%)	24 (20%)	11 (26%)	11 (21%)	2 (8%)
Former, No.(%)	149 (61%)	66 (54%)	83 (68%)	29 (67%)	35 (66%)	19 (73%)
Current, No.(%)	30 (12%)	15 (12%)	15 (12%)	3 (7%)	7 (13%)	5 (19%)
Cigarette Pack-Years						
Mean (SD)	26 (29)	17 (23)	35 (32)	26 (27)	36 (35)	45 (30)
Median (Range)	16 (0-116)	7 (0-114)	31 (0-116)	20 (0-105)	33 (0-116)	45 (0-96)
Alcohol Drinks per Week						
Mean (SD)	18 (26)	15 (27)	21 (24)	18 (23)	22 (25)	23 (25)
Median (Range)	7 (0-199)	6 (0-199)	14 (0-155)	7 (0-90)	18 (0-155)	19 (0-113)

Univariate analysis revealed that significantly fewer demethylated copies of NKp46 were detected in HNSCC blood than in control blood ($p < 0.0001$, Figure 39), indicative of a diminished NK cell compartment in the peripheral blood of HNSCC patients. There was no significant univariate association observed between the measured number of demethylated NKp46 copies and age, gender, HPV16 (E6 and/or E7) serology, cigarette smoking, alcohol consumption, or body mass index. There was no significant difference in the number of demethylated *NKp46* copies detected in patients with oral, pharyngeal, and laryngeal tumors.

To determine whether the observed association between NK cells and case status was attributable to systemic chemotherapy or other treatments, the number of demethylated NKp46 copies detected in case blood samples drawn within one month of diagnosis was compared to those drawn more than one month after diagnosis, and no significant difference was observed.

The NKp46 MS-qPCR measurements from cancer-free control blood samples were used to determine suitable cutoffs for NKp46 demethylation tertiles. The proportion of total HNSCC cases decreased significantly with increasing demethylation tertile ($p > 0.001$, Figure 37), indicating that HNSCC patients are more likely to have depressed levels of NK cells in their peripheral blood. The trend held true independent of the case stratification by HPV16 (E6 and/or E7) serology, or time of blood drawing within a month of diagnosis or earlier. Multivariate logistic regression controlling for age, gender, cigarette smoking, alcohol consumption, BMI, and HPV16 (E6 and/or E7) serology confirmed increased HNSCC risk for individuals in the lower two normal NKp46 demethylation tertiles (Table 25), strongly indicating that lower levels of NK cells in the peripheral blood are significantly associated with HNSCC.

Table 25 Logistic regression of HNSCC risk

NKp46 demethylation tertile	Crude		Adjusted*	
	OR (95% CI)	p-value	OR (95% CI)	p-value
1st (lowest)	4.3 (2.2, 9.0)	$5.0x10^{-5}$	5.6 (2.0,17.4)	0.002
2nd (middle)	2.8 (1.4, 6.0)	0.006	4.9 (1.8, 16.1)	0.004
3rd (highest)	Reference		Reference	

*Unconditional multivariate model controlling for age, gender, smoking, drinking, BMI and HPV16 (E6 and/or E7) serology

Example 36: Application of the methodology to mRNA data

The statistical methods described herein for determining changes the distribution of white blood cells among different subpopulations are applicable to mRNA expression profiles with the following considerations. A mathematical consideration is that mRNA is typically analyzed on a logarithmic scale, yet the assumptions of the methods herein involve linearity on an arithmetic scale, since the mixing coefficients are assumed to act linearly on absolute numbers of nucleic acid molecules; thus, the proposed methods would require analysis of untransformed fluorescence intensities, for which skewed distributions would result in numerical instabilities. A biological consideration is absence of a linear relationship between cell number and mRNA copies, since proteins may be translated as a consequence of an initial burst of mRNA transcription upon cellular development, followed by significant mRNA degradation. In contrast, one would expect the average beta value provided by Illumina bead-array products, as well as similarly constructed quantities from other platforms to scale in proportion to the actual fraction of methylated nucleic acids with a biologically reasonable assumption of two DNA molecules per cell.

An example of an application of methods herein is shown using mRNA data. The validation data set S_0 was obtained from Watkins NA et al., 2009, Blood 113: e1-e9, in which the Illumina Human-6 v2 Expression BeadChip was used to characterize the mRNA expresion profile of eigt types of blood cells: B cells, granulocytes, erythroblasts, megakaryocytes, monocytes, natural killer cells, CD4+ T cells, and CD8+ T cells. For this analysis erythroblasts (nucleated progenitors of red blood cells) and megakaryocytes (progenitors of platelets) were removed. The target data set S_1 was obtained from Showe MK et al., 2009, Cancer Res 69: 9202-10, in which the same mRNA expression platform was used to characterize expression differences in isolated mononuclear cells between nonsmall cell lung cancer (NSCLC) cases and controls having non-cancer lung disease, adjusting for age, sex and smoking. In addition, data was presented from 18 matched case samples, pre- and post-operative.

The same methodology was used as for the DNA methylation data sets herein, ordering the 46,693 transcripts by F statistic according to their ability to distinguish six types of leukocytes. Of the 100 transcripts having the largest F statistics it was observed that 86 overlapped with the transcripts in Showe MK et al., 2009, Cancer Res 69: 9202-10. Thus the remainder of the analysis was carried out using the 86 overlapping loci. In the analyses, untransformed data (i.e. using either the normalized fluorescence intensities or 2 raised to the power of the normalized $\log_2$ intensities) were used. Application of the constrained projection in Examples 1 and 5 resulted in an average percentage estimates consistent with mononuclear cells (i.e. a subfraction with most granulocytes removed): 3.3% B cell, 3.4% granulocyte, 18.1% monocyte, 29.5 % NK cell, 11.6 CD4+ T cell, and 2.2 % CD8+ T cell.

Table 26 presents results from 137 NSCLC cases and 91 controls, adjusted for age, sex, and smoking status. Table 27 presents results from 18 matched pre-operative and post-operative samples from NSCLC cases, where the analyzed outcome was the difference in untransformed expression (post-operative expression minus pre-operative expression), and coefficients displayed correspond to the intercept of $\mathbf{B}_1$ (analogous to a paired t-test). Perturbations in T cell distribution were consistent with known immunological changes resulting from NSCLC (Ginns LC et al., 1982, Am Rev Respir Dis 23: 265—9; Mazzoccoli G et al., 1999, In Vivo 13: 205-9), as well as with age and smoking. The perturbations and coefficient signs were reasonable; the magnitudes were potentially biased. For example, the estimates corresponding to granulocyte distribution were much larger than expected given the relatively small number of granulocytes present in a monouclear subfraction. Thus, the methods herein were determined to be suitable for application to mRNA data sets.

Table 26. White blood cell distribution comparing cases to controls in NSCLC mRNA data set

	Est	SE_2	p-value
Case Status			
B Cell	0.8	4.15	0.8511
Granulocyte	-34.6	9.48	0.0003
Monocyte	17.9	9.58	0.0613
NK	1.3	5.18	0.8095
T Cell (CD4+)	24.9	9.01	0.0057
T Cell (CD8+)	-15.2	9.03	0.0931
Age (decades)			
B Cell	-0.7	1.36	0.5824
Granulocyte	-7.9	3.45	0.0218
Monocyte	-6.5	2.76	0.0180
NK	-4.0	1.80	0.0255

T Cell (cd4+)	13.0	2.89	0.0000
T Cell (CD8+)	8.3	2.96	0.0052
Sex (male)			
B Cell	0.1	2.66	0.9827
Granulocyte	-34.8	6.41	0.0000
Monocyte	6.8	5.44	0.2091
NK	-7.8	3.32	0.0193
T Cell (CD4+)	21.1	5.39	0.0001
T Cell (CD8+)	13.2	5.76	0.0223
Former Smoker			
B Cell	1.6	3.97	0.6821
Granulocyte	17.2	8.25	0.0375
Monocyte	6.1	7.84	0.4368
NK	2.7	5.19	0.6103
T Cell (CD4+)	-11.3	8.02	0.1578
T Cell (CD8+)	-20.3	8.28	0.0141
Current Smoker			
B Cell	3.4	5.21	0.5183
Granulocyte	31.6	11.26	0.0049
Monocyte	17.8	10.49	0.0907
NK	5.4	6.93	0.4373
T Cell (CD4+)	-21.8	10.25	0.0337
T Cell (CD8+)	-41.2	11.10	0.0002

Est = Regression coefficient estimate (× 100%)
SE_2 = Double-bootstrap standard error (× 100 %).

Table 27. White blood cell distribution comparing matched pre-operative and post-operative cases in NSCLC mRNA data set

	Est	SE_2	p-value
B Cell	-10.7	5.55	0.0543
Granulocyte	-19.4	11.16	0.0826
Monocyte	-13.4	10.43	0.1987
NK	6.3	7.15	0.3794
T Cell (CD4+)	-11.3	10.57	0.2859
T Cell (CD8+)	48.8	11.33	0.0000

Est = Regression coefficient estimate (× 100%)
SE_2 = Double-bootstrap standard error (× 100%).

Example 37: An array for high-throughput DNA methylation analysis

An array for performing DNA methylation analysis in a high-throughput manner was made using VeraCode microbeads (Illumina, San Diego, CA USA) and DNA sequences of

regions in 96 different genes, each sequence having one CpG dinucleotide shown within square brackets (Figure 40) and used to determine methylation status of the gene. Veracode beads are cylindrical glass microbeads 240 microns in length by 28 microns in diameter with a surface suitable for attaching DNA, RNA, protein, antibody and other ligands for performing bioassays. For performing DNA methylation analysis various CpG specific DNA oligomers were attached to these beads. Each microbead is inscribed with a high-density holographic code (24-bit), allowing development of very large numbers of bead types. When a laser is shone at the high density codes of the beads they emit a signal specific to the code and the signal is detected by a CCD camera. The fluorescence of the bead indicates whether the particular CpG site carried by the bead is demethylated. The result is compared with the fluorescence readout obtained from DNA from a purified leukocyte sample. A VeraCode array is a collection of beads, each carrying a DNA oligomer specific for either the methylated or the unmethylated form of a particular CpG locus, distributed into different wells of a micro titer plate. A user selects all or a subset of nucleotide sequences containing CpG sites in a gene or genes of interest for attaching to VeraCode beads to have a custom designed VeraCode array particularly advantageous for the user's analysis.

To ascertain which 96 CpGs would give optimal precision for all of the white blood cell (WBC) types the following procedure was followed. The Infinium HumanMethylation 27K data corresponding to all of the Magnetic activated cell sorting (MACS sorted leukocyte DNA were assembled in the methylation module of GenomeStudio, and the quality of the data was assessed by calculating Mahalanobis distances. All 47 samples yielded acceptable data. A matrix of β-values was generated with rows defined by microarray CpG locus and columns defined by sample identification. A corresponding matrix indicating cellular phenotypes was also generated, with rows defined by sample identification (in precisely the same order as the columns in the corresponding matrix) and columns defining the cell lineage(s) to which each cell lineage belongs.

A linear mixed effects (LME) model was applied to the Illumina Infinium HumanMethylation27 WBC lineage as the fixed effect and beadchip plate as the random effect. The fixed effect groups were: Pan-T cell, CD4+ T cell, CD8+ T cell, Pan-NK cell, $CD56^{dim}$ NK cell, $CD56^{bright}$ NK cell, B cell, granulocyte, neutrophil, eosinophil, and monocyte. Across all gene loci, this model generated coefficients for each fixed effect group indicating relative estimates of DNA methylation for each of the different cell types. Collapsing categories accounted for the hierarchical relationships among cell lineages and a linear transformation was applied to convert coefficient estimates to estimated mean value per cell type, resulting in a

matrix $\tilde{\mathbf{B}}_0$ of mean values, each row corresponding to a CpG locus and each column corresponding to a cell type. The model also generated an F-statistic for each locus that indicates how significantly different DNA methylation was between the cell types.

A stochastic search algorithm was then employed to select the differentially methylated regions (DMRs) that work best in concert on a custom microarray to distinguish leukocyte lineages, and would therefore be the most effective at quantifying immune cell types in a biological sample. The objective was to ascertain which 96 CpGs would give optimal precision for all of the WBC types.

The stochastic search algorithm was designed to maximize precision of estimated cellular fractions, under the assumption that the variance-covariance of the fraction estimates is proportional to $(\tilde{\mathbf{B}}_0^{\mathrm{T}}\tilde{\mathbf{B}}_0)^{-1}$. To optimize precision for a single individual cell type, the corresponding diagonal element of $(\tilde{\mathbf{B}}_0^{\mathrm{T}}\tilde{\mathbf{B}}_0)^{-1}$ was minimized; to optimize a set of cell types, the sum of the corresponding diagonal elements was minimized.

The general strategy was as follows. The engine is a stochastic search algorithm that starts with an initial set of CpGs, which is the beginning choice for the "current" set. On each iteration a randomly chosen CpG from the current set is switched out with a randomly chosen CpG from the remaining (unselected) CpGs, and precision is compared between the current set and the "candidate" set. If the candidate set gives better precision then the switch is accepted. Otherwise it is rejected. Ideally, by the end of the algorithm, the acceptance rate should be 0%.

The algorithm was run for 50,000 iterations starting with the 500 CpGs having the best F statistics. This was repeated ten times with different random number seeds each time. Then, the algorithm was run for 50,000 iterations starting with the CpGs having the 500 largest absolute effect sizes (coefficients generated by the LME model) for the WBC types. This was also repeated ten times with different random number seeds each time. Next all 20 runs were compared and the algorithm run for 50,000 iterations starting with the 500 most frequently chosen CpGs from the previous 20 runs. This was repeated five times with different random number seeds each time. Finally, a run was performed for 750,000 iterations starting with the 96 most frequently chosen CpGs from the previous five runs.

Example 38: Mediation analysis for estimating effects of an exposure or phenotype on measured DNA methylation

A method is described for conducting a mediation analysis to estimate the effects of an exposure or to estimate the effects of a specific phenotype on measured DNA methylation along two paths: through changes in WBC distribution, and directly, unmediated by changes in WBC

distribution. Most Epigenome-wide association scans (EWAS) have attempted to estimate the marginal effect ($\boldsymbol{\beta}$, depicted in Figure 41, panel A) on measured DNA methylation, which are effects not adjusted for WBC distribution. However, a significant portion of the effect on DNA methylation is mediated through changes in WBC distribution as shown in Figure 41, panel B. Of interest in EWAS studies is $\boldsymbol{\alpha}$, the direct effect adjusted for WBC distribution. Estimating this effect requires estimation of two other quantities, $\boldsymbol{\Gamma}$, the effect of exposure or phenotype on WBC distribution, and ξ, the effect of WBC distribution on methylation. If y is the DNA methylation measured for subject i at a particular CpG site (j, subscript suppressed for clarity), $\mathbf{z}_i$ is a $p \times 1$ matrix of covariates for subject i (including the exposure or phenotype of interest), and $\boldsymbol{\omega}_i$ is the subject-specific WBC distribution estimated using constrained projection in the manner described in Example 1 then $y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha} + \boldsymbol{\omega}_i^{\mathrm{T}}\xi + e_i$, where e_i is a zero-mean error. Additionally, the effect of exposure/phenotype on WBC distribution can be modeled as $\boldsymbol{\omega}_i = \boldsymbol{\Gamma}\mathbf{z} + \mathbf{u}_i$, where $\mathbf{u}_i$ is a zero-mean error vector. It is noted that $\boldsymbol{\alpha}$ is a $p \times 1$ vector, and K cell types are assumed, so that $\boldsymbol{\omega}_i$ is a $K \times 1$ vector, $\boldsymbol{\Gamma}$ is a $K \times p$ matrix, and ξ is a $K \times 1$ vector. It follows that $y = \mathbf{z}_i^{\mathrm{T}}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}^{\mathrm{T}}\xi) + \mathbf{u}_i^{\mathrm{T}}\xi + e_i$, so that the marginal effect $\boldsymbol{\beta}$ is the $p \times 1$ vector $\boldsymbol{\alpha} + \Gamma^{\mathrm{T}}\xi$. Estimation proceeds first by computing $\hat{\boldsymbol{\Gamma}} = \left(\sum_{i=1}^{n} \boldsymbol{\omega}_i \mathbf{z}_i\right)\left(\sum_{i=1}^{n} \mathbf{z}_i^{\mathrm{T}}\mathbf{z}_i\right)^{-1}$, then computing $\hat{\mathbf{u}}_i = \boldsymbol{\omega}_i - \boldsymbol{\Gamma}\mathbf{z}_i$, $\mathbf{r}_i = (\mathbf{z}_i^{\mathrm{T}}, \hat{\mathbf{u}}_i^{\mathrm{T}})^{\mathrm{T}}$, $\hat{\xi} = \left(\sum_{i=1}^{n} \mathbf{r}_i^{\mathrm{T}}\mathbf{r}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{r}_i y_i\right)$, extracting $\hat{\xi}$ as the last K components of $\hat{\xi}$ and obtaining $\hat{\boldsymbol{\alpha}}$ by subtracting $\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\hat{\xi}$ from the first p components of $\hat{\xi}$. Statistical inference is achieved by permutation. Specifically, the null distributions of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Gamma}}$ are obtained by permuting the exposure or phenotype of interest within $\mathbf{z}$ (only the components representing the covariate to be tested), and the null distribution of $\hat{\xi}$ is obtained by permuting the subject assignments corresponding to $\boldsymbol{\omega}_i$. Adjustments for multiple comparisons are achieved by nesting within each permutation a loop that estimates $\hat{\boldsymbol{\alpha}}_j$, $\hat{\boldsymbol{\Gamma}}_j$, and $\hat{\xi}_j$ for each individual CpG, with adjusted p-values obtained by comparing the maximum absolute values of $\hat{\boldsymbol{\alpha}}_j$, $\hat{\boldsymbol{\Gamma}}_j$, and $\hat{\xi}_j$ (over all CpGs) to the corresponding statistics computed from each individual permutation. For comparison purposes, a similar permutation test can be applied for the marginal coefficient $\boldsymbol{\beta}$.

This method to a data set consisting of n=205 control subjects in a bladder cancer case/control study (Karagas MR et al., 1998, Environ Health Perspect 106: 1047-1050). Four separate analyses were performed: (1) the phenotype of interest was age; (2) the exposure of

interest was current smoker status; (3) the exposure of interest was toenail arsenic; and (4) the exposure of interest was reported use of hair dye. Sex was included as a covariate in all analyses, and age was included in (2)-(4).

The relationship between $\hat{\alpha}$ and $\hat{\beta}$ for the covariate of interest over all autosomal CpGs is shown in Figure 42. Dots represents overall methylation as indicated by the first component of the coefficient vector $\hat{\boldsymbol{\beta}}$, corresponding to the intercept (light=low, black=moderate, dark=high). The diagonal straight line represents the identity ($\hat{\alpha}=\hat{\beta}$). The curve depicts a loess fit to the scatter plot. In all cases there is an S-shaped relationship that shows attenuation of effect ($\hat{\alpha}$ tends to be smaller than $\hat{\beta}$). Table 28 shows the multiple-comparisons adjusted p-values for each coefficient corresponding to the covariate of interest (β, α, γ) and overall WBC distribution effect on DNA methylation (ξ), obtained by permutation test using 5000 permutations. As shown in the table, significance of α may be greater than, less than, or equal to the significance of β. Remarkably, in every case, the covariate of interest shows a strongly significant association with WBC distribution. It is noted that WBC shows significant overall association with DNA methylation.

Table. 28. Multiple-comparisons adjusted p-values

Exposure/Phenotype	β	α	γ	ξ
Age	0.0358	0.0838	<0.0002	0.0100
Current Smoker	0.0326	0.0200	<0.0002	0.0134
Toenail Arsenic	0.1054	0.0512	<0.0002	0.0148
Dye Use	0.2614	0.2570	<0.0002	0.0102

Example 39: Comparison of methods herein for estimating fractions of blood cell types with non-negative matrix factorization (NNMF)

The methods herein are predicated on the relationship $E(\mathbf{Y}_i) = \sum_{l=0}^{d_0} \mathbf{b}_{0l}\omega_{il}$, where $\mathbf{Y}_i$ is a vector of DNA methylation measurements obtained for subject i, d_0 is the number of blood cell types to be assayed, ω_{il} are the fractions of each blood cell type corresponding to subject i, and $\mathbf{b}_l$ is the vector of methylation fractions corresponding to blood cell type l; the methods herein provide techniques for estimating the fractions ω_{il} assuming the values of $\mathbf{b}_l$ have been obtained from an external validation data set. In contrast, non-negative matrix factorization

(NNMF) could be used to estimate ω_{il} and $\mathbf{b}_l$ simultaneously in absence of an external validation set. In the context of NNMF, the d_0 vectors $\boldsymbol{\omega}_{\bullet l}$ are considered "factors", and the d_0 vectors (assumed to represent individual methylation profiles) are considered "basis vectors" and the number of factors d_0 must be provided to the NNMF algorithm.

Using the 12 experimental samples described in Example 5 NNMF was compared to methods herein (Examples 1-3). Highest ranking 100 and 500 pseudo-DMRs were selected on the basis of informativeness as in Example 4; for each choice, the constrained projection described in Examples 1 and 5 was used to impute specific cell distributions, then NNMF was performed assuming four, five, and six factors (i.e. factor values assumed to represent the fractions ω_{il} for one cell type l). The *nmf* function in the R package *NMF* was used with default settings. Since NNMF requires random inputs, NNMF was applied 100 times, each with different randomly generated starting values according to the default settings of the *nmf* function. Six cases were considered, *viz.*, 100 CpGs and 500 CpGs for each of four, five and six factors. For each of the 100 runs in each of the six cases, the fitted factors $\boldsymbol{\omega}_{\bullet l}$ (values of which were assumed to correspond to fractions ω_{il}) were correlated to expected fractions of B cells, T cells, monocytes, and granulocytes, and for each specific cell type, the factor with the maximum correlation to that type was assigned to it. Then, for each cell type in each case, the median correlation with assigned factor was tabulated. Table 29 below reports these median values, and Table 30 reports the correlation between expected fraction and the fraction observed using methods herein. A comparison of these tables demonstrates that, though NNMF can achieve high correlation with expected cell fraction if the pseudo-DMRs are known in advance, the methods described herein in Examples 1-4 still achieves higher correlation. In addition, NNMF occasionally fails to match known cell types to imputed cell types in a monomorphic manner. Table 31 reports the percentage of runs for which at least two different cell types were matched via NNMF to the same factor.

It is expected that NNMF would behave less favorably than methods described herein (Examples 1-4), since NNMF requires the estimation of (n + M) F unknown parameters (where n = # of target samples, M = # of CpGs, and F = # of factors) and methods herein require the estimation of only n K unknown parameters, where K < F and K is the number of known cell types.

Table 29. Median correlation for two different sets of CpG containing sequences

100 CpGs

	Factors = 4	Factors = 5	Factors = 6
B cells	0.998	0.996	0.996
T cells	0.988	0.989	0.990
Monocytes	0.832	0.900	0.927
Granulocytes	0.967	0.954	0.963

500 CpGs

	Factors = 4	Factors = 5	Factors = 6
B cells	0.998	0.996	0.996
T cells	0.985	0.993	0.990
Monocytes	0.798	0.896	0.879
Granulocytes	0.943	0.977	0.970

Table 30. Correlation between expected fraction and the fraction observed using methods herein.

	100 DMRs	500 DMRs
B cells	1.000	1.000
T cells	0.998	0.997
Monocytes	1.000	1.000
Granulocytes	0.997	0.999

Table 31. Percentage of runs for which at least two different cell types were matched to the same factor

Factors	DMRs = 100	DMRs = 500
4	4	2
5	0	1
6	0	0

Example 40: Quantitation of T cell, Treg and CD16+CD56^{dim} NK cell numbers by CD3Z, FoxP3 and NKp46 methylation assays, respectively using droplet digital PCR

A droplet digital PCR technique was used to quantitate T cell, Treg and CD16+CD56^{dim} NK cell numbers using CD3Z, FoxP3 and NKp46 methylation assays described in Examples 15 and 30. Digital PCR (dPCR) is a refinement of conventional PCR methods and is used to directly quantify and clonally amplify nucleic acids. dPCR and traditional PCR differ in method

of measuring nucleic acid amounts, as dPCR is more precise. The two PCR methods differ in that the sample is separated into a large number of partitions in dPCR, and the reaction in each partition is carried out individually. This separation produces a more reliable collection and sensitive measurement of nucleic acid amounts.

Isolated and purified T cells and Tregs were serially diluted, and copies of each of the targets were quantified as measures of cell numbers. Bisulfite converted DNA from whole blood, isolated human T-cells and Treg cells and from NK cells was quantified using the emulsion partitioning method of BioRad QX100™ Droplet Digital™ PCR (ddPCR™) system. This system creates portioned PCR reaction using water-in-oil droplets for performing high-throughput digital PCR. The QX100 droplet generator partitions samples into 20,000 nanoliter-sized droplets. After PCR using a thermal cycler, droplets from the samples were streamed in single file on a reader (QX100 droplet reader). The PCR-positive and PCR-negative droplets were counted to obtain quantification of target DNA in digital form. Results are shown in Figures 43-46 as dot plots of fluorescence intensities of the droplets, with each point on the plot representing a single droplet. The horizontal lines are cutoffs between "positive" and "negative" droplets for each sample. A measure of concentration of the target sequence (demethylated CD3Z, Fox3P or NKp46) in copies per microliter was obtained as readout from the system. Dividing target sequence concentration by total DNA concentration obtained by C-less PCR yielded the percent of total DNA that was positive for the target DNA region (Figures 45-46).

Figures 43 and 44 show that successful amplification and detection of CD3Z and Foxp3 DMRs, respectively were obtained. Panel A of Figures 43 and 44 show dot plots indicating distinguishing of positive droplets and negative droplets. Panel B of Figures 43 and 44 show the calculated absolute numbers of positive PCR droplets. Results obtained from dilution of standard purified T cells shows correspondence of quantities of CD3Z and FoxP3 genes with extent of dilution and hence validity of dPCR as a detection method for methylation based assay of immune cell identity. Other partitioning approaches have been developed that employ microfluidic manipulation and results similar to the data obtained herein are expected from the use of such other methods of partitioning. Figure 45 shows quantitation of purified NK cells under different conditions and Figure 46 shows quantitation of whole blood and of purified leukocyte subsets by measuring demethylated NKp46 DMR described in Example 30.

What is claimed is:

1. A method for assessing a disease condition in a subject, comprising:

measuring a CD3Z positive T lymphocyte cell number in a sample from the subject by analyzing methylation in the sample of at least one CpG dinucleotide (CpG) in gene CD3Z or in an orthologous or a paralogous gene thereof, wherein an amount of a demethylated C of the at least one CpG in the sample is a measure of CD3+ T lymphocyte cell number; and

comparing the amount of the demethylated C in the sample from the subject with that in positive control samples from patients with the disease condition, and with that in negative control samples from healthy subjects, wherein the disease condition is selected from: an autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, immunosuppression and a cancer.

2. The method according to claim 1, wherein assessing a disease condition comprises at least one of: monitoring, diagnosing, prognosing, and measuring response to therapy by comparing the measured CD3+ T lymphocyte cell numbers in the subject after therapy to that in the patients with the disease condition and in the healthy subjects.

3. The method according to claim 1, wherein the sample a fresh sample.

4. The method according to claim 1, wherein the sample an archival sample.

5. The method according to claim 1, wherein the amount of the demethylated C of the at least one CpG in the CD3Z gene in the sample is at least about 80%, at least about 90%, or at least about 95% of the total amount of the CpG in CD3Z genes in the sample.

6. The method according to claim 1, wherein analyzing methylation of the CD3Z gene further comprises amplifying by Polymerase Chain Reaction (PCR) using primer pairs specific for amplification of specific demethylated CpG loci.

7. The method according to claim 1, wherein analyzing methylation of the CD3Z gene further comprises a method selected from the group of: Pyrosequencing, Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE), Methylation-sensitive single stranded conformation analysis (MS-SSCA), High resolution melting analysis (HRM), and digital PCR methods comprising emulsion and nanofluidic partitioning.

8. The method according to claim 6 wherein amplification by PCR comprises monitoring quantitative PCR in real time using a MethyLight assay or digital PCR.

9. The method according to claim 7 wherein Methylation-sensitive single-nucleotide primer extension further comprises:

chemically converting lymphocyte derived whole genomic DNA with bisulfite;

amplifying chemically converted whole genomic DNA;

enzymatically fragmenting resulting amplified DNA;

hybridizing fragmented DNA to methylation sensitive CpG locus specific DNA oligomers; and

labeling by single-base extension using labeled nucleotides.

10. The method according to claim 1, further comprising analyzing methylation of differentially methylated regions (DMRs) of gene FOXP3 using primer pairs for amplification of specific loci of demethylated CpG.

11. The method according to claim 10 further comprising:

determining a ratio of CpG demethylation of FOXP3 gene DMR to the CpG demethylation of CD3Z gene DMR, wherein the sample is a tumor infiltrate, and wherein the ratio is an index of T regulatory cell number to the total T cell number in the infiltrate; and

providing a diagnosis of a pathological grade of the cancer, wherein the index of T regulatory cell number to the total T cell number in the tumor infiltrate correlates with the grade of the cancer.

12. The method according to claim 11 wherein the cancer is selected from: a glioma; an ovarian cancer; and a head and neck squamous cell cancer (HNSCC).

13. The method according to claim 1 further comprising prognosing survival of a patient having or needing a diagnosis of glioma or HNSCC, wherein the amount of demethylation of CD3Z gene DMR as a percent of total DNA greater than a median value in a sample population of subjects correlates with a prognosis of poor survival.

14. A kit for measuring CD3+ T lymphocyte and FOXP3+ T regulatory cell numbers, by analyzing methylation of CpG positions in CD3Z and FOXP3 genes, the kit comprising

sequencing and PCR primers specific for the CD3Z and the FOXP3 gene DMRs and instructions for analyzing and comparing methylation of the CpG positions of a subject in need of diagnosis of a disease with that of control subjects.

15. A method for assessing a disease condition by estimating an alteration in proportions of types of leukocytes in a sample from a subject, the method comprising:

measuring a DNA methylation profile for each type of leukocyte and for unfractionated cells, wherein DNA methylation profiles are obtained for a plurality of CpG loci, and obtaining the status of an individual CpG locus by amplifying DNA from each of the types of leukocyte and from the unfractionated cells, wherein amplifying comprises hybridizing methylation sensitive locus-specific DNA oligomers corresponding to each CpG locus;

ordering CpG loci by ability to distinguish types of leukocytes, wherein the ordering of the CpG loci determines differentially methylated DNA regions (DMRs), wherein obtaining DMRs comprises statistically minimizing introduction of bias in amount of total methylation status of a large number of CpG loci obtained from the unfractionated cells by employing a Bayesian treatment utilizing prior probabilities of the methylation status at each individual locus, thereby identifying a plurality of CpG loci to include in the measurement, wherein an amount of CpG loci distinguishes DMR signatures among the types of leukocytes and minimizes bias;

obtaining DNA methylation profiles comprising DMRs from the types of leukocytes, wherein the DNA methylation profiles comprise validating measures of relative amounts of the types of leukocytes, and obtaining DNA methylation profiles of the unfractionated cells as surrogate measures of relative amounts of each type of leukocyte in the unfractionated cells;

employing an analog of a measurement error model wherein a DNA methylation surrogate y is reverse formulated with respect to the disease outcome *z*, as

$$y=f(z),$$

wherein y denotes a multivariate random variable representing a methylation profile, z denotes a disease outcome or state, and *f* denotes a probability distribution; y, z, and leukocyte distribution, ω are related by the estimator equations,

$$E(y|\omega)=g(\omega), \text{ and}$$

under an assumption $E=(z|\omega,y) = E(z|\omega)$, wherein E denotes an expectation of a random variable and ω denotes a subject specific distribution of leukocytes; and,

comparing relative amounts of each type of leukocyte in the sample from the subject with those in a control sample, thereby providing an assessment of the disease condition.

16. The method according to claim 15, wherein the locus-specific DNA oligomers are linked to an array selected from the group of: a glass slide array; a quartz slide array; a fiber optic bundle array, a planar slide array, a micro-well array; a multi-well dish array; a digital PCR array; and a bead array having beads located at known addressable locations on the array.

17. The method according to claim 15, wherein assessing a disease condition comprises at least one of: monitoring, diagnosing, prognosing, and measuring response to therapy of the disease condition.

18. The method according to claim 15, further comprising analyzing sensitivity for correcting bias, wherein the correcting bias is unrelated to measurement error and is related to errors arising from unprofiled cell types and non-cell mediated profile differences.

19. The method according to claim 15 wherein fractionated leukocyte types comprise at least one selected from: CD19+ B lymphocytes, CD15+ granulocytes, CD14+ monocytes, CD56+ Natural Killer cells, and CD3+ T lymphocytes.

20. The method according to claim 17 wherein the disease condition is Head and Neck Squamous Cell Carcinoma (HNSCC).

21. The method according to claim 1, wherein the inherited disease is an aneuploidy.

22. The method according to claim 21, wherein the aneuploidy is selected from trisomy 21, Turner's syndrome, and Klinefelter's syndrome.

23. The method according to claim 15, wherein the control sample is taken from the subject at a different point in time, for prognosis of the course of the disease condition in the subject.

24. The method according to claim 15, further comprising after employing the measurement model, comparing the distribution of leukocytes to the relative amounts in the control sample as a normal standard, wherein the normal standard is a statistical measure obtained from a plurality of disease-free subjects.

25. The method according to claim 10, further comprising providing a diagnosis of immunosuppression due to smoking in a currently smoking subject by:

determining a ratio of CpG demethylation of FOXP3 gene DMR to the CpG demethylation of CD3Z gene DMR in blood in the currently smoking subject, wherein the ratio comprises an index of T regulatory cell number to the total T cell number; and

providing a diagnosis of immunosuppression in the currently smoking subject, wherein the value of the index of T regulatory cell number to the total T cell number in the currently smoking subject, greater than the average value in a sample population of currently non-smoking subjects correlates with immunosuppression due to smoking.

26. The method according to claim 25, wherein the subject has cancer, an infection or need of a transplant.

27. A method of predicting a methylation class membership in a bodily fluid sample of a subject for assessing disease status of the subject, wherein the methylation class membership corresponds to an epigenetic signature of a plurality of leukocyte types, the method comprising:

measuring amounts of DNA methylation in each of a plurality of leukocyte type populations to determine differentially methylated regions (DMRs);

ranking leukocyte DMRs for each leukocyte type according to statistical strength of association of the DMR with each leukocyte type;

randomly dividing a data set of control subjects and subjects with a disease into groups having substantially the same numbers of control subjects and subjects with the disease to obtain a training set and a testing set;

clustering samples in the training set using a defined number of highest ranked leukocyte DMRs to determine clustering solutions, wherein a clustering solution corresponds to the methylation class membership; and

predicting the methylation class membership for subjects within the testing set by applying the clustering solutions obtained from the training set to the highest ranked leukocyte DMRs in the testing set, wherein clinical utility of the predicted methylation class membership is determined by testing association of the predicted methylation class membership with the disease status of the subject.

