

METHODS AND COMPOSITIONS FOR PREDICTING AND/OR MONITORING CARDIOVASCULAR DISEASE AND TREATMENTS THEREFOR

CROSS REFERENCE TO RELATED APPLICATIONS

5 This application claims the benefit of priority under 35 U.S.C. § 119(e) to U.S. Application No. 63/074,878 filed on September 4, 2020. This document is incorporated by reference herein in its entirety.

TECHNICAL FIELD

10 This disclosure generally relates to methods and compositions related to predicting cardiovascular disease (CVD) in an individual such as, for example, coronary heart disease (CHD).

BACKGROUND

15 Cardiovascular disease (CVD), and particularly coronary heart disease (CHD), is the most common type of heart disease and was responsible for over 360,000 deaths in the United States in 2017. In order to decrease this toll, a number of risk estimators have been developed to better identify those with or at risk for CHD. Beginning with the Framingham Risk Score (FRS) and more recently, the ASCVD Pooled Cohort Equation (PCE), these tools capture variance in key physiological parameters, such as serum lipid levels, known to be
20 associated with risk for CVD, including CHD.

 Despite the magnitude of these efforts, current risk estimators often lack in sensitivity and specificity. As a result, there is a need for alternative stratification approaches for CVD.

SUMMARY

25 Methods and compositions for predicting the incidence or risk of cardiovascular disease (CVD) are provided. For example, methods and compositions for predicting the one-year, three-year or five-year incidence of coronary heart disease (CHD) are described herein. The general principals apply to other windows of incidence (e.g., one-month, six-month,

two-year, or ten-year) as well as the incidence or prevalence of other types of CVD including, without limitation, CHD, stroke, arrhythmia, cardiac arrest, and congestive heart failure. Specifically, methods and compositions for determining the methylation status of at least one CpG locus and at least one single nucleotide polymorphism (SNP) are described.

5 In one aspect, kits for determining methylation status of at least one CpG dinucleotide and a genotype of at least one single-nucleotide polymorphism (SNP) are provided. Such kits typically include at least one first nucleic acid primer at least 8 nucleotides in length that is complementary to a bisulfite-converted nucleic acid sequence comprising a first CpG dinucleotide at a GC locus selected from the group consisting of cg00300879, cg09552548,
10 and cg14789911 or at a second CpG dinucleotide in linkage disequilibrium with the first CpG dinucleotide at a GC locus selected from the group consisting of cg00300879, cg09552548, and cg14789911, wherein the linkage disequilibrium has a value of $R > 0.3$, wherein the at least one first nucleic acid primer detects a methylated or unmethylated CpG dinucleotide, and at least one second nucleic acid primer at least 8 nucleotides in length that
15 is complementary to a DNA sequence or a bisulfite-converted DNA sequence of a first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 or a second SNP in linkage disequilibrium with the first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, wherein the linkage disequilibrium has a value of $R > 0.3$.

20 In some embodiments, the at least one first nucleic acid primer detects the unmethylated CpG dinucleotide. In some embodiments, the at least one first nucleic acid primer detects the methylated CpG dinucleotide.

 In some embodiments, the kits described herein further including at least a third nucleic acid primer at least 8 nucleotides in length that is complementary to a nucleic acid
25 sequence upstream of the CpG dinucleotide. In some embodiments, the kits further include at least a third nucleic acid primer at least 8 nucleotides in length that is complementary to a nucleic acid sequence downstream of the CpG dinucleotide.

 In some embodiments, the at least one first nucleic acid primer comprises one or more nucleotide analogs. In some embodiments, the at least one first nucleic acid primer
30 comprises one or more synthetic or non-natural nucleotides.

In some embodiments, the kits described herein further include a solid substrate to which the at least one first nucleic acid primer is bound. In some embodiments, the substrate is a polymer, glass, semiconductor, paper, metal, gel or hydrogel. In some embodiments, the solid substrate is a microarray or microfluidics card.

5 In some embodiments, the kits described herein further include a detectable label.

In another aspect, methods of determining the presence of biomarkers associated with predicting CHD in a biological sample from a patient is provided. Such methods typically include (a) providing a first portion of the biological sample and a second portion of the biological sample, wherein the nucleic acid from at least the first portion is bisulfite
10 converted; (b) contacting the first portion of the biological sample with a first oligonucleotide primer at least 8 nucleotides in length that is complementary to a sequence that comprises a first CpG dinucleotide at a GC locus selected from the group consisting of cg00300879, cg09552548, and cg14789911, or a second CpG dinucleotide in linkage disequilibrium with the first CpG dinucleotide at a GC locus selected from the group
15 consisting of cg00300879, cg09552548, and cg14789911, wherein the linkage disequilibrium has a value of $R > 0.3$, wherein the first nucleic acid primer detects a methylated or unmethylated CpG dinucleotide; and (c) contacting the second portion of the biological sample with a nucleic acid primer at least 8 nucleotides in length that is complementary to a DNA sequence or a bisulfite-converted DNA sequence of a first SNP selected from the group
20 consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 or a second SNP in linkage disequilibrium with the first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, wherein the linkage disequilibrium has a value of $R > 0.3$. Generally, the percentage of methylation of the CpG dinucleotide at the GC locus selected from the group consisting of cg00300879, cg09552548,
25 and cg14789911, and the identity of the nucleotide at the first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 or the second SNP in linkage disequilibrium with the first SNP are biomarkers associated with the incidence of CHD.

30 In some embodiments, the biological sample is selected from the group consisting of blood and saliva.

In some embodiments, the at least one first nucleic acid primer detects the unmethylated CpG dinucleotide. In some embodiments, the at least one first nucleic acid primer detects the methylated CpG dinucleotide.

5 In some embodiments, the at least one first nucleic acid primer comprises one or more nucleotide analogs. In some embodiments, the at least one first nucleic acid primer comprises one or more synthetic or non-natural nucleotides.

In some embodiments, the window of incidence is three years.

10 In still a further aspect, methods of determining the presence of a biomarker associated with CHD in a patient sample are provided. Such methods typically include (a) isolating nucleic acid sample from the patient sample, (b) performing a genotyping assay on a first portion of the nucleic acid sample to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from Appendix C to obtain genotype data; and/or (c) bisulfite converting the nucleic acid in a second portion of the nucleic acid and
15 performing methylation assessment on the second portion of the nucleic acid sample to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data; and (d) inputting the genotype data from step (b) and/or methylation data from step (c) into an algorithm that accounts for at least one SNP main effect and/or at least one CpG main effect and/or at least
20 one interaction effect, wherein the algorithm is a machine learning algorithm capable of accounting for linear and non-linear effects.

In some embodiments, the at least one interaction effect is selected from the group consisting of a gene-environment interaction (SNPxCpG) effect, a gene-gene interaction (SNPxSNP) effect, and an environment-environment interaction (CpGxCpG) effect. In some
25 embodiments, the at least one interaction effect is a gene-environment interaction effect (SNPxCpG) between a CpG site from Appendix A or a CpG site that is collinear ($R > 0.3$) with a CpG site from Appendix A and a SNP from Appendix C or a SNP within moderate linkage disequilibrium ($R > 0.3$) from a SNP from Appendix C. In some embodiments, the at least one interaction effect is an environment-environment interaction effect (CpGxCpG)
30 between at least two CpG sites from Appendix A.

In some embodiments, one or both of the at least two CpG sites are collinear ($R > 0.3$) with one or both of the at least two CpG sites from Appendix A. In some embodiments, the at least one interaction effect is a gene-gene interaction effect (SNPxSNP) between at least two SNPs from Appendix C. In some embodiments, one or both of the at least two SNPs are collinear ($R > 0.3$) with one or both of the at least two SNPs from Appendix C.

In some embodiments, the biological sample is a saliva sample.

In another aspect, systems for determining methylation status of at least one CpG dinucleotide and a genotype of at least one single-nucleotide polymorphism (SNP) are provided. Such systems typically include: a nucleic acid isolation module configured to isolate a nucleic acid sample from a subject sample; a genotyping assay module configured to perform a genotyping assay on a first portion of the nucleic acid sample to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from Appendix C to obtain genotype data; a methylation assay module configured to bisulfite convert the nucleic acid in a second portion of the nucleic acid and perform a methylation assessment on a second portion of the nucleic acid sample to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data; and an identification system configured to account for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect based on the genotype data from step (b) and/or methylation data from step (c).

In some embodiments, such systems further include an output module configured to provide an output based on an identification by the identification system, wherein the identification accounts for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect based on the genotype data from step (b) and/or methylation data from step (c).

In some embodiments, the algorithm is a machine learning algorithm capable of accounting for linear and non-linear effects.

In yet another aspect, non-transitory computer-readable media storing instructions executable by a processing device to perform operations are provided. Such operations typically include accounting for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect based on genotype data and/or methylation data,

wherein: (i) the genotype data is based on a genotyping assay on a first portion of a nucleic acid sample isolated from a subject sample to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from Appendix C to obtain the genotype data; and (ii) the methylation data is based on a methylation assay on a bisulfite converted nucleic acid in a second portion of the nucleic acid sample to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data.

In some embodiments, the operations further include providing an output based on the accounting. Representative outputs, without limitation, include one or more of storing a report based on the accounting to another non-transitory computer-readable medium, modifying a display based on the accounting, triggering an audible alert based on the accounting, triggering a haptic or vibratory alert based on the accounting, triggering the printing of a report based on the accounting, or triggering the delivery of a therapeutic based on the accounting.

The integrated genetic-epigenetic model described herein provides several advantages and benefits. The first is the overall sensitivity across cohorts. On average, in the Intermountain Healthcare (IM) cohort, the typical risk calculators accurately identify 5 out of 10 individuals at high risk for an incident event compared to the integrated genetic-epigenetic model described herein, which accurately identifies 7 out of 10 individuals. The second is with respect to the performance of standard risk calculators by gender. On average, in the IM cohort, the typical risk calculators accurately identify 5 of 10 men and 4 of 10 women at risk for an incident event. On average, in the IM cohort, the integrated genetic-epigenetic tool described herein accurately identifies 7 of 10 men and 7 of 10 women at risk for an incident event. Thus, the integrated genetic-epigenetic model described herein does not exhibit gender gap in its ability to identify men and women at risk for an incident event.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the methods and compositions of matter belong. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the methods and compositions

of matter, suitable methods and materials are described below. In addition, the materials, methods, and examples are illustrative only and not intended to be limited to predicting incident CHD. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety.

5

DESCRIPTION OF DRAWINGS

FIG. 1 is a graph showing the distribution of the number of incident cases over three years in the Framingham Heart Study Offspring cohort.

FIG. 2 is a graph showing the distribution of the number of incident cases over three years in the Intermountain Healthcare cohort.

10

FIG. 3 is a graph showing the ROC curves of the integrated genetic-epigenetic model for three-year incidence CHD risk assessment in the FHS training, FHS test, IM validation and IM test sets.

15

FIG. 4 is a graph showing the average AUC of the baseline integrated genetic-epigenetic model compared to models with only SNPs, only DNA methylation loci and the addition of conventional risk factors and Polygenic Risk Score.

FIG. 5 shows a Kaplan-Meier survival curve of the high and low risk groups.

FIG. 6 shows a Kaplan-Meier survival curve for high, intermediate and low prognostic scores.

20

FIG. 7 shows the correlation ($r=0.94$) between digital PCR and array DNA methylation values for cg00300879.

FIG. 8 is a block diagram of an example cardiovascular disease classification system.

FIG. 9 is a flow diagram of an example process for cardiovascular disease classification.

25

FIG. 10 is a block diagram of example computing devices.