28. The method according to claim 27, wherein the highest ranked leukocyte DMRs is shown in Table 21, wherein each DMR is identified by chromosomal location and gene name , and the

defined number of highest ranked leukocyte DMRs is selected from: at least 10, at least 20, at least 30, at least 40 and 50.

29. The method according to claim 27, wherein the bodily fluid sample is a fresh sample.

30. The method according to claim 27, wherein the bodily fluid is an archival sample.

31. The method according to claim 27, wherein the bodily fluid sample is blood.

32. The method according to claim 27, wherein the methylation class membership of the subject in the testing set is predicted using a naïve Bayes classifier.

33. The method according to claim 27, wherein testing the association of the predicted methylation class with the disease status comprises using receiver operating characteristic curves (ROC) and the corresponding area under each curve.

34. The method according to claim 27, further comprising at least one of: diagnosing; monitoring; prognosing; and measuring response to therapy of the disease status of the subject.

35. The method according to claim 27 wherein the leukocyte types are selected from the group of: Natural killer cells, B Cells, CD4+ T cells, CD8+ T cells, granulocytes, and monocytes.

36. The method according to claim 27, wherein the disease is one of: head and neck squamous cell carcinoma (HNSCC), ovarian cancer and bladder cancer.

37. An array for estimating proportions of leukocyte types in a sample from a mammal for assessing a disease condition of the mammal by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the array comprising: a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, wherein the surface at each location is attached to a DNA probe having a specific nucleotide sequence, wherein the DNA probe having the specific nucleotide sequence hybridizes to a DNA sequence of a methylated form or an ummethylated form of a CpG dinucleotide in a sequence of a gene of the plurality of genes in the sample, wherein the array is selected from having: at least 16 probes, at least 64 probes, at least 96 probes, and at least 384 probes.

38. The array according to claim 37, wherein the plurality of DNA probes has nucleotide sequences that hybridize with a respective plurality of 96 different nucleotide sequences occurring in the plurality of genes.

39. The array according to claim 38, wherein the plurality of 96 nucleotide sequences comprises SEQ ID NO: 1 to SEQ ID NO: 96.

40. The array according to claim 37, wherein the addressable locations are wells of a substrate, wherein the substrate is selected from: glass slide; quartz slide; fiber optic bundle and planar silica slides.

41. The array according to claim 40, wherein the plurality of surfaces comprises particles added to the wells.

42. The array according to claim 40, wherein the surfaces comprise interior walls and sides of the wells.

43. The array according to claim 37, wherein the addressable locations are defined spots on a glass slide.

44. The array according to claim 41, wherein the particles are microbeads labeled with a code.

45. The array according to claim 41, wherein the particles are microbeads identifiable with inscribed holographic code.

46. The array according to claims 37, wherein the disease condition is selected from: autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, immunosuppression, and a cancer.

47. A method for estimating proportions of types of leukocytes in a sample from a subject for assessing a disease condition of the subject by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the method comprising:

providing an array having a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, wherein the surface at each location is attached to a DNA probe having a specific nucleotide sequence;

reacting genomic DNA in the sample with a bisulfite reagent to convert unmethylated cytosine residues to uracil;

hybridizing resulting bisulfite treated genomic DNA with the array to obtain resulting hybridized probes on the array, wherein the DNA probes hybridize to a DNA sequence of each of a methylated form and an ummethylated form of a sequence having a CpG dinucleotide in a gene for each of the plurality of genes; and

detecting the methylation status of each of the CpG dinucleotides in each sequence, thereby estimating proportions of types of leukocyte in the sample from the subject for assessing the disease condition of the subject.

48. The method according to claim 47, wherein detecting the methylation status of the CpG dinucleotide sequence further comprises:

extending each hybridized probe of the resulting hybridized probes on the array by primer extension to obtain a resulting primer extension product;

ligating the resulting primer extension product to an oligonucleotide complementary to the DNA sequence of a 3' region of the gene to obtain a resulting template for PCR on the array; and,

amplifying by PCR and measuring amount of resulting PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

49. The method according to claim 48, wherein amplifying by PCR further comprises:

using primers pairs having a 5' primer specific to each of the methylated or the unmethylated form of the CpG dinucleotide containing gene, and a 3'primer specific to the gene containing the CpG dinucleotide, thereby obtaining a first PCR product;

amplifying the first PCR product with differentially labeled 5' primers specific for each of the methylated and the unmethylated form of the CpG dinucleotide sequence containing gene, and a common 3' primer, thereby obtaining a differentially labeled second PCR product, and hybridizing the second PCR product to the CpG dinucleotide containing gene for measuring amount of the second PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

50. The method according to claim 47, wherein detecting the methylation status of the CpG dinucleotide sequence comprises extending the resulting hybridized probes on the array by single base primer extension with a labeled nucleotide.

51. The method according to claim 47, wherein the array comprises at least 16 probes, at least 64, at least 96 probes or at least 384 probes.

52. The method according to claim 47, wherein the plurality of probes on the array hybridizes with a respective plurality of 96 different sequences occurring in the plurality of genes.

53. The method according to claim 52, wherein each probe on the array is complementary to nucleotide sequences having SEQ ID NO: 1 to SEQ ID NO: 96.

54. The method according to claim 47, wherein the disease condition assessed is selected from: an autoimmune disease, an allergy, a transplant rejection, obesity, an inherited disease, and a cancer.

55. The method according to claim 47, wherein assessing the disease condition using the array comprises at least one of: monitoring, diagnosing, prognosing, and measuring response to therapy by comparing estimated proportions of types of leukocytes of the subject after therapy to proportions of leukocytes from a healthy subject.

56. The method according to claims 47, wherein the sample is fresh.

57. The method according to claim 47, wherein the sample is archival.

58. The method according to claim 47, wherein leukocyte types comprise at least one selected from: CD19+ B lymphocytes, CD15+ granulocytes, CD14+ monocytes, CD56+ Natural Killer cells, and CD3+ T lymphocytes.

59. A kit for estimating proportions of leukocyte types in a sample from a subject by analyzing differential methylation of CpG dinucleotides in a plurality of genes of the sample, the kit comprising:

an array comprising: a plurality of DNA probes attached to a plurality of surfaces at known addressable locations on the array, wherein the surface at each location is attached to a DNA probe having a specific nucleotide sequence, wherein the DNA probe having the specific nucleotide sequence hybridizes to a DNA sequence of a methylated form or an ummethylated form of a CpG dinucleotide in a sequence of a gene of the plurality of genes in the sample,

wherein the array is selected from having: at least 16 probes, at least 64 probes, at least 96 probes, and at least 384 probes;

primers and reagents for detecting the hybridized probes and for detecting the reaction products derived from the hybridized probes; and

instructions for using the array with a bisulfite reagent, thereby providing an estimation of proportions of leukocyte types in the sample.

60. The kit according to claim 59, wherein the probes hybridize with a respective plurality of 96 different DNA sequences occurring in the plurality of genes.

61. The kit according to claim 59 wherein, the probes have nucleotide sequences complementary to SEQ ID NO: 1 to SEQ ID NO: 96.

62. The kit according to claim 59, wherein the instructions comprise methods for:

reacting genomic DNA in the sample with the bisulfite reagent to convert unmethylated cytosine residues to uracil;

hybridizing resulting bisulfite treated genomic DNA with probes immobilized to the surfaces to obtain resulting hybridized probes on the array, wherein the DNA probes hybridize to a DNA sequence of each of a methylated form and an ummethylated form of a CpG dinucleotide sequence in a gene of the plurality of genes; and

detecting the methylation status of the CpG dinucleotide sequence, thereby estimating proportions of leukocyte types in the sample from the subject for assessing the disease condition.

63. The kit according to claim 62, wherein the instructions further comprise:

extending each hybridized probe of the resulting hybridized probes on the array by primer extension to obtain a resulting primer extension product;

ligating the resulting primer extension product to an oligonucleotide complementary to the DNA sequence of a 3' region of the gene to obtain a resulting template for PCR on the array; and

amplifying by PCR and measuring amount of resulting PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

64. The kit according to claim 63, wherein the instructions further describe methods for amplifying by PCR comprising:

amplifying the resulting template on the array using primers pairs comprising a 5' primer specific to each of the methylated or the unmethylated form of the CpG dinucleotide containing gene, and a 3'primer specific to the gene containing the CpG dinucleotide, thereby resulting in a first PCR product;

amplifying the resulting first PCR product with differentially labeled 5' primers that specifically amplify either the methylated or the unmethylated form of the CpG dinucleotide sequence containing gene, and a common 3' primer, resulting in a differentially labeled second PCR product, and hybridizing the second PCR product to the CpG dinucleotide containing gene for measuring amount of the second PCR product, thereby detecting the methylation status of the CpG dinucleotide sequence.

65. The kit according to claim 62, wherein the instructions further describe methods for detecting the methylation status of the CpG dinucleotide sequence by extending the resulting hybridized probes on the array by single base primer extension with a labeled nucleotide.

66. A method of treating a subject for a disease condition, wherein the subject is a human patient and wherein the disease condition is a cancer, the method comprising:

obtaining signatures comprising differentially methylated regions (DMRs) from types of leukocytes in a blood sample of the patient, the types of leukocytes comprising at least one selected from: CD19+ B lymphocyte, CD15+ granulocyte, CD14+ monocyte, $CD56^{dim}$ Natural Killer cell, $CD56^{bright}$ Natural Killer cell, and CD3+ T lymphocyte; and from a healthy control human subject not having the cancer;

comparing a signature for a specific type of leukocyte in the patient with that in the healthy subject, wherein the signature for the specific type of leukocyte is an indication of amount of cells of the specific type of leukocyte circulating in blood, and wherein a decreased amount of the cells of the specific type of leukocyte circulating in the blood of the patient compared to the healthy subject is an indicium of the cancer; and,

administering a composition comprising the cells of the type of leukocyte to the patient, thereby increasing the amount of the cells of the type of leukocyte in the patient and treating the cancer.

67. The method according to claim 66, wherein the leukocyte type cell is the $CD56^{dim}$ Natural Killer cell.

68. The method according to claim 66 or 67, wherein the cancer is head and neck squamous cell carcinoma (HNSCC).

69. The method according to claim 67, wherein the DMR signature specific for $CD56^{dim}$ Natural Killer cells comprises at least one CpG dinucleotide in a region near the promoter of gene *NKp46*.

70. The method according to claim 67, wherein the DMR signature specific for $CD56^{dim}$ Natural Killer cells comprises a CpG dinucleotide in a region near the promoter of the gene *NKp46*, wherein the methylation status of the CpG dinucleotide is quantified by methylation specific quantitative polymerase chain reaction (MS-qPCR) using primers and probes having SEQ ID NOs: 116-118 and 97-99.

71. The method according to claim 67, wherein the DMR signature specific for $CD56^{dim}$ Natural Killer cells is a CpG dinucleotide in a region near the promoter of the gene *NKp46*, wherein the methylation status of the CpG dinucleotide is quantified by digital PCR comprising emulsion and nanofluidic partitioning using primers and probes having SEQ ID NOs: 116-118 and 97-99.

72. The method according to claim 66, wherein the blood sample is archival.

73. The method according to claim 66, wherein the blood sample is fresh.

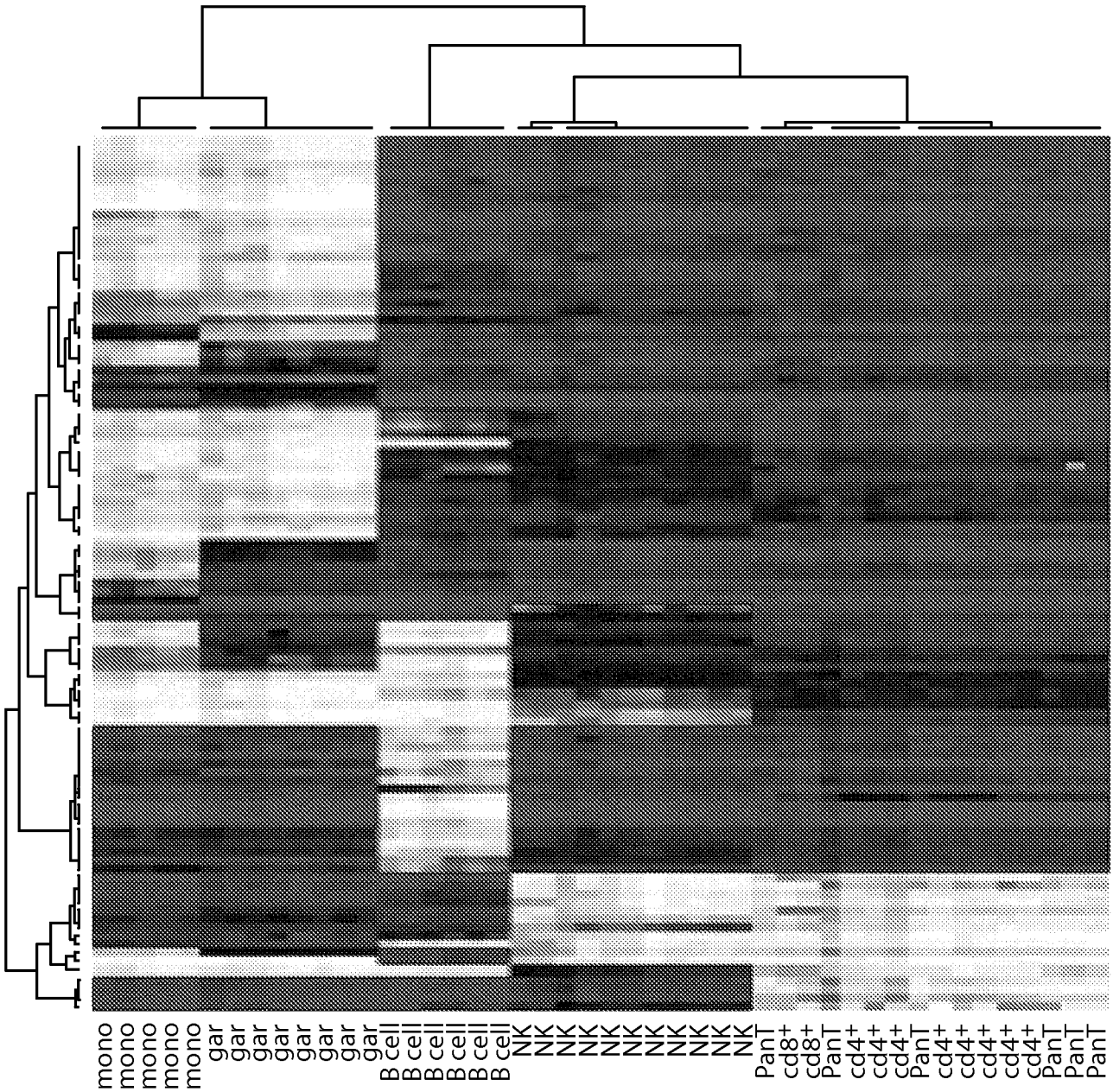

mono
mono
mono
mono
mono
gar
gar
gar
gar
gar
gar
gar
B cell
B cell
B cell
B cell
B cell
NK
NK
NK
NK
NK
NK
NK
NK
NK
PanT
cd8+
cd8+
PanT
cd4+
cd4+
cd4+
PanT
cd4+
cd4+
cd4+
cd4+
cd4+
PanT
PanT
PanT

Figure 1

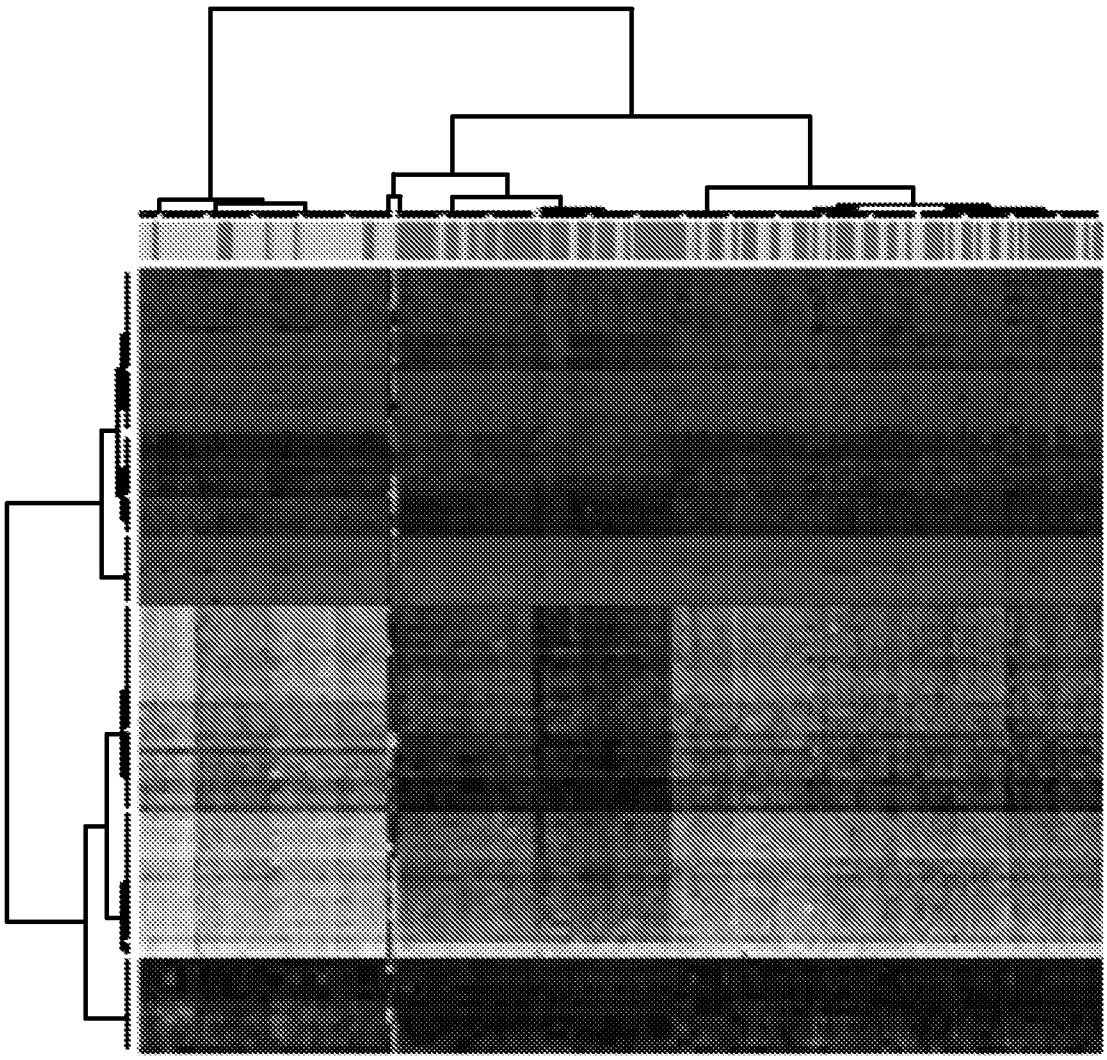
Expected

Monocyte	90	80	60	40	20	10	0	0	0	0	0	0
B Cell	10	20	40	60	80	90	0	0	0	0	0	0
NK	0	0	0	0	0	0	0	0	0	0	0	0
T Cell	0	0	0	0	0	0	10	20	40	60	80	90
Granulocyte	0	0	0	0	0	0	90	80	60	40	20	10

Observed

Monocyte	86	77	57	37	18	9	0.5	0	0	0	0	2
B Cell	10	20	37	56	74	86	0	0	0	0	0	0
NK	2	2	2	0	0	0	1	3	6	5	5	2
T (CD8+)	0	0	0	0	0	0	2	0	0	1	9	2
T (CD4+)	0	0	0	0	0	0	13	23	40	56	66	83
Granulocyte	3	4	7	10	10	7	82	73	54	38	20	10

Figure 2



Heatmap of HNSCC data (S_1)

Figure 3

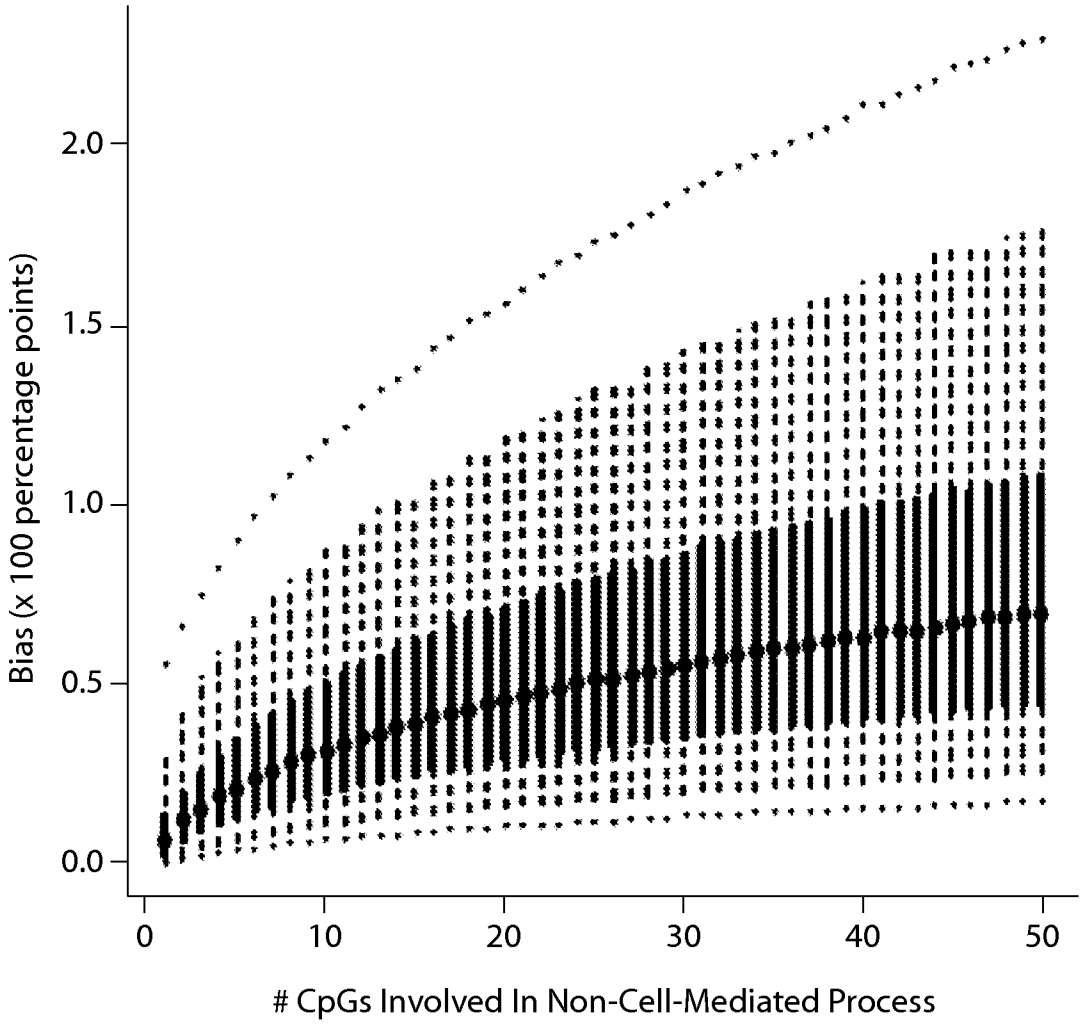

Bias (x 100 percentage points)
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
CpGs Involved In Non-Cell-Mediated Process

Figure 4

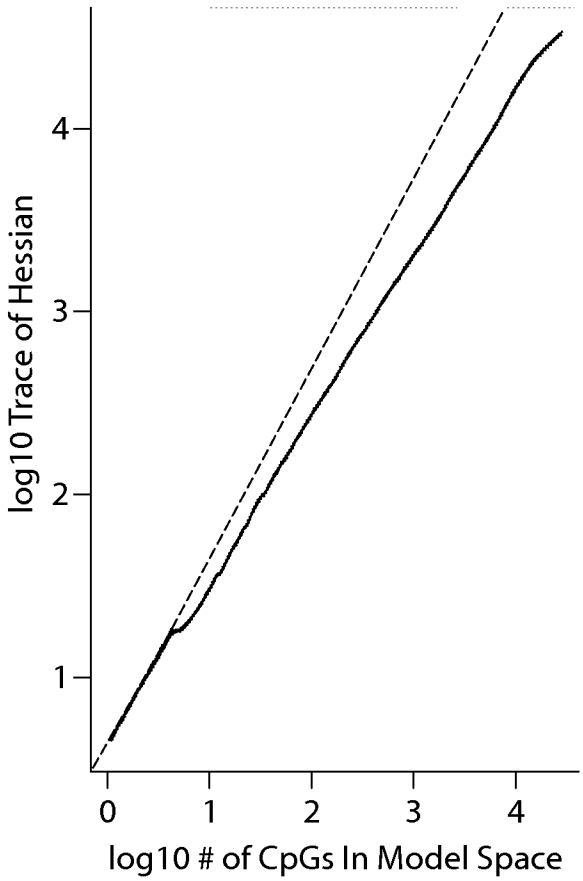


Figure 5A

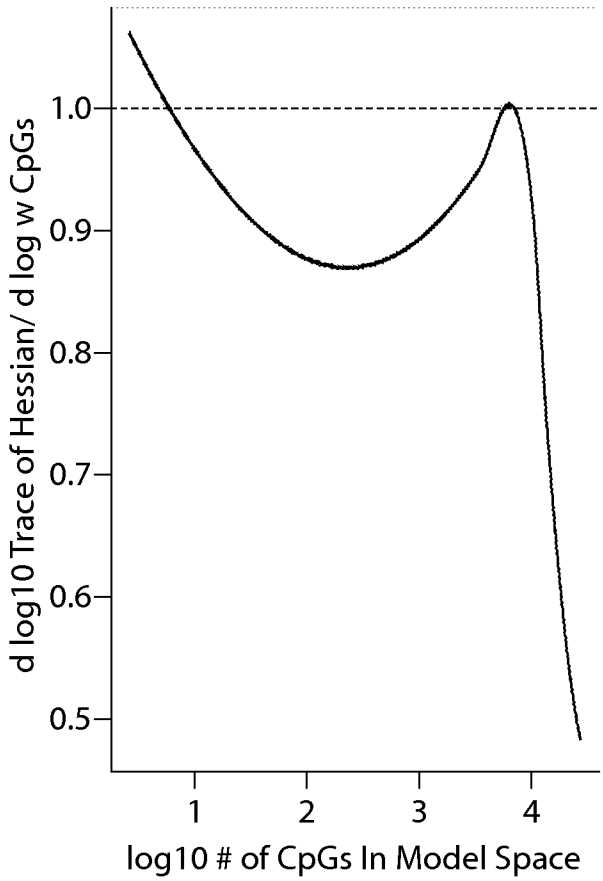
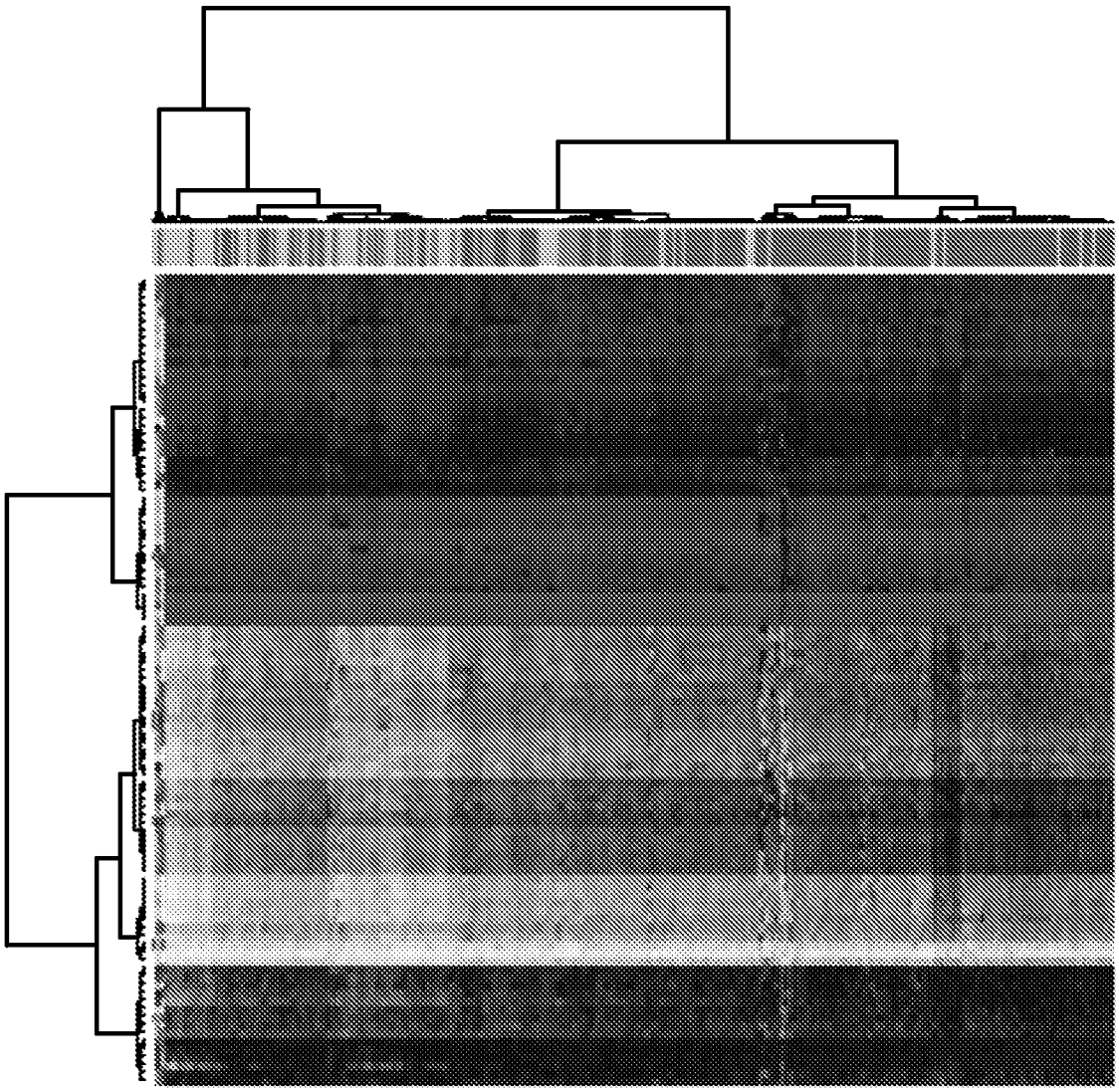
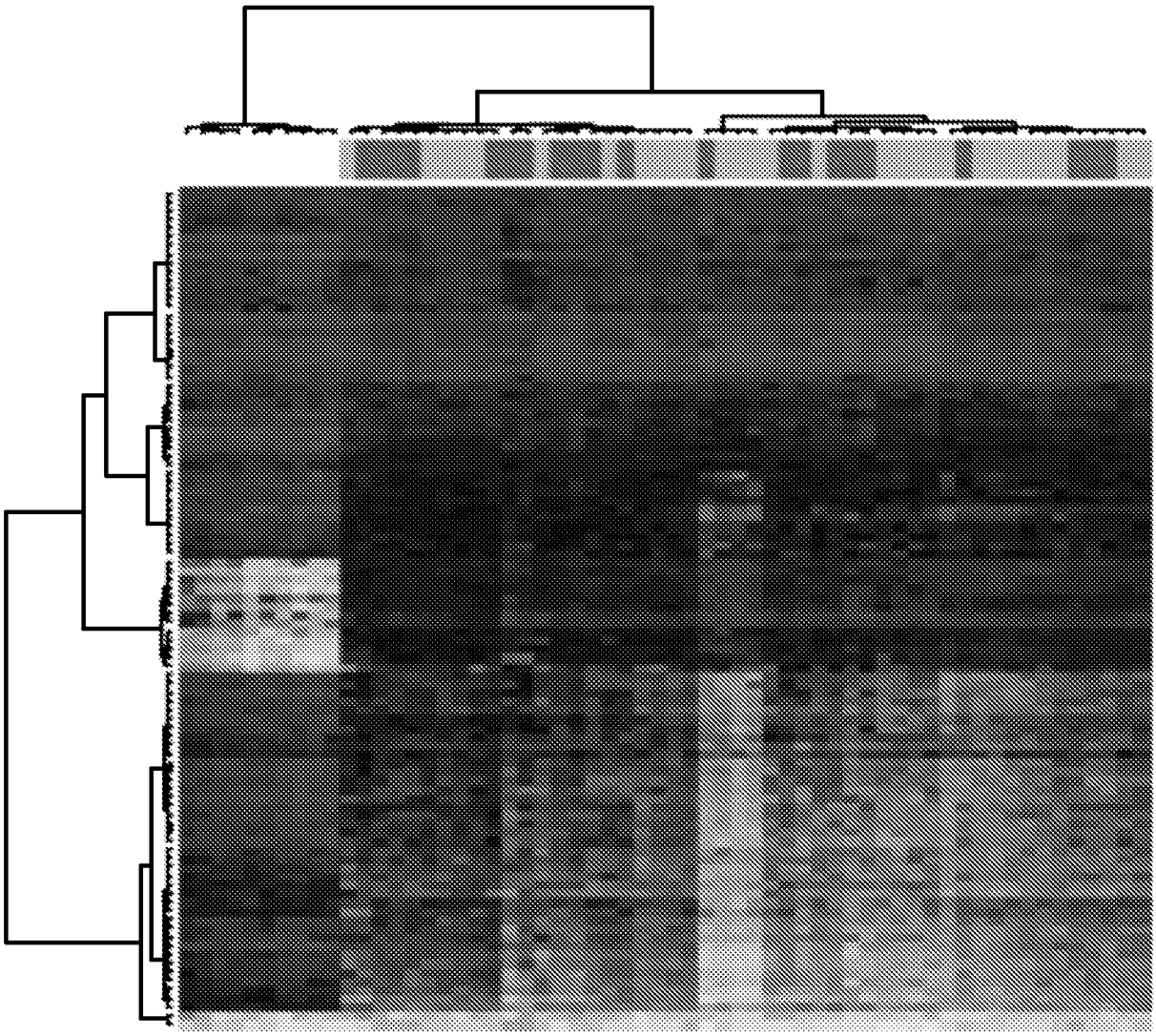