FIG. 11A-11C are graphs showing the relationship of change in increases in cg05575921 methylation seen in response to smoking cessation to changes in methylation at each of the three loci associated with cardiac risk between study entry and study exit (3 months) in the 20 subjects who had biochemically verified smoking cessation. FIG. 11A is a plot of the change of methylation status at cg14789911 with respect to the change of methylation status at cg05575921. FIG. 11A shows the relationship between increases in

30

methylation at cg05575921 seen in response to smoking cessation and changes in methylation at cg14789911 between study entry to study exit (3 months) in the 20 subjects who had biochemically verified smoking cessation. A negative Δ indicates an increase in methylation at the marker associated with cardiac risk. FIG. 11B is a plot of the change of methylation status at cg09552548 with respect to the change of cg05575921 methylation. FIG. 11B shows the relationship between increases of cg05575921 methylation seen in response to smoking cessation and changes in methylation at cg09552548 between study entry to study exit (3 months) in the 20 subjects who had biochemically verified smoking cessation. A negative Δ indicates an increase in methylation at the marker associated with cardiac risk. FIG. 11C is a plot of the change of cg00300879 with respect to the change of cg05575921 methylation. The change illustrated in FIG. 11C is significant after Bonferroni correction (Adj R² 0.26, $p < 0.04$). FIG. 11C shows the relationship between increases of cg05575921 methylation seen in response to smoking cessation and changes in methylation at cg00300879 between study entry to study exit (3 months) in the 20 subjects who had biochemically verified smoking cessation. A negative Δ indicates an increase in methylation at the marker associated with cardiac risk.

DETAILED DESCRIPTION

Recent risk prediction strategies have taken advantage of the rapid advancements in assessing genome-wide genetic or transcriptional variation. Though each of these approaches have had some success, their clinical impact has been limited. In particular, those relying only on genetic information have a clear ceiling in predictive capacity, are potentially sensitive to ethnic stratification, and, because genotype is static, cannot be used to monitor changes in disease status.

Recent advances in genome-wide epigenetic profiling techniques have raised the possibility that DNA methylation assessments of peripheral blood DNA may serve as a mechanism for more accurate prediction of cardiovascular disease or mortality. Prediction models that only account for epigenetic signatures, however, fail to account for confounding genetic variation, which affects the vast majority of the environmentally responsive methylome. This may result in models that lack robustness with respect to generalizability, especially in different ethnic groups.

As a result, we have developed a highly sensitive, clinically implementable integrated genetic-epigenetic risk assessment tool capable of identifying those at risk of cardiovascular disease (e.g., having a heart attack or sudden cardiac death) within one year, three years or five years. As shown herein, the methylation status of one or more particular CpG dinucleotides in combination with the genotype at one or more particular loci (e.g., CH3xSNP) can be used to predict the incidence (e.g., one-year, three-year, five-year) of cardiovascular disease (CVD) including coronary heart disease (CHD).

As described herein, biomarkers described herein can be used in the diagnosis and prognosis of cardiovascular diseases and events. The terms “marker” and “biomarker” can be used interchangeably. As used herein, a biomarker generally refers to a measurable or detectable biological moiety (e.g., the presence or amount of a protein, a genetic and/or histological component). As described in more detail below, the biomarkers used herein typically are associated with cardiovascular disease.

DNA Methylation

DNA does not exist as naked molecules in the cell. For example, DNA is associated with proteins called histones to form a complex substance known as chromatin. Chemical modifications of the DNA or the histones alter the structure of the chromatin without changing the nucleotide sequence of the DNA. Such modifications are described as “epigenetic” modifications of the DNA. Changes to the structure of the chromatin can have a profound influence on gene expression. If the chromatin is condensed, factors involved in gene expression may not have access to the DNA, and the genes will be switched off. Conversely, if the chromatin is “open,” the genes can be switched on. Some important forms of epigenetic modification are DNA methylation and histone deacetylation.

DNA methylation is a chemical modification of the DNA molecule itself and is carried out by an enzyme called DNA methyltransferase. Methylation can directly switch off gene expression by preventing transcription factors binding to promoters. A more general effect is the attraction of methyl-binding domain (MBD) proteins. These are associated with further enzymes called histone deacetylases (HDACs), which function to chemically modify histones and change chromatin structure. Chromatin-containing acetylated histones are open and accessible to transcription factors, and the genes are potentially active. Histone

deacetylation causes the condensation of chromatin, making it inaccessible to transcription factors and causing the silencing of genes.

CpG islands are short stretches of DNA in which the frequency of the CpG sequence is higher than other regions. The “p” in the term CpG indicates that cysteine (“C”) and guanine (“G”) are connected by a phosphodiester bond. CpG islands are often located around promoters of housekeeping genes and many regulated genes. At these locations, the CG sequence is not methylated. By contrast, the CG sequences in inactive genes are usually methylated to suppress their expression.

As used herein, the term “methylation status” means the determination whether a certain target DNA, such as a CpG dinucleotide, is methylated or is unmethylated. As used herein, the term “CpG dinucleotide repeat motif” means a series of two or more CpG dinucleotides positioned in a DNA sequence.

About 56% of human genes and 47% of mouse genes are associated with CpG islands. Often, CpG islands overlap the promoter and extend about 1000 base pairs downstream into the transcription unit. Identification of potential CpG islands during sequence analysis helps to define the extreme 5’ ends of genes, something that is notoriously difficult with cDNA-based approaches. The methylation of a CpG island can be determined by a skilled artisan using any method suitable to determine such methylation. For example, the skilled artisan can use a bisulfite reaction-based method for determining such methylation.

The present disclosure provides methods to determine the nucleic acid methylation of one or more loci in a subject in order to predict the three-year clinical course and eventual outcome of subjects having CVD.

Genetic screening (also called genotyping or molecular screening) can be broadly defined as testing to determine if a subject has a genetic marker that either causes a disease state or is “linked” to the genetic component causing the disease state. Linkage refers to the phenomenon that DNA sequences which are close together in the genome have a tendency to be inherited together. Two sequences may be linked because of some selective advantage of co-inheritance. More typically, however, two polymorphic sequences are co-inherited because of the relative infrequency with which meiotic recombination events occur within the region between the two polymorphisms. The co-inherited polymorphic alleles are said to

be in “linkage disequilibrium” with one another because, in a given population, they tend to either both occur together or else not occur at all in any particular member of the population. Indeed, where multiple polymorphisms in a given chromosomal region are found to be in linkage disequilibrium with one another, they define a quasi-stable genetic “haplotype.” In contrast, recombination events occurring between two polymorphic loci cause them to become separated onto distinct homologous chromosomes. If meiotic recombination between two physically linked polymorphisms occurs frequently enough, the two polymorphisms will appear to segregate independently and are said to be in linkage equilibrium.

It would be understood that linkage disequilibrium can be quantitated (using, for example, the Pearson correlation (R) or co-inheritance of alleles (D')). For example, a low level of linkage can be reflected in a correlation (e.g., R value) of about 0.1 or less, a moderate level of linkage is reflected in a R value of about 0.3, while a high level of linkage is reflected in a R value of 0.5 or greater. It also would be understood that, when referring to methylation (i.e. CpGs), collinearity (with an R value) is used as a determination of the linear strength of the association between two CpGs (e.g., a low level of collinearity can be reflected by an R value of about 0.1 or less; a moderate level of collinearity can be reflected by an R value of about 0.3; and a high level of collinearity can be reflected by an R value of about 0.5 or greater).

In particular, in certain embodiments of the disclosure, the methods may be practiced as follows. A sample, such as a blood sample, is taken from a subject. In certain embodiments, a single cell type, e.g., lymphocytes, basophils, or monocytes isolated from the blood, may be isolated for further testing. The DNA is harvested from the sample and examined to determine the methylation of one or more loci. For example, the DNA of interest can be treated with bisulfite to deaminate unmethylated cytosine residues to uracil. Since uracil base pairs with adenosine, thymidines are incorporated into subsequent DNA strands in the place of unmethylated cytosine residues during subsequent PCR amplifications. Next, the target sequence is amplified by PCR, and probed with a loci-specific probe. Depending on the particular sequence of the probe used, only the methylated or unmethylated DNA will bind to the probe.

Methods of determining the subject nucleic acid profile are well known to a skilled artisan and include any of the well-known detection methods. Various PCR methods are described, for example, in *PCR Primer: A Laboratory Manual*, Dieffenbach & Dveksler, Eds., Cold Spring Harbor Laboratory Press, 1995. Other methods include, but are not limited to, nucleic acid quantification, restriction enzyme digestion, DNA sequencing, hybridization technologies, such as Southern Blotting, amplification methods such as Ligase Chain Reaction (LCR), Nucleic Acid Sequence Based Amplification (NASBA), Self-sustained Sequence Replication (SSR or 3SR), Strand Displacement Amplification (SDA), and Transcription Mediated Amplification (TMA), Quantitative PCR (qPCR), or other DNA analyses, as well as RT-PCR, in vitro translation, Northern blotting, and other RNA analyses. In another embodiment, hybridization on a microarray is used.

Single Nucleotide Polymorphism (SNP) Genotyping

Traditional methods for the screening of heritable diseases have depended on either the identification of abnormal gene products (e.g., sickle cell anemia) or an abnormal phenotype (e.g., mental retardation). With the development of simple and inexpensive genetic screening methodology, it is now possible to identify polymorphisms that indicate a propensity to develop disease, even when the disease is of polygenic origin.

Single nucleotide polymorphism (SNP) genotyping measures genetic variations of SNPs between members of a species. A SNP is a single base pair change at a specific locus, usually consisting of two alleles (where the rare allele frequency is >1%). SNPs are very common. Because SNPs are conserved during evolution, they have been proposed as markers for use in quantitative trait loci (QTL) analysis and in association studies in place of microsatellites. Many different SNP genotyping methods are known, including hybridization-based methods (such as Dynamic allele-specific hybridization, molecular beacons, and SNP microarrays) enzyme-based methods (including restriction fragment length polymorphism, PCR-based methods, flap endonuclease, primer extension, 5'-nuclease, and oligonucleotide ligation assay), other post-amplification methods based on physical properties of DNA (such as single strand conformation polymorphism, temperature gradient gel electrophoresis, denaturing high performance liquid chromatography, high-resolution melting of the entire amplicon, use of DNA mismatch-binding proteins, SNPlex and surveyor

nuclease assay), and sequencing (such as “next generation” sequencing). See, e.g., US Patent No. 7,972,779.

A plurality of alleles at a locus can arise from one or more polymorphisms in a region of a gene that encodes a polypeptide or in a regulatory control sequence that affects
5 expression of the polypeptide, such as a promoter or polyadenylation sequence.

Alternatively, alleles can arise from one or more polymorphisms at a locus distal to a gene that encodes a polypeptide or in a regulatory control sequence. A polymorphism can affect a polypeptide at a transcriptional or a translational level (e.g., a polypeptide’s transcription rate, translation rate, degradation rate, and/or activity). Allelic differences can be
10 characterized in a sample from a single subject or from a plurality of subjects using methods that are known to a skilled artisan. Such methods can include, but are not limited to, measuring the potential for a polynucleotide sequence to be expressed and/or measuring an amount of an encoded polypeptide. Methods are available that can detect proteins or nucleic acids directly or indirectly, and assay methods are specifically contemplated to include
15 screening for the presence of particular sequences or structures of nucleic acids or polypeptides using, e.g., any of various known microarray technologies.

It will be fully appreciated by the skilled artisan that the allele need not have previously been shown to have had any link or association with the disorder phenotype. Instead, an allele and a pathogenic environmental risk factor can interact to predict a
20 predisposition to a disorder phenotype even when neither the allele nor the risk factor bears any direct relation to the disorder phenotype.

Genetic screening (also called genotyping or molecular screening) can be broadly defined as testing to determine if a subject has mutations (or alleles or polymorphisms) that either cause a disease state or are “linked” to the mutation causing a disease state. Linkage
25 refers to the phenomenon that DNA sequences which are close together in the genome have a tendency to be inherited together. Two sequences may be linked because of some selective advantage of co-inheritance. More typically, however, two polymorphic sequences are co-inherited because of the relative infrequency with which meiotic recombination events occur within the region between the two polymorphisms. The co-inherited polymorphic alleles are
30 said to be in “linkage disequilibrium” with one another because, in a given population, they tend to either both occur together or else not occur at all in any particular member of the

population. Indeed, where multiple polymorphisms in a given chromosomal region are found to be in linkage disequilibrium with one another, they define a quasi-stable genetic “haplotype.” In contrast, recombination events occurring between two polymorphic loci cause them to become separated onto distinct homologous chromosomes. If meiotic recombination between two physically linked polymorphisms occurs frequently enough, the two polymorphisms will appear to segregate independently and are said to be in linkage equilibrium.

It would be understood that linkage disequilibrium can be quantitated (using, for example, the Pearson correlation (R) or co-inheritance of alleles (D')). For example, a low level of linkage can be reflected in a correlation (e.g., R value) of about 0.1 or less, a moderate level of linkage is reflected in a R value of about 0.3, while a high level of linkage is reflected in a R value of 0.5 or greater. It also would be understood that, when referring to methylation (i.e. SNPs), collinearity (with an R value) is used as a determination of the linear strength of the association between two SNPs (e.g., a low level of collinearity can be reflected by an R value of about 0.1 or less; a moderate level of collinearity can be reflected by an R value of about 0.3; and a high level of collinearity can be reflected by an R value of about 0.5 or greater).

While the frequency of meiotic recombination between two markers is generally proportional to the physical distance between them on the chromosome, the occurrence of “hot spots” as well as regions of repressed chromosomal recombination can result in discrepancies between the physical and recombinational distance between two markers. Thus, in certain chromosomal regions, multiple polymorphic loci spanning a broad chromosomal domain may be in linkage disequilibrium with one another, and thereby define a broad-spanning genetic haplotype. Furthermore, where a disease-causing mutation is found within or in linkage with this haplotype, one or more polymorphic alleles of the haplotype can be used as a diagnostic or prognostic indicator of the likelihood of developing the disease. This association between otherwise benign polymorphisms and a disease-causing polymorphism occurs if the disease mutation arose in the recent past, so that sufficient time has not elapsed for equilibrium to be achieved through recombination events. Therefore, identification of a haplotype that spans or is linked to a disease-causing mutational change serves as a predictive measure of an individual's likelihood of having inherited that disease-

causing mutation. Such prognostic or diagnostic procedures can be utilized without necessitating the identification and isolation of the actual disease-causing lesion. This is significant because the precise determination of the molecular defect involved in a disease process can be difficult and laborious, especially in the case of multifactorial diseases.

5 The statistical correlation between a disorder and a polymorphism does not necessarily indicate that the polymorphism directly causes the disorder. Rather the correlated polymorphism may be a benign allelic variant which is linked to (i.e., in linkage disequilibrium with) a disorder-causing mutation that has occurred in the recent evolutionary past, so that sufficient time has not elapsed for equilibrium to be achieved through
10 recombination events in the intervening chromosomal segment. Thus, for the purposes of diagnostic and prognostic assays for a particular disease, detection of a polymorphic allele associated with that disease can be utilized without consideration of whether the polymorphism is directly involved in the etiology of the disease. Furthermore, where a given benign polymorphic locus is in linkage disequilibrium with an apparent disease-causing
15 polymorphic locus, still other polymorphic loci which are in linkage disequilibrium with the benign polymorphic locus are also likely to be in linkage disequilibrium with the disease-causing polymorphic locus. Thus, these other polymorphic loci will also be prognostic or diagnostic of the likelihood of having inherited the disease-causing polymorphic locus. A broad-spanning haplotype (describing the typical pattern of co-inheritance of alleles of a set
20 of linked polymorphic markers) can be targeted for diagnostic purposes once an association has been drawn between a particular disease or condition and a corresponding haplotype. Thus, the determination of an individual's likelihood for developing a particular disease or condition can be made by characterizing one or more disease-associated polymorphic alleles (or even one or more disease-associated haplotypes) without necessarily determining or
25 characterizing the causative genetic variation.

 Many methods are available for detecting specific alleles at polymorphic loci. Certain methods for detecting a specific polymorphic allele will depend, in part, upon the molecular nature of the polymorphism. For example, the various allelic forms of the polymorphic locus may differ by a single base-pair of the DNA. Such single nucleotide
30 polymorphisms (or SNPs) are major contributors to genetic variation, comprising some 80% of all known polymorphisms, and their density in the genome is estimated to be on average 1

per 1,000 base pairs. SNPs are most frequently bi-allelic, or occurring in only two different forms (although up to four different forms of an SNP, corresponding to the four different nucleotide bases occurring in DNA, are theoretically possible). Nevertheless, SNPs are mutationally more stable than other polymorphisms, making them suitable for association studies in which linkage disequilibrium between markers and an unknown variant is used to map disease-causing mutations. In addition, because SNPs typically have only two alleles, they can be genotyped by a simple plus / minus assay rather than a length measurement, making them more amenable to automation.

In one embodiment, allelic profiling can be accomplished using a nucleic acid microarray. The genetic testing field is rapidly evolving and, as such, the skilled artisan will appreciate that a wide range of profiling tests exist, and will be developed, to determine the allelic profile of individuals in accord with the disclosure.

Nucleic Acids and Polypeptides

The term “nucleic acid” refers to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form, made of monomers (nucleotides) containing a sugar, phosphate and a base that is either a purine or pyrimidine. Unless specifically limited, the term encompasses nucleic acids containing known analogs of natural nucleotides that have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues. The terms “nucleic acid,” “nucleic acid molecule,” or “polynucleotide” are used interchangeably and may also be used interchangeably with gene, cDNA, DNA and/or RNA encoded by a gene.

The term “nucleotide sequence” refers to a polymer of DNA or RNA which can be single-stranded or double-stranded, optionally containing synthetic, non-natural or altered nucleotide bases capable of incorporation into DNA or RNA polymers. A DNA molecule or

polynucleotide is a polymer of deoxyribonucleotides (A, G, C, and T), and an RNA molecule or polynucleotide is a polymer of ribonucleotides (A, G, C and U).

5 A “gene,” for the purposes of the present disclosure, includes a DNA region encoding a gene product, as well as all DNA regions, which regulate the production of the gene product, whether or not such regulatory sequences are adjacent to coding and/or transcribed sequences. The term “gene” is used broadly to refer to any segment of nucleic acid associated with a biological function. Genes include coding sequences and/or the regulatory sequences required for their expression. Accordingly, a gene includes, but is not necessarily limited to, promoter sequences, terminators, translational regulatory sequences such as
10 ribosome binding sites and internal ribosome entry sites, enhancers, silencers, insulators, boundary elements, replication origins, matrix attachment sites and locus control regions. For example, “gene” refers to a nucleic acid fragment that expresses mRNA, functional RNA, or specific protein, including regulatory sequences. “Functional RNA” refers to sense RNA, antisense RNA, ribozyme RNA, siRNA, or other RNA that may not be translated but
15 yet has an effect on at least one cellular process. “Genes” also include non-expressed DNA segments that, for example, form recognition sequences for other proteins. “Genes” can be obtained from a variety of sources, including cloning from a source of interest or synthesizing from known or predicted sequence information, and may include sequences designed to have desired parameters.

20 “Gene expression” refers to the conversion of the information, contained in a gene, into a gene product. It refers to the transcription and/or translation of an endogenous gene, heterologous gene or nucleic acid segment, or a transgene in cells. In addition, expression refers to the transcription and stable accumulation of sense (mRNA) or functional RNA. Expression may also refer to the production of protein. The term “altered level of
25 expression” refers to the level of expression in transgenic cells or organisms that differs from that of normal or untransformed cells or organisms.

A gene product can be the transcriptional product of a gene (*e.g.*, mRNA, tRNA, rRNA, antisense RNA, ribozyme, structural RNA or any other type of RNA) or a protein produced by translation of an mRNA. Gene products also include RNAs that are modified,
30 by processes such as capping, polyadenylation, methylation, and editing, and proteins modified by, for example, methylation, acetylation, phosphorylation, ubiquitination, ADP-

ribosylation, myristilation, and glycosylation. The term “RNA transcript” refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a complementary copy of the DNA sequence, it is referred to as the primary transcript; a RNA sequence derived from post-transcriptional processing of the primary transcript is referred to as the mature RNA. “Messenger RNA” (mRNA) refers to the RNA that lacks introns and that can be translated into protein by the cell. “cDNA” refers to a single- or a double-stranded DNA that is complementary to and derived from mRNA. “Functional RNA” refers to sense RNA, antisense RNA, ribozyme RNA, siRNA, or other RNA that may not be translated but yet has an effect on at least one cellular process.

A “coding sequence” or a sequence that “encodes” a polypeptide is a nucleic acid molecule that is transcribed (in the case of DNA) and/or translated (in the case of mRNA) into a polypeptide *in vivo* when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5’ (amino) terminus and a translation stop codon at the 3’ (carboxy) terminus. A coding sequence can include, but is not limited to, cDNA from viral, prokaryotic or eukaryotic mRNA, genomic DNA sequences from viral (e.g., DNA viruses and retroviruses) or prokaryotic DNA, and synthetic DNA sequences. A transcription termination sequence can be located 3’ to the coding sequence.

“Regulatory sequences” and “suitable regulatory sequences” each refer to nucleotide sequences located upstream (5’ non-coding sequences), within, or downstream (3’ non-coding sequences) of a coding sequence, and which influence the transcription, RNA processing or stability, or translation of the associated coding sequence. Regulatory sequences include enhancers, promoters, translation leader sequences, introns, and polyadenylation signal sequences. They include natural and synthetic sequences as well as sequences that may be a combination of synthetic and natural sequences.

Certain embodiments of the disclosure encompass isolated or substantially purified nucleic acid compositions. In the context of the present disclosure, an “isolated” or “purified” DNA molecule or RNA molecule is a DNA molecule or RNA molecule that exists apart from its native environment and is, therefore, not a product of nature. An isolated DNA molecule or RNA molecule may exist in a purified form or may exist in a non-native environment such as, for example, a transgenic host cell. For example, an “isolated” or

“purified” nucleic acid molecule is substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized. In one embodiment, an “isolated” nucleic acid is free of sequences that naturally flank the nucleic acid (i.e., sequences located at the 5' and 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived.

By “fragment” is intended a polypeptide consisting of only a part of the intact full-length polypeptide sequence and structure. The fragment can include a C-terminal deletion, an N-terminal deletion, and/or an internal deletion of the native polypeptide. A fragment of a protein will generally include at least about 5-100 contiguous amino acid residues of the full-length molecule (e.g., at least about 15-25 contiguous amino acid residues of the full-length molecule, at least about 20-50 or more contiguous amino acid residues of the full-length molecule, or any integer between 5 amino acids and the full-length sequence).

“Naturally occurring” is used to describe a composition that can be found in nature as distinct from being artificially produced. For example, a nucleotide sequence present in an organism, which can be isolated from a source in nature and which has not been intentionally modified by a person in the laboratory, is naturally occurring.

A “5' non-coding sequence” refers to a nucleotide sequence located 5' (upstream) to the coding sequence. 5' non-coding sequences are present in the fully processed mRNA upstream of the initiation codon and may affect processing of the primary transcript to mRNA, mRNA stability or translation efficiency. A “3' non-coding sequence” refers to nucleotide sequences located 3' (downstream) to a coding sequence and may include polyadenylation signal sequences and other sequences encoding regulatory signals capable of affecting mRNA processing or gene expression.

A “promoter” refers to a nucleotide sequence, usually upstream (5') to its coding sequence, which directs and/or controls the expression of the coding sequence by providing the recognition for RNA polymerase and other factors required for proper transcription. “Promoter” can include a minimal promoter that is a short DNA sequence comprised of a TATA-box and other sequences that serve to specify the site of transcription initiation, to which regulatory elements are added for control of expression. “Promoter” also can refer to a nucleotide sequence that includes a minimal promoter plus one or more regulatory elements

(e.g., enhancers) that are capable of controlling the expression of a coding sequence or functional RNA. Promoters may be derived in their entirety from a native sequence, or be composed of different elements derived from different promoters found in nature, or even be comprised of synthetic DNA sequences. A promoter may also contain DNA sequences that are involved in the binding of protein factors that control the effectiveness of transcription initiation in response to physiological or developmental conditions. “Constitutive expression” refers to expression using a constitutive promoter. “Conditional” and “regulated expression” refer to expression controlled by a regulated promoter.

An “enhancer” is a DNA sequence that can stimulate promoter activity. An enhancer may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue specificity of a promoter. Enhancers often are capable of operating in both orientations, and are capable of functioning even when moved either upstream or downstream from the promoter. Both enhancers and other regulatory elements within a promoter bind sequence-specific DNA-binding proteins that mediate their effects.