Figure 5B



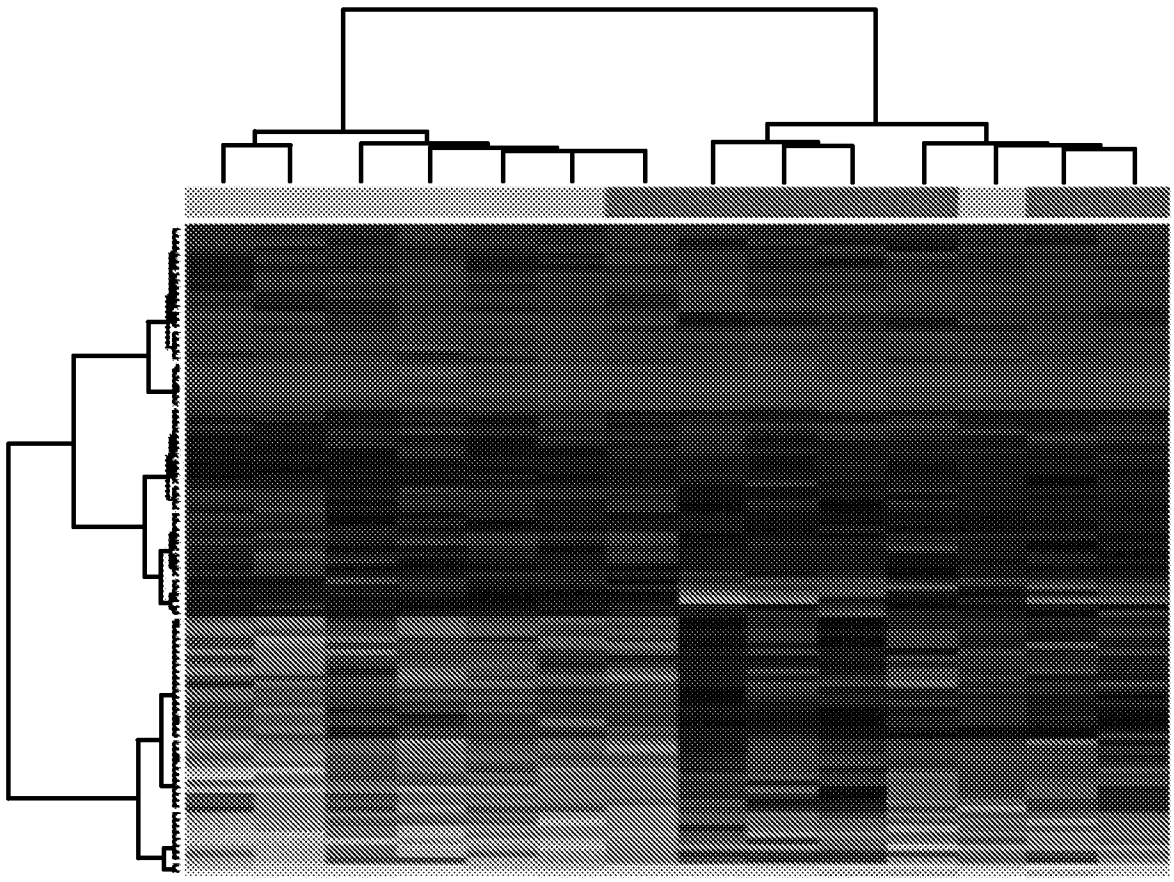
Heatmap of Ovarian cancer data (S_1)

Figure 6



Heatmap of Down Syndrome data (S_1)

Figure 7



Heatmap of Obesity data (S_1)

Figure 8

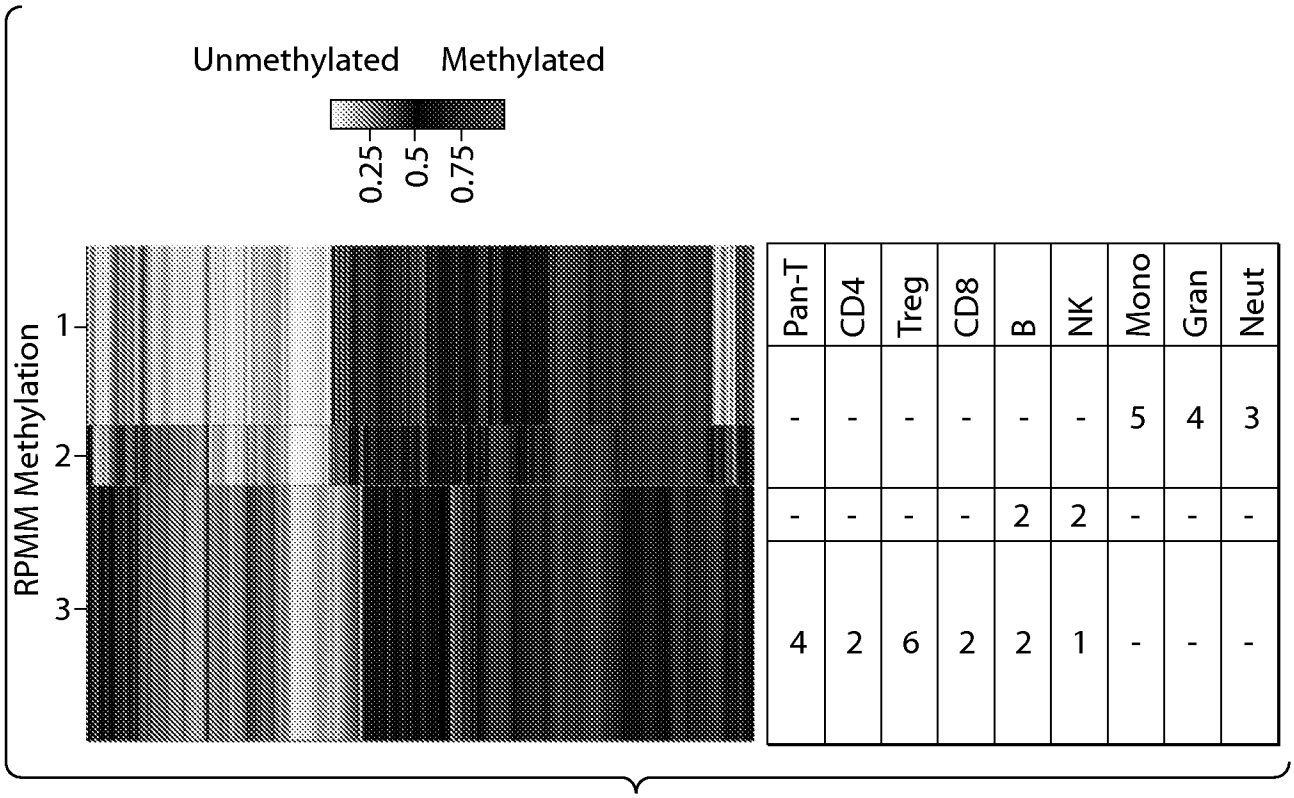

Unmethylated
Methylated
0.25
0.5
0.75
RPMM Methylation
1
2
3
Pan-T
CD4
Treg
CD8
B
NK
Mono
Gran
Neut
- - - - - - 5 4 3
- - - - 2 2 - - -
4 2 6 2 2 1 - - -

Fig. 9

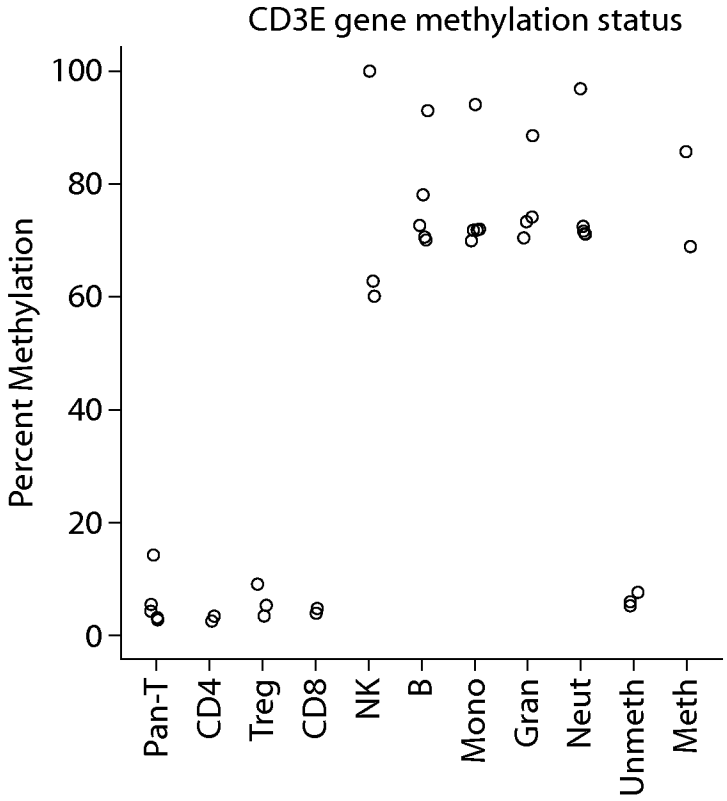


Figure 10A

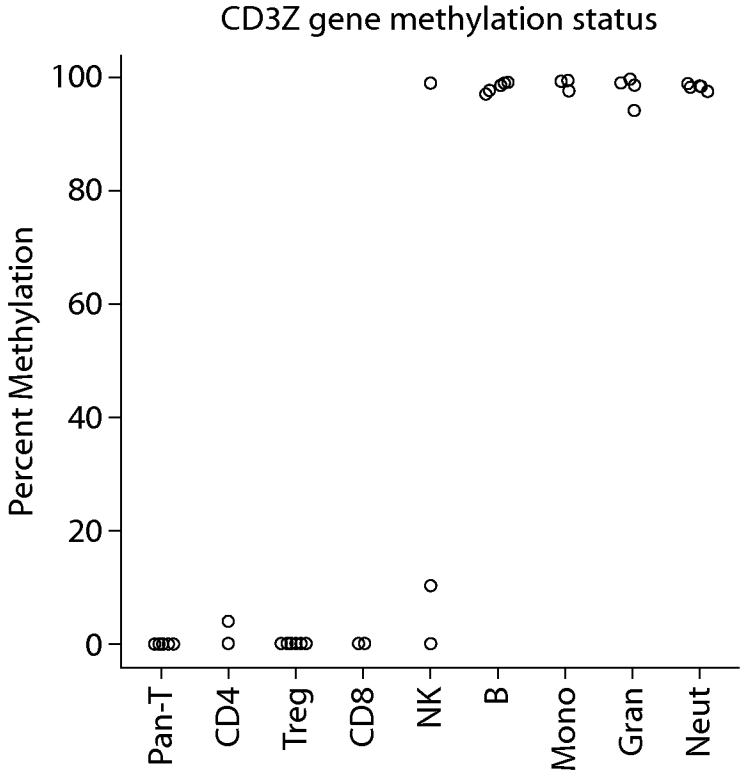


Figure 10B

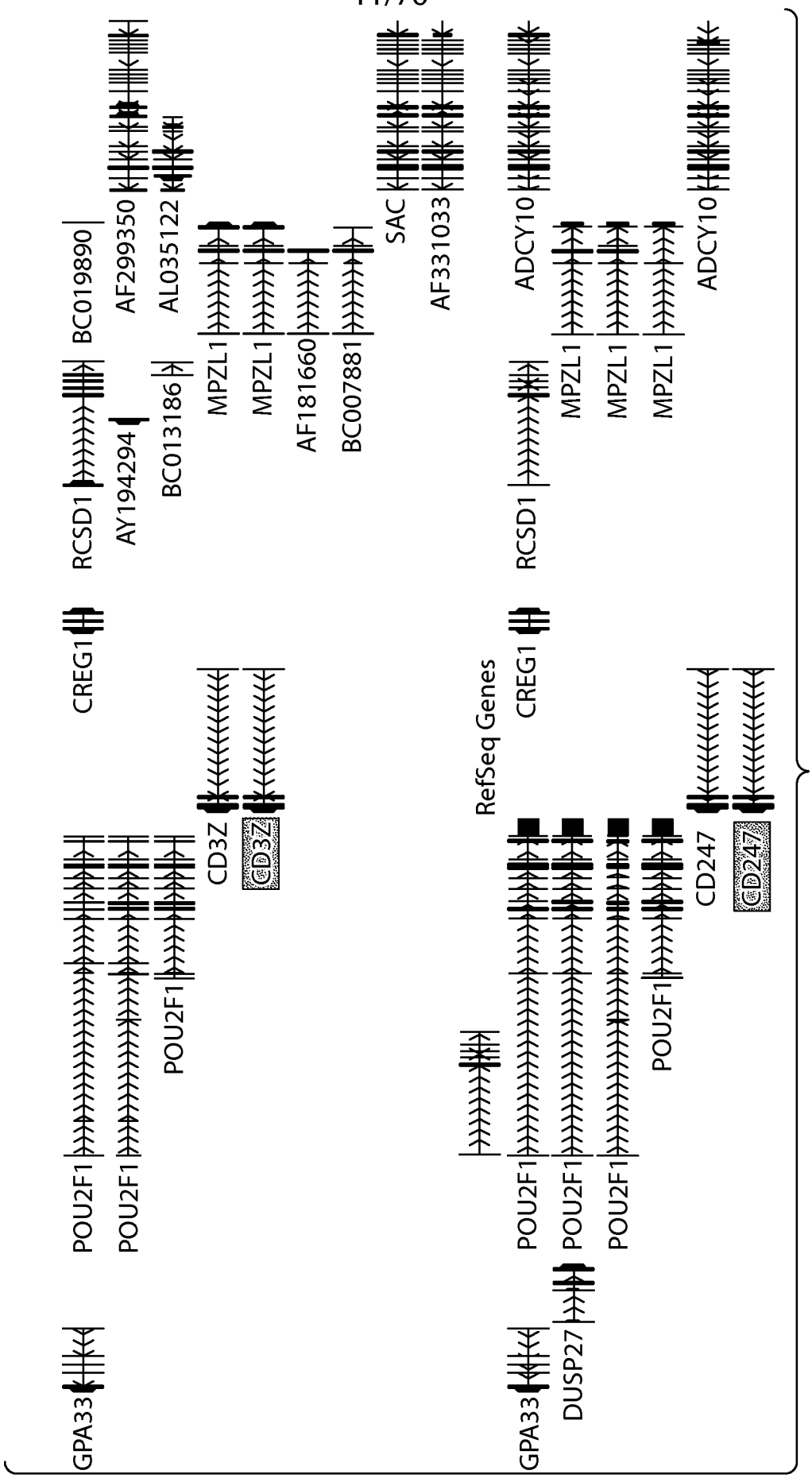

GPA33
POU2F1
POU2F1
CREG1
RCSD1
AY194294
BC013186
POU2F1
MPZL1
MPZL1
AF181660
BC007881
CD3Z
CD3Z
BC019890
AF299350
AL035122
SAC
AF331033
ADCY10
ADCY10
RefSeq Genes
GPA33
DUSP27
POU2F1
POU2F1
POU2F1
CREG1
RCSD1
MPZL1
MPZL1
MPZL1
POU2F1
CD247
CD247

Figure 11

FOXP3, CD3Z and C-Less Primer Sequences.
Underlined letters are "C" within CpG motifs.

Primer name	Chromosome location	Wildtype Sequence 5'-3'	Bisulfite sequence	Amplicon bp
FOXP3 qMSP Forward	chrX:49004142-49004170	CATCTGGGCCCTGTTGTCACAGCCCCCG (SEQ ID NO: 108)	TATTTGGGTTTTGTTGTTATAGTTTTTG (SEQ ID NO: 106)	109 (Lower Strand)
FOXP3 qMSP Reverse	chrX:49004227-49004250	CGACACCACGGAGGAAGAGAAGAG (SEQ ID NO: 109)	CTCTTCTCTTCCTCCATAATATCA (SEQ ID NO: 107)	(Lower Strand)
FOXP3 qMSP Probe-Deme	chrX:49004200-49004218	ATGGCGGCCGGATGCGCCG (SEQ ID NO: 110)	CAACACATCCAACCACCAT (SEQ ID NO: 105)	(Lower Strand)
Reference: Examples 13-21				
CD3ZU5 Forward	chr1:165754293-165754313	GGATGGCCGCGGTGAAAAGCG (SEQ ID NO: 111)	GGATGGTTGTGGTGAAAAGTG (SEQ ID NO: 100)	117 (Lower Strand)
CD3ZU3 Reverse	chr1:165754386-165754409	CGGTTAGGAGAAAAGGAGTCTCTG (SEQ ID NO: 112)	CAAAAACTCCTTTTCTCCTAACCA (SEQ ID NO: 101)	(Lower Strand)
CD3ZUProbe	chr1:165754365 165754383	CTGAGGCAGCGGTGGCCGG (SEQ ID NO: 113)	CCAACCACCACTACCTCAA (SEQ ID NO: 102)	(Lower Strand)
Reference: Examples 13-21				
C-Less Forward	chr20:19199387-19199412	TTGTATGTATGTGAGTGTGGGAGAGA (SEQ ID NO: 97)	TTGTATGTATGTGAGTGTGGGAGAGA (SEQ ID NO: 97)	69 (Lower Strand)
C-Less Reverse	chr20:19199434-19199455	GGAAGAGAAGGGGTGGAAGAAA (SEQ ID NO: 114)	TTTCTTCCACCCCTTCTCTTCC (SEQ ID NO: 98)	(Lower Strand)
C-Less Probe	chr20:19199414-19199431	ATAGAGTTAGAGGGGGAG (SEQ ID NO: 115)	CTCCCCCTCTAACTCTAT (SEQ ID NO: 99)	(Lower Strand)
Reference: Daniel J. Weisenberger, Mihaela Campan, et al., Nucleic Acids Research 2005, 33:6823-36.				

Figure 12

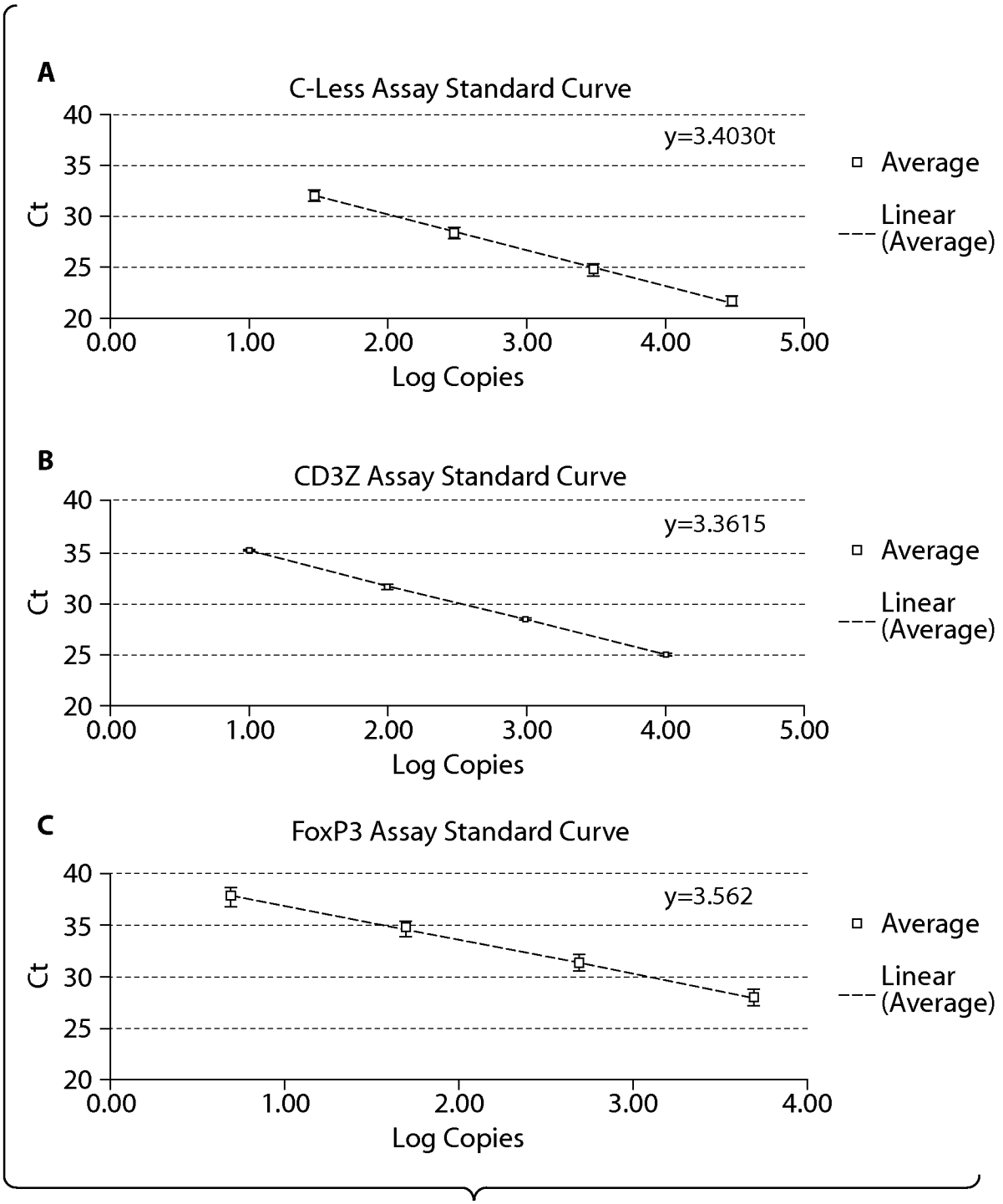

A
C-Less Assay Standard Curve
y=3.4030t
Ct
40
35
30
25
20
0.00
1.00
2.00
3.00
4.00
5.00
Log Copies
Average
Linear (Average)
B
CD3Z Assay Standard Curve
y=3.3615
Ct
40
35
30
25
20
0.00
1.00
2.00
3.00
4.00
5.00
Log Copies
Average
Linear (Average)
C
FoxP3 Assay Standard Curve
y=3.562
Ct
40
35
30
25
20
0.00
1.00
2.00
3.00
4.00
Log Copies
Average
Linear (Average)

Figure 13

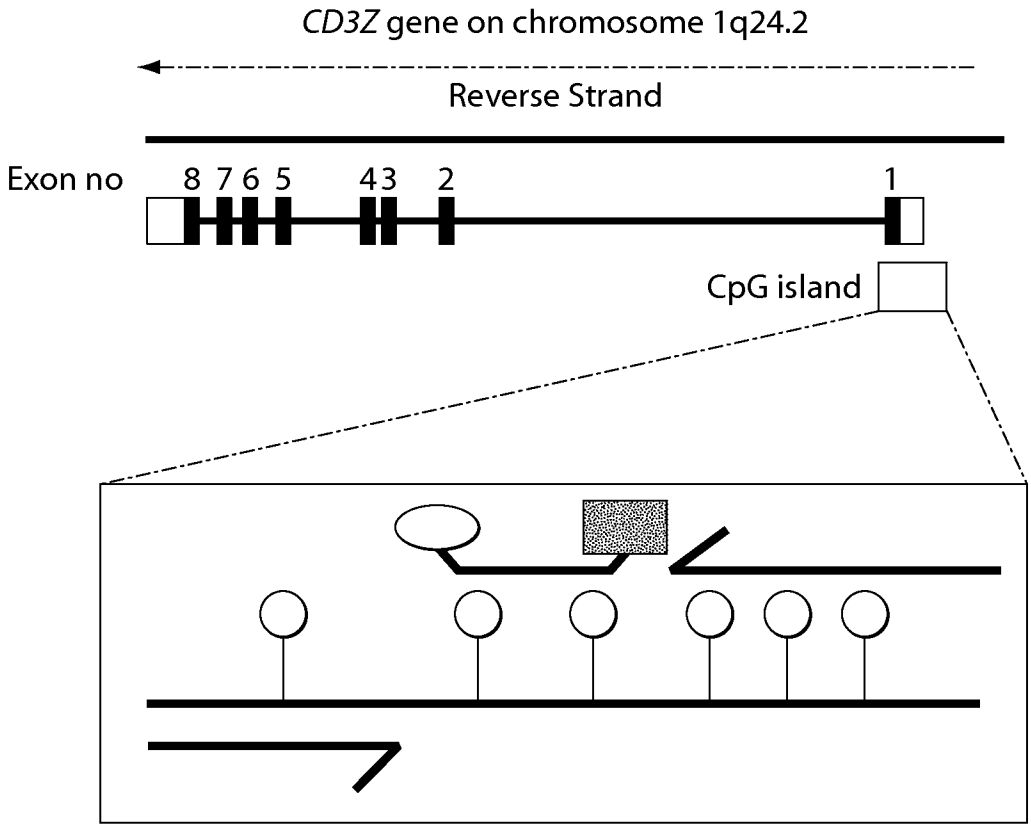


Figure 14A

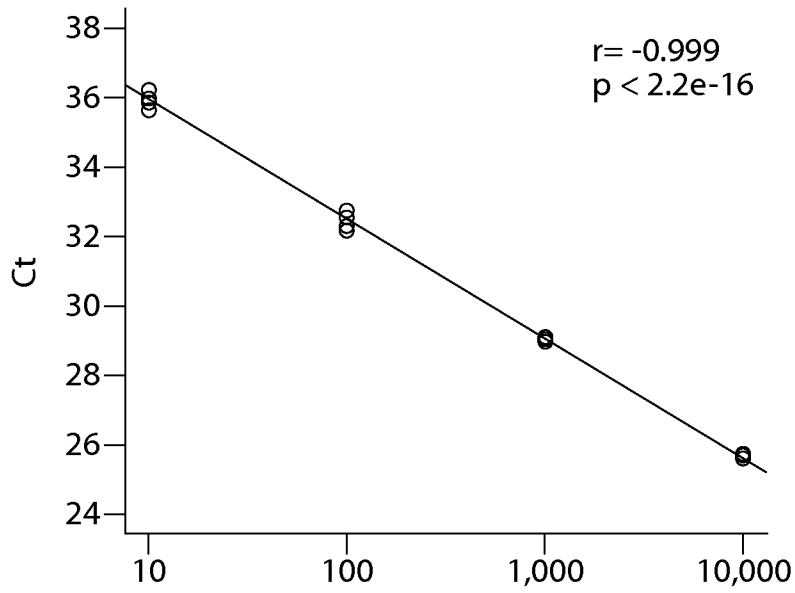


Figure 14B

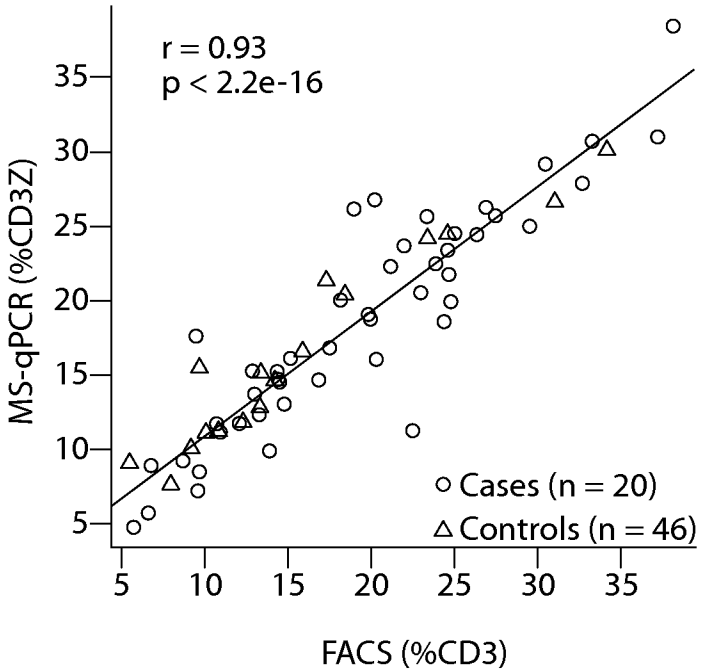


Figure 14C

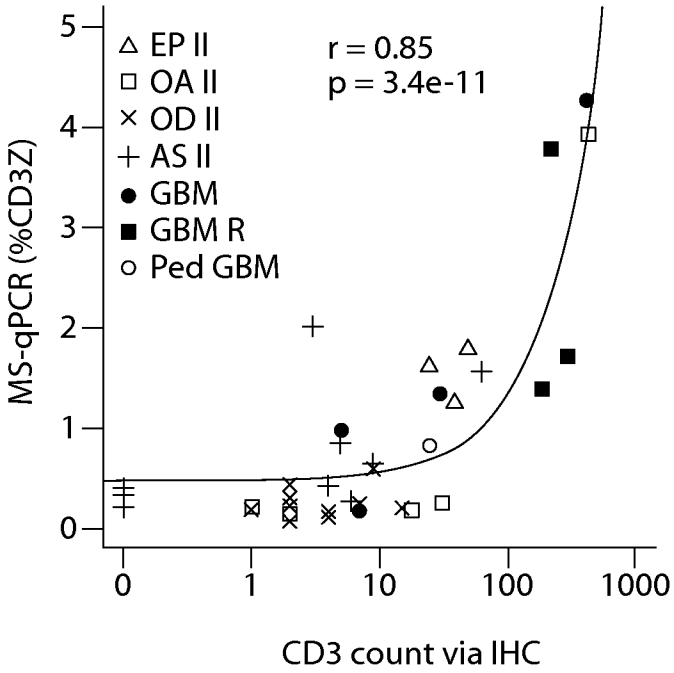


Figure 14D

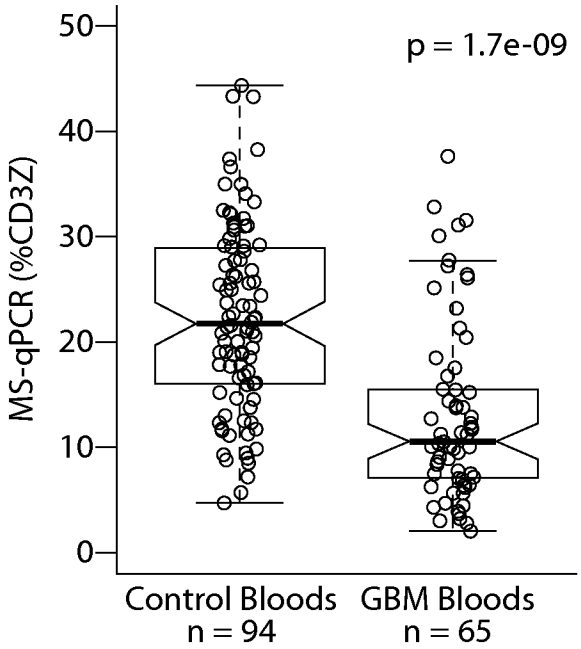


Figure 15A

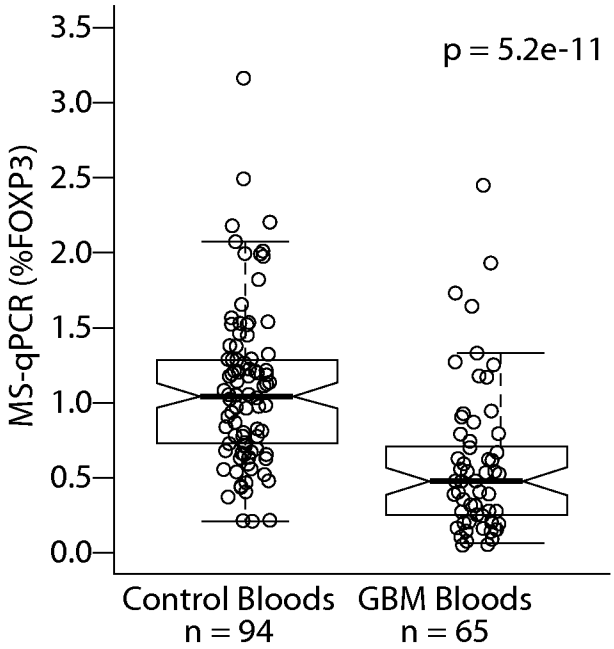


Figure 15B

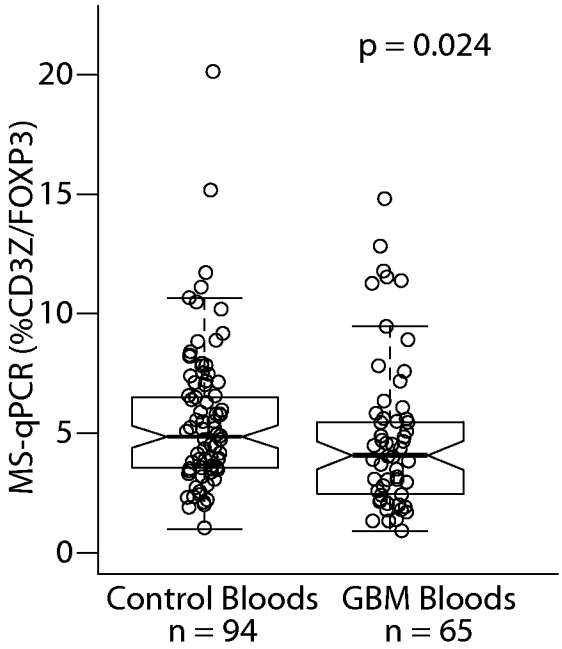


Figure 15C

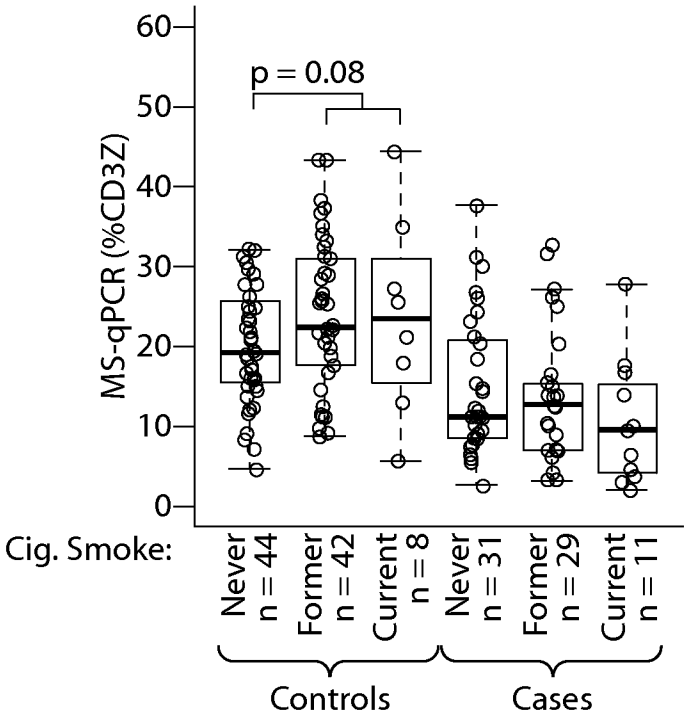


Figure 16A

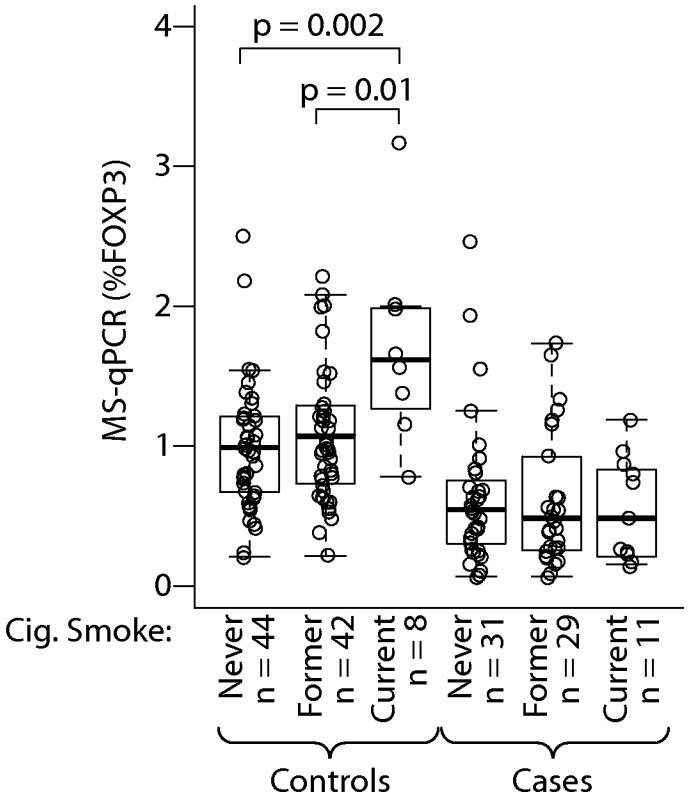


Figure 16B

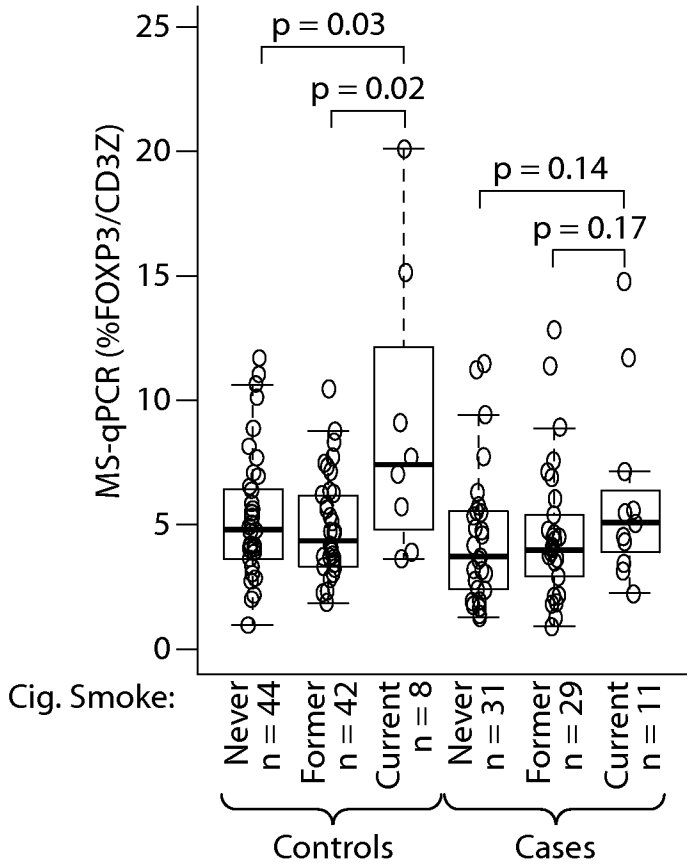

p = 0.03
p = 0.02
p = 0.14
p = 0.17
MS-qPCR (%FOXP3/CD3Z)
25
20
15
10
5
0
Cig. Smoke:
Never n = 44
Former n = 42
Current n = 8
Never n = 31
Former n = 29
Current n = 11
Controls
Cases

Figure 16C

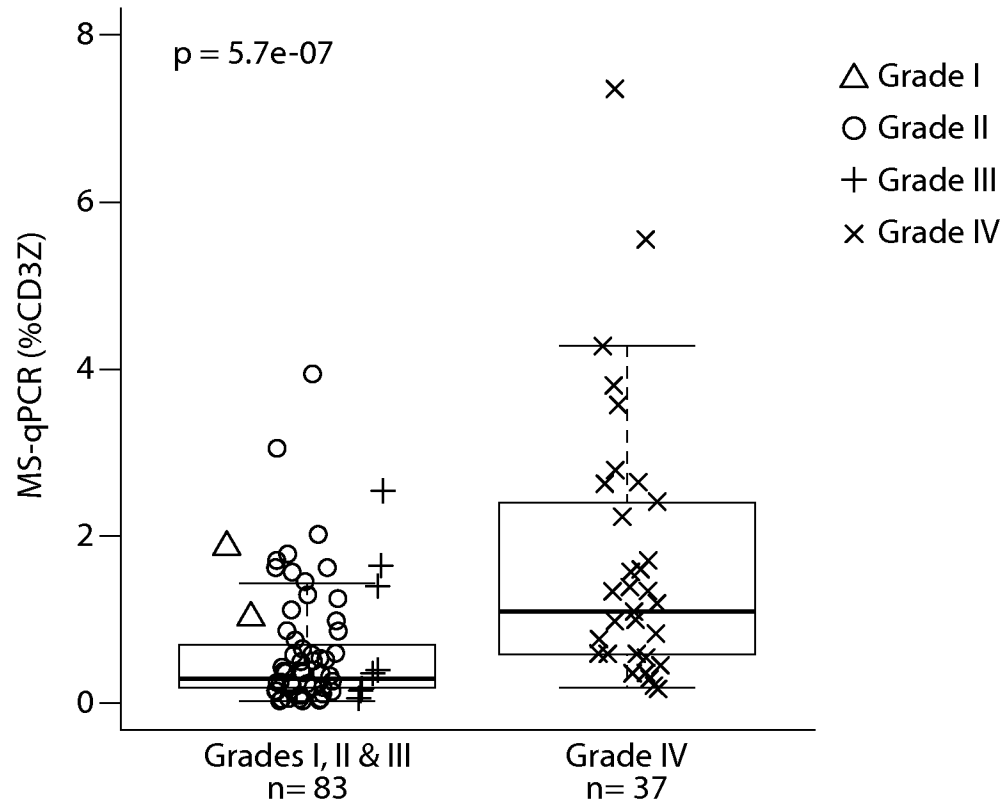


Figure 17A

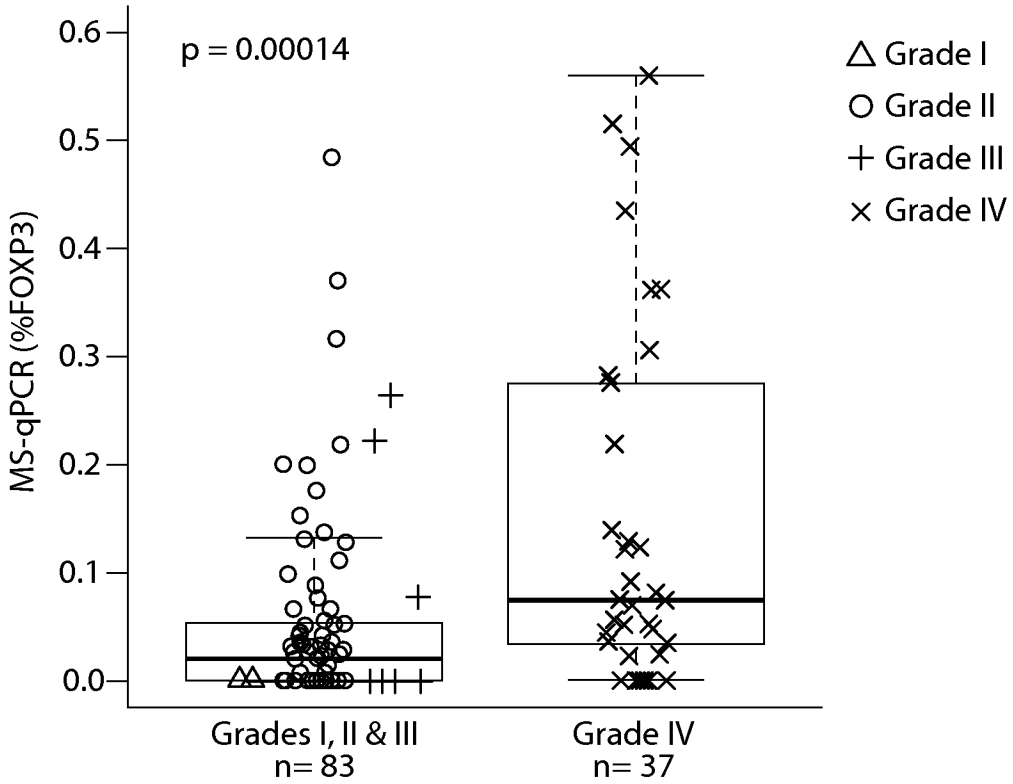


Figure 17B

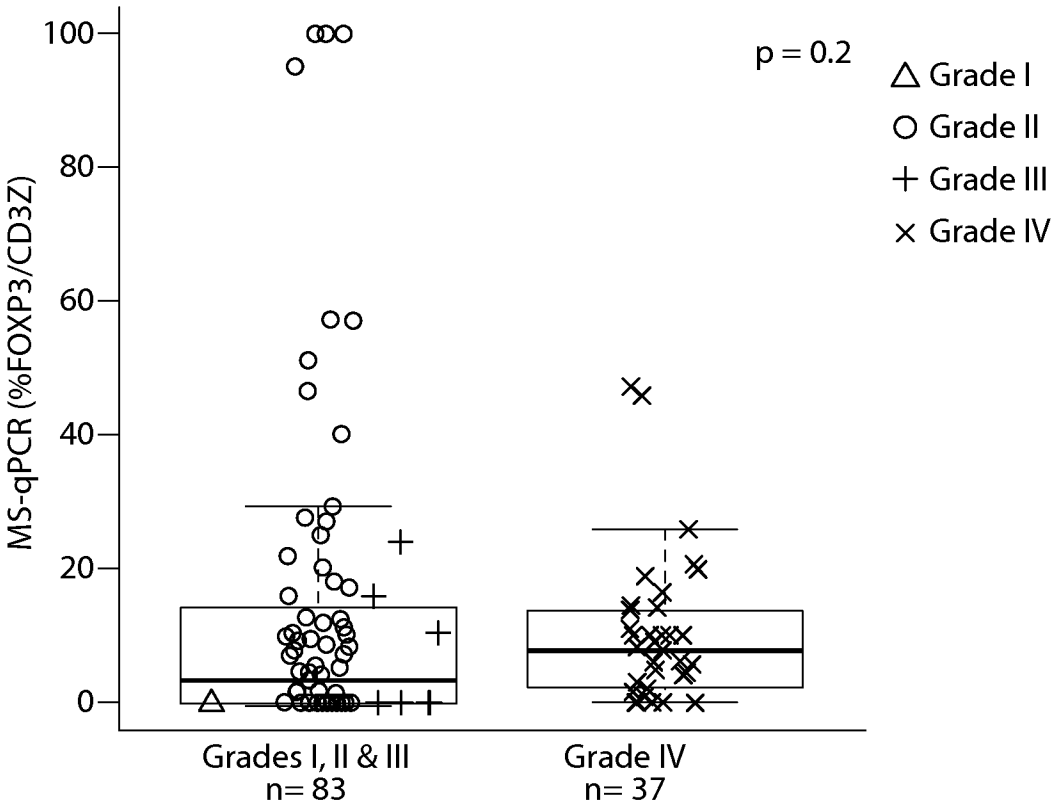

100
80
60
40
20
0
MS-qPCR (%FOXP3/CD3Z)
p = 0.2
Grade I
Grade II
Grade III
Grade IV
Grades I, II & III
n= 83
Grade IV
n= 37

Figure 17C

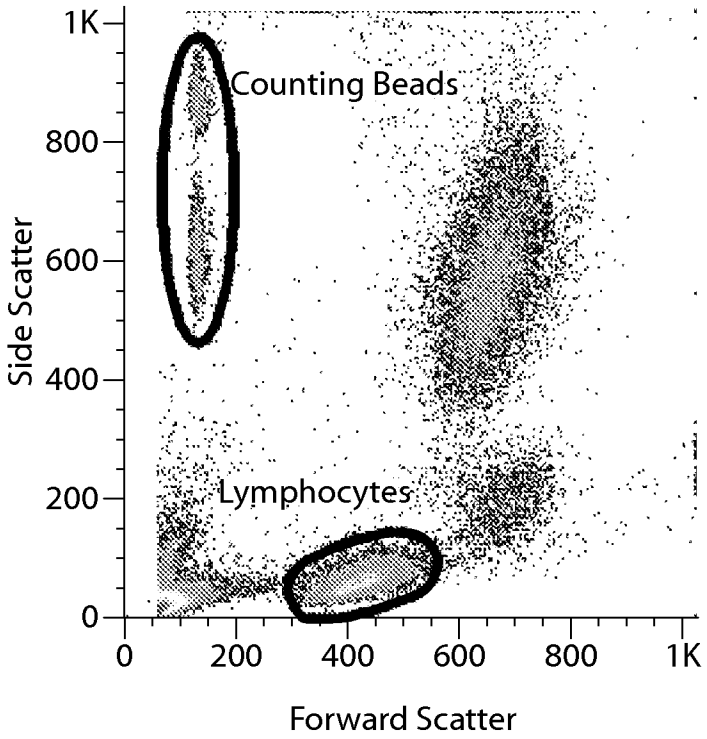


Figure 18A

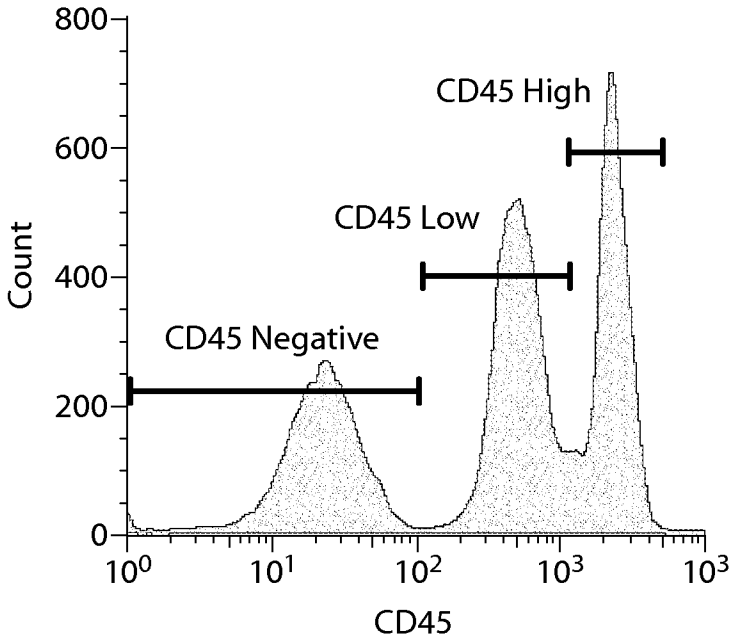


Figure 18B

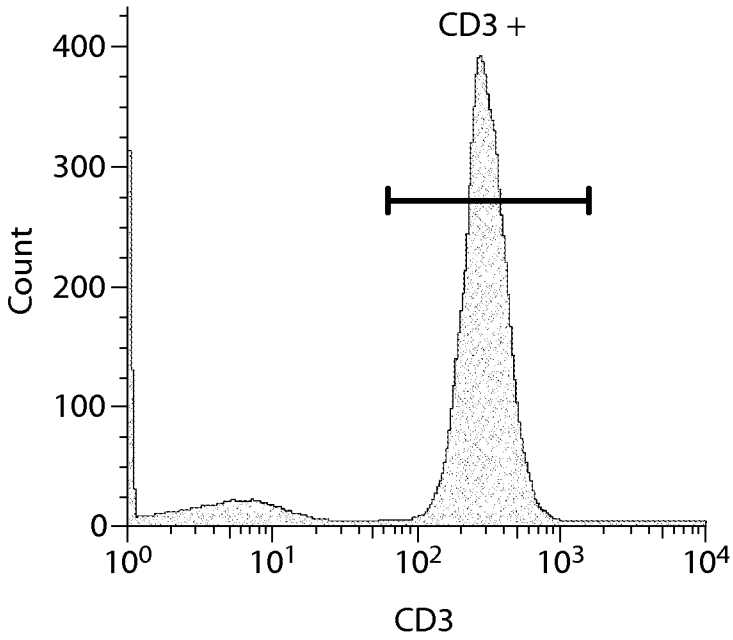


Figure 18C

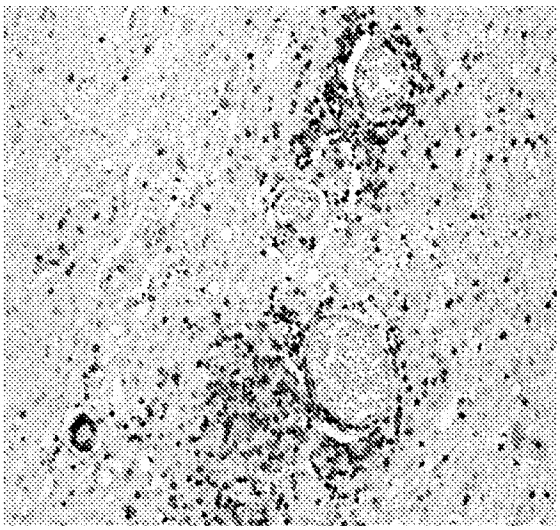


Figure 19A

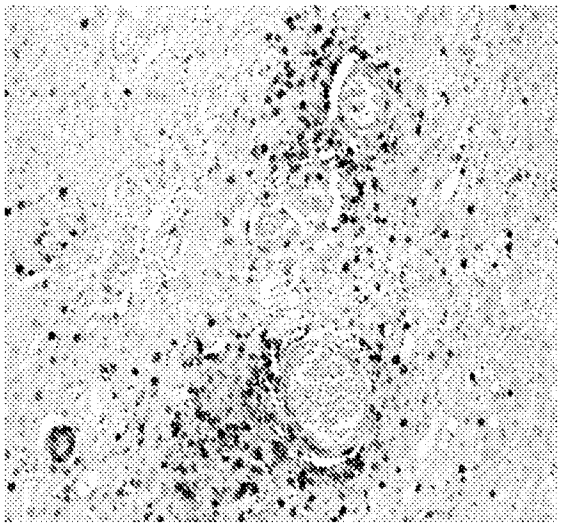


Figure 19B

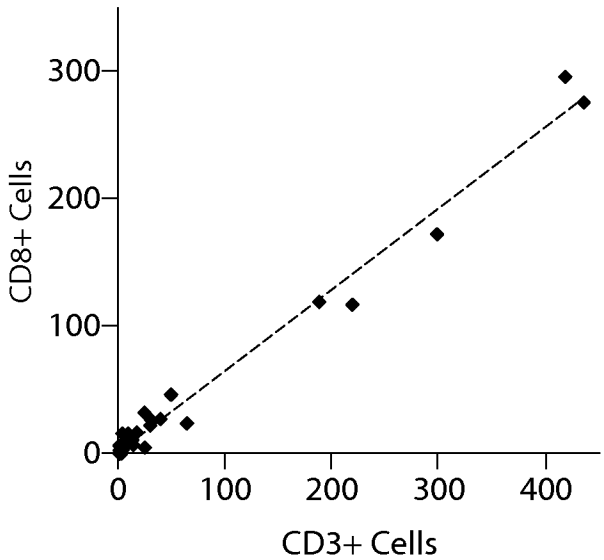


Figure 19C

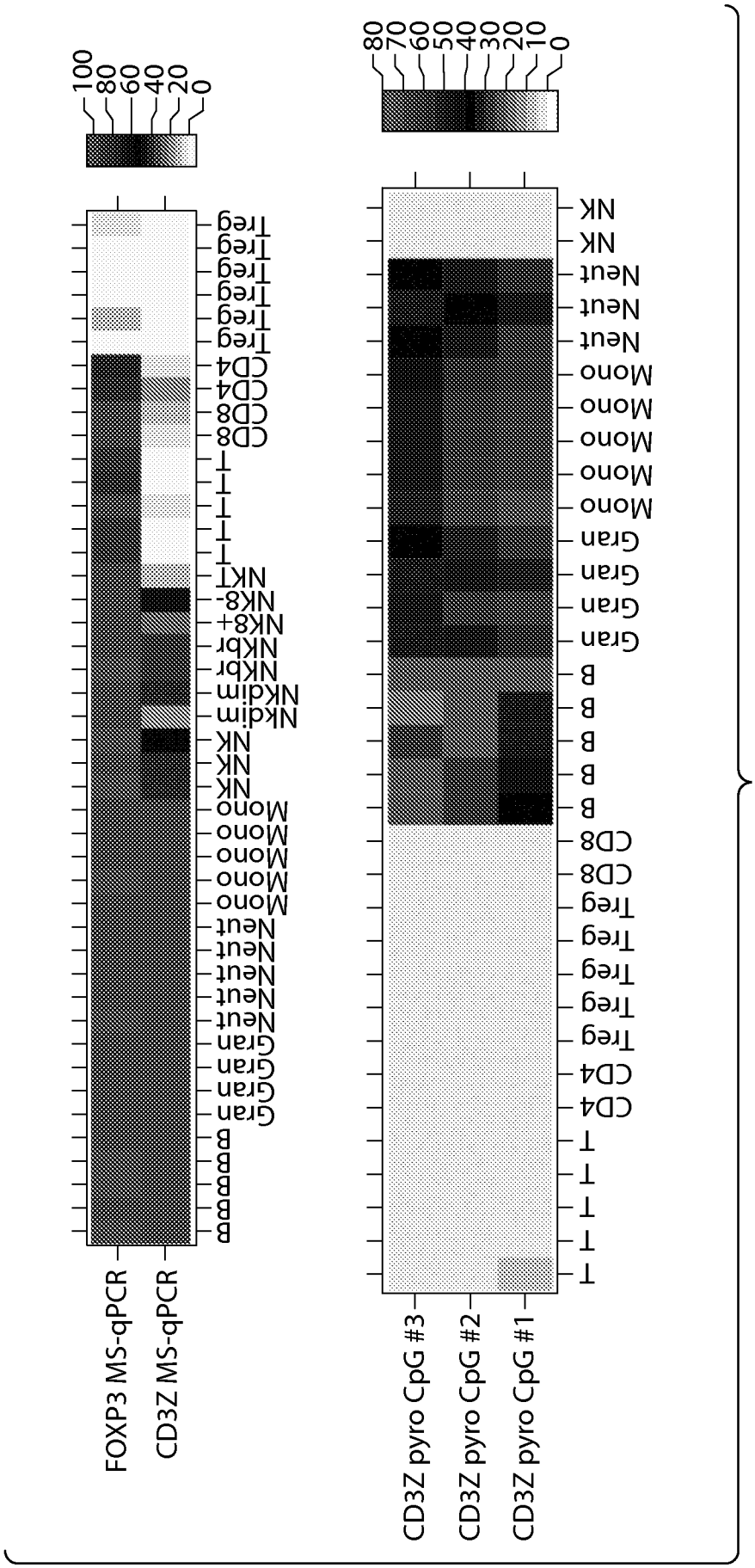

100
80
60
40
20
0
Treg
Treg
Treg
Treg
Treg
Treg
CD4
CD4
CD8
CD8
T
T
T
T
NKT
NK8-
NK8+
NKbr
NKbr
NKdim
NKdim
NK
NK
NK
Mono
Mono
Mono
Mono
Mono
Neut
Neut
Neut
Neut
Neut
Gran
Gran
Gran
B
B
B
B
B
FOXP3 MS-qPCR
CD3Z MS-qPCR
80
70
60
50
40
30
20
10
0
NK
NK
Neut
Neut
Neut
Mono
Mono
Mono
Mono
Mono
Gran
Gran
Gran
Gran
B
B
B
B
B
CD8
CD8
Treg
Treg
Treg
Treg
Treg
CD4
CD4
T
T
T
T
T
CD3Z pyro CpG #3
CD3Z pyro CpG #2
CD3Z pyro CpG #1

Figure 20

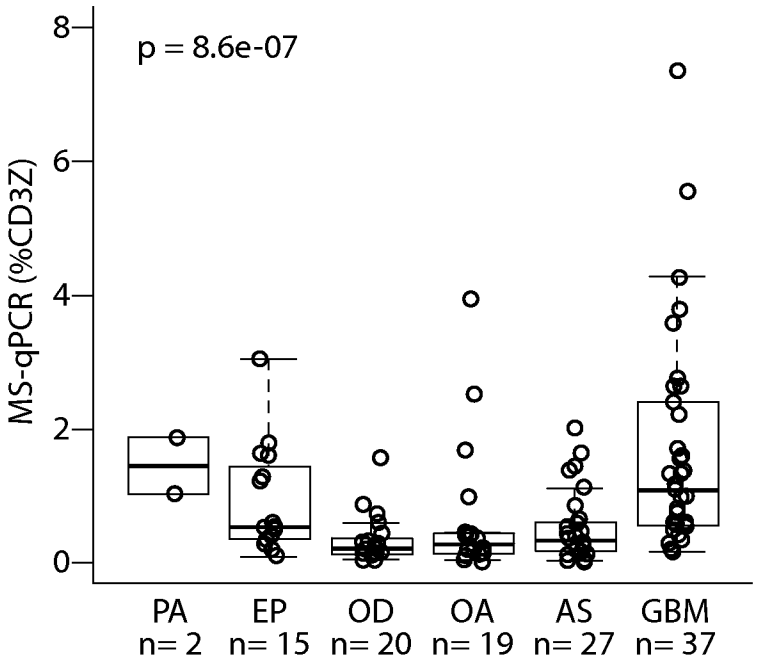


Figure 21A

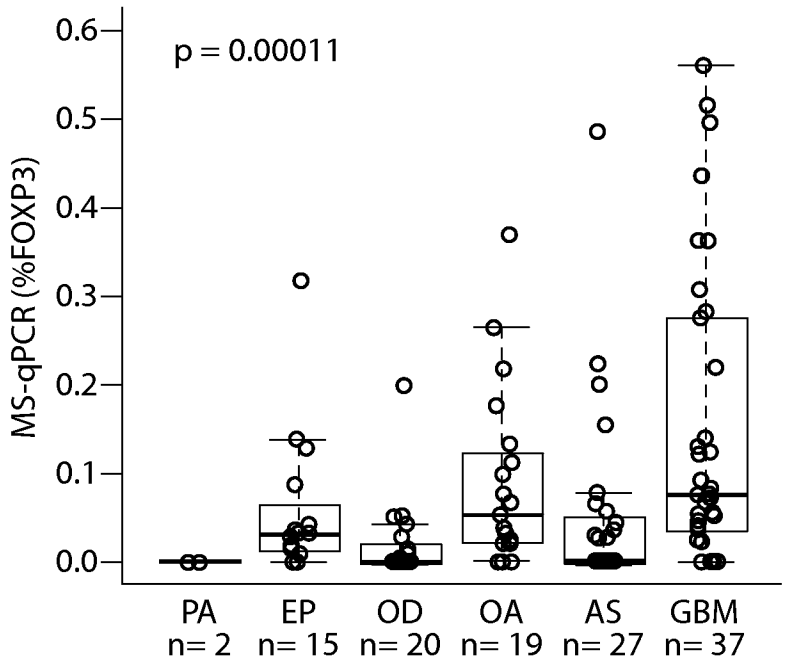


Figure 21B

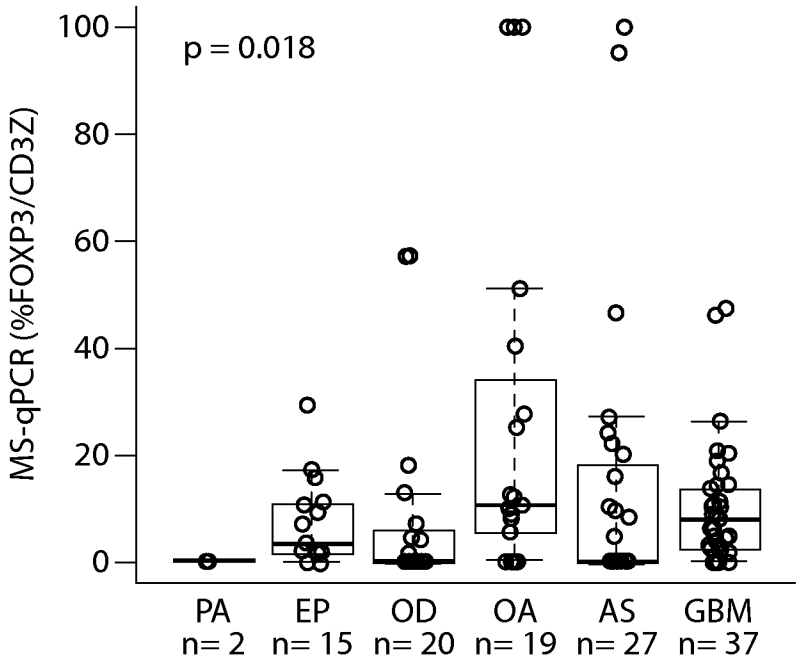

p = 0.018
MS-qPCR (%FOXP3/CD3Z)
100
80
60
40
20
0
PA
n= 2
EP
n= 15
OD
n= 20
OA
n= 19
AS
n= 27
GBM
n= 37

Figure 21C

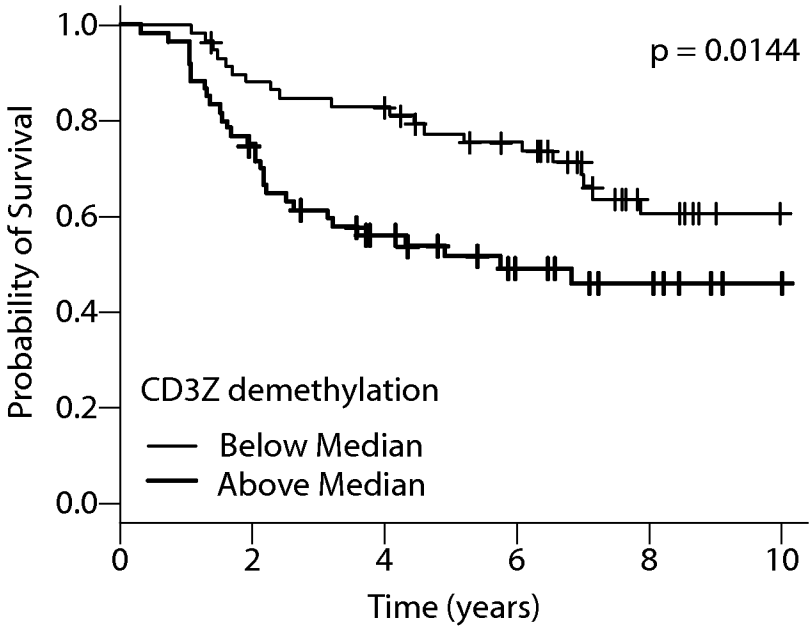


Figure 22A

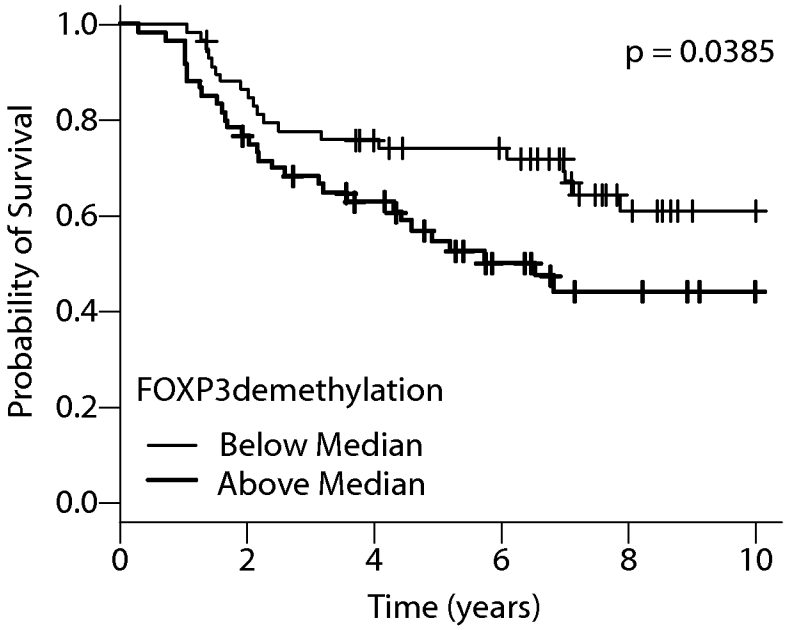


Figure 22B

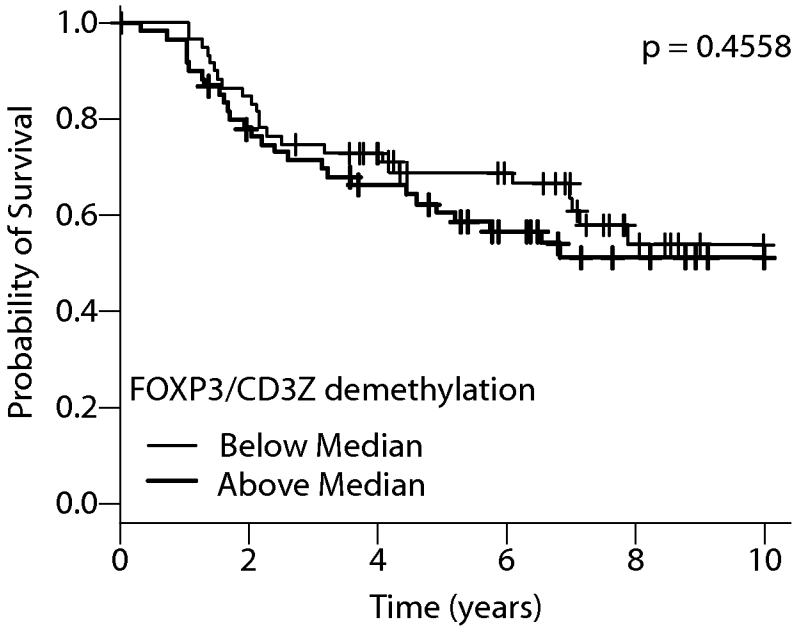

1.0
0.8
0.6
0.4
0.2
0.0
Probability of Survival
p = 0.4558
FOXP3/CD3Z demethylation
Below Median
Above Median
0
2
4
6
8
10
Time (years)