“Operably-linked” refers to the association of nucleic acid sequences on a single nucleic acid fragment so that the function of one of the sequences is affected by another. For example, a regulatory DNA sequence is said to be “operably linked to” or “associated with” a DNA sequence that codes for an RNA or a polypeptide if the two sequences are situated such that the regulatory DNA sequence affects expression of the coding DNA sequence (i.e., that the coding sequence or functional RNA is under the transcriptional control of the promoter). Coding sequences can be operably-linked to regulatory sequences in sense or antisense orientation.

“Expression” refers to the transcription and/or translation of an endogenous gene, heterologous gene or nucleic acid segment, or a transgene in cells. In addition, expression refers to the transcription and stable accumulation of sense (mRNA) or functional RNA. Expression may also refer to the production of protein. The term “altered level of expression” refers to a level of expression in cells or organisms that differs from that of normal cells or organisms.

For sequence comparison, typically one sequence acts as a reference sequence to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are input into a computer, and sequence algorithm program parameters

are designated. The sequence comparison algorithm then calculates the percent sequence identity for the test sequence(s) relative to the reference sequence, based on the designated algorithm parameters.

5 The following terms are used to describe the sequence relationships between two or more nucleic acids or polynucleotides: (a) "reference sequence," (b) "comparison window," (c) "sequence identity," (d) "percentage of sequence identity," and (e) "as is for sequence comparison. A reference sequence may be a subset or the substantial identity." As used herein, "reference sequence" is a defined sequence used as a entirety of a specified sequence; for example, as a segment of a full-length cDNA or gene sequence, or the
10 complete cDNA or gene sequence. As used herein, "comparison window" makes reference to a contiguous and specified segment of a polynucleotide sequence, wherein the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Generally, the comparison window is
15 at least 20 contiguous nucleotides in length, and optionally can be 30, 40, 50, 100, or longer. Those of skill in the art understand that, to avoid a high similarity to a reference sequence due to inclusion of gaps in the polynucleotide sequence, a gap penalty is typically introduced and is subtracted from the number of matches.

Methods of alignment of sequences for comparison are well-known in the art. Thus,
20 the determination of percent identity between any two sequences can be accomplished using a mathematical algorithm. Non-limiting examples of such mathematical algorithms are the algorithm of Myers and Miller (Myers and Miller, CABIOS, 4, 11 (1988)); the local homology algorithm of Smith et al. (Smith et al., Adv. Appl. Math., 2, 482 (1981)); the homology alignment algorithm of Needleman and Wunsch (Needleman and Wunsch, JMB,
25 48, 443 (1970)); the search-for-similarity-method of Pearson and Lipman (Pearson and Lipman, Proc. Natl. Acad. Sci. USA, 85, 2444 (1988)); the algorithm of Karlin and Altschul (Karlin and Altschul, Proc. Natl. Acad. Sci. USA, 87, 2264 (1990)), modified as in Karlin and Altschul (Karlin and Altschul, Proc. Natl. Acad. Sci. USA 90, 5873 (1993)).

Computer implementations of these mathematical algorithms can be utilized for
30 comparison of sequences to determine sequence identity. Such implementations include, but are not limited to: CLUSTAL in the PC/Gene program (available from Intelligenetics,

Mountain View, Calif.); the ALIGN program (Version 2.0) and GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Version 8 (available from Genetics Computer Group (GCG), 575 Science Drive, Madison, Wis., USA).

Alignments using these programs can be performed using the default parameters. The CLUSTAL program is well described by Higgins et al. (Higgins et al., CABIOS, 5, 151 (1989)); Corpet et al. (Corpet et al., Nucl. Acids Res., 16, 10881 (1988)); Huang et al. (Huang et al., CABIOS, 8, 155 (1992)); and Pearson et al. (Pearson et al., Meth. Mol. Biol., 24, 307 (1994)). The ALIGN program is based on the algorithm of Myers and Miller, *supra*. The BLAST programs of Altschul et al. (Altschul et al., J. Mol. Biol., 215, 403 (1990)) are based on the algorithm of Karlin and Altschul, *supra*.

Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length “W” in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. “T” is referred to as the neighborhood word score threshold. These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters “M” (reward score for a pair of matching residues; always >0) and “N” (penalty score for mismatching residues; always <0), and for amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when the cumulative alignment score falls off by the quantity “X” from its maximum achieved value, the cumulative score goes to zero or below due to the accumulation of one or more negative-scoring residue alignments, or the end of either sequence is reached.

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences. One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a test nucleic acid sequence is considered similar to a reference sequence if the smallest sum probability in a comparison of

the test nucleic acid sequence to the reference nucleic acid sequence is less than about 0.1, less than about 0.01, or even less than about 0.001.

To obtain gapped alignments for comparison purposes, Gapped BLAST (in BLAST 2.0) can be utilized. Alternatively, PSI-BLAST (in BLAST 2.0) can be used to perform an iterated search that detects distant relationships between molecules. When utilizing BLAST, Gapped BLAST, PSI-BLAST, the default parameters of the respective programs (e.g., BLASTN for nucleotide sequences, BLASTX for proteins) can be used. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix. Alignment may also be performed manually by inspection.

For purposes of the present disclosure, comparison of nucleotide sequences for determination of percent sequence identity to the promoter sequences disclosed herein may be made using the BlastN program (version 1.4.7 or later) with its default parameters or any equivalent program. By “equivalent program” is intended any sequence comparison program that, for any two sequences in question, generates an alignment having identical nucleotide or amino acid residue matches and an identical percent sequence identity when compared to the corresponding alignment generated by the program.

As used herein, “sequence identity” or “identity” in the context of two nucleic acid or polypeptide sequences makes reference to a specified percentage of residues in the two sequences that are the same when aligned for maximum correspondence over a specified comparison window, as measured by sequence comparison algorithms or by visual inspection. When percentage of sequence identity is used in reference to proteins it is recognized that residue positions which are not identical often differ by conservative amino acid substitutions, where amino acid residues are substituted for other amino acid residues with similar chemical properties (e.g., charge or hydrophobicity) and, therefore, do not change the functional properties of the molecule. When sequences differ in conservative substitutions, the percent sequence identity may be adjusted upwards to correct for the conservative nature of the substitution. Sequences that differ by such conservative substitutions are said to have “sequence similarity” or “similarity.” Means for making this

adjustment are well known to those of skill in the art. Typically, this involves scoring a conservative substitution as a partial rather than a full mismatch, thereby increasing the percentage sequence identity. Thus, for example, where an identical amino acid is given a score of 1 and a non-conservative substitution is given a score of zero, a conservative substitution is given a score between zero and 1. The scoring of conservative substitutions is calculated, e.g., as implemented in the program PC/GENE (Intelligenetics, Mountain View, Calif.).

As used herein, “percent sequence identity” means the value determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison, and multiplying the result by 100 to yield the percentage of sequence identity.

The term “substantial identity” of polynucleotide sequences means that a polynucleotide comprises a sequence that has at least 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, or 94%, or even at least 95%, 96%, 97%, 98%, 99% or 100% sequence identity, compared to a reference sequence using one of the alignment programs described herein using standard parameters. One of skill in the art will recognize that these values can be appropriately adjusted to determine corresponding identity of proteins encoded by two nucleotide sequences by taking into account codon degeneracy, amino acid similarity, reading frame positioning, and the like. Substantial identity of amino acid sequences for these purposes normally means sequence identity of at least 70% (e.g., 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%), at least 80% (e.g., 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%), at least 90% (e.g., 91%, 92%, 93%, or 94%), or even at least 95% (e.g., 96%, 97%, 98%, 99%, or 100%).

The term “substantial identity” in the context of a peptide indicates that a peptide comprises a sequence with at least 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%,

80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, or 94%, or even 95%, 96%, 97%, 98% or 99%, sequence identity to the reference sequence over a specified comparison window. In certain embodiments, optimal alignment is conducted using the homology alignment algorithm of Needleman and Wunsch (Needleman and Wunsch, J. Mol. Biol., 48, 443 (1970)). An indication that two peptide sequences are substantially identical is that one peptide is immunologically reactive with antibodies raised against the second peptide. Thus, a peptide is substantially identical to a second peptide, for example, where the two peptides differ only by a conservative substitution. Thus, the disclosure also provides nucleic acid molecules and peptides that are substantially identical to the nucleic acid molecules and peptides presented herein.

Another indication that nucleotide sequences are substantially identical is if two molecules hybridize to each other under stringent conditions. Hybridization of nucleic acids is discussed in more detail below.

Oligonucleotide Primers and Probes

As used herein, “primer,” “probe,” and “oligonucleotide” are used interchangeably. The term “nucleic acid probe” or a “probe specific for” a nucleic acid refers to a nucleic acid sequence that has at least about 80%, e.g., at least about 90%, e.g., at least about 95% contiguous sequence identity or homology to the nucleic acid sequence encoding the targeted sequence of interest. A probe (or oligonucleotide or primer) of the disclosure is at least about 8 nucleotides in length (e.g., at least about 8-50 nucleotides in length, e.g., at least about 10-40, e.g., at least about 15-35 nucleotides in length). The oligonucleotide probes or primers of the disclosure may comprise at least about eight nucleotides at the 3' of the oligonucleotide that have at least about 80%, e.g., at least about 85%, e.g., at least about 90%, e.g., at least about 95% contiguous identity to the targeted sequence of interest.

Primer pairs are useful for determination of the nucleotide sequence of a particular SNP using PCR. The pairs of single-stranded DNA primers can be annealed to sequences within or surrounding the SNP in order to prime amplifying DNA synthesis of the SNP itself. The first step of the process involves contacting a biological sample obtained from a subject, which sample contains nucleic acid, with at least one primer to form a hybridized DNA. The oligonucleotide primers that are useful in the methods of the present disclosure can be any

primer comprised of about 8 bases up to about 80 or 100 bases or more. In one embodiment of the present disclosure, the primers are between about 10 and about 20 bases.

The primers themselves can be synthesized using techniques that are well known in the art. Generally, the primers can be made using oligonucleotide synthesizing machines that are commercially available.

The primers or probes of the present disclosure can be labeled using techniques known to those of skill in the art. For example, the labels used in the assays of disclosure can be primary labels (where the label comprises an element that is detected directly) or secondary labels (where the detected label binds to a primary label, e.g., as is common in immunological labeling). An introduction to labels (also called “tags”), tagging or labeling procedures, and detection of labels is found in Polak and Van Noorden (1997) Introduction to Immunocytochemistry, second edition, Springer Verlag, N.Y. and in Haugland (1996) Handbook of Fluorescent Probes and Research Chemicals, a combined handbook and catalogue Published by Molecular Probes, Inc., Eugene, Oreg. Primary and secondary labels can include undetected elements as well as detected elements. Useful primary and secondary labels in the present disclosure can include spectral labels such as fluorescent dyes (e.g., fluorescein and derivatives such as fluorescein isothiocyanate (FITC) and Oregon Green™, rhodamine and derivatives (e.g., Texas red, tetramethylrhodamine isothiocyanate (TRITC), etc.), digoxigenin, biotin, phycoerythrin, AMCA, CyDyes™, and the like), radiolabels (e.g., ^3H , ^{125}I , ^{35}S , ^{14}C , ^{32}P , ^{33}P), enzymes (e.g., horse-radish peroxidase, alkaline phosphatase) spectral colorimetric labels such as colloidal gold or colored glass or plastic (e.g., polystyrene, polypropylene, latex) beads. The label may be coupled directly or indirectly to a component of the detection assay (e.g., the labeled nucleic acid) according to methods well known in the art. As indicated above, a wide variety of labels may be used, with the choice of label depending on sensitivity required, ease of conjugation with the compound, stability requirements, available instrumentation, and disposal provisions.

In general, a detector that monitors a probe-substrate nucleic acid hybridization is adapted to the particular label that is used. Typical detectors include spectrophotometers, phototubes and photodiodes, microscopes, scintillation counters, cameras, film and the like, as well as combinations thereof. Examples of suitable detectors are widely available from a variety of commercial sources known to persons of skill. Commonly, an optical image of a

substrate comprising bound labeled nucleic acids is digitized for subsequent computer analysis.