Figure 22C

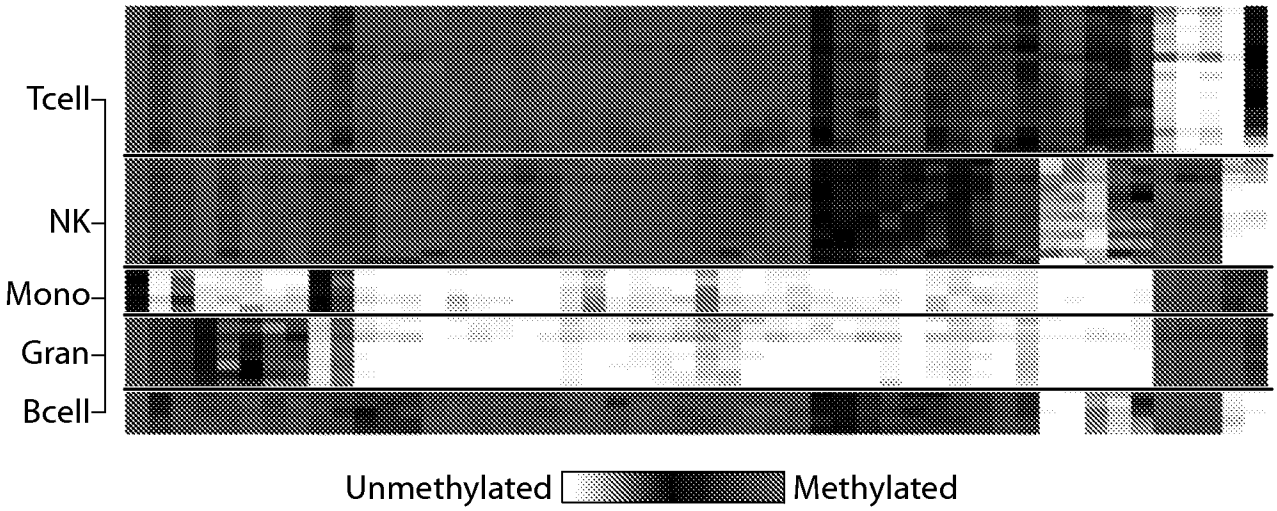


Figure 23A

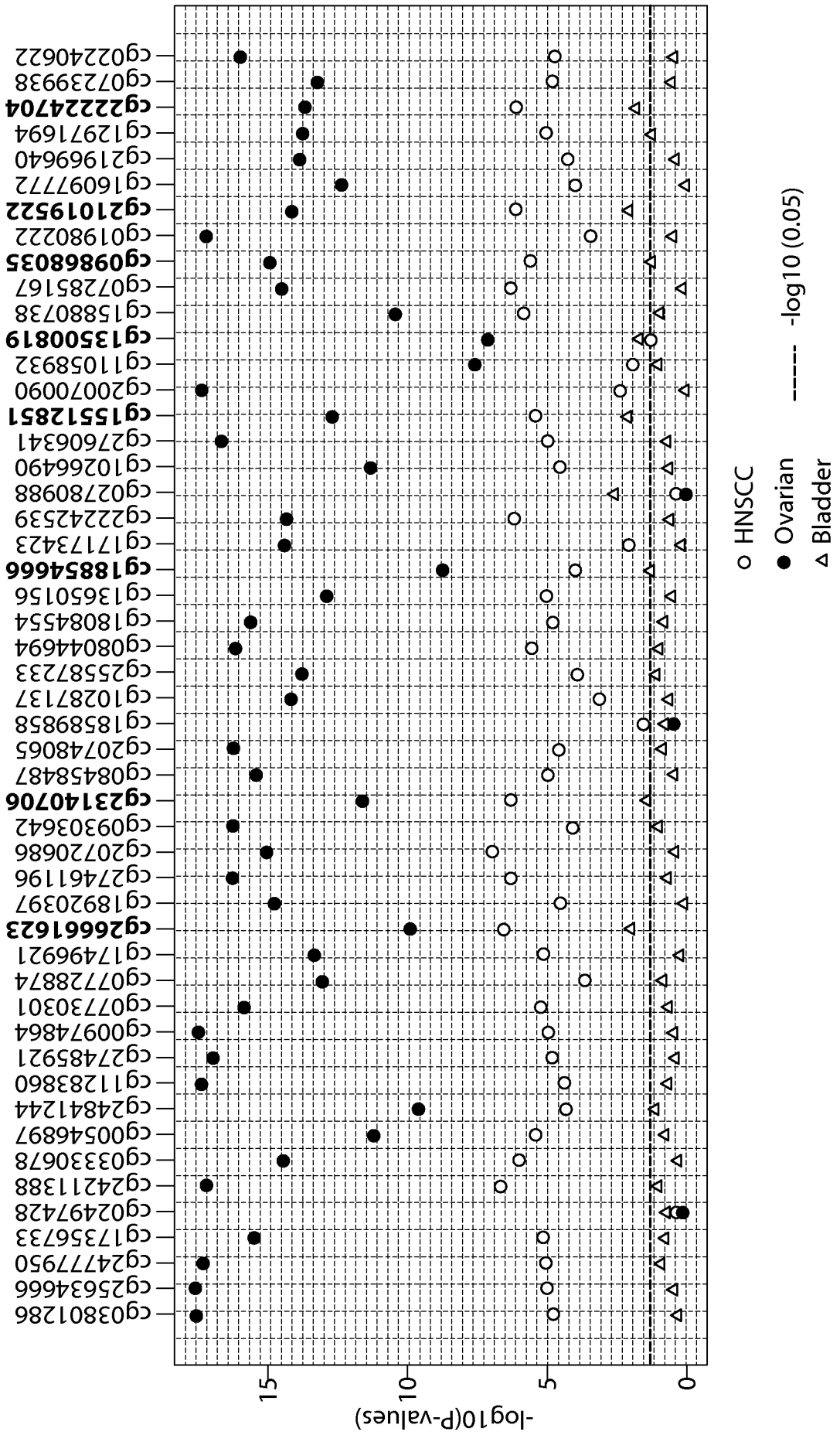

cg03801286
cg25634666
cg24777950
cg17356733
cg02497428
cg24211388
cg03330678
cg00546897
cg24841244
cg11283860
cg27485921
cg00974864
cg07730301
cg07728874
cg17496921
cg26661623
cg18920397
cg27461196
cg20720686
cg09303642
cg23140706
cg08458487
cg20748065
cg18589858
cg10287137
cg25587233
cg08044694
cg18084554
cg13650156
cg18854666
cg17173423
cg22242539
cg02780988
cg10266490
cg27606341
cg15512851
cg20070090
cg11058932
cg13500819
cg15880738
cg07285167
cg09868035
cg01980222
cg21019522
cg16097772
cg21969640
cg12971694
cg22224704
cg07239938
cg02240622
15
10
5
0
-log10(P-values)
HNSCC
Ovarian
Bladder
-log10 (0.05)

Figure 23B

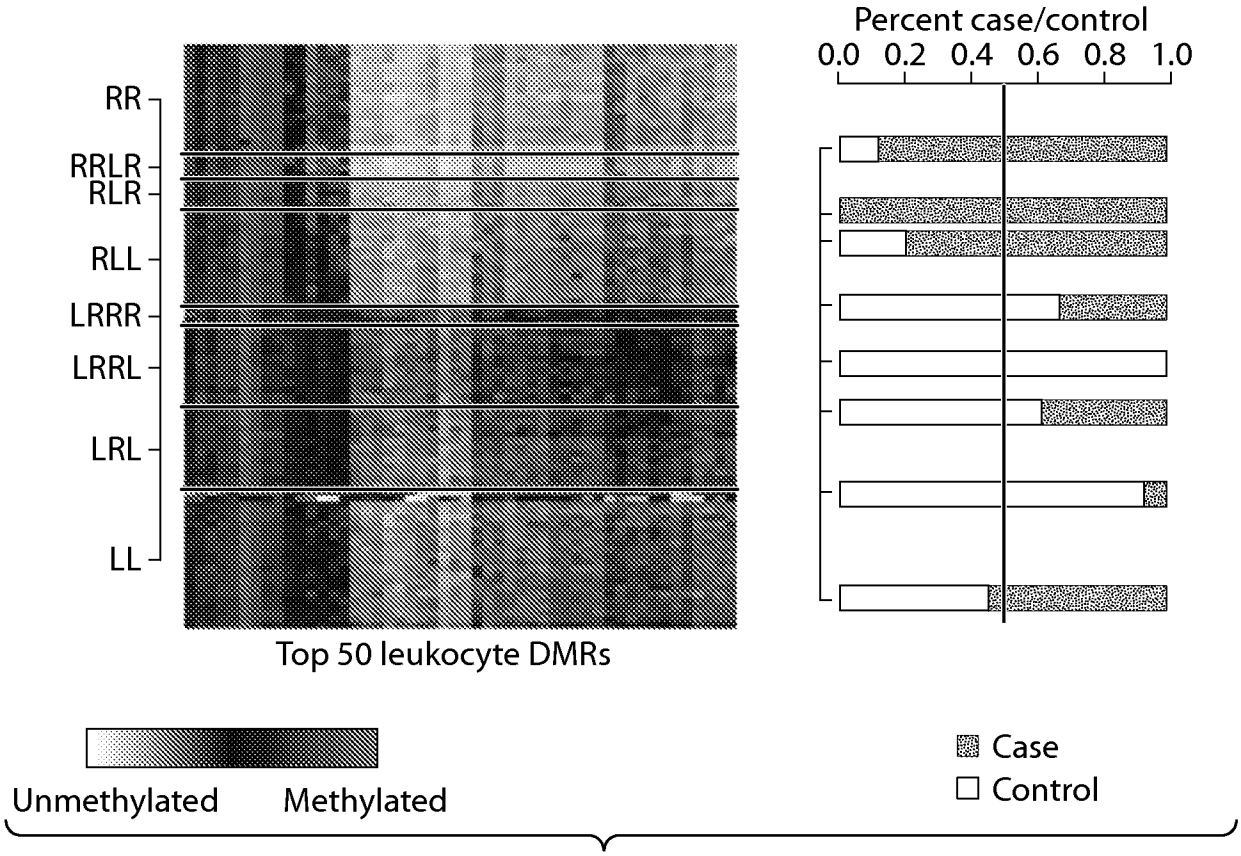


Figure 24A

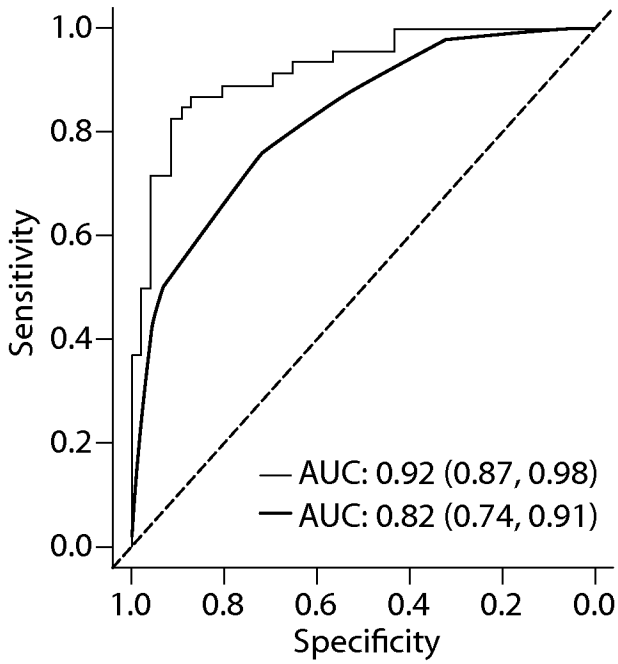


Figure 24B

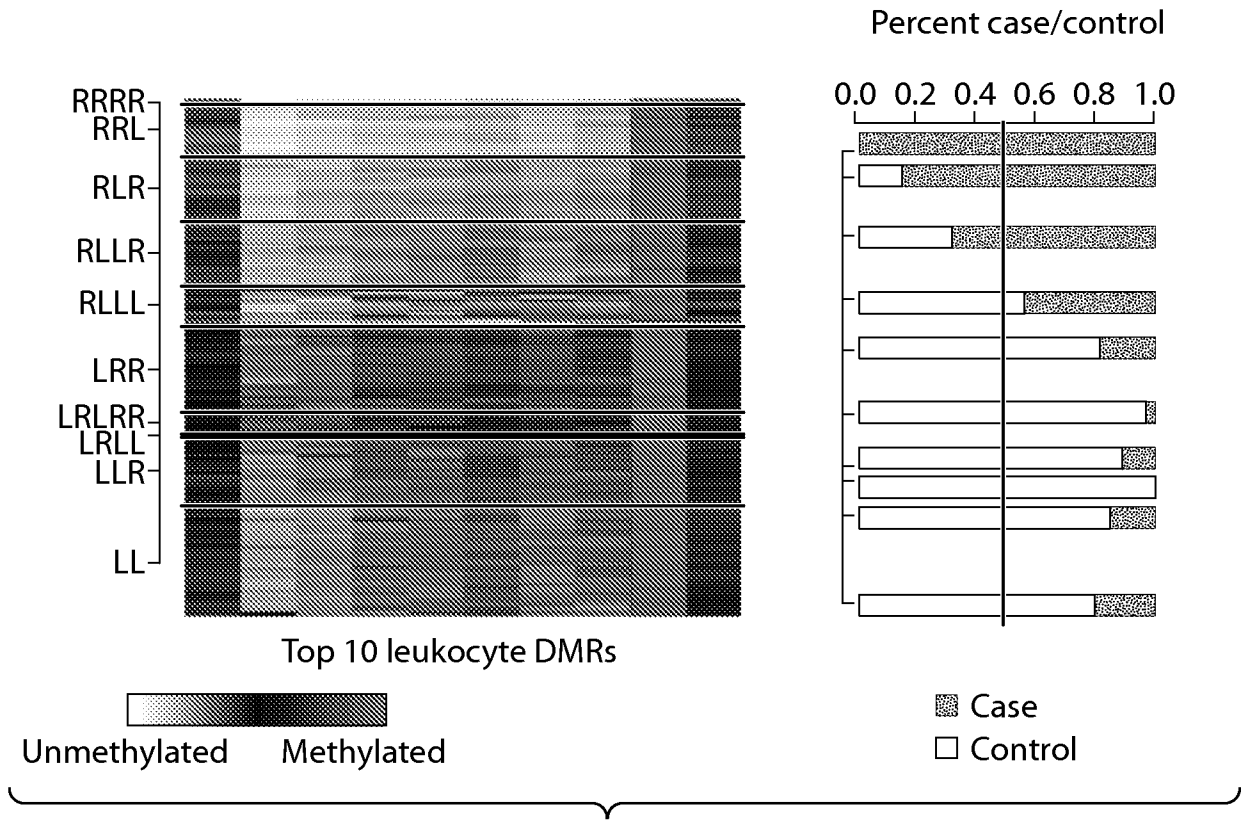


Figure 25A

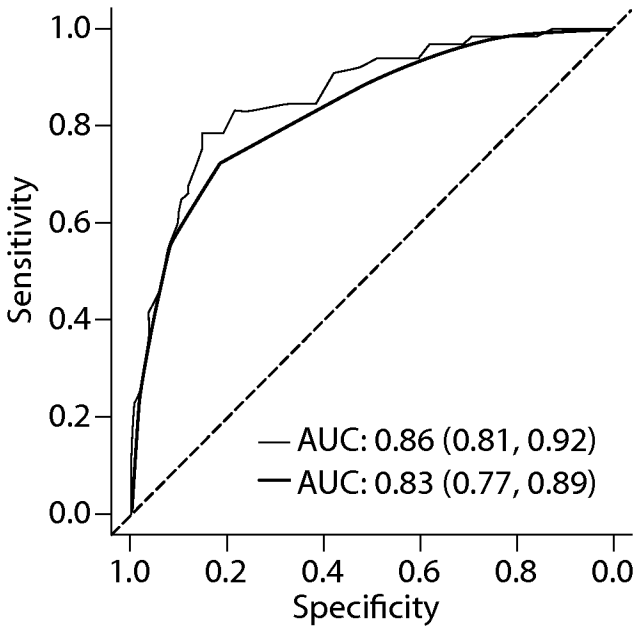


Figure 25B

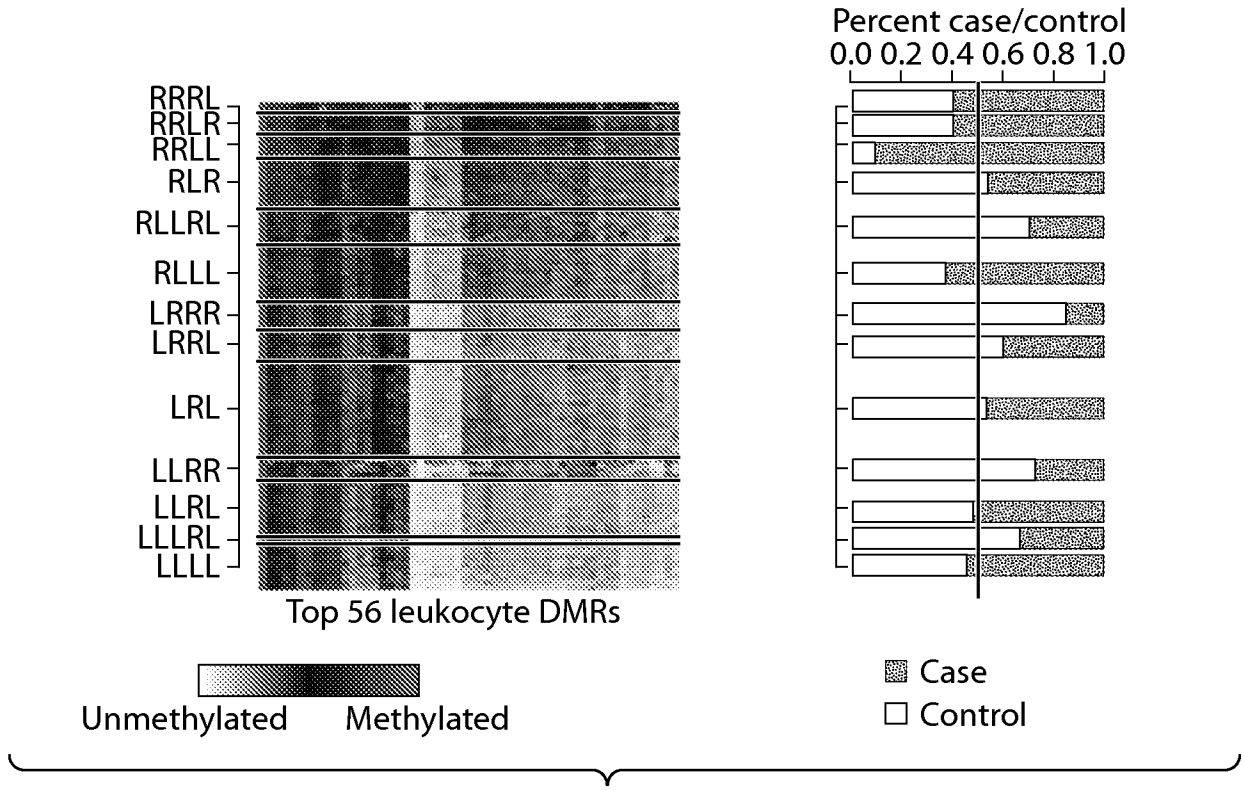


Figure 26A

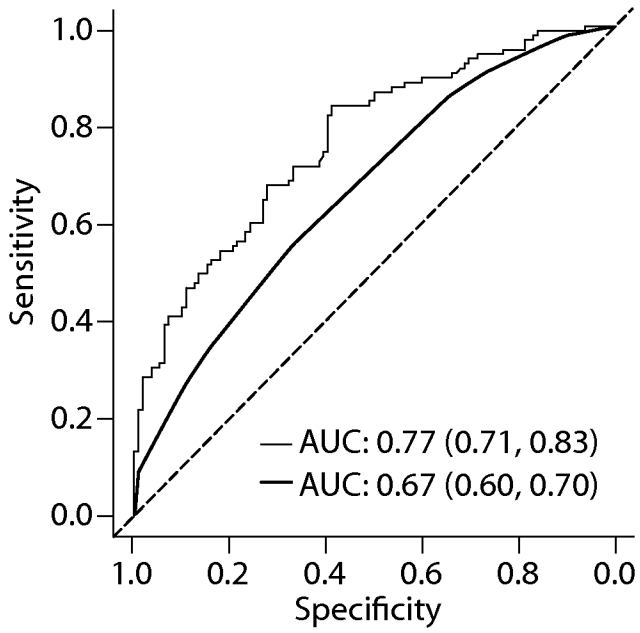


Figure 26B

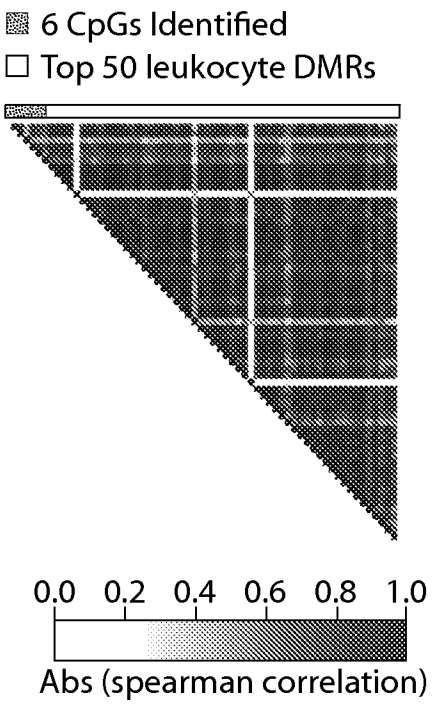


Figure 27A

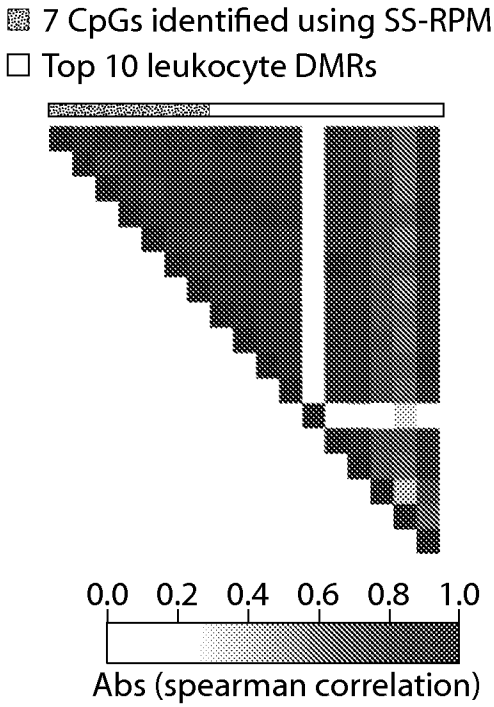


Figure 27B

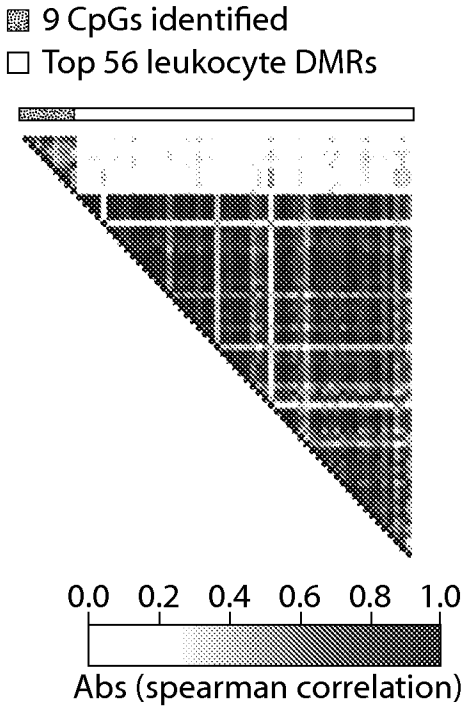


Figure 27C

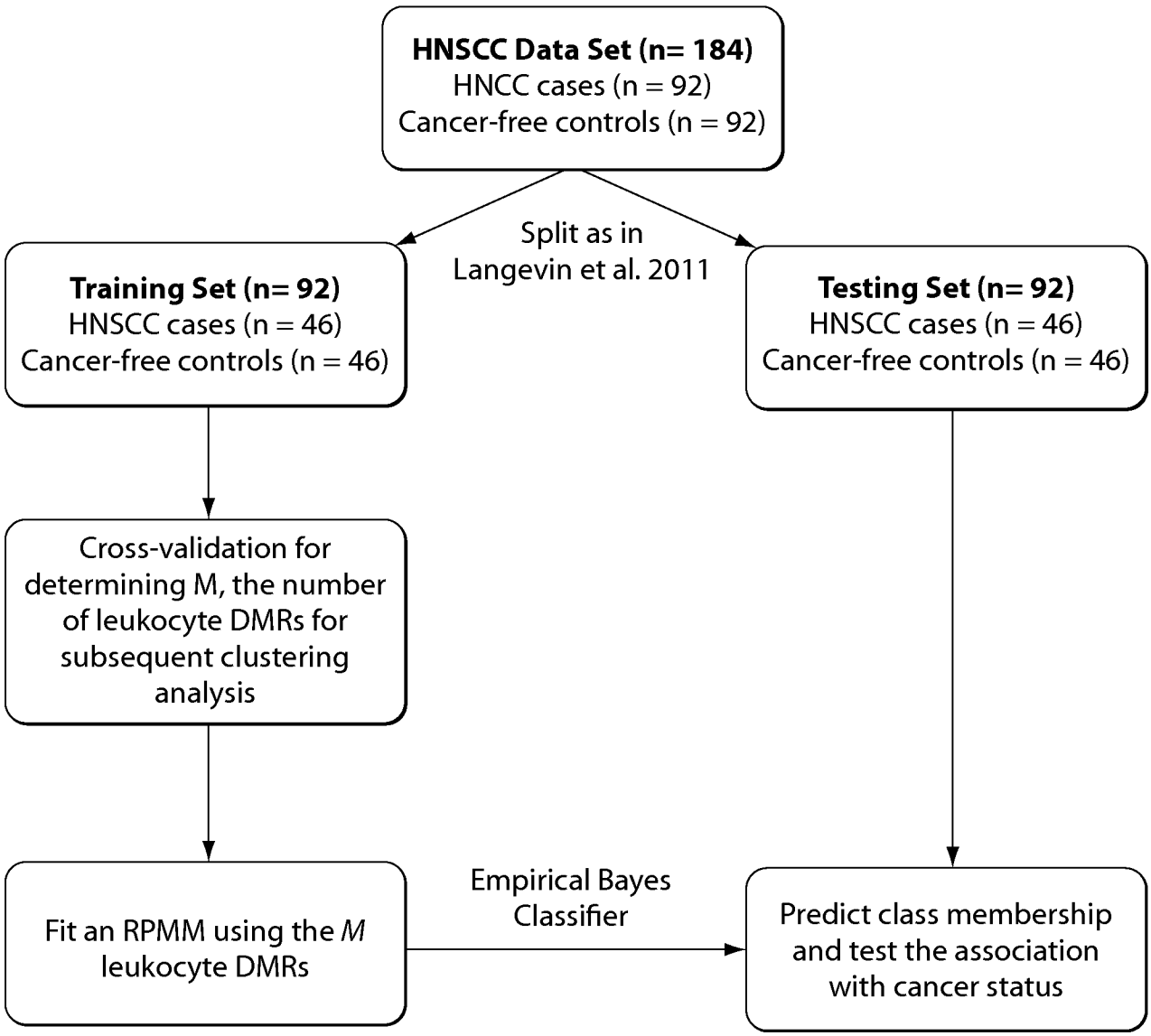

HNSCC Data Set (n= 184)
HNCC cases (n = 92)
Cancer-free controls (n = 92)
Split as in
Langevin et al. 2011
Training Set (n= 92)
HNSCC cases (n = 46)
Cancer-free controls (n = 46)
Testing Set (n= 92)
HNSCC cases (n = 46)
Cancer-free controls (n = 46)
Cross-validation for determining M, the number of leukocyte DMRs for subsequent clustering analysis
Fit an RPMM using the M leukocyte DMRs
Empirical Bayes Classifier
Predict class membership and test the association with cancer status

Figure 28

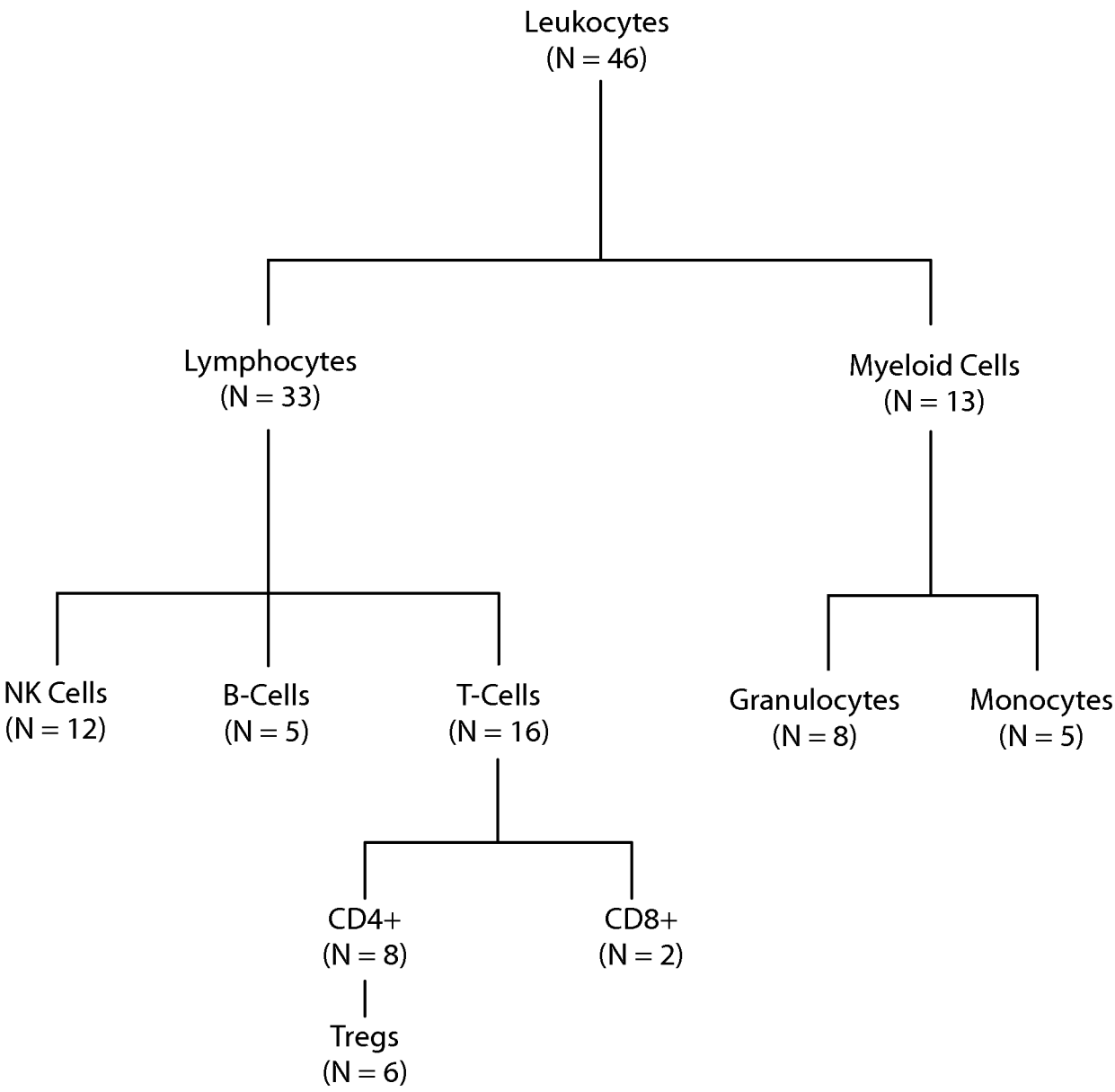

Leukocytes
(N = 46)
Lymphocytes
(N = 33)
Myeloid Cells
(N = 13)
NK Cells
(N = 12)
B-Cells
(N = 5)
T-Cells
(N = 16)
Granulocytes
(N = 8)
Monocytes
(N = 5)
CD4+
(N = 8)
CD8+
(N = 2)
Tregs
(N = 6)

Figure 29

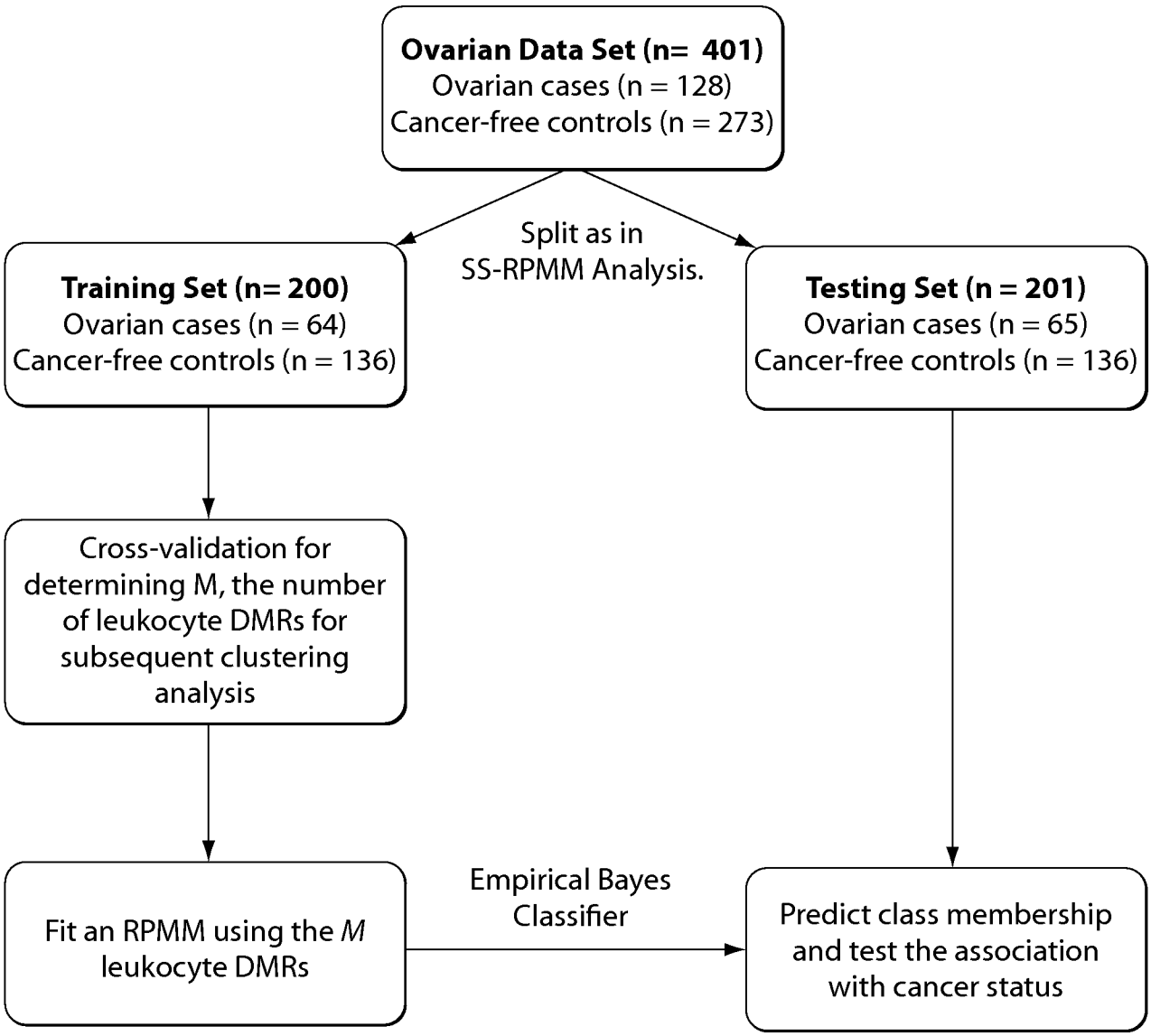

Ovarian Data Set (n= 401)
Ovarian cases (n = 128)
Cancer-free controls (n = 273)
Split as in
SS-RPMM Analysis.
Training Set (n= 200)
Ovarian cases (n = 64)
Cancer-free controls (n = 136)
Testing Set (n = 201)
Ovarian cases (n = 65)
Cancer-free controls (n = 136)
Cross-validation for determining M, the number of leukocyte DMRs for subsequent clustering analysis
Fit an RPMM using the M leukocyte DMRs
Empirical Bayes Classifier
Predict class membership and test the association with cancer status

Figure 30

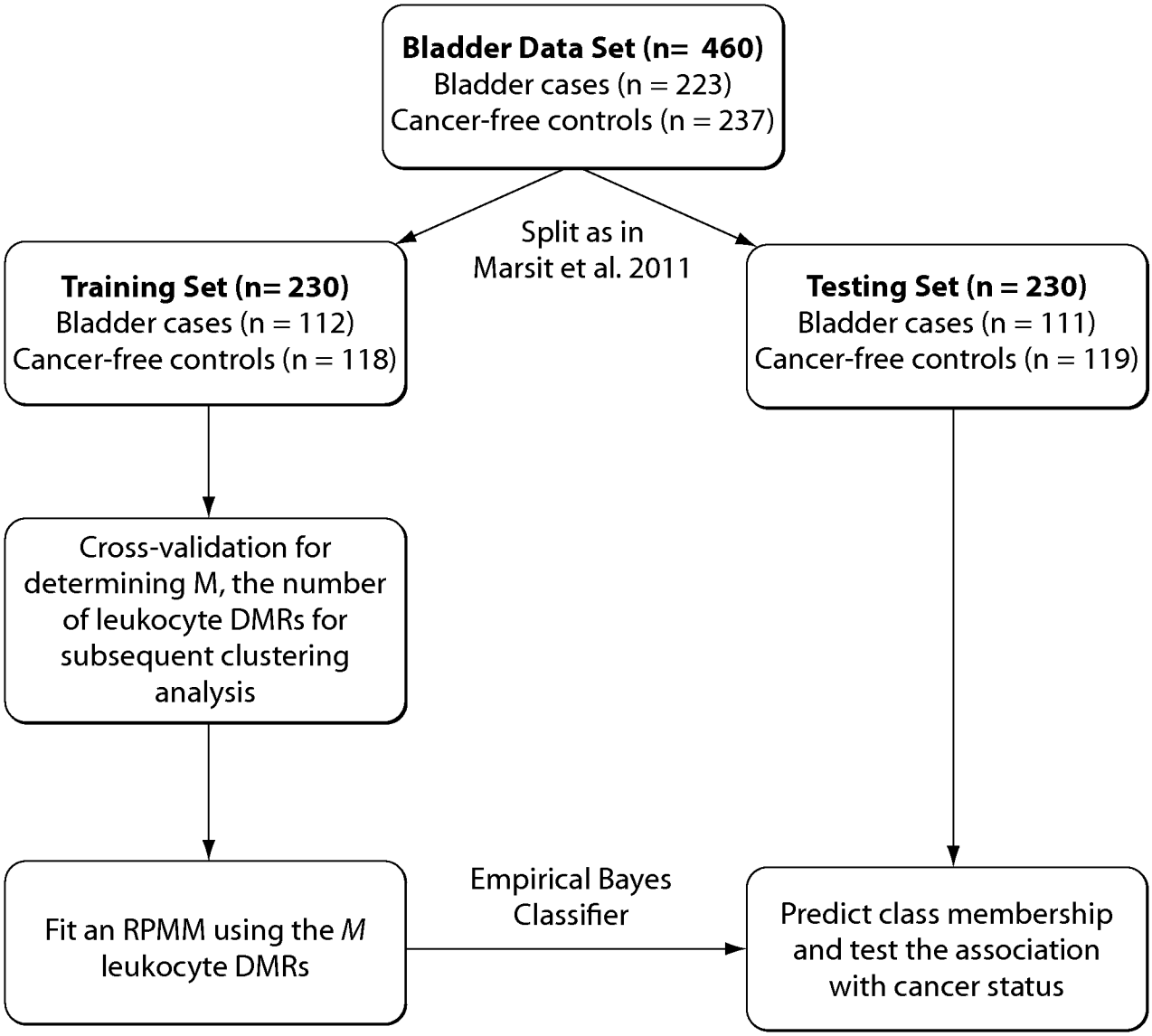

Bladder Data Set (n= 460)
Bladder cases (n = 223)
Cancer-free controls (n = 237)
Split as in
Marsit et al. 2011
Training Set (n= 230)
Bladder cases (n = 112)
Cancer-free controls (n = 118)
Testing Set (n = 230)
Bladder cases (n = 111)
Cancer-free controls (n = 119)
Cross-validation for determining M, the number of leukocyte DMRs for subsequent clustering analysis
Empirical Bayes Classifier
Fit an RPMM using the M leukocyte DMRs
Predict class membership and test the association with cancer status

Figure 31

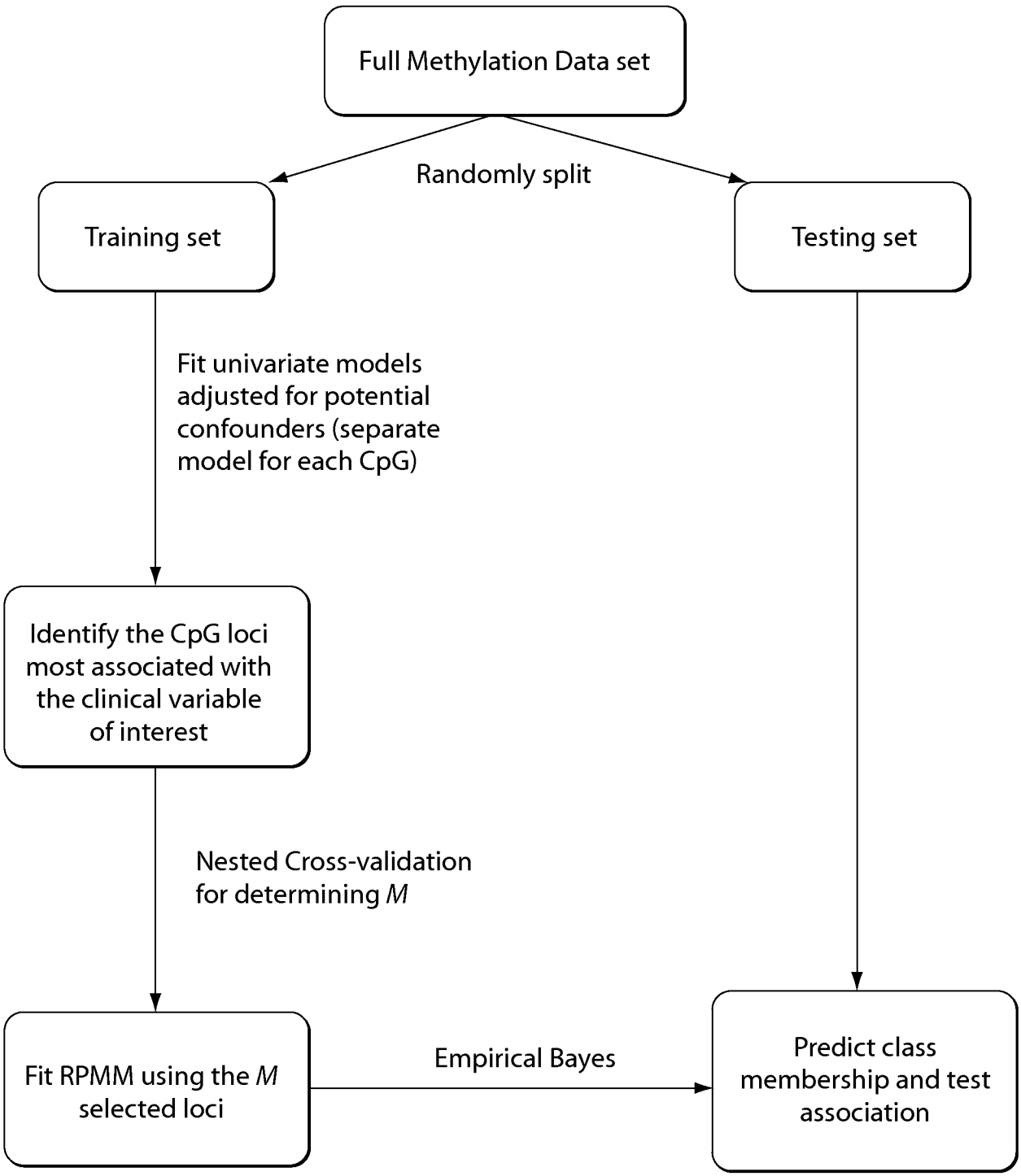

Full Methylation Data set
Randomly split
Training set
Testing set
Fit univariate models adjusted for potential confounders (separate model for each CpG)
Identify the CpG loci most associated with the clinical variable of interest
Nested Cross-validation for determining M
Fit RPMM using the M selected loci
Empirical Bayes
Predict class membership and test association

Figure 32

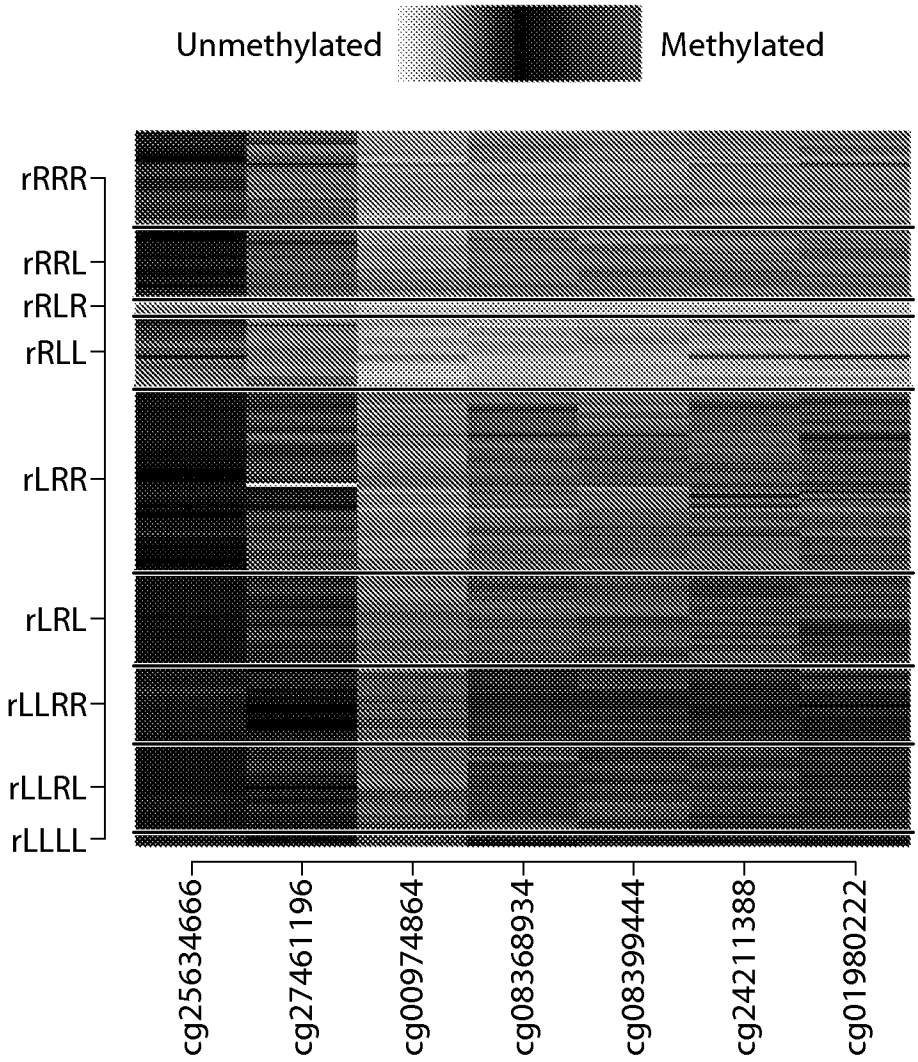

Unmethylated
Methylated
rRRR
rRRL
rRLR
rRLL
rLRR
rLRL
rLLRR
rLLRL
rLLLL
cg25634666
cg27461196
cg00974864
cg08368934
cg08399444
cg24211388
cg01980222
Loci used for determining subtypes (M=7)

Figure 33A

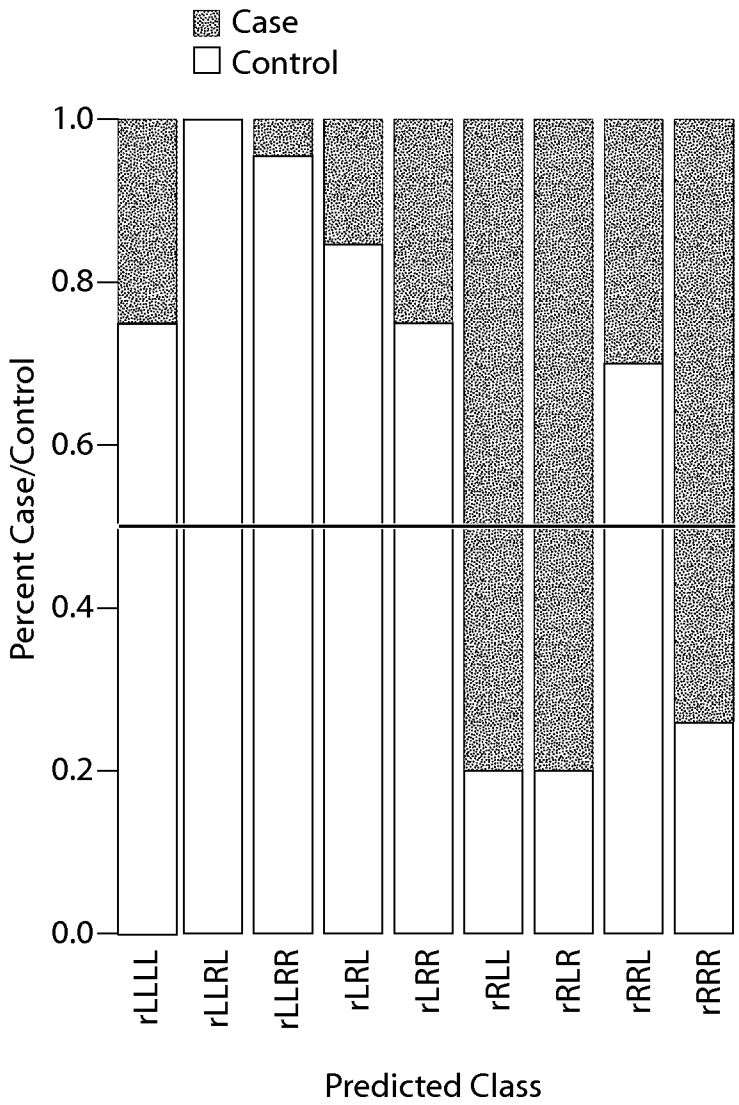

Case
Control
1.0
0.8
0.6
0.4
0.2
0.0
Percent Case/Control
rLLLL
rLLRL
rLLRR
rLRL
rLRR
rRLL
rRLR
rRRL
rRRR
Predicted Class

Figure 33B

Figure 33

C

CpG Name	Chromosome	MapInfo	Entrez ID	Gene Symbol
cg00974864	1	159867677	2215	FCGR3B
cg01980222	6	41238895	54209	TREM2
cg08368934	16	56258956	222487	GPR97
cg08399444	12	13139815	83445	GSG1
cg24211388	6	31690816	199	AIF1
cg25634666	11	71524436	2352	FOLR3
cg27461196	19	40321946	5348	FXYD1

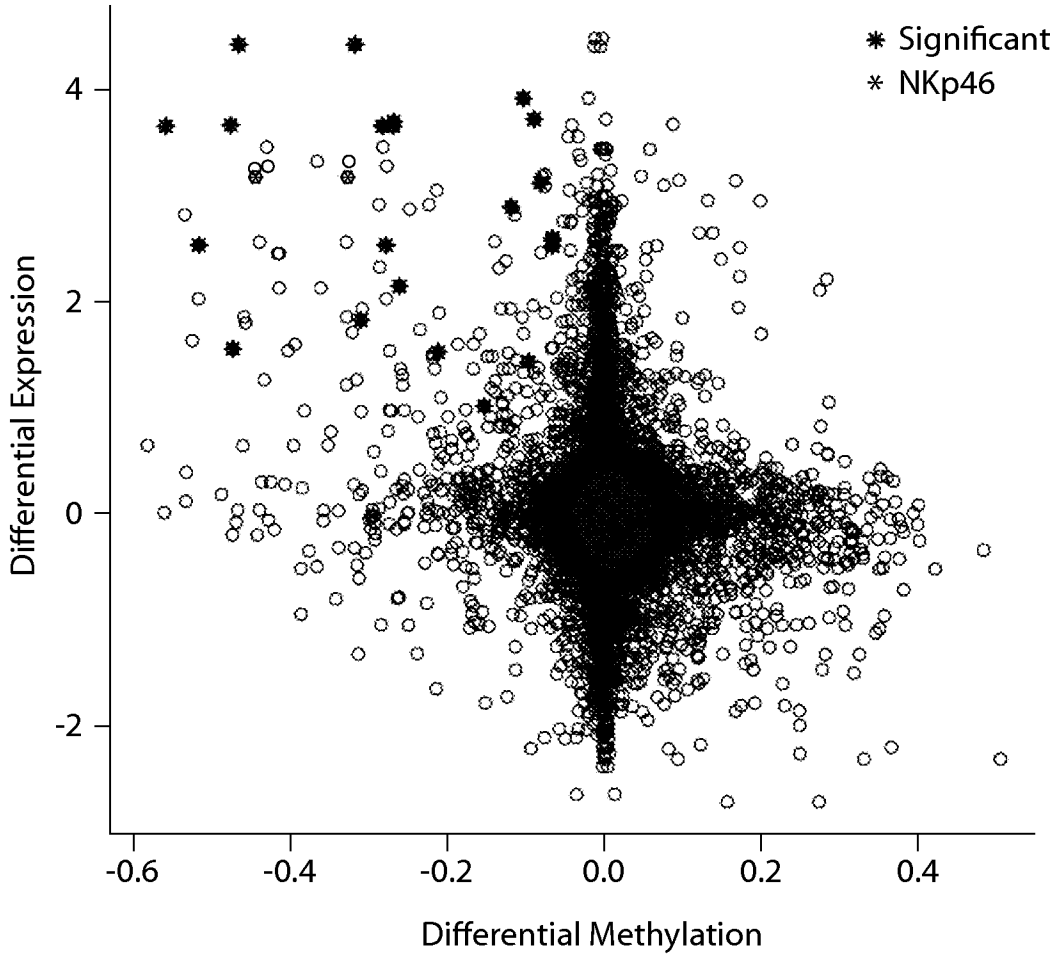
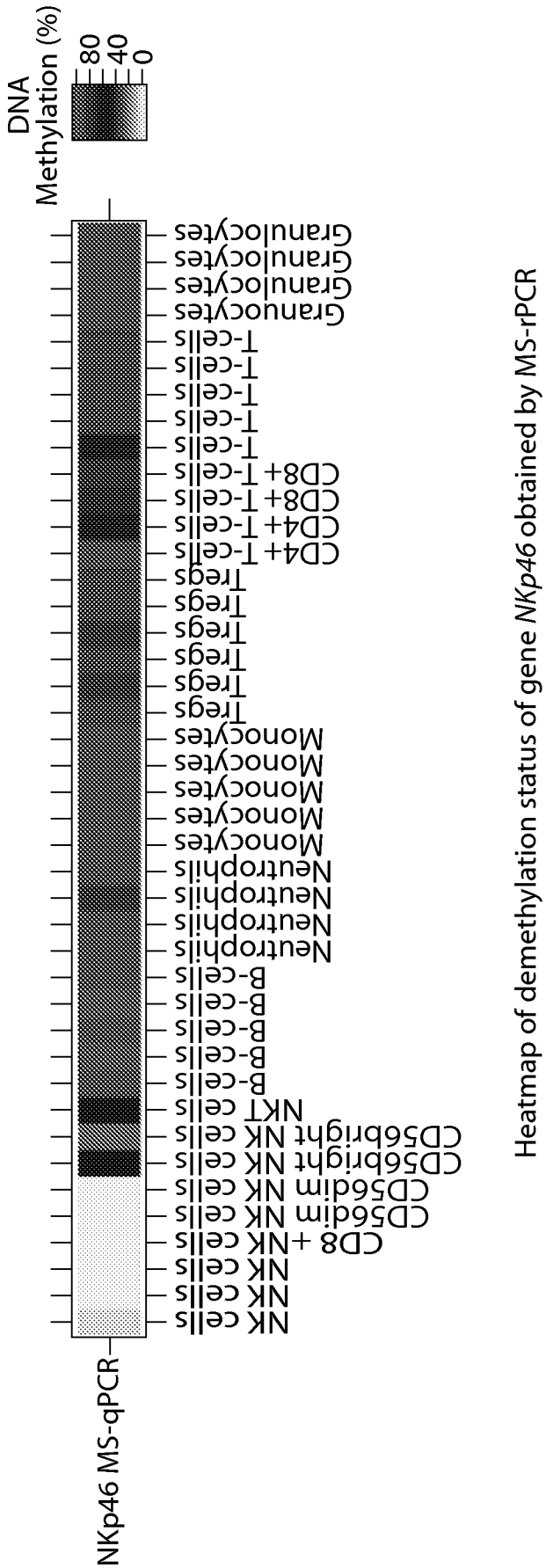

Significant
NKp46
Differential Expression
4
2
0
-2
-0.6
-0.4
-0.2
0.0
0.2
0.4
Differential Methylation

Figure 34



Heatmap of demethylation status of gene *Nkp46* obtained by MS-rPCR

Figure 35

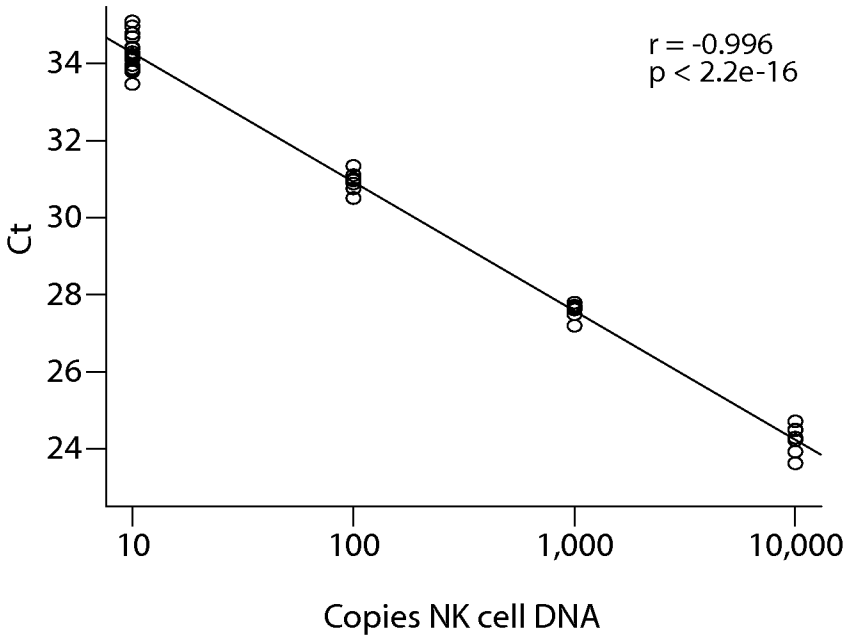


Figure 36

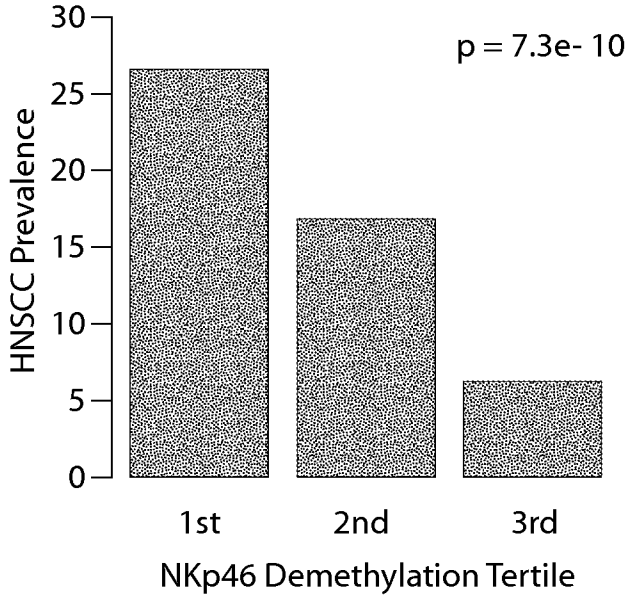
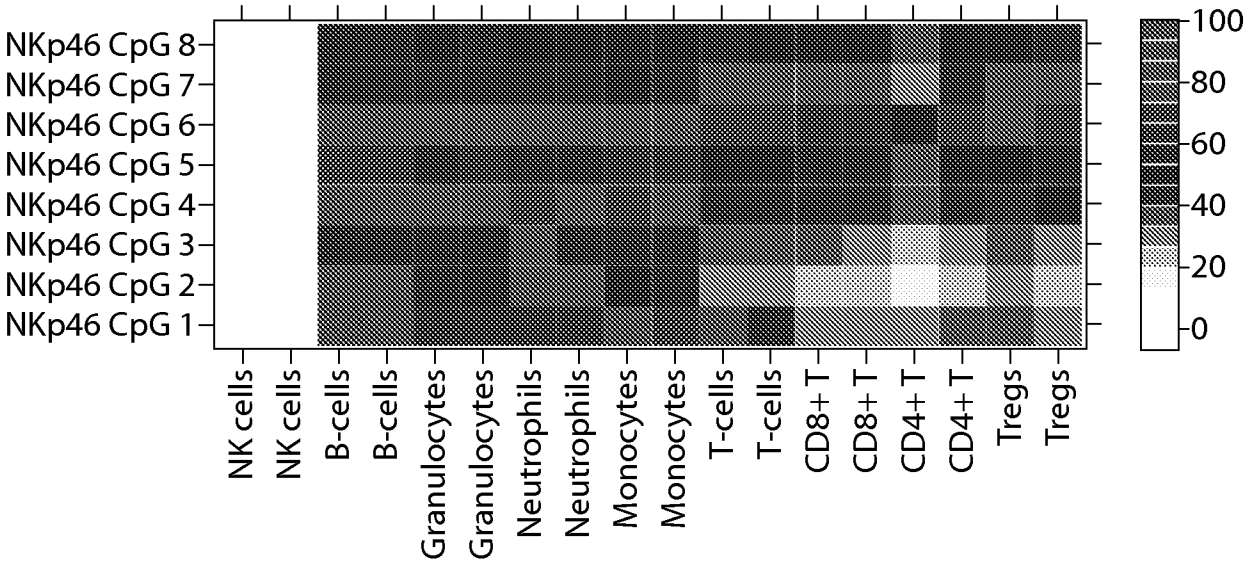


Figure 37



Heatmap of demethylation status of gene *NKp46* obtained by bisulfite pyrosequencing

Figure 38

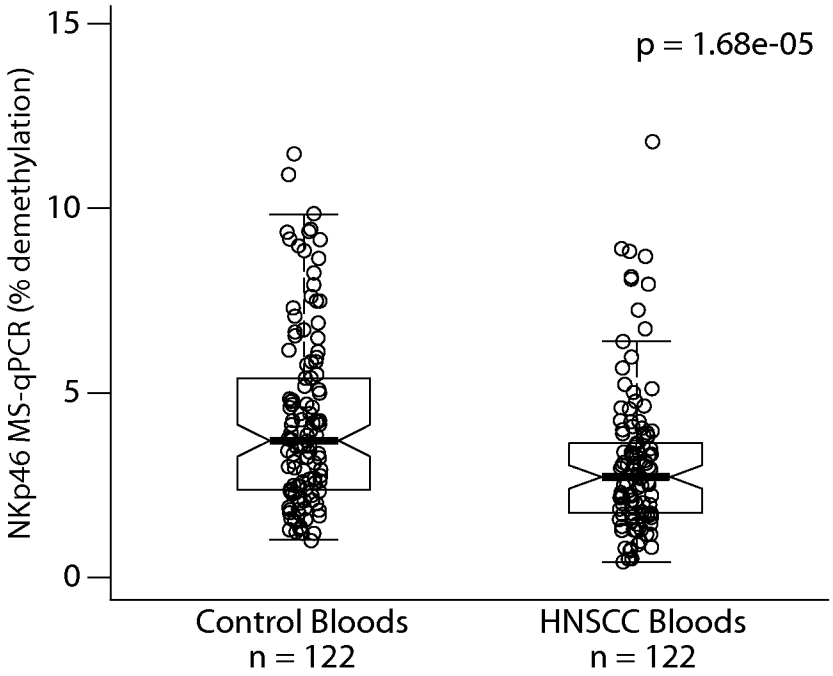


Figure 39

SEQ ID NO: 1
Chromosome: 1
Coordinate: 170894886
Source: NCBI
Gene: FASLG
Accession Id.:NM_000639.1
GCCAGCCCCAGCAAACGGTTTTACTTCTTCTCAGTCCTGTAGAGGCTGAGGTGTCAAGGA
[CG]GGACCCTGTTGCTGACTGCTCAAGAGGAGGCAAGCTGGATCTCTCTTATAGAGTTTC
CAT

SEQ ID NO: 2
Chromosome: 6
Coordinate: 37080005
Source: NCBI
Gene: FGD2
Accession Id.: NM_173558.2
TCTCCACTGAACCTGGTCTGTGTCCTGGAGTGCAGGGCCTGAGCCTCGGTGCTCTTGGTA
[CG]TGAAGTGCCTGGAACAGCTTCACGCTCCAGTAACTGTGAACCCCTGGGGCCTCACAT
GCC

SEQ ID NO: 3
Chromosome: 12
Coordinate: 14930292
Source: NCBI
Gene: MGP
Accession Id.: NM_000900.2
GATCATGTGTTTGTGGCAACTTCCTCTGTGGGCTTTTGCCCAGGTCTGTCCCCAAGCATA
[CG]ATGGCCAAAACTTCTGCACCAGAGCAGCATCCTGTGTAACACAGTCAGGTCCAGCAG
TTA

SEQ ID NO: 4
Chromosome: 1
Coordinate: 110108114
Source: NCBI
Gene: EPS8L3
Accession Id.: NM_024526.2
AGCTTCTCTGCTAGTGGCCACAGGCAGAGCCTGCCTTTGATGAGGTTACAGAGGCAGCCA
[CG]CCTGTGCTCTTTGGACTCTGGTGGGTGGGGAGGCTTCCTGGTCACTAACCGCTCAAC
ATC

SEQ ID NO: 5
Chromosome: 21
Coordinate: 44056660
Source: NCBI
Gene: LOC284837
Accession Id.: NM_194310.1
CACCCACGGGGCCAGGCTGGCACAACGCCCCAACGCTGCAATCCTGGGAAGAGTCACACG
[CG]CCCTCCCGGGACCCACGTGACCATCAAGGGAGTGTGGAGGACACATCCCTCGGGGGT
GAC

Figure 40-1

SEQ ID NO: 6
Chromosome: 11
Coordinate: 47237946
Source: NCBI
Gene: NR1H3
Accession Id.: NM_005693.1
TGGCTTCCTCTCTGAGGATGCAGCTGCTGCCTCCCTGGGCCTGGCTGCTTGCATCCTGTG
[CG] CCTGGCTTCCACAGCTCCACCGCAGAGTCTGAGCTCCAAAAGAGAGGCACAAGGGGG
TCT

SEQ ID NO: 7
Chromosome: 9
Coordinate: 126992712
Source: NCBI
Gene: PPP6C
Accession Id.: NM_002721.3
TTTTTGTAGCTACAGATAGAATAGAGATCTTTGTCTATTTTGTTCACGAAGTACTGCAAG
[CG] CCTTGAGCTGTACCTGGCACATCTTTGTTGCTCAGCAAAGAGTGGTTGGATAAACGA
ACG

SEQ ID NO: 8
Chromosome: 21
Coordinate: 36458793
Source: NCBI
Gene: DOPEY2
Accession Id.: NM_005128.2
ACCAAATGAGAGAGCCATTTGGGGATACAAATCATTCCCAACCCAGAGCTACAGAAGACA
[CG] TGTCCACAAACACAACTGTGCACAAACTCACTGGGCAATCCTGTTCAAATATTTAGC
AAG

SEQ ID NO: 9
Chromosome: 19
Coordinate: 40727715
Source: NCBI
Gene: NIFIE14
Accession Id.: NM_032635.2
TCTCTACGGACTTCCTCGGTGATACCCACTCGTCCATCTTCGATGCTAAGGCCGGCATTG
[CG] CTCAATGACAATTTCGTGAAGCTCATTTCATGGTAAGGGGGAAGGAGCTGGAGACTT
AGA

SEQ ID NO: 10
Chromosome: 2
Coordinate: 242450682
Source: NCBI
Gene: PDCD1
Accession Id.: NM_005018.1
CGCGCCAGCTGTCAGGCGGTTTCTAGCCTCGCTTCGGTTATTTTAAGCTGATGAGCCTGA
[CG] CATCTCATCACTAATATCAGCAGTTTCATTTCTCCTGTTTTCCATTCGCTGTAATAA
AAT