Labels include those that use (1) chemiluminescence (using Horseradish Peroxidase and/or Alkaline Phosphatase with substrates that produce photons as breakdown products) with kits being available, *e.g.*, from Molecular Probes, Amersham, Boehringer-Mannheim, and Life Technologies/Gibco BRL; (2) color production (using both Horseradish Peroxidase and/or Alkaline Phosphatase with substrates that produce a colored precipitate) (kits available from Life Technologies/Gibco BRL, and Boehringer-Mannheim); (3) hemifluorescence using, *e.g.*, Alkaline Phosphatase and the substrate AttoPhos (Amersham) or other substrates that produce fluorescent products, (4) fluorescence (*e.g.*, using Cy-5 (Amersham), fluorescein, and other fluorescent labels); (5) radioactivity using kinase enzymes or other end-labeling approaches, nick translation, random priming, or PCR to incorporate radioactive molecules into the labeled nucleic acid. Other methods for labeling and detection will be readily apparent to one skilled in the art.

Fluorescent labels can be used and have the advantage of requiring fewer precautions in handling, and being amenable to high-throughput visualization techniques (optical analysis including digitization of the image for analysis in an integrated system comprising a computer). Preferred labels are typically characterized by one or more of the following: high sensitivity, high stability, low background, low environmental sensitivity and high specificity in labeling. Fluorescent moieties, which can be incorporated into a label, generally are known including Texas red, digoxigenin, biotin, 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilenin, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidazolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, calicylate, strophanthidin, porphyrins, triarylmethanes, flavin and many others. Many fluorescent labels are commercially available from the SIGMA Chemical Company (Saint Louis, MO), Molecular Probes, R&D systems (Minneapolis, MN), Pharmacia LKB Biotechnology (Piscataway, NJ), CLONTECH Laboratories, Inc. (Palo Alto, CA), Chem Genes Corp., Aldrich Chemical Company (Milwaukee, WI), Glen Research, Inc., GIBCO

BRL Life Technologies, Inc. (Gaithersburg, MD), Fluka ChemicaBiochemika Analytika (Fluka Chemie AG, Buchs, Switzerland), and Applied Biosystems™ (Foster City, CA), as well as many other commercial sources known to one of skill.

Means of detecting and quantifying labels are well known to those of skill in the art.

5 Thus, for example, when the label is a radioactive label, means for detection include a scintillation counter or photographic film as in autoradiography; and when the label is optically detectable, typical detectors include microscopes, cameras, phototubes, photodiodes and many other detection systems that are widely available.

Oligonucleotide primers or probes may be prepared having any of a wide variety of
 10 base sequences according to techniques that are well known in the art. Suitable bases for preparing an oligonucleotide primer or probe may be selected from naturally occurring nucleotide bases such as adenine, cytosine, guanine, uracil, and thymine; and non-naturally occurring or “synthetic” nucleotide bases such as 7-deaza-guanine 8-oxo-guanine, 6-mercaptoguanine, 4-acetylcytidine, 5-(carboxyhydroxyethyl)uridine, 2'-O-methylcytidine, 5-
 15 carboxymethylamino-methyl-2-thioridine, 5-carboxymethylaminomethyluridine, dihydrouridine, 2'-O-methylpseudouridine, β ,D-galactosylqueosine, 2'-O-methylguanosine, inosine, N6-isopentenyladenosine, 1-methyladenosine, 1-methylpseudouridine, 1-methylguanosine, 1-methylinosine, 2,2-dimethylguanosine, 2-methyladenosine, 2-methylguanosine, 3-methylcytidine, 5-methylcytidine, N6-methyladenosine, 7-
 20 methylguanosine, 5-methylaminomethyluridine, 5-methoxyaminomethyl-2-thiouridine, β ,D-mannosylqueosine, 5-methoxycarbonylmethyluridine, 5-methoxyuridine, 2-methylthio-N6-isopentenyladenosine, N-((9- β -D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine, N-((9- β -D-ribofuranosylpurine-6-yl)N-methyl-carbamoyl)threonine, uridine-5-oxyacetic acid methylester, uridine-5-oxyacetic acid, wybutosine, pseudouridine,
 25 queosine, 2-thiocytidine, 5-methyl-2-thiouridine, 2-thiouridine, 2-thiouridine, 5-Methyluridine, N-((9-beta-D-ribofuranosylpurine-6-yl)carbamoyl)threonine, 2'-O-methyl-5-methyluridine, 2'-O-methyluridine, wybutosine, and 3-(3-amino-3-carboxypropyl)uridine. Any oligonucleotide backbone may be employed, including DNA, RNA (although RNA is less preferred than DNA), modified sugars such as carbocycles, and sugars containing 2'
 30 substitutions such as fluoro and methoxy. The oligonucleotides may be oligonucleotides wherein at least one, or all, of the internucleotide bridging phosphate residues are modified

phosphates, such as methyl phosphonates, methyl phosphonotioates, phosphoroinorpholidates, phosphoropiperazidates and phosphoramidates (for example, every other one of the internucleotide bridging phosphate residues may be modified as described). The oligonucleotide may be a “peptide nucleic acid” such as described in Nielsen *et al.*,
5 Science, 254:1497-1500 (1991).

As used herein, a “single base pair extension probe” is a nucleic acid that selectively recognizes a single nucleotide polymorphism (i.e., either the A or the G of an A/G polymorphism). Generally, these probes take the form of a DNA primer (e.g., as in PCR primers) that are modified so that incorporation of the primer releases a fluorophore. One
10 example of this is a Taqman[®] probe that uses the 5' exonuclease activity of the enzyme Taq Polymerase for measuring the amount of target sequences in the samples. TaqMan[®] probes consist of a 18-22 bp oligonucleotide probe, which is labeled with a reporter fluorophore at the 5' end, and a quencher fluorophore at the 3' end. Incorporation of the probe molecule into a PCR chain (which occurs because the probe set is contained in a mixture of PCR primers)
15 liberates the reporter fluorophore from the effects of the quencher. The primer must be able to recognize the target binding site. Some primer extension probes can be “activated” directly by DNA polymerase without a full PCR extension cycle.

The only requirement is that the oligonucleotide probe should possess a sequence at least a portion of which is capable of binding to a known portion of the sequence of the DNA
20 sample. The nucleic acid probes provided by the present disclosure are useful for a number of purposes.

Methods of Detecting Nucleic Acids

A. Amplification

25 According to the methods of the present disclosure, the amplification of DNA present in a biological sample may be carried out by any means known to the art. Examples of suitable amplification techniques include, but are not limited to, polymerase chain reaction (including, for RNA amplification, reverse-transcriptase polymerase chain reaction), ligase chain reaction, strand displacement amplification, transcription-based amplification, self-
30 sustained sequence replication (or “3SR”), the Qbeta replicase system, nucleic acid

sequence-based amplification (or “NASBA”), the repair chain reaction (or “RCR”), and boomerang DNA amplification (or “BDA”).

The bases incorporated into the amplification product can be natural or modified bases (modified before or after amplification), and the bases can be selected to optimize subsequent electrochemical detection steps.

Polymerase chain reaction (PCR) can be carried out in accordance with known techniques. See, e.g., U.S. Patent Numbers 4,683,195; 4,683,202; 4,800,159; and 4,965,188. In general, PCR involves, first, treating a nucleic acid sample (e.g., in the presence of a heat stable DNA polymerase) with one oligonucleotide primer for each strand of the specific sequence to be detected under hybridizing conditions so that an extension product of each primer is synthesized that is complementary to each nucleic acid strand, with the primers sufficiently complementary to each strand of the specific sequence to hybridize therewith so that the extension product synthesized from each primer, when it is separated from its complement, can serve as a template for synthesis of the extension product of the other primer, and then treating the sample under denaturing conditions to separate the primer extension products from their templates if the sequence or sequences to be detected are present. These steps are cyclically repeated until the desired degree of amplification is obtained. Detection of the amplified sequence may be carried out by adding, to the reaction product, an oligonucleotide probe capable of hybridizing to the reaction product (e.g., an oligonucleotide primer or probe of the present disclosure), the probe carrying a detectable label, and then detecting the label in accordance with known techniques. Various labels that can be incorporated into or operably linked to nucleic acids are well known in the art, such as radioactive, enzymatic, and florescent labels. Where the nucleic acid to be amplified is RNA, amplification may be carried out by initial conversion to DNA by reverse transcriptase in accordance with known techniques.

Strand displacement amplification (SDA) can be carried out in accordance with known techniques. For example, SDA can be carried out with a single amplification primer or a pair of amplification primers, with exponential amplification being achieved with the latter. In general, SDA amplification primers comprise, in the 5' to 3' direction, a flanking sequence (the DNA sequence of which is noncritical), a restriction site for the restriction enzyme employed in the reaction, and an oligonucleotide sequence (e.g., an oligonucleotide

primer or probe as described herein) that hybridizes to the target sequence to be amplified and/or detected. The flanking sequence, which serves to facilitate binding of the restriction enzyme to the recognition site and provides a DNA polymerase priming site after the restriction site has been nicked, can be about 15 to 20 nucleotides in length. The restriction site is functional in the SDA reaction. For example, the oligonucleotide primer or probe portion can be about 13 to 15 nucleotides in length.

Ligase chain reaction (LCR) also can be carried out in accordance with known techniques. In general, the reaction is carried out with two pairs of oligonucleotide probes: one pair binds to one strand of the sequence to be detected; the other pair binds to the other strand of the sequence to be detected. Each pair together completely overlaps the strand to which it corresponds. The reaction is carried out by, first, denaturing (e.g., separating) the strands of the sequence to be detected, then reacting the strands with the two pairs of oligonucleotide probes in the presence of a heat stable ligase so that each pair of oligonucleotide probes is ligated together, then separating the reaction product, and then cyclically repeating the process until the sequence has been amplified to the desired degree. Detection then can be carried out in like manner as described above with respect to PCR.

According to the methods described herein, a particular SNP at a particular locus can be detected. Techniques that are useful in the methods described herein include, but are not limited to, direct DNA sequencing, PFGE analysis, allele-specific oligonucleotide (ASO), dot blot analysis and denaturing gradient gel electrophoresis, and are well known to a skilled artisan.

There are several methods that can be used to detect DNA sequence variation. Direct DNA sequencing, either manual sequencing or automated fluorescent sequencing can detect sequence variation. Another approach is the single-stranded conformation polymorphism assay (SSCA). This method does not detect all sequence changes, especially if the DNA fragment size is greater than 200 bp, but can be optimized to detect most DNA sequence variation. The reduced detection sensitivity is a disadvantage, but the increased throughput possible with SSCA makes it an attractive, viable alternative to direct sequencing for mutation detection on a research basis. The fragments that have shifted mobility on SSCA gels then can be sequenced to determine the exact nature of the DNA sequence variation. Other approaches based on the detection of mismatches between the two complementary

DNA strands include clamped denaturing gel electrophoresis (CDGE), heteroduplex analysis (HA) and chemical mismatch cleavage (CMC). Once a sequence change has been identified, an allele specific detection approach such as allele specific oligonucleotide (ASO) hybridization can be utilized to rapidly screen large numbers of other samples for that same sequence change (e.g., mutation, polymorphism). Such a technique can utilize probes that are labeled with gold nanoparticles to yield a visual color result.

Detection of SNPs can be accomplished by sequencing the desired target region using techniques well known in the art. Alternatively, sequences can be amplified directly from a genomic DNA preparation from subject tissue using known techniques. The DNA sequence of the amplified sequences then can be determined.

There are several well known methods for a more complete, yet still indirect, test for confirming the presence of a mutant allele: 1) single stranded conformation analysis (SSCA); 2) denaturing gradient gel electrophoresis (DGGE); 3) RNase protection assays; 4) allele-specific oligonucleotides (ASOs); 5) the use of proteins which recognize nucleotide mismatches, such as the *E. coli* mutS protein; and/or 6) allele-specific PCR. For allele-specific PCR, primers are used that hybridize at their 3' ends to a particular allele. If the particular mutation is not present, an amplification product is not observed. Amplification Refractory Mutation System (ARMS) can also be used. Insertions and deletions of genes can also be detected by cloning, sequencing and amplification. In addition, restriction fragment length polymorphism (RFLP) probes for the gene or surrounding marker genes can be used to score alteration of an allele or an insertion in a polymorphic fragment. Other techniques for detecting insertions and deletions as known in the art can be used.