Figure 40-2

SEQ ID NO: 11
Chromosome: 10
Coordinate: 44815977
Source: NCBI
Gene: ZNF22
Accession Id.: NM_006963.3
CATTTACAGGAAAGGAACCAAGGCTCAGAGAAGAAATGTGCTCCGTTCACCGTGTGTAAG
[CG]GACGACCCAGAATTGGAATGGTTCTTTGTGGCTCCAAAGTCTGATTTCAACACACCC
CTT

SEQ ID NO: 12
Chromosome: 1
Coordinate: 26728010
Source: NCBI
Gene: RPS6KA1
Accession Id.: NM_001006665.1
GGGAGACCTGACCTGAAAGGACCCCCTTCAAGTGATAGGGCAGAGCACAGATTGCAAAAA
[CG]CATATTAAGAAATCACTCTTGGCCGGGCGCGGTGGCTCATGCCTGTAATCCCAGCAC
TTT

SEQ ID NO: 13
Chromosome: 2
Coordinate: 108971950
Source: NCBI
Gene: EDAR
Accession Id.: NM_022336.1
TCCCAGGCAGCCCTGCTGGCCCCTAAGGACATAGAGTACCTGCTTCTGAGAGGGCTGCCA
[CG]GTGGCCACCTGTGAAGCCTGTCACCCAGAACTGGATGGTACCTGACTTTCTTCATAG
ACC

SEQ ID NO: 14
Chromosome: 1
Coordinate: 54293020
Source: NCBI
Gene: TMEM59
Accession Id.: NM_004872.3
TTTAACATTCCTAACAGACTACATTTTGCAAAAGAATAACAACGAAGGGGACTTGTCCTA
[CG]AGCTAGCACACATGGTGTAAACCGGAGTAATTACAAGGGTGTAGCAGGGGTGTGCAG
AAA

SEQ ID NO: 15
Chromosome: 5
Coordinate: 176869969
Source: NCBI
Gene: DOK3
Accession Id.: NM_024872.1
GGCCTCCCTTGACCACTCCACGCTGTCCGAGAGCTCAAAGGCCCTCACGGTATACACTCA
[CG]CTGGGCATCCAGTCCACATGGGACCCACAGCCCTGAATGGCCCCAACCACGTGAGTG
TGG

Figure 40-3

SEQ ID NO: 16
Chromosome: 6
Coordinate: 43720625
Source: NCBI
Gene: C6orf206
Accession Id.: NM_152732.2
CCTCCTCCCCACGGAGGCCTAGGCATCAGCCCCCTCCCTCATCCTTTCCAGAGTTTGGGA
[CG] GGATGTCTTCAGTTGCCACGGCCACAGTATGGCTTCCCCTACAGTTAGGCTACAGTT
GGG

SEQ ID NO: 17
Chromosome: 22
Coordinate: 37044362
Source: NCBI
Gene: CSNK1E
Accession Id.: NM_001894.4
AGGGAAGATACGGCTATTATAGAAGTGACTCCTCCCAGGAACTGTGCTTCCGGGATTGGA
[CG] CAGGGCCTCAGGCATTTTGCGTGTCCACAGTCACAACTGTGTGAATATAGGTGTGTC
ATA

SEQ ID NO: 18
Chromosome: 5
Coordinate: 139705861
Source: NCBI
Gene: HBEGF
Accession Id.: NM_001945.1
GTCGGCATCGGTGCGTGTTGGTCAGGGGTCTGGGCGGGTGTCTGATGCGGCCTGGCCTCT
[CG] CCCGCAGTTCTCTCGGCACTGGTGACTGGCGAGAGCCTGGAGCGGCTTCGGAGAGGG
CTA

SEQ ID NO: 19
Chromosome: 10
Coordinate: 44815441
Source: NCBI
Gene: ZNF22
Accession Id.: NM_006963.3
TGGCTTGGTGGTTATAGCAGTGGAAGTGTTGAAAGTGGCTTGATAATGAATATATTTTAA
[CG] ATGAAGCCGACAGAATTTGTGGATAAATCACAGGTGAATTTTGGATGAAAAAAAGAA
GAG

SEQ ID NO: 20
Chromosome: 20
Coordinate: 57016289
Source: NCBI
Gene: CTSZ
Accession Id.: NM_001336.2
CCCAAAGGAAGATTCCACTTGGCGCAGGCATCAGGAGTTATCCAATGTGACTTCCAAAGA
[CG] CCTTGAAAAGGTTTTCTGCTAACGAAACTCTTCTTAGTCAAATGAGGAACCAAAAGC
AGA

Figure 40-4

SEQ ID NO: 21
Chromosome: 14
Coordinate: 22658002
Source: NCBI
Gene: CEBPE
Accession Id.: NM_001805.2
GCCCGGGGAGCTAGGGGACATGTGTGAGCATGAGGCCTCCATTGACCTCTCCGCCTACAT
[CG]AGTCTGGGGAAGAGCAGCTTCTCTCCGATCTCTTTGCCGTGAAGCCAGCGCCTGAGG
CCA

SEQ ID NO: 22
Chromosome: 12
Coordinate: 8167985
Source: NCBI
Gene: CLEC4A
Accession Id.: NM_016184.2
TGAGTATGTCTGGGAAACACAAGAGTCCCAGAAGATTGAGTGGCCTGCAGATACGCATTA
[CG]GGGTGTACATTTGTATTGTGGAGAAGAAAAGATTTTGTGCCACTCTCTTCAGCCTCC
ACT

SEQ ID NO: 23
Chromosome: 6
Coordinate: 6534074
Source: NCBI
Gene: LY86
Accession Id.: NM_004271.3
CAGCTGCAGTGGAGGCGGCGGTGGGAAAGCCTGGCCCACACACGTGGTCTGTAGCGACAG
[CG]GCTTGGAAGTGCTCTACCAGAGTTGCGGTAAGCCCTTGCAGTACACCCATGTGTGTT
TAT

SEQ ID NO: 24
Chromosome: 16
Coordinate: 66120919
Source: NCBI
Gene: FAM65A
Accession Id.: NM_024519.2
CAGCCCCTTGCCCAAACCAGTTCTGCAGAGAGCCCAGGCCCGGCTGTTGCAGGAACCTTG
[CG]CCAACCTCCATTTCCAGGGAAAAGCTCCGTTCCCCGACAAGGACGATCCTCTCCGGC
TTC

SEQ ID NO: 25
Chromosome: 15
Coordinate: 38388759
Source: NCBI
Gene: PLCB2
Accession Id.: NM_004573.1
TTTCTGTGCTGGGAATTCCCTTAGCTCCAGCCTCCACTGGGCAGTTTATTATCTTAATTC
[CG]CATGAAGAGTGTCCTCCCTCACCCTCCACCCTGCCCTGGACCAGACCTCCAGCCGCG
ACA

Figure 40-5

SEQ ID NO: 26
Chromosome: 15
Coordinate: 48345209
Source: NCBI
Gene: HDC
Accession Id.: NM_002112.1
GAGCTAAGGTCAAAGAAAGAACCCTTTAAATAAAGGGCCCACACTGGCTGCCAGGGAGTG
[CG] CAGGACTGGCAAGAGGGAAGCCGGGCTGCTCCACGCCTTTCACGCCTTCCACCTCCT
GCG

SEQ ID NO: 27
Chromosome: 17
Coordinate: 77868065
Source: NCBI
Gene: CD7
Accession Id.: NM_006137.6
CCGTCCTCGTAGTAAATGATGTCTTGGGGCTGTGGCCCGAGCTGCCTCAGGTAGATCCCA
[CG] CAGGCCCCCGCTGGTGGAGCAGGTGATGTTGACGGAGGCTCCCACGGGGACAGTCGT
GCA

SEQ ID NO: 28
Chromosome: 16
Coordinate: 21572639
Source: NCBI
Gene: IGSF6
Accession Id.: NM_005849.1
GGAGAGACACAAGGCCTGGGAGCCGCTTTCCTGGCCTGCCGTGCAGCTGAGGCACTGGCA
[CG] CAGCCTAAGCCAGGCACACTTGCCCATGCCCTGGAATGGAGAGCCAGTGACCCAGAG
TAG

SEQ ID NO: 29
Chromosome: 7
Coordinate: 100667471
Source: NCBI
Gene: CLDN15
Accession Id.: NM_138429.1
GCTGATGCTGGGGGTGACTCTGCCAAACAGCTACTGGCGAGTGTCCACTGTGCACGGGAA
[CG] TCATCACCACCAACACCATCTTCGAGAACCTCTGGTTTAGCTGTGCCACCGACTCCC
TGG

SEQ ID NO: 30
Chromosome: 4
Coordinate: 47830991
Source: NCBI
Gene: TXK
Accession Id.: NM_003328.1
GAGGAAAGGATCATGGTAGCCCCTTCTGCGGGGAGCACACAACAGTCTTCAGTTCTTCTG
[CG] GTGCTCTACTCACAAAAACACATCTTTCAACTGAAATCATAGTTCGCTCAAGATGTT
TCT

Figure 40-6

SEQ ID NO: 31
Chromosome: 4
Coordinate: 8251449
Source: NCBI
Gene: SH3TC1
Accession Id.: NM_018986.2
TGGCCCGGAGGGCACCCGGGCAGAGACGGAAGAAATTGCACGTGAGCGTTTGTGTGCATA
[CG] TGTGCCTGTCCATGTGTGCACACACTTGTGCTTGTGAGTCTCTGTGTGCCCATGCAT
ATG

SEQ ID NO: 32
Chromosome: 3
Coordinate: 139810060
Source: NCBI
Gene: FAIM
Accession Id.: NM_018147.2
AAGTGTGGGTAGGAGCCGGCCGCTGGCCCCGCTCTGGGCTAGACGGTGGGGACATACTGG
[CG] GGCAAAAACAGCCCTGTGCCTGCTCTGCAGCTATGGGGAAGGAATATCTGTTCATGG
CAG

SEQ ID NO: 33
Chromosome: 1
Coordinate: 111848223
Source: NCBI
Gene: ADORA3
Accession Id.: NM_020683.5
TTGGTGGCAGCCTCCTAACCTTAGCCAGAACTATTCCTGCTAAGTTCTTGCACGAGTTGA
[CG] CTTTGCTGAGCACAGCCGATACCCAGCCTTTGCAGCAAAGATCCTTGGTCAAAGGGA
TAA

SEQ ID NO: 34
Chromosome: 12
Coordinate: 121753810
Source: NCBI
Gene: GPR109A
Accession Id.: NM_177551.3
CCAGAAAGTGATCCTGCAGATGGTGCCGATTCATGAGTGCGGCTAGTGAGTCCGATGGAG
[CG] CCTCGCCTAGTGAATGCTCCAGCAAGGAGGTGTGTGTCTGTGTGGTGAACGTGTGGT
TCC

SEQ ID NO: 35
Chromosome: 17
Coordinate: 72827828
Source: NCBI
Gene: 39333
Accession Id.: NM_006640.2
GACAATGCTACTTCAGTTTGGAGCACAAACATATGATCAGCACATGGAAATGTGGTAATT
[CG] GATGCATTCGTGATTGCAACAGATTGAAGAAATTAGACCAGACAAAGAGTGTTTTTA
GAG

Figure 40-7

SEQ ID NO: 36
Chromosome: 4
Coordinate: 16509831
Source: NCBI
Gene: LDB2
Accession Id.: NM_001290.2
CTCCAGCCAGGACCCTTCACAACCTGATTGCTAAGCTTGTTAGCATAGAGGTGGTCTAAC
[CG] CTACATGAGCCGCTCACCCCTGACAACCACACTGTTGTAATGTATCAGAAATGTTGA
TTA

SEQ ID NO: 37
Chromosome: 19
Coordinate: 40512021
Source: NCBI
Gene: CD22
Accession Id.: NM_001771.1
CACGCGGAAACAGGTAAAAATCATTTTGCTTTTATTTTGCATTCAACAAGCAAGTTATTA
[CG] GAACAGCAGTTATGGGCCAGGCATACCTCCCAGAGCTGGGAACACAGTGGGGACCTC
CCT

SEQ ID NO: 38
Chromosome: 19
Coordinate: 59739533
Source: NCBI
Gene: FLJ00060
Accession Id.: NM_033206.1
ACCTTGTGATCCGCCCACGTCAGCCTCCCAAAGTGCTGGGATTACAGGCGTAAGCCACAG
[CG] CCCAGCCTCGCTGTTCTTATCTTGGCAGCAGATTCCGAATGTCGGCTGGTGCCCCTG
TCA

SEQ ID NO: 39
Chromosome: 9
Coordinate: 134926722
Source: NCBI
Gene: CEL
Accession Id.: NM_001807.2
AGGCTGGATGGTGACACTTCCACACCCTTGAGTGGGACTGCCTTGTGCTGCTCTGGGATT
[CG] CACCCAGCTTGGACTACCCGCTCCACGGGCCCCAGGAAAAGCTCGTACAGATAAGGT
CAG

SEQ ID NO: 40
Chromosome: 8
Coordinate: 11388980
Source: NCBI
Gene: BLK
Accession Id.: NM_001715.2
CAGAGTTAGCAAACCTCCATGCTGACTCTACAAGGTAATTTGCCCTGCCGTGTGGACAAA
[CG] CTGCAGATCTCATGGAGAGGGCTTGGGCTCTGCCATGTGCCATCTGTGTGCACCAGG
GCA

Figure 40-8

SEQ ID NO: 41
Chromosome: 6
Coordinate: 41230059
Source: NCBI
Gene: TREML1
Accession Id.: NM_178174.2
GCCCATGGCTGGGCAGAGAAATGTCAACTCCTGGGCTTGCCTGGGCACTGATGCAGCATT
[CG] CCTGAGGGCAGGAAACATCTGCCTCAGAAAGTCACTTGGGGTGGGAGAAAGGAAATG
ATG

SEQ ID NO: 42
Chromosome 3
Coordinate: 48240150
Source: NCBI
Gene: CAMP
Accession Id.: NM_004345.3
ATTGCCCAGGTCCTCAGCTACAAGGAAGCTGTGCTTCGTGCTATAGATGGCATCAACCAG
[CG] GTCCTCGGATGCTAACCTCTACCGCCTCCTGGACCTGGACCCCAGGCCCACGATGGT
GAG

SEQ ID NO: 43
Chromosome: 6
Coordinate: 32892233
Source: NCBI
Gene: HLA-DOB
Accession Id.: NM_002120.2
GAGAAACAACCTGCAGTAGGCTGGGTCACAGAGGCAATCTGTGATTTTTTGGTCAGGACA
[CG] GAAACAAATCTCAGTTGGGGTATATGTGGACAAATGAAACTGGAAACAAAGGTTGCT
CCT

SEQ ID NO: 44
Chromosome: 20
Coordinate: 34707347
Source: NCBI
Gene: SLA2
Accession Id.: NM_032214.2
ACGCCCGTCCTAGTCCCATCTCAGGTGCGCACTTGCTGTGTGACTTTGGGCCCCTCTCTG
[CG] CTGCAGTCAGACTCCAAAGTCAGGAACGTGAGGGCTACCATCTCTCAAGACATTTCA
GCT

SEQ ID NO: 45
Chromosome: 19
Coordinate: 6621390
Source: NCBI
Gene: TNFSF14
Accession Id.: NM_003807.2
GTGACTCAGGTGGCAAGTGCAGTGGGGAGCCCCCAGCTTTCCCTTCTTGGATGCTTCATT
[CG] CTTGGGGCCACCAAATATCGACTGAGGACTTTCTGCCCATGCCAGGCTCTGCTCTCG
GTG

Figure 40-9

SEQ ID NO: 46
Chromosome: 3
Coordinate: 189379299
Source: NCBI
Gene: FLJ42393
Accession Id.: NM_207488.1
CTAGGAAACTTCTTCCATATATCATAAACAGAGACCAGTATTACAATACTTCACCCACTG
[CG] CCAATTTGGCTTTCATGTCTGTTTCCTGTGTCGATCACAACATCCTAGACAGCCCAA
ACA

SEQ ID NO: 47
Chromosome: 17
Coordinate: 43861946
Source: NCBI
Gene: SCAP1
Accession Id.: NM_003726.2
GCTCTCCAGGGGGCTGCGAGGGGCTCATGGGATCCCCATGGGCCAAGGCCAGGTGGTTGA
[CG] TGAGTTTTTGTCAGTGCGAAAACCCCAGCCCTCCCTTTATCACCCTGCAGACGTCTA
GGG

SEQ ID NO: 48
Chromosome: 17
Coordinate: 22823100
Source: NCBI
Gene: KSR1
Accession Id.: NM_014238.1
TGCCCCCAAGCAGGCCGGGACTGCCAGGCTTTACATCAGAGAACTGAGTTTCAGTTACCA
[CG] GTGAAGGCTGACAGCACAGAGCACAGTTCCGTGCAAATCAAGACACATTTCCCAAGT
CCC

SEQ ID NO: 49
Chromosome: 15
Coordinate: 66285305
Source: NCBI
Gene: CALML4
Accession Id.: NM_033429.1
CCACCCTTAAAGTCCTCAGAAGGTGGGAACTGAACTGGCACAGGATGGGAACCGGCTGTG
[CG] CTGGCCACTTGATTTTGCCAGCTGCCCTGTAATTCAGCTGGTGAGGAAACTGAGGCA
CAG

SEQ ID NO: 50
Chromosome: 9
Coordinate: 130007893
Source: NCBI
Gene: CIZ1
Accession Id.: NM_012127.2
TGATAGCCAATTAGGCTTGGGGACCTGCATGCCCAGCCCCTGCCTTCCTGGAGCCCATGA
[CG] CAGGGGCCATCCCTGACCACAGCAGATTTCATCGAGTACTTGCTTGTTGAGTGGTGG
AGC

Figure 40-10

SEQ ID NO: 51
Chromosome: 8
Coordinate: 71479423
Source: NCBI
Gene: NCOA2
Accession Id.: NM_006540.2
CTCCAGAGGGGATGGAGAGGCGCGACTGTGGGAGCTGGAAGGGGCACCACCCGGCAATTG
[CG] GGATAAAGCAAATGCTGCACACAGAGTGTGAAACTTAACCTGGTTGAGAATTTTCGG
CAC

SEQ ID NO: 52
Chromosome: 17
Coordinate: 24393906
Source: NCBI
Gene: PIPOX
Accession Id.: NM_016518.2
CTGACCTCACCACCCACCAGGGAGGTGGGTCTTATTCTGGGCATCGTGCCAAGTTCTTAG
[CG] GGGCCCTCTAGAATCTCTAAAGCAAATCAGGCTGAAGAGGGGAAAACCAGCAGGGGG
AGG

SEQ ID NO: 53
Chromosome: 9
Coordinate: 136949449
Source: NCBI
Gene: FCN1
Accession Id.: NM_002003.2
GGGTTGTTACCAGCTTTTAGGGACCAGAAAACCCAGGTCTGTCTCACCTGGACATGTGTC
[CG] CAGCCTGGGCAGGCAGGTTCTTGATATGCAGGAACAAGACTAGCAGGACAGCGAGCC
CCC

SEQ ID NO: 54
Chromosome: 19
Coordinate: 43995615
Source: NCBI
Gene: LGALS4
Accession Id.: NM_006149.2
TCCCTTGCCAGCTTCCCTGGTGACCAGCCAGGACCCAAATCACCTGGGTCCCCTCCCCTA
[CG] CCCTCCTGCAAAGAGGAAGTGCTCATGAACTTCGGCCCTGCCAGGGCCTTATCAGAG
CCC

SEQ ID NO: 55
Chromosome: 17
Coordinate: 70039110
Source: NCBI
Gene: CD300LB
Accession Id.: NM_174892.1
AGGCTGAGAAGGAGCAGAGCAGGGGGCAGCCACATGGCTCTGCCTTCCCGGCTCCTCGTC
[CG] CCTGATCTGCAACCAGTGGCAAATGCAGATCCCAGATGCACTCTGGAAGTTCTGCCT
GAG

Figure 40-11

SEQ ID NO: 56
Chromosome: 12
Coordinate: 6423750
Source: NCBI
Gene: TNFRSF7
Accession Id.: NM_001242.3
ACCAACTGGGAGGAAGCTTAAATAGCCTTGTCTCAATTGAGGTCTGGTTTGATGGCCAAA
[CG] AGTTTGCTACAGAATGCTCAGAATTGCAAGCAAGGGGTGTAGAGCTGCCTCTCTTCT
GTC

SEQ ID NO: 57
Chromosome: 11
Coordinate: 59979867
Source: NCBI
Gene: MS4A1
Accession Id.: NM_152866.2
AGTTTGTTTCTCAAGCACACTGGGAGGGTGAGTGGTGTAGTCCAGGCCTGAAGATGAAAT
[CG] CTGATAGACATCAGGTGACAGGAAATCAGTAGCTTCTGCTACCTTGGGCTTCGCTCC
AAT

SEQ ID NO: 58
Chromosome: 22
Coordinate: 43451388
Source: NCBI
Gene: PRR5
Accession Id.: NM_015366.3
AAGCCCAGCTGCTGGCTGATAAATATTTTATCACTGCTCACAGAGCAGTCCCCAGGAAGG
[CG] CCTGCATCCTCCAAGCCCACAGAGCACCCCTTCCTGCCCGGACAGAAGGAACTGGCC
AGG

SEQ ID NO: 59
Chromosome: 5
Coordinate: 118718932
Source: NCBI
Gene: TNFAIP8
Accession Id.: NM_014350.1
AAGGCCCTTTGGAGTAACTGCAGCAATGAGTGCCCGGGCTGTGCTTGGAGTACCAGTGCT
[CG] CCCGGGGCTATACTGAATGAGTAAGCAGCCCCGTCTGCTTTTGCTGTGCAAAGGTAA
GGG

SEQ ID NO: 60
Chromosome: 17
Coordinate: 25729371
Source: NCBI
Gene: CPD
Accession Id.: NM_001304.3
GAAATCTGCCTAATGAGGGTCGAGGCCAGCACACACAGGGACCTATTTGCAGTAAAACAA
[CG] TGGGGTGACGCCTAAGAAATAGACAACATTAACACAAAGGGAGCCTACTACGTAGCA
CAT

Figure 40-12

SEQ ID NO: 61
Chromosome: 19
Coordinate: 15451532
Source: NCBI
Gene: PGLYRP2
Accession Id.: NM_052890.2
TGGCCAGCAGCGGCTACAGAGCCGTCACTATGGGGAGGGACAGGACTTGAGGGGTTGCCT
[CG]GTCCACCTCACTGGAGAATGGGCAGAGTTTATGGAGTCTGAACCACCTGGTCTCCAG
GCC

SEQ ID NO: 62
Chromosome: 1
Coordinate: 24386999
Source: NCBI
Gene: IL28RA
Accession Id.: NM_170743.2
TTTCTGACCCCGAAGGCTGTGGTGTTCACCTGGACAGCAGTAGCTTCCCAGTAAGGCACA
[CG]CCACGACGCGCAATATTATGCGGCCCTTTAGGAGGACGTTGCCGAATGGTGTGTATC
GAC

SEQ ID NO: 63
Chromosome: 11
Coordinate: 67534528
Source: NCBI
Gene: ALDH3B1
Accession Id.: NM_000694.2
AGTGGGCCAGCAGTCGGGCCAGAGTCCAGCTCAGCAACTCCGGGTTACAGGCAGCCCAGG
[CG]GGCCTAGCCACCGGCAGCTGCACTCAGAGGCCACTGTGTCCTGGCTGAGCTCATCTG
CCT

SEQ ID NO: 64
Chromosome: 11
Coordinate: 3077976
Source: NCBI
Gene: OSBPL5
Accession Id.: NT_009237.17
CGGCAAAGCCACGCTCACCTTCCTGAACCGAGCCGAGGATTACACCCTTACCATGCCCTA
[CG]CCCACTGCAAAGGTGAGAGGCTCAGCCACACACTCCGAGGGCAGAGCCAGGCTCTGT
GAG

SEQ ID NO: 65
Chromosome: 19
Coordinate: 15252927
Source: NCBI
Gene: BRD4
Accession Id.: NM_058243.1
GAGCTGCCTCGGCGCACGGCCACTGGCCCGGCTCCAGGCGGCGCAGTCTGGCTGATGACA
[CG]AGCGCTGTTCTCACCAGCTGCCTGAGCCAGTCAGATGGAAAAGTAATCCTATTTGTG
CTT

Figure 40-13

SEQ ID NO: 66
Chromosome: 16
Coordinate: 56258956
Source: NCBI
Gene: GPR97
Accession Id.: NM_170776.3
TAAAGACAACAATTTCACAGCTCTGATGATCAGAAATGATGTAATGGCCACAGGCGGCTC
[CG] CCTGCGTCATCCATGATTTCATCACACACCTCGGGAGGCTCAGGGTGACAGACAGTG
CAT

SEQ ID NO: 67
Chromosome: 12
Coordinate: 13139815
Source: NCBI
Gene: GSG1
Accession Id.: NM_031289.1
AAACCAAAGGGACTTGGAGTGCAGATGGCATCCTTCGGTTCTTCCAGACAAGCTGCAAGA
[CG] CTGACCATGGCCAAGGTAACCGGCTTCCCCTCCTATTGCTCAAAGGATGCAGTCTAC
AGC

SEQ ID NO: 68
Chromosome: 10
Coordinate: 81699171
Source: NCBI
Gene: SFTPD
Accession Id.: NM_003019.4
AGAAGTGGACACAGCAGGTCTTGGCTCTTAAGATCTGCAGTTGTGAGTTCCTTTTGCAAT
[CG] CTGTAGGTCATTGTGCAACCTGCTGGTCTCTGGACTCCTGATTTCTAGACATCTATA
AAA

SEQ ID NO: 69
Chromosome: 20
Coordinate: 58063710
Source: NCBI
Gene: FLJ33860
Accession Id.: NM_173644.1
CCCTTCAAAGCCCGCCTTCTTGCCGTGTGATGCTGCCTGGGCCAGCAGGGCAGGTCACCA
[CG] CTGTCTCTTCAAAGCAGCTCGCTCATGCCCACAGCGCTGGGCACAAGGGCAGCCACG
AGC

SEQ ID NO: 70
Chromosome: 2
Coordinate: 73723682
Source: NCBI
Gene: NAT8
Accession Id.: NM_003960.2
TCGGGTACAAGAGCATGAATTTGGGCCTCCCCAACATCTGCAGTGCAAAATATTTAACAA
[CG] GGTGTGGCACAGCCTCTGACCAACAGCCAGAACACACACAAGCCACACACAGCCATG
CCT

Figure 40-14

SEQ ID NO: 71
Chromosome: 14
Coordinate: 22375620
Source: NCBI
Gene: MMP14
Accession Id.: NM_004995.2
TTTTCCGGTTTTTGATCTTTCTTCTGCTTAGTCGGCGAACTGGGGTCTGGTTCCCTCTCT
[CG] CTCTCTCCTCTGGTCCCTCCCTTCTCCCACAGCCTCTCCTCCGTCCCCGCCCCAGTG
CCC

SEQ ID NO: 72
Chromosome: 17
Coordinate: 43977490
Source: NCBI
Gene: HOXB2
Accession Id.: NM_002145.2
CCAGGCCCAGACGAGCGATTGGCGGAGGCCGGTCCCGTGACCACGAATTCCCTGTAATTT
[CG] CTGGAGTCCTGGGTTTAATAGAGAGAGTCCCCATACGCTTGTATTTATCAGCAATAT
ACA

SEQ ID NO: 73
Chromosome: 11
Coordinate: 33870664
Source: NCBI
Gene: LMO2
Accession Id.: NM_005574.2
CTCACATGACCTGGATTGGAACTTGGCTCAGCCACTGACTAGCCAGACAAACTCAAATAA
[CG] TACACAGCTTCTCAGCACCTCACCCCCTCATTCATAAAAAACAGGGACAATGGTACC
CAC

SEQ ID NO: 74
Chromosome: 10
Coordinate: 124729456
Source: NCBI
Gene: C10orf89
Accession Id.: NM_153336.1
CTGTCTGCATGCACTGGAATTGAGGTCTGTGGATGTGCCTTTCCTGACAATATTTCTTCA
[CG] CTTGCTGCCCACTGGTGCTGTGAGGGCACAATAACGAATGTTTACTTTGCCCTTGCA
CTC

SEQ ID NO: 75
Chromosome: 9
Coordinate: 128925551
Source: NCBI
Gene: ANGPTL2
Accession Id.: NM_012098.2
TGAGCTAATTAATACTAGTAATCTACCTGCAACAGCTGCAGCGAGGACTCTGTGAGGTCA
[CG] TGGGAAGGAGCTTGGCACAGTGTCAAGGACGCCTCCTTGAACTGAGCTTAGGACTCT
GGA

Figure 40-15

SEQ ID NO: 76
Chromosome: 22
Coordinate: 35586846
Source: NCBI
Gene: NCF4
Accession Id.: NM_013416.2
GGAAGTGGACCCTCGGGTGCCAGGTTTGCAGGAATCCACTTCCTTGATGTCAGTCCTTGG
[CG] CCAAGCCTCAGTTGGGTATCAGAAGCCTTGCTCCATCAGAGATGGGGTCCCAGCCAT
CAG

SEQ ID NO: 77
Chromosome: 20
Coordinate: 61962518
Source: NCBI
Gene: C20orf135
Accession Id.: NM_080622.2
CTGCCAGCAACAGCAGTGACCTTCTGGGGCGGGTCCTGCCTGGCTGGGGTTCCTCTTTCT
[CG] CTCCTGGGTCGAGCCCCCACTCCCAGGCTGCGCCTCCCTCTTTTCTGGAGAGGTATC
TTT

SEQ ID NO: 78
Chromosome: 20
Coordinate: 23419693
Source: NCBI
Gene: CST8
Accession Id.: NM_005492.2
TGTGGGTCCTGGTCTGCGGTCTCTCTTGCCCCTCTGAGTCCACGCCCTGCAGGGAGGTTA
[CG] CTTTGTGATGTAATTCAGCACCTGTGTCTTGTCCCAGTGAGGACATCTCCCACTTGC
CAG

SEQ ID NO: 79
Chromosome: 13
Coordinate: 98028005
Source: NCBI
Gene: STK24
Accession Id.: NM_003576.2
CTTTGGTTACCGAAAACAGCCCGGCTGGGACTGCTGGGCTGGGAACTTAGCTAAGCAGTG
[CG] GAGGCTGAACCCCACCATCTCTGGGATCCGCAGCAAATCAGAAGCCCCCACCCACGA
TAA

SEQ ID NO: 80
Chromosome: 1
Coordinate: 54786297
Source: NCBI
Gene: ACOT11
Accession Id.: NM_015547.2
TGGGGGTGCCTGGAGTTTGGCTGGGGCTGGGTGCCCAGTGGGCGGGCACAGGCCCCTTGA
[CG] TGGCTGTGGCCTAGCTGGCAGCCTCGTCCTTCCTCTCCGCTAGGCGGGCACTGGAGC
TTT

Figure 40-16

SEQ ID NO: 81
Chromosome: 17
Coordinate: 59437937
Source: NCBI
Gene: ICAM2
Accession Id.: NM_000873.2
CTTCCCAGCTTCCTCTGCCTGGATTCTTAGAGGCCTGGGGTCCTAGAACGAGCTGGTGCA
[CG] TGGCTTCCCAAAGATCTCTCAGATAATGAGAGGAAATGCAGTCATCAGTTTGCAGAA
GGC

SEQ ID NO: 82
Chromosome: 11
Coordinate: 72606702
Source: NCBI
Gene: P2RY2
Accession Id.: NM_002564.2
AGGGGCGGGACAGGGGTAGGGTGGCGCGGTGGCTGGGCGCAAAGGTCCCGCAGTGGGCCA
[CG] CAGGCACCGGGCTGACCTGGCAAAACTTTGGCGTCTCTGAAAACCTCTGGTAACCAG
CTC

SEQ ID NO: 83
Chromosome: 7
Coordinate: 100077108
Source: NCBI
Gene: TFR2
Accession Id.: NM_003227.2
TGCCTGCCAGGACTGATAAGGGGCCCTCCTAGGGCTCCCACAAACGGTTTATCGGTTTAT
[CG] CTGGGGGACAGCCTGCAGGCTTCAGGAGGGGACACAAGCATGGAGCGGCTTTGGGGT
CTA

SEQ ID NO: 84
Chromosome: 16
Coordinate: 11256179
Source: NCBI
Gene: SOCS1
Accession Id.: NT_010393.15
CCTCCGCGACTACCTGAGCTCCTTCCCCTTCCAGATTTGACCGGCAGCGCCCGCCGTGCA
[CG] CAGCATTAACTGGGATGCCGTGTTATTTTGTTATTACTTGCCTGGAACCATGTGGGT
ACC

SEQ ID NO: 85
Chromosome: 11
Coordinate: 122214681
Source: NCBI
Gene: CRTAM
Accession Id.: NM_019604.2
GACACACACTATAATGATCCTTTCTATACTCCTTAGCCATTGAACGAGAGATCAAATAAA
[CG] CAGTAACATCCCTCAGATGCATGATTTGAGCATGGCTTGGAAAGTATTAGCAGTTAC
CTG

Figure 40-17

SEQ ID NO: 86
Chromosome: 2
Coordinate: 106047843
Source: NCBI
Gene: ECRG4
Accession Id.: NM_032411.1
GTTGAACAGGCCAGTTACTGGGATGCAGTTCTGCGTTTCCCTTGGGTCTCACCTTAACAT
[CG] CTCGCTGAAGTGTGCCCAGATTACAGAGCGGGCAAAGGGAAGCAGTGGTTTTGCTCA
CAG

SEQ ID NO: 87
Chromosome: 15
Coordinate: 56217683
Source: NCBI
Gene: AQP9
Accession Id.: NM_020980.2
TTGGTTTTTTTCAAGAGATGAGAAAAGAGATGTGCCAGTTGTGTTGCCAAATCACAGTGA
[CG] GGCCCTGGTCCAGAAAAGATTTTCATGTTACACAATTGCAGGCTTCTGATTTTTTTT
TCT

SEQ ID NO: 88
Chromosome: 7
Coordinate: 94374723
Source: NCBI
Gene: PPP1R9A
Accession Id.: XM_371933.2
CCTGGGGCAAGGCCCCTTCCTGTTCGGGTGTTGGCTCCGGAACTTGGTTCTGGGGCTGAC
[CG] CTGCTGGGGCCCCACTTAGTCTGAGTCTGCAGTTAACTCCGTGACCCCAAGGCATCC
AAG