In the first three methods (SSCA, DGGE and RNase protection assay), a new electrophoretic band appears. SSCA detects a band that migrates differentially because the sequence change causes a difference in single-strand, intramolecular base pairing. RNase protection involves cleavage of the mutant polynucleotide into two or more smaller fragments. DGGE detects differences in migration rates of mutant sequences compared to wild-type sequences, using a denaturing gradient gel. In an allele-specific oligonucleotide assay, an oligonucleotide is designed which detects a specific sequence, and the assay is performed by detecting the presence or absence of a hybridization signal. In the mutS assay,

the protein binds only to sequences that contain a nucleotide mismatch in a heteroduplex between mutant and wild-type sequences.

Mismatches, according to the present disclosure, are hybridized nucleic acid duplexes in which the two strands are not 100% complementary. Lack of total homology may be due to deletions, insertions, inversions or substitutions. Mismatch detection can be used to detect point mutations in the gene or in its mRNA product. While these techniques are less sensitive than sequencing, they are simpler to perform on a large number of samples. An example of a mismatch cleavage technique is the RNase protection method. The riboprobe and either mRNA or DNA isolated from the tumor tissue are annealed (hybridized) together and subsequently digested with the enzyme RNase A that is able to detect some mismatches in a duplex RNA structure. If a mismatch is detected by RNase A, it cleaves at the site of the mismatch. Thus, when the annealed RNA preparation is separated on an electrophoretic gel matrix, if a mismatch has been detected and cleaved by RNase A, an RNA product will be seen which is smaller than the full length duplex RNA for the riboprobe and the mRNA or DNA. The riboprobe need not be the full length of the mRNA or gene but can be a segment of either. If the riboprobe includes only a segment of the mRNA or gene, it will be desirable to use a number of these probes to screen the whole mRNA sequence for mismatches.

In similar fashion, DNA probes can be used to detect mismatches, through enzymatic or chemical cleavage. Alternatively, mismatches can be detected by shifts in the electrophoretic mobility of mismatched duplexes relative to matched duplexes. With either riboprobes or DNA probes, the cellular mRNA or DNA that might contain a mutation can be amplified using PCR before hybridization.

B. Hybridization

The phrase “hybridizing specifically to” refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (e.g., total cellular) DNA or RNA. “Bind(s) substantially” refers to complementary hybridization between a primer or probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target nucleic acid sequence.

Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. However, stringent conditions encompass temperatures in the range of about 1°C to about 20°C, depending upon the desired degree of stringency as otherwise qualified herein. Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the polypeptides they encode are substantially identical. This may occur, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code. One indication that two nucleic acid sequences are substantially identical is when the polypeptide encoded by the first nucleic acid is immunologically cross reactive with the polypeptide encoded by the second nucleic acid.

“Stringent conditions” are those that (1) employ low ionic strength and high temperature for washing, for example, 0.015 M NaCl / 0.0015 M sodium citrate (SSC); 0.1% sodium lauryl sulfate (SDS) at 50°C, or (2) employ a denaturing agent such as formamide during hybridization, e.g., 50% formamide with 0.1% bovine serum albumin / 0.1% Ficoll / 0.1% polyvinylpyrrolidone / 50 mM sodium phosphate buffer at pH 6.5 with 750 mM NaCl, 75 mM sodium citrate at 42°C. Another example is use of 50% formamide, 5 x SSC (0.75 M NaCl, 0.075 M sodium citrate), 50 mM sodium phosphate (pH 6.8), 0.1% sodium pyrophosphate, 5x Denhardt's solution, sonicated salmon sperm DNA (50 µg/ml), 0.1% SDS, and 10% dextran sulfate at 42°C, with washes at 42°C in 0.2 x SSC and 0.1% SDS. Other examples of stringent conditions are well known in the art.

“Stringent hybridization conditions” and “stringent hybridization wash conditions” in the context of nucleic acid hybridization experiments such as Southern and Northern hybridizations are sequence dependent, and are different under different environmental parameters. Longer sequences hybridize specifically at higher temperatures. The thermal melting point (T_m) is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched primer or probe sequence. Specificity is typically the function of post-hybridization washes, the critical factors being the ionic strength and temperature of the final wash solution. For DNA-DNA hybrids, the T_m can be approximated from the equation of Meinkoth and Wahl (1984); $T_m = 81.5^\circ\text{C} + 16.6 (\log M) + 0.41 (\%GC) - 0.61 (\% \text{ form}) - 500/L$; where M is the molarity of monovalent cations, %GC is the percentage of guanosine and cytosine nucleotides in the DNA, % form is the

percentage of formamide in the hybridization solution, and L is the length of the hybrid in base pairs. T_m is reduced by about 1°C for each 1% of mismatching; thus, T_m , hybridization, and/or wash conditions can be adjusted to hybridize to sequences of the desired identity. For example, if sequences with >90% identity are sought, the T_m can be decreased 10°C. Generally, stringent conditions are selected to be about 5°C lower than the T_m for the specific sequence and its complement at a defined ionic strength and pH. However, severely stringent conditions can utilize a hybridization and/or wash at 1, 2, 3, or 4°C lower than the T_m ; moderately stringent conditions can utilize a hybridization and/or wash at 6, 7, 8, 9, or 10°C lower than the T_m ; low stringency conditions can utilize a hybridization and/or wash at 11, 12, 13, 14, 15, or 20°C lower than the T_m . Using the equation, hybridization and wash compositions, and desired temperature, those of ordinary skill will understand that variations in the stringency of hybridization and/or wash solutions are inherently described. If the desired degree of mismatching results in a temperature of less than 45°C (aqueous solution) or 32°C (formamide solution), the SSC concentration can be increased so that a higher temperature can be used. Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the T_m for the specific sequence at a defined ionic strength and pH.

An example of highly stringent wash conditions is 0.15 M NaCl at 72°C for about 15 minutes. An example of stringent wash conditions is a 0.2 x SSC wash at 65°C for 15 minutes. Often, a high stringency wash is preceded by a low stringency wash to remove background signal. An example of a medium stringency wash for a duplex of, e.g., more than 100 nucleotides, is 1 x SSC at 45°C for 15 minutes. For short nucleotide sequences (e.g., about 10 to 50 nucleotides), stringent conditions typically involve salt concentrations of less than about 1.5 M, less than about 0.01 to 1.0 M, Na ion concentration (or other salts) at pH 7.0 to 8.3, and the temperature is typically at least about 30°C and at least about 60°C for long oligonucleotides (e.g., >50 nucleotides). Stringent conditions also can be achieved by the addition of destabilizing agents such as formamide. In general, a signal to noise ratio of 2x (or higher) than that observed for an unrelated oligonucleotide in the particular hybridization assay indicates detection of a specific hybridization. Nucleic acids that do not hybridize to each other under stringent conditions are still substantially identical if the

proteins that they encode are substantially identical. This can occur, e.g., when a copy of a nucleic acid is created using the maximum codon degeneracy permitted by the genetic code.

Very stringent conditions can be equal to the T_m for a particular oligonucleotide. An example of stringent conditions for hybridization of complementary nucleic acids that have
5 more than 100 complementary residues on a filter in a Southern or Northern blot is 50% formamide, e.g., hybridization in 50% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.1 x SSC at 60 to 65°C. Exemplary low stringency conditions include hybridization with a buffer solution of 30 to 35% formamide, 1 M NaCl, 1% SDS (sodium dodecyl sulphate) at 37°C, and a wash in 1x to 2x SSC (20x SSC = 3.0 M NaCl/0.3 M trisodium citrate) at 50 to
10 55°C. Exemplary moderate stringency conditions include hybridization in 40 to 45% formamide, 1.0 M NaCl, 1% SDS at 37°C, and a wash in 0.5x to 1x SSC at 55 to 60°C.

“Northern analysis” or “Northern blotting” is a method used to identify RNA sequences that hybridize to a known probe such as an oligonucleotide, DNA fragment, cDNA or fragment thereof, or RNA fragment. The probe can be labeled with a radioisotope
15 such as ^{32}P , by biotinylation or with an enzyme. The RNA to be analyzed can be usually electrophoretically separated on an agarose or polyacrylamide gel, transferred to nitrocellulose, nylon, or other suitable membrane, and hybridized with the probe, using standard techniques well known in the art.

Nucleic acid sample may be contacted with an oligonucleotide in any suitable manner
20 known to those skilled in the art. For example, the DNA sample may be solubilized in solution, and contacted with the oligonucleotide by solubilizing the oligonucleotide in solution with the DNA sample under conditions that permit hybridization. Suitable conditions are well known to those skilled in the art. Alternatively, the DNA sample may be solubilized in solution with the oligonucleotide immobilized on a solid support, whereby the
25 DNA sample may be contacted with the oligonucleotide by immersing the solid support having the oligonucleotide immobilized thereon in the solution containing the DNA sample.

The term “substrate” refers to any solid support to which an oligonucleotide may be attached. The substrate material may be modified, covalently or otherwise, with coatings or functional groups to facilitate binding of oligonucleotides. Suitable substrate materials
30 include polymers, glasses, semiconductors, papers, metals, gels and hydrogels among others. Substrates may have any physical shape or size, e.g., plates, strips, or microparticles. The

term “spot” refers to a distinct location on a substrate to which oligonucleotides of known sequence are attached. A spot may be an area on a planar substrate, or it may be, for example, a microparticle distinguishable from other microparticles. The term “bound” means affixed to the solid substrate. A spot is “bound” to the solid substrate when it is affixed in a particular location on the substrate for purposes of the screening assay.

In certain embodiments of the present disclosure, the substrate is a polymer, glass, semiconductor, paper, metal, gel or hydrogel. In certain embodiments of the present disclosure, a kit can further include a solid substrate and at least one control oligonucleotide, wherein the at least one control oligonucleotide is bound onto the substrate in a distinct spot.

In certain embodiments of the present disclosure, the solid substrate is a microarray. An “array” or “microarray” is used synonymously herein to refer to a plurality of primers or probes attached to one or more distinguishable spots on a substrate. A microarray may include a single substrate or a plurality of substrates, for example a plurality of beads or microspheres. A “copy” of a microarray contains the same types and arrangements of primer or probes.

Methods for Detecting Cardiovascular Disease

Better risk assessment for cardiovascular disease is the first step toward more effective prevention. Those identified as being at higher risk (e.g., PPV of 69% for CHD incidence within three years) can be followed up promptly for further testing such as with coronary calcium or angiography, and more aggressive interventions. Conversely, those at lower risk (e.g., NPV of 99% for CHD incident within three years) can be re-tested periodically and monitored to ensure continued prevention due to the dynamic nature of DNA methylation. Compared to the integrated genetic-epigenetic model, overall, conventional risk factors-based calculators were considerably less sensitive, less generalizable, and also depicted a gender gap in performance. In contrast, the integrated genetic-epigenetic model described herein has the ability to capture and better understand the complex nature of CVD via three angles, genetics (inherited risk that is static), DNA methylation (acquired risk that is dynamic) and the genetic confounding of methylation signatures (i.e., $G + M + G \times M$).

The present disclosure provides a method for determining whether a subject has a likelihood of having a CVD incidence within, for example, three years, by determining methylation status of a CpG dinucleotide repeat or CpG dinucleotide repeat motif region, where the methylation status of the CpG dinucleotide is associated with the incidence of CVD. However, the same principals apply to other windows of incidence as well as to the assessment of both the prevalence and incidence of a number of different types of CVD including, without limitation, CHD, stroke, arrhythmia, cardiac arrest, congestive heart failure, atherosclerotic cardiovascular disease (ASCVD) and its associated cardiovascular events (CVE) including, for example, obstructive coronary artery disease (CAD), myocardial infarction (MI), stroke, and cardiovascular death (CVD). In certain embodiments, the method determines the methylation status of a plurality (e.g., any integer between 1 and 10,000, such as at least 100) of CpG dinucleotides and/or SNPs.