SEQ ID NO: 89
Chromosome: 2
Coordinate: 128174869
Source: NCBI
Gene: SFT2D3
Accession Id.: NM_032740.3
TGCTTCTCTATTCTGTTCTCAGTTTCGGCCACAGGCCTGGCAACATCCTTGACTCCTTCG
[CG] CCCCTTGTCCAAGACTCGGTGCTGCTGTCCCATGTGTTTGGTGTCACTCTCGTGCTC
TGG

SEQ ID NO: 90
Chromosome: 1
Coordinate: 27822930
Source: NCBI
Gene: FGR
Accession Id.: NM_005248.1
ATTGGAGCCGGTGGCCACGGCCAAGGAGGATGCTGGCCTGGAAGGGGACTTCAGAAGCTA
[CG] GGGCAGCAGACCACTATGGGCCTGACCCCACTAAGGCCCGGCCTGCATCCTCATTTG
CCC

Figure 40-18

SEQ ID NO: 91
Chromosome: 19
Coordinate: 60109459
Source: NCBI
Gene: NCR1
Accession Id.: NM_004829.3
TGAAGGAAGGACTCACGCTGCTGGGCGCTGATCCTCTGACTCAGACACAGCCCTGGAAGA
[CG] GGAGTAATGAGACCTGTTGCCTCCCAGGCACACCGTGATCCCATTCCCCTTCCACGC
CAG

SEQ ID NO: 92
Chromosome: 11
Coordinate: 117720322
Source: NCBI
Gene: CD3G
Accession Id.: NM_000073.1
TGGAGCCAGTCTAGCTGCTGCACAGGCTGGCTGGCTGGCTGGCTGCTAAGGGCTGCTCCA
[CG] CTTTTGCCGGAGGACAGAGACTGACATGGAACAGGGGAAGGGCCTGGCTGTCCTCAT
CCT

SEQ ID NO: 93
Chromosome: 11
Coordinate: 117718540
Source: NCBI
Gene: CD3D
Accession Id.: NM_000732.3
AGGGCAGCTCTCACCCAGGCTGATAGTTCGGTGACCTGGCTTTATCTACTGGATGAGTTC
[CG] CTGGGAGATGGAACATAGCACGTTTCTCTCTGGCCTGGTACTGGCTACCCTTCTCTC
GCA

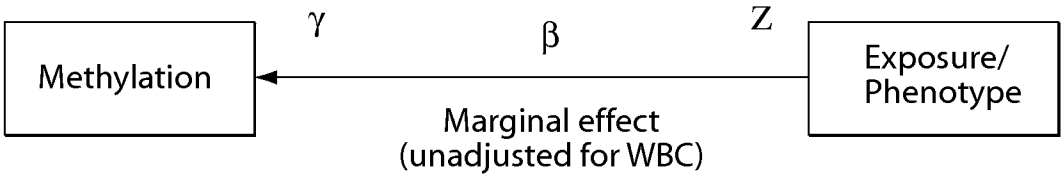
SEQ ID NO: 94
Chromosome: 1
Coordinate:
Source: NCBI
Gene: CD3Z
Accession Id.: NM_000734.2
CTGCCTCCCAGCCTCTTTCTGAGGGAAAGGACAAGATGAAGTGGAAGGCGCTTTTCACCG
[CG] GCCATCCTGCAGGCACAGTTGCCGATTACAGGTAGGGCCGACGTGTCGACGGCAGGG
AAC

SEQ ID NO: 95
Chromosome: X
Coordinate:
Source: NCBI
Gene: FOXP3
Accession Id.: NM_014009.2
TTGGATTATTAGAAGAGAGAGGTCTGCGGCTTCCACACCGTACAGCGTGGTTTTTCTTCT
[CG] GTATAAAAGCAAAGTTGTTTTTGATACGTGACAGTTTCCCACAAGCCAGGCTGATCC
TTT

Figure 40-19

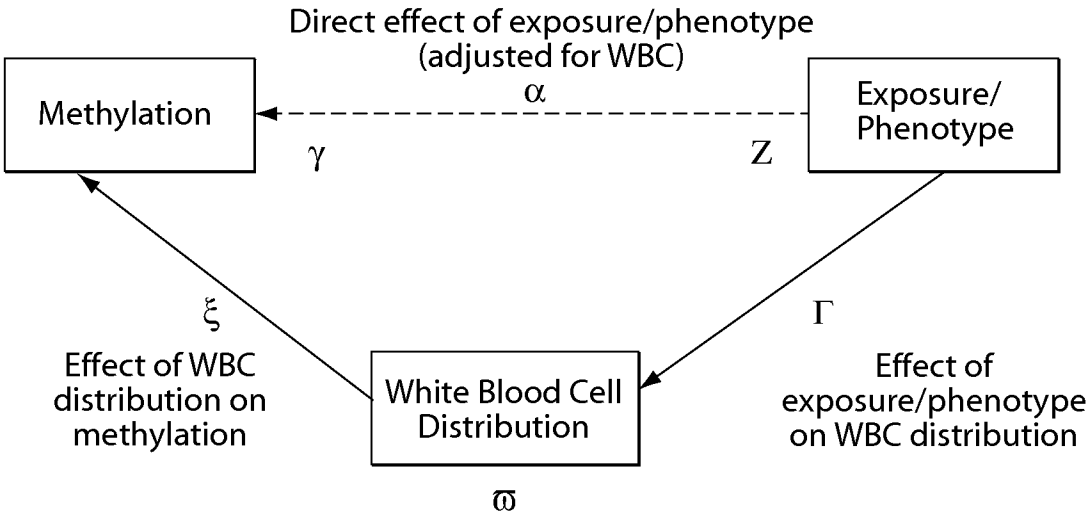
SEQ ID NO: 96
Chromosome: X
Coordinate:
Source: NCBI
Gene: FOXP3
Accession Id.: NM_014009.2
TCGATGAAGCCCGGCGCATCCGGCCGCCATGACGTCAATGGCGGAAAAATCTGGGCAAGT
[CG]GGGGCTGTGACAACAGGGCCCAGATGCAGACCCCGATATGAAAACATAATCTGTGTC
CCA

Figure 40-20



Different sources of effects on measured DNA methylation

Figure 41A



Different sources of effects on measured DNA methylation

Figure 41B

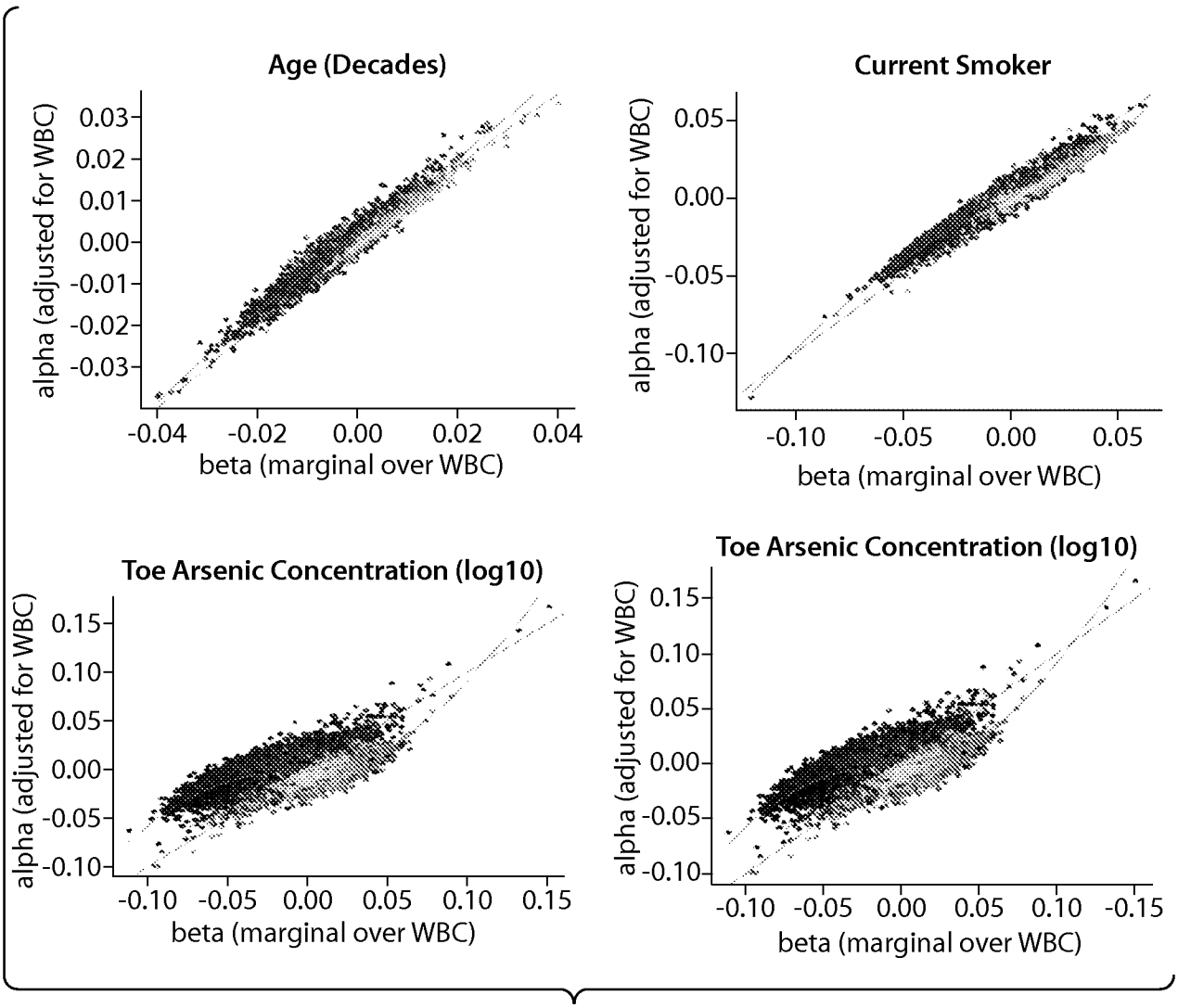

Age (Decades)
alpha (adjusted for WBC)
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04 -0.02 0.00 0.02 0.04
beta (marginal over WBC)
Current Smoker
alpha (adjusted for WBC)
0.05
0.00
-0.05
-0.10
-0.10 -0.05 0.00 0.05
beta (marginal over WBC)
Toe Arsenic Concentration (log10)
alpha (adjusted for WBC)
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.10 -0.05 0.00 0.05 0.10 0.15
beta (marginal over WBC)
Toe Arsenic Concentration (log10)
alpha (adjusted for WBC)
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.10 -0.05 0.00 0.05 0.10 -0.15
beta (marginal over WBC)

Figure 42

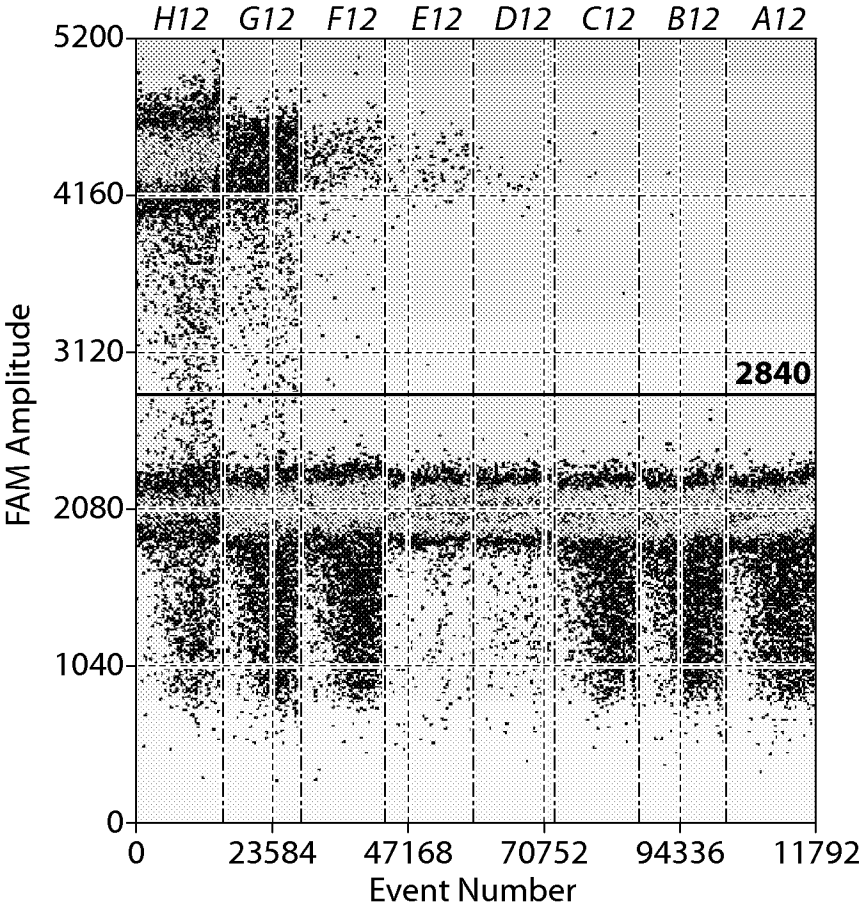

H12
G12
F12
E12
D12
C12
B12
A12
5200
4160
3120
2080
1040
0
2840
FAM Amplitude
0
23584
47168
70752
94336
11792
Event Number

Figure 43A

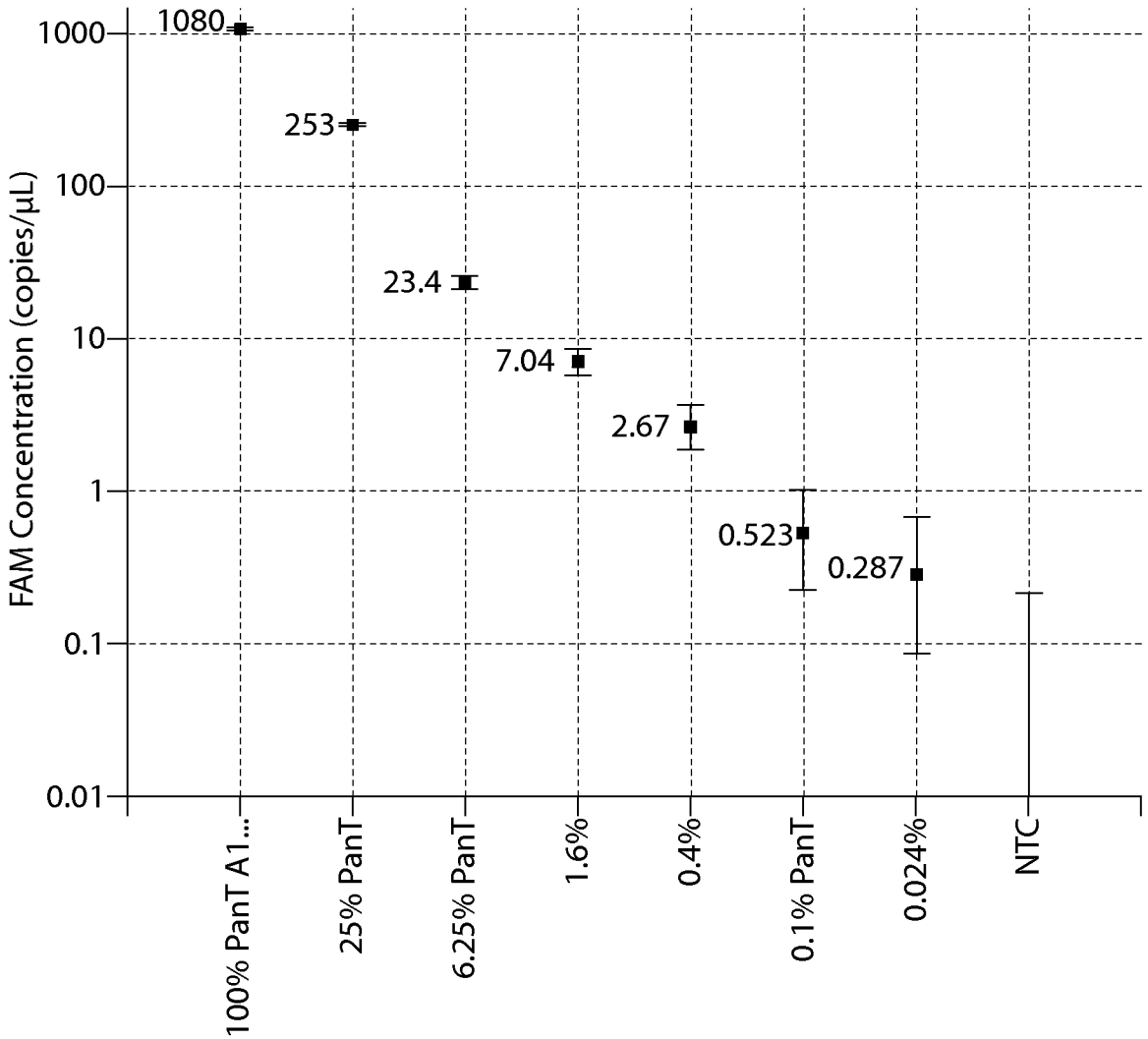

FAM Concentration (copies/µL)
1000
100
10
1
0.1
0.01
1080
253
23.4
7.04
2.67
0.523
0.287
100% PanT A1...
25% PanT
6.25% PanT
1.6%
0.4%
0.1% PanT
0.024%
NTC

Figure 43B

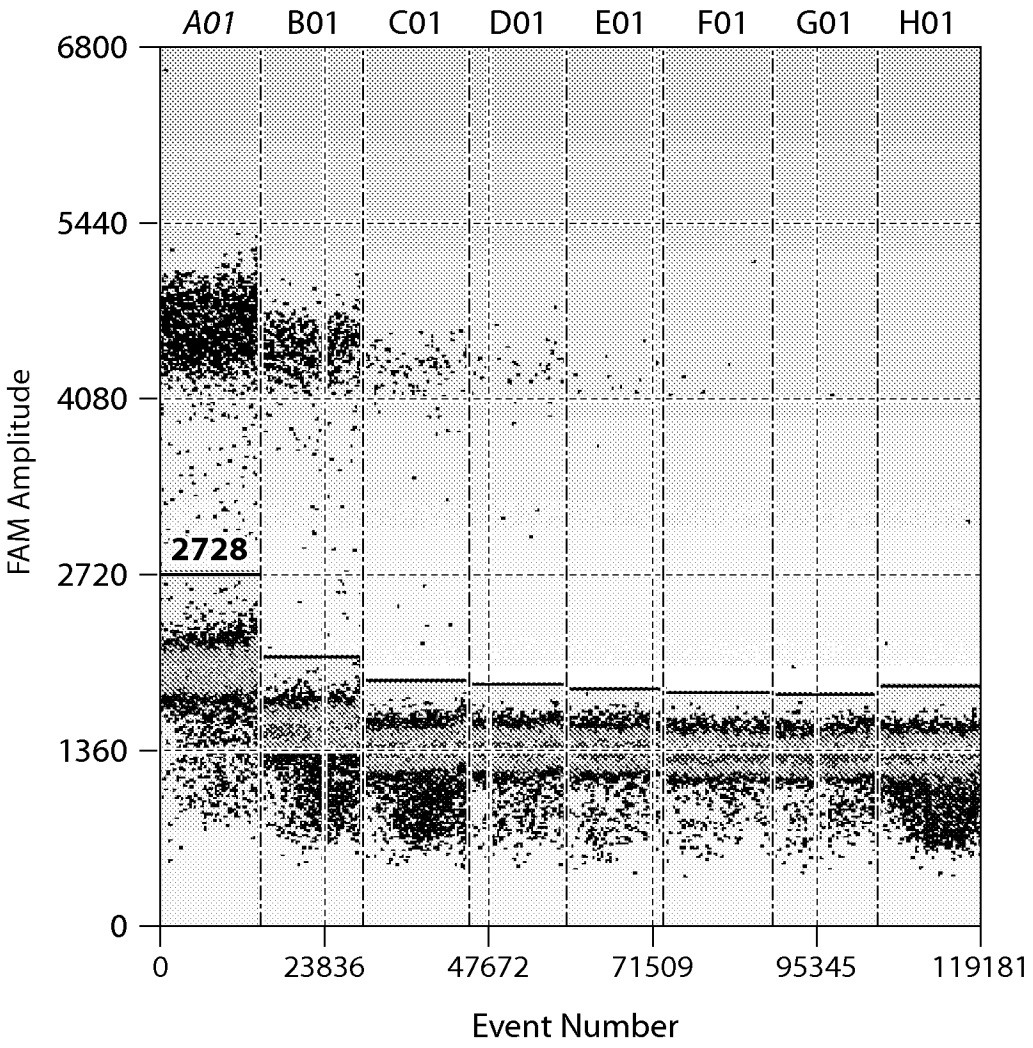

A01
B01
C01
D01
E01
F01
G01
H01
6800
5440
4080
2720
1360
0
FAM Amplitude
2728
0
23836
47672
71509
95345
119181
Event Number

Figure 44A

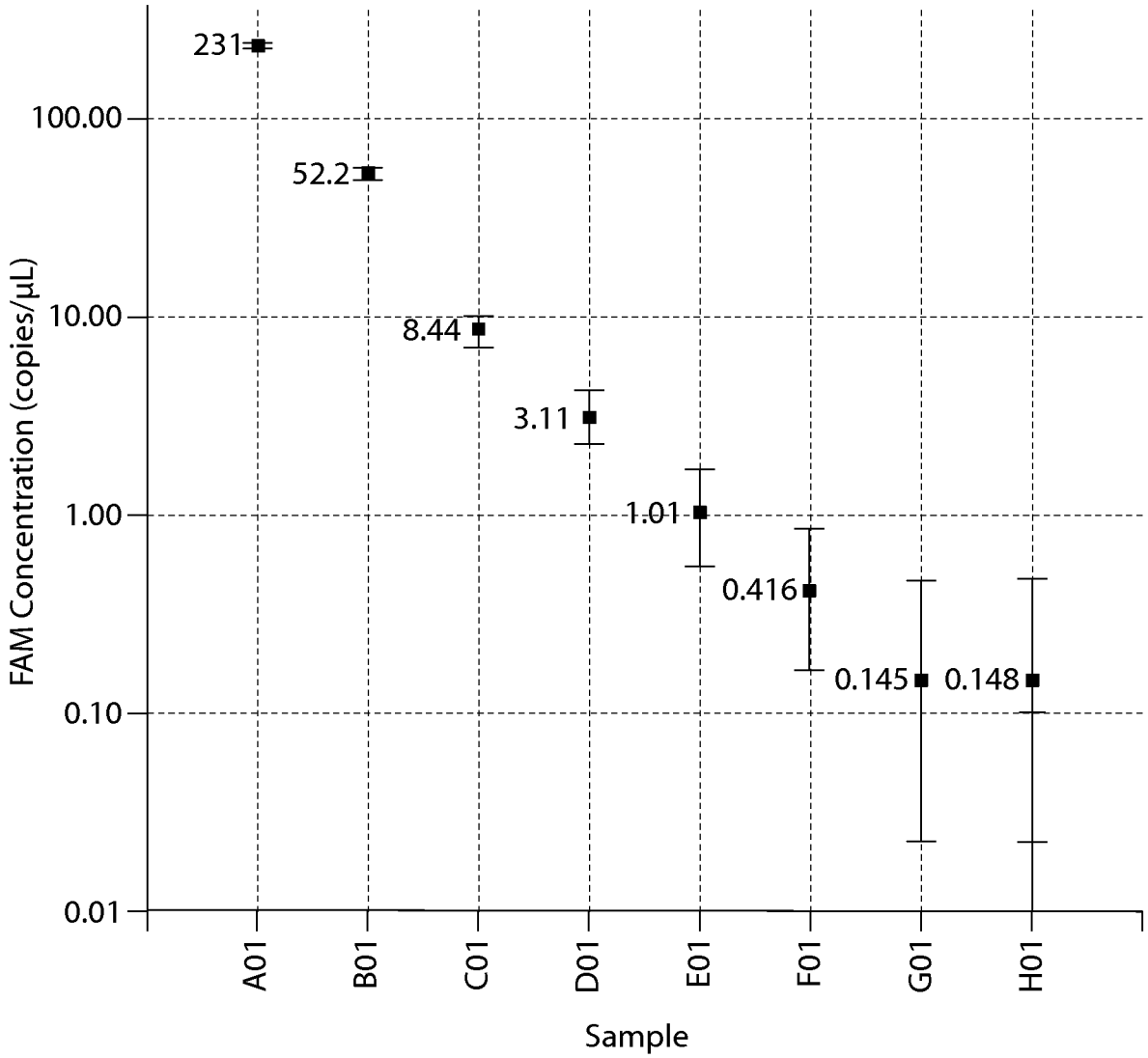

231
52.2
8.44
3.11
1.01
0.416
0.145
0.148
100.00
10.00
1.00
0.10
0.01
FAM Concentration (copies/µL)
A01
B01
C01
D01
E01
F01
G01
H01
Sample

Figure 44B

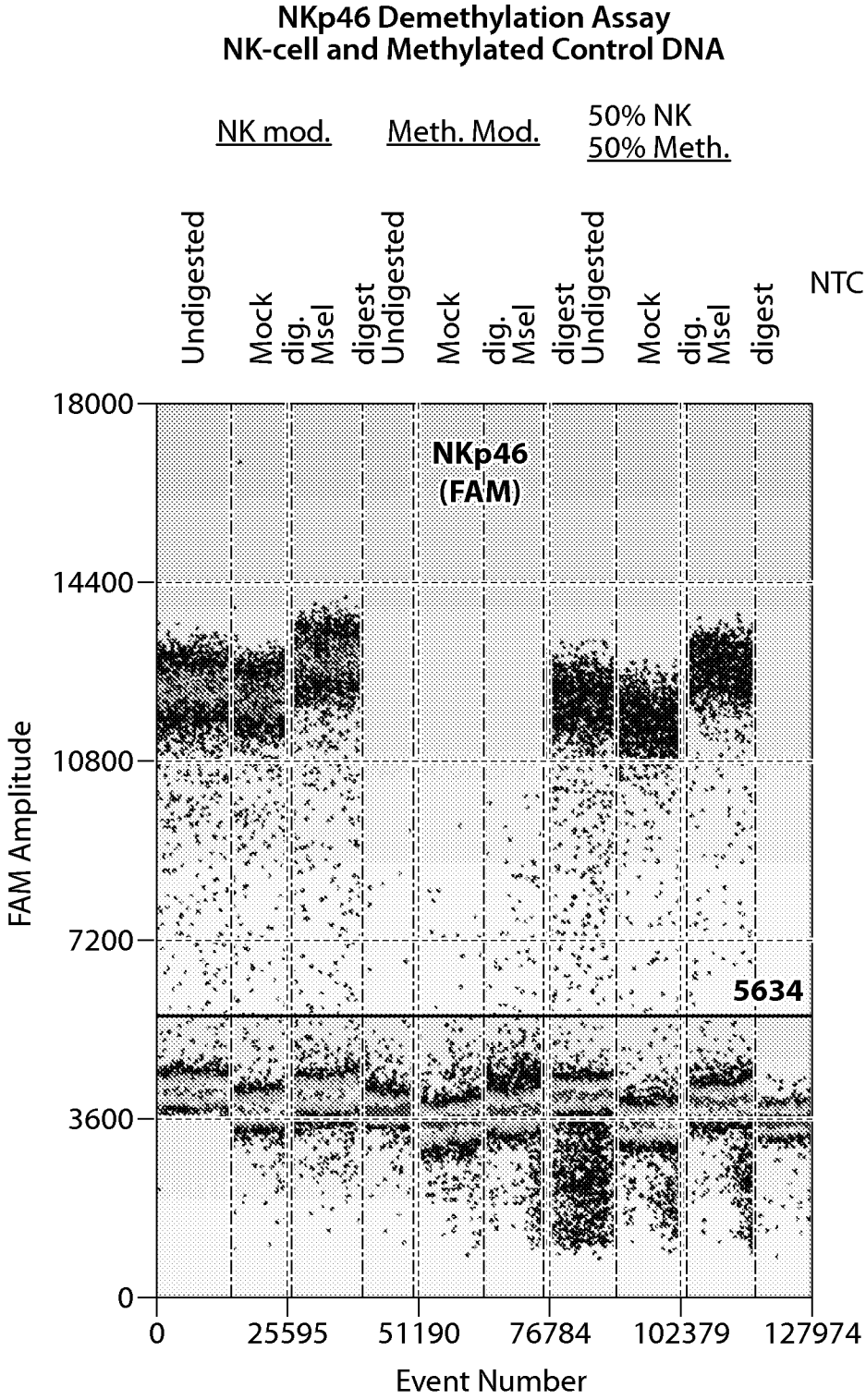

NKp46 Demethylation Assay
NK-cell and Methylated Control DNA
NK mod.
Meth. Mod.
50% NK
50% Meth.
Undigested
Mock dig. MseI digest
Undigested
Mock dig. MseI digest
Undigested
Mock dig. MseI digest
NTC
NKp46 (FAM)
5634
FAM Amplitude
18000
14400
10800
7200
3600
0
0
25595
51190
76784
102379
127974
Event Number

Figure 45A

NKp46 Demethylation Assay
NK-cell and Methylated Control DNA

Sample	Restriction Enzyme Digestion	C-less (FAM) conc. (copies/ul)	NKp46 (FAM) conc. (copies/ul)	NKp46 / C-less
NK2510	None	524	403	0.769
NK2510	Mock	506	352	0.696
NK2510	MseI	467	353	0.756
Methylated Control	None	432	0.769	0.002
Methylated Control	Mock	320	1.04	0.003
Methylated Control	Mse I	466	4.76	0.010
50/50 (NK/Meth)	None	515	181	0.351
50/50 (NK/Meth)	Mock	339	146	0.431
50/50 (NK/Meth)	MseI	432	134	0.310
NTC	None	0	0	N/A

Figure 45B

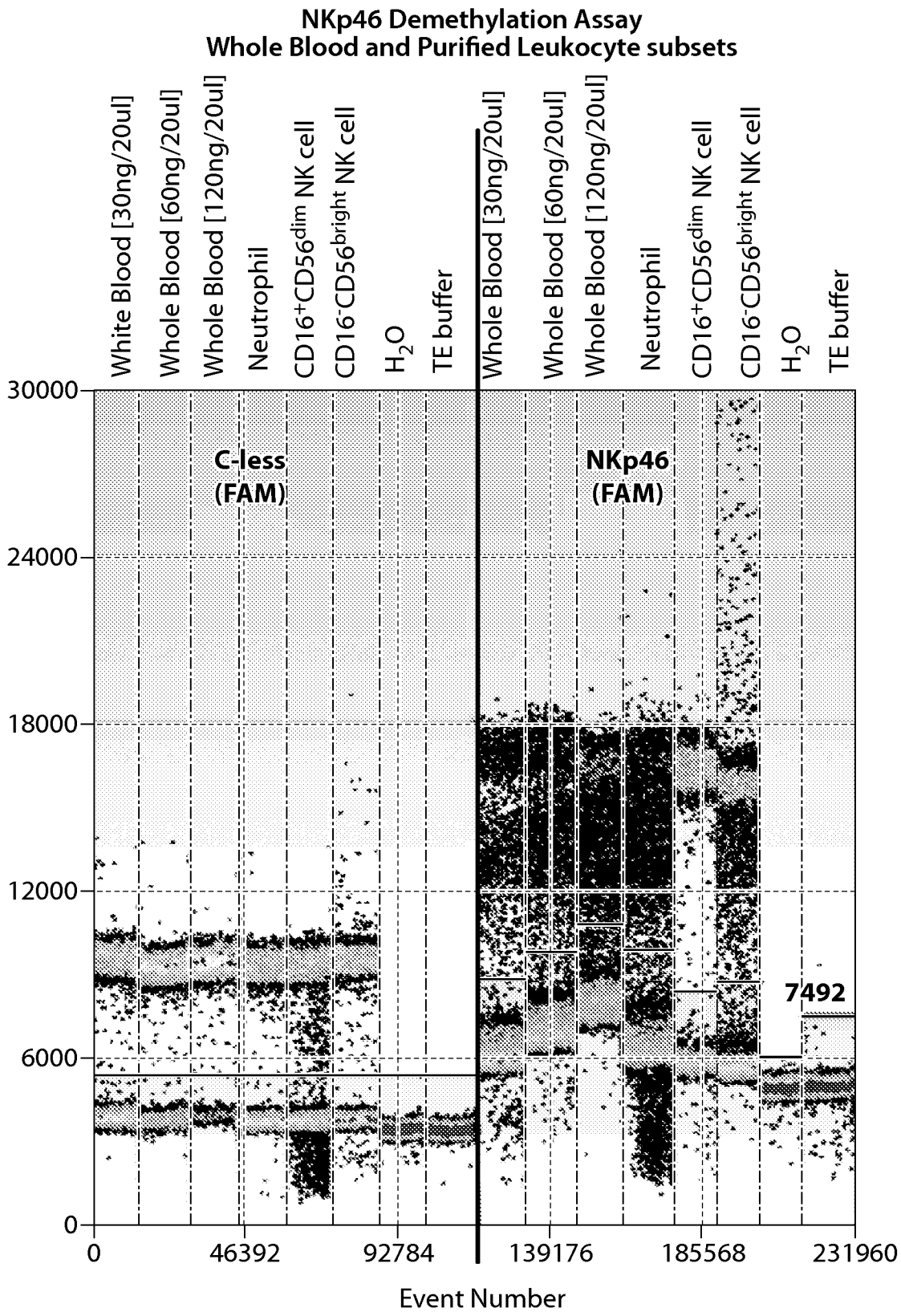

NKp46 Demethylation Assay
Whole Blood and Purified Leukocyte subsets
White Blood [30ng/20ul]
Whole Blood [60ng/20ul]
Whole Blood [120ng/20ul]
Neutrophil
CD16+CD56dim NK cell
CD16-CD56bright NK cell
H2O
TE buffer
Whole Blood [30ng/20ul]
Whole Blood [60ng/20ul]
Whole Blood [120ng/20ul]
Neutrophil
CD16+CD56dim NK cell
CD16-CD56bright NK cell
H2O
TE buffer
C-less (FAM)
NKp46 (FAM)
30000
24000
18000
12000
6000
0
7492
0
46392
92784
139176
185568
231960
Event Number

Figure 46A

Sample	NKp46 (FAM) copies/ul	C-less (FAM) copies/ul	NKp46 / C-less
Whole Blood DNA [30ng/20ul]	150	724	0.207
Whole Blood DNA [60ng/20ul]	268	1450	0.185
Whole Blood DNA [120ng/20ul]	463	3070	0.151
Neutrophil DNA	325	858	0.379
$CD16^{+}CD56^{dim}$ NK cell Cell DNA	720	779	0.924
$CD16^{-}CD56^{bright}$ NK cell Cell DNA	523	776	0.674

Figure 46B