As used herein, a "biological sample" encompasses essentially any sample type obtained from a subject that can be used in a method as described herein. The biological sample may be any bodily fluid, tissue or any other sample from which clinically relevant biomarker levels may be determined. "Biological samples" also can encompass cells in culture, cell supernatants, cell lysates, blood, serum, plasma, urine, cerebral spinal fluid, biological fluid, and tissue samples. Various techniques and reagents find use in the methods of the present disclosure. In one embodiment of the disclosure, blood samples, or samples derived from blood, e.g. plasma, circulating, peripheral, lymphocytes, etc., are assayed for the presence of one or more SNPs and/or the methylation status of one or more CpG dinucleotides. A biological sample also can be saliva. Typically, a biological sample that contains nucleic acids is provided and tested. Biological samples can be derived from subjects using well known techniques such as finger prick, venipuncture, lumbar puncture, fluid sample such as saliva or urine, or tissue biopsy and the like.

As used herein, the term "healthy" means that a subject does not manifest a particular condition, and is no more likely than at random to be susceptible to a particular condition.

Prevalence is defined by the American Psychological Association (APA) as the "the total number or percentage of cases (e.g., of a disease or disorder) existing in a population" (APA Dictionary of Psychology, (American Psychological Association, Washington, DC, 2007)). In some instances, point prevalence is used to describe the prevalence of cases at a

discrete point of time, and period prevalence is used to describe the number of cases that exist for a period of time (e.g., a month, a year). Prevalence typically is expressed as a rate per population unit (e.g., number of cases per 100,000 people) instead of an absolute number or a percent.

5 Similarly, incidence is defined by the APA as “the rate of occurrence of new cases of a given event or condition (e.g., a disorder, disease, symptom, or injury) in a particular population in a given period” of time (APA Dictionary of Psychology, (American Psychological Association, Washington, DC, 2007)). As used herein, the term “incidence” is defined as a tendency or susceptibility for a subject to manifest a condition, in this case, CVD
10 (e.g., CHD). In some instances, the period of time can be a year or less than a year; in some instances, the period of time can be longer than a year (e.g., two years, five years, ten years).

Diagnosis is defined by the APA as the “process of identifying and determining the nature of a disease or disorder by its signs and symptoms, through the use of assessment techniques (e.g., tests and examinations) and other available evidence” (APA Dictionary of
15 Psychology, (American Psychological Association, Washington, DC, 2007)). A diagnosis can refer to the present time period, or to a time period in the past or the future.

Likewise, prognosis is defined by the APA as “a prediction of the course, duration, severity, and outcome of a condition, disease, or disorder” (APA Dictionary of Psychology, (American Psychological Association, Washington, DC, 2007)). A prognosis can be made,
20 for example, over a period of one month, six months, one year, five years, ten years, or longer.

Risk assessment is defined as a “study of a subject done for the purpose of trying to determine the probability that that person will develop a particular disease or, if the disease is already present, the probability that the person will suffer exacerbation of it or death from it”
25 (Youngson, 2005, Collins Dictionary of Medicine). In some instances, risk assessment is based on conditions or events and not on disease. In some instances, a risk assessment is determined over a period of time (e.g., months, years).

Biomarkers are described herein that can be used in methods (e.g., predictive or prognostic) of detecting the incidence (e.g., one-year, three-year, five-year) of CVD in a
30 subject. Such methods typically include providing a biological sample from the subject; contacting DNA from the biological sample with bisulfite under alkaline conditions;

contacting the bisulfite-treated DNA with at least one first oligonucleotide primer at least 8 nucleotides in length that is complementary to a sequence that comprises a CpG dinucleotide (e.g., at a GC locus referred to as cg00300879, cg09552548, and cg14789911, or another biomarker from Appendix A); and determining the methylation status of the CpG
5 dinucleotide. It would be understood that the at least one first oligonucleotide probe can detect either the unmethylated CpG dinucleotide or the methylated CpG dinucleotide. Such a method can further include determining the genotype of a single nucleotide polymorphism (SNP) (e.g., rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, or another biomarker from Appendix C) or a second SNP in linkage disequilibrium with the first SNP.
10 As described herein, methylation of one or more particular CpG dinucleotides and the presence of one or more particular SNPs can be used to predict the three-year incidence of CHD in the subject.

In some embodiments, the method further comprises contacting the bisulfite-treated DNA with at least one second oligonucleotide probe at least 8 nucleotides in length that is
15 complementary to a sequence that comprises a CpG dinucleotide, where the at least one second oligonucleotide probe detects either the unmethylated CpG dinucleotide or the methylated CpG dinucleotide, whichever is not detected by the at least one first oligonucleotide probe.

In some embodiments, the ratio of methylated CpG dinucleotides to unmethylated
20 CpG dinucleotides in the biological sample can be determined as a part of the methods described herein. Determining the ratio of methylated CpG dinucleotides to unmethylated CpG dinucleotides can allow for a risk or outcome to be estimated or determined.

It would be appreciated that determining the methylation status of the one or more CpG dinucleotides and determining the presence (or absence) of a SNP can utilize any
25 number of techniques, such as, for example, amplifying and/or sequencing steps. Amplifying and sequencing are well known techniques in the art and are used routinely to determine both the methylation status of a particular sequence and the presence / absence of a SNP.

Methods of determining the presence of biomarkers associated with the three-year incidence of CHD in a biological sample from a subject are provided. A similar approach
30 can be used for any other form of CVD as well. Such methods typically include providing a first portion of the biological sample and contacting DNA from the first portion with bisulfite

under alkaline conditions. The bisulfite-treated first portion can be contacted with a first oligonucleotide probe that is at least 8 nucleotides in length and that is complementary to a sequence that comprises a CpG dinucleotide (detected, e.g., at a CG locus referred to as cg00300879, cg09552548, and cg14789911, or another biomarker from Appendix A), and a
5 second portion of the biological sample can be contacted with a nucleic acid probe at least 8 nucleotides in length that is complementary to a SNP (e.g., rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, or another biomarker from Appendix C).

As described herein, the percentage of methylation of the CpG dinucleotide at one or more of the GC loci designated cg00300879, cg09552548, and cg14789911 (or at a CpG
10 dinucleotide that is in linkage disequilibrium with one or more of such CpG dinucleotides) and the identity of the nucleotide at one or more SNPs designated rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 (or at a SNP that is in linkage disequilibrium with one or more of such SNPs) are biomarkers associated with CVD and can be used to predict the likelihood that an individual will develop CVD and/or prognosticate as to the severity of the
15 disease or the outcome for the individual.

While the effects of the indicated loci on whether or not an individual will develop CVD is a complex relationship, the following trends can be associated with CVD: decreasing methylation of the CpG dinucleotide at loci cg09552548 or cg14789911; increasing methylation of the CpG dinucleotide at loci cg00300879; the presence of a G nucleotide at
20 SNP rs11716050, the presence of a G nucleotide at SNP rs6560711, the presence of a G nucleotide at SNP rs3735222, the presence of a C nucleotide at SNP rs6820447, and the presence of a G nucleotide at SNP rs9638144.

In addition to the SNP and CpG biomarkers identified herein, one or more clinical indicators can be used to aid in either or both diagnostics and prognostics. Without
25 limitation, such clinical indicators can include demographics (e.g., age, sex, race); vital signs (e.g., heart rate (beats/min), systolic BP (mm Hg), diastolic BP (mm Hg)); medical history (e.g., smoking, atrial fibrillation/flutter, hypertension, coronary heart disease, myocardial infarction, heart failure, peripheral artery disease, COPD, diabetes (type 1 or type 2), CVA/TIA, chronic kidney disease, hemodialysis, angioplasty (peripheral or coronary), stent
30 (peripheral or coronary), CABG, percutaneous coronary intervention); medications (ACE-I/ARB, beta blocker, aldosterone antagonist, loop diuretics, nitrates, CCB, statin, aspirin,

warfarin, clopidogrel); echocardiographic results (e.g., LVEF (%), RSVP (mm Hg)); stress test results (e.g., ischemia on scan, ischemia on ECG); angiography results (e.g., $\geq 70\%$ coronary stenosis in ≥ 2 vessels, $\geq 70\%$ coronary stenosis in ≥ 3 vessels); and/or lab measures (e.g., sodium, blood urea nitrogen (mg/dL), creatinine (mg/dL), eGFR (median, 5 CKDEPI), total cholesterol (mg/dL), LDL cholesterol (mg/dL), glycohemoglobin (%), glucose (mg/dL), HGB (mg/dL)).

Kits for Detecting Cardiovascular Disease

10 In a further embodiment of the disclosure, articles of manufacture and kits containing probes, oligonucleotides or antibodies are provided. Such articles of manufacture can be used in the methods described herein. An article of manufacture can include one or more containers with, for example, a label. Suitable containers include, for example, bottles, vials, and test tubes. The containers can be formed from a variety of materials such as glass or plastic. The container can hold a composition that includes one or more agents that are 15 effective for practicing the methods described herein. The label on the container indicates that the composition can be used for a specific application. The kit of the disclosure will typically comprise the container described above and one or more other containers comprising materials desirable from a commercial and user standpoint, including buffers, diluents, filters and package inserts with instructions for use.

20 In certain embodiments, the present disclosure provides a kit for determining the methylation status of at least one CpG dinucleotide and the presence of at least one single-nucleotide polymorphism (SNP). In certain embodiments, a kit as described herein may contain a number of primers that is any integer between 1 and 10,000, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, . . . 9997, 9998, 9999, 10,000. As used herein, the term “nucleic acid primer” or 25 “nucleic acid probes” or “oligonucleotide” encompasses both DNA and RNA sequences. In certain embodiments, the primers or probes may be physically located on a single solid substrate or on multiple substrates.

A kit as described herein can include at least one first nucleic acid primer (e.g., at least 8 nucleotides in length) that is complementary to a bisulfite-converted nucleic acid 30 sequence comprising a CpG dinucleotide (detected, e.g., at a GC locus referred to as cg00300879, cg09552548, and cg14789911), and at least one second nucleic acid primer

(e.g., at least 8 nucleotides in length) that is complementary to a SNP (e.g., rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144). The at least one first nucleic acid primer can detect the methylated or unmethylated CpG dinucleotide.

5 It would be appreciated that any of the nucleic acid primers, probes or oligonucleotides described herein can include one or more nucleotide analogs and/or one or more synthetic or non-natural nucleotides.

10 It also would be appreciated that the kits described herein can include a solid substrate. In some embodiments, one or more of the nucleic acid primers can be bound to the solid support. Examples of solid supports include, without limitation, polymers, glass, semiconductors, papers, metals, gels or hydrogels. Additional examples of solid supports include, without limitation, microarrays or microfluidics cards.

15 It also would be appreciated that any of the kits described herein can include one or more detectable labels. In some embodiments, one or more of the nucleic acid primers can be labeled with the one or more detectable labels. Representative detectable labels include, without limitation, an enzyme label, a fluorescent label, and a colorimetric label.

Algorithm for Predicting the Incidence of Cardiovascular Disease

20 Any number of algorithms that can capture linear effects (e.g., linear regression) or both linear and non-linear effects (e.g., Random Forest, Gradient Boosting, Neural Networks (e.g., deep neural network, extreme learning machine (ELM)), Support Vector Machine, Hidden Markov model) can be used in the methods described herein. See, for example, McKinney et al., 2011, Appl. Bioinform., 5(2):77-88; Gunther et al., 2012, BMC Genet., 13:37; and Ogutu et al., 2011, BMC Proceedings, 5(Suppl 3):S11. Any type of machine learning algorithm or deep learning neural network algorithm (tuned or non-tuned) capable of capturing linear and/or non-linear contribution of traits for the prediction can be used. In some instances, a combination of algorithms (e.g., a combination or ensemble of multiple algorithms that capture linear and/or non-linear contributions of traits) is used.

30 Simply by way of example, Random Forest™ is a popular machine learning algorithm created by Breiman & Cutler for generating “classification trees” (see, for example, “stat.berkeley.edu/~breiman/RandomForests/cc_home.htm” on the World Wide Web). Using standard machine learning and predictive modeling techniques, a diagnostic

classifier algorithm was written to be implemented in R and Python programming languages (though it can be implemented in many other programming languages), according to well described guidelines by Breiman & Cutler. A diagnostic classifier algorithm was generated using data from at least two traits (T) and the diagnosis of interest from that population. To
5 determine the output (e.g., diagnosis) for a new individual, one simply determines values for the at least two traits (T) and inputs that information into an algorithm (e.g., the diagnostic classifier algorithm described herein or another algorithm discussed above) that is capable of capturing the linear and non-linear contributions of the traits.

As described herein, the inputs are at least one genotype (e.g., SNP) and the
10 methylation status of at least one CpG dinucleotide, and the outcome can represent a positive or a negative probability for the incidence (e.g., one-year, three-year, five-year) of CVD. The Traits (T) used to determine the outcome can represent the methylation status of at least one CpG dinucleotide or at least one genotype (e.g., of a SNP), but Traits (T) also can correspond to at least one interaction (e.g., between methylation status and genotype
15 (CpGxSNP), between the methylation status of two different sites (CpGxCpG) or between two different genotypes (SNPxSNP)). It would be appreciated that any such interactions can be visualized using partial dependence plots.

FIG. 8 is a block diagram of an example coronary heart disease classification system
800. In some embodiments, the system 800 can perform monitoring and/or prediction of
20 coronary heart disease. For example, the system 800 can be used to perform one or more of the example processes described herein.

In the illustrated example, a subject 801 provides a subject sample 802. In some
embodiments, the subject sample 802 can be a blood sample, a saliva sample, a mucus
sample, a urine or stool sample, or any other appropriate biological sample from the subject
25 801. In some embodiments, medical personnel 803 (e.g., a doctor, a nurse, a lab technician, a caregiver) may assist the subject 801 with obtaining the subject sample 802. In some
embodiments, the subject 801 may obtain the subject sample 802 from herself or himself
(e.g., by using a portable blood sampling device or a home collection kit).

A nucleic acid isolation module 810 isolates a nucleic acid sample 812 from the
30 subject sample 802. In some embodiments, the nucleic acid isolation module 810 can be a manual, semi-automated, or automatic process that perform or more of cell lysis, removal of

contaminating proteins, deactivating DNAases and/or RNAases, and recovery of DNA and/or RNA. For example, the nucleic acid isolation module 810 can be a part of an automated process or analysis device configured to isolate the nucleic acid sample 812 from the subject sample 802. In another example, the nucleic acid isolation module 810 can be part of one or more of the example kits described in this document, to be used by a human user such as the medical personnel 803.

A genotyping assay module 820 receives a portion 814a of the nucleic acid sample 812. The genotyping assay module 820 is configured to perform a genotyping assay on the portion 814a of the nucleic acid sample 812 to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from Appendix C to determine, identify, or otherwise obtain a collection of genotype data 822. In some embodiments, the genotyping assay module 820 can be a manual, semi-automated, or automatic process. For example, genotyping assay module 820 can be a part of an automated process or analysis device configured to perform a genotyping assay on the portion 814a. In another example, the genotyping assay module 820 can be part of one or more of the example kits described in this document, to be used by a human user such as the medical personnel 803 or a laboratory technician.

A methylation assay module 830 receives a portion 814b of the nucleic acid sample 812. The methylation assay module 830 is configured to bisulfite convert the nucleic acid in the portion 814b of the nucleic acid sample 812 and perform methylation assessment on the portion 814b of the nucleic acid sample 812 to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to determine, identify, or otherwise obtain a collection of methylation data 832.

An identification system 840 is configured to receive the collection of genotype data 822 and the collection of methylation data 832, and identify one or more predetermined traits or characteristics of the subject 801 based on a diagnostic classifier algorithm module 842. The diagnostic classifier algorithm module 842 is configured to account for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect. In some embodiments, the diagnostic classifier algorithm module 842 can perform one or more of the algorithms described herein that may indicate the presence of disease (e.g., diagnostic

indicators) or a propensity to develop disease (e.g., predict). For example, the identification system may be configured to identify genetic and/or environmental characteristics that determines the presence of or the likelihood of a subject developing disease (e.g., cardiovascular disease), even when the disease is of polygenic origin. In some
5 implementations, the diagnostic classifier algorithm module 842 can be a machine learning algorithm capable of accounting for linear and non-linear effects.

The identification system 840 provides an output 850 based on the diagnostic and/or prognostic indicators provided by the diagnostic classifier algorithm module 842. In some embodiments, the identification system 840 can include an output module configured to
10 provide the output 850. In some implementations, the output 850 can be an identification of one or more diseases that the subject 801 may already have. For example, the output 850 may indicate that traits that are indicative of the presence of cardiovascular disease were found in the subject 801. In some implementations, the output 850 can be an indication of a likelihood that the subject 801 may develop a disease within a predetermined time frame (e.g., the
15 subject 801 may have a 43% chance of developing cardiovascular disease within 3 years, the subject 801 may have a 77% of having a heart attack within 2 years). In some implementations, the output 850 can include therapeutic and/or preventative recommendations based on the diagnostic and/or prognostic indicators provided by the diagnostic classifier algorithm module 842. For example, in response to an identification or
20 prediction of a diabetic or cardiac condition in the subject 801, the output 850 may include a recommendation to consult with the medical personnel 803, identify possible dietary or lifestyle changes by the subject 801 to address or avoid the condition, identify potential treatments and/or remedies for the subject 801 to consider in consultation with the medical personnel 803, or combinations of these and/or any other appropriate information based on
25 the output of the algorithm(s) of the diagnostic classifier algorithm module 842.

In the illustrated example, the output 850 is provided in various formats. The information provided by the output 850 can be formatted into a message 860 that is provided to the subject 801 and/or to the medical personnel 803. In some implementations, the message 860 can be formatted as a report (e.g., a word processing file, a portable document
30 format file) that is at least temporarily stored on a non-transitory storage medium (e.g., a hard drive, a FLASH memory), where it can be retrieved by the subject 801 and/or the medical

personnel 803 for review. In some implementations, the message 860 can be formatted as an electronic message (e.g., an email, a text message, an instant message) that is transmitted to the subject 801 and/or the medical personnel 803 for review. In some implementations, the message 860 can be a printed report. For example, the output 850 can be provided to a
5 printing system that is configured to generate a hard copy report based on the output 850. Subsequent automated or manual processing systems can package the report as a letter or other parcel that can be sent for physical delivery to the subject 801 and/or to the medical personnel 803 (e.g., the system 800 can create a paper printout of the results and mail them through postal mail).

10 A treatment device 870 can be configured to receive the diagnostic and/or prognostic indicators provided by the output 850 and provide therapy and/or treatment based on the diagnostic and/or prognostic indicators. For example, the output 850 may indicate that the subject 801 has a high likelihood of suffering cardiac arrest within the next two years, and the treatment device 870 may be a drug (e.g., a tablet or capsule) or an implantable drug
15 delivery system that reacts by identifying or by receiving configuration settings for an appropriate dosage of a statin, acetylsalicylic acid (aspirin), an anti-inflammatory drug, a blood thinner, or combinations of these and/or any other appropriate therapeutic and/or preventative substances. In some embodiments, the treatment device 870 can be configured to also include one or more of the nucleic acid isolation module 810, the genotyping assay
20 module 820, the methylation assay module 830, or the identification system 840.

A storage system 880 is configured to store the output 850. For example, the information included in the output 850 can be stored temporarily, for a predetermined period of time, or substantially permanently in a database, in a file, or as any other appropriate collection of data. In some embodiments, the storage system 880 can store the output 850 in a
25 non-transitory storage medium (e.g., a hard drive, a FLASH memory). For example, the output 850 may include some or all of the collection of genotype data 822, the collection of methylation data 832, and/or the output 850 in a personal health record that the subject 801 can store or carry with them. In some embodiments, the storage system 880 can store the output 850 as a physical medium, for example, the storage system 880 can include a printer that can
30 generate a paper report based on the output 850, and/or store the report as a hard copy that can be physically filed away for later retrieval.

An input/output device 882 is physical device configured to display or otherwise present an output that is perceptible to humans (e.g., the subject 801, the medical personnel 803). For example, the input/output device 882 may be an electronic display device in a doctor's office. The system 800 may process the subject sample 802, and then alter the configuration of pixels onscreen to modify the information displayed by the input/output device 882 based on the output 850 (e.g., a screen can be updated to display an identified diagnosis and/or prognosis for the subject 801 to the medical personnel). In another example, the input/output device 882 can be configured to provide audible (e.g., spoken output) and/or tactile (e.g., braille, haptic, vibratory) output that modifies or otherwise transforms the output 850 into a physical and/or tangible output (e.g., to convey the diagnostic and/or prognostic indicators in a manner that is perceptible to a user who is sight-challenged). In another example, the input/output device 882 can be configured to alter, transform, or modify a physical characteristic of a physical structure or medium based on the output 850.

A user device 884 (e.g., a computer, a smartphone, a tablet computer, a computerized terminal) is configured to display, emit, or otherwise present one or more outputs that are perceptible to a human user, such as the subject 801 and/or the medical personnel 803. For example, the user device 884 can receive the output 850 (e.g., as data, as the message 860) and provide an alert to the user and/or provide an output (e.g., display a report, read a report aloud) based on the output 850. In some embodiments, the user device 884 can include one or more of the storage device 880 or the input/output device 882. In some embodiments, the user device 882 can be part of the treatment device 870. In some embodiments, the user device 884 can be configured to include one or more of the nucleic acid isolation module 810, the genotyping assay module 820, the methylation assay module 830, or the identification system 840.

In some implementations, some or all of the system 800 may be reused to provide additional information. For example, the system 800 may be used to gather an initial set of health information for the subject 801 and/or identify information that can assist the medical personnel 803 with an initial diagnosis/prognosis. Later, the patient 801 may be re-examined using the system 800, for example, to determine the effectiveness of prescribed medical and/or lifestyle strategies over time. Since the collection of genotype data 822 does not change over time for an individual person, the system 800 may refrain from performing the

functions of the genotyping assay module 820 again. In such examples, the methylation assay module 830 may be used to generate an updated version of the collection of methylation data 832, and the updated collection of methylation data 832 can be provided to the identification system 840 for processing along with the collection of genotype data 822 that was previously generated. In some implementations, the subject sample 802 can be collected on a periodic basis and processed based on the existing collection of genotype data 822 and updated collections of methylation data 832 to produce updated outputs 850 that can be used to provide ongoing monitoring of one or more conditions identified for the subject 801.

FIG. 9 is a flow diagram of an example process 900 for cardiovascular disease classification. In some implementations, the process 900 can be some or all of the example processes described above. In some implementations, the process 900 can be the process performed by some or all of the example system 800 of FIG. 8.

At 910, a nucleic acid sample is isolated from a subject sample. For example, the example nucleic acid isolation module 810 can be configured to isolate and/or substantially purify nucleic acid compositions from the example subject sample 802 to produce the example nucleic acid sample 812.

At 920, a genotyping assay is performed on a first portion of the nucleic acid sample to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from Appendix C to obtain genotype data. For example, the example genotyping assay module 820 could be used to analyze the example portion 814a of the nucleic acid sample 812 to produce the example collection of genotype data 822.

At 930, a second portion of the nucleic acid sample is bisulfite converted, and a methylation assessment is performed on the second portion of the nucleic acid sample to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data. For example, the example methylation assay module 830 can be used to process the portion 814b of the nucleic acid sample 812 to produce the example collection of methylation data 832.

At 940, the genotype data from step 920 and/or methylation data from step 930 is input into an algorithm. For example, the example collection of genotype data 822 and the

example collection of methylation data 832 are input into the example identification system 840 and processed using the example diagnostic classifier algorithm module 842.

At 950, at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect are accounted for. For example, the example diagnostic classifier algorithm module 842 can be configured to account for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect. In some implementations, the diagnostic classifier algorithm module 842, can be a machine learning algorithm capable of accounting for linear and non-linear effects.

At 960, an output is provided. For example, the example identification system 840 can provide the example output 850.

At 970 another nucleic acid sample is isolated from another sample from the subject. For example, the example nucleic acid isolation module 810 can be configured to isolate and/or substantially purify nucleic acid compositions from another sample to produce another example nucleic acid sample. Since the collection of genotype data 822 from a subject does not change over time, the newly-produced nucleic acid sample can be used to obtain methylation data 832, which is used along with the existing collection of genotype data 822 to provide an updated output (e.g., to perform a checkup on the subject 801 at a later point in time). In some implementations, this abbreviated process can be performed on a periodic or semi-periodic basis to provide ongoing monitoring of one or more medical conditions identified for the subject 801.

FIG. 10 is a block diagram of example computing devices 1000, 1050 that may be used to implement the systems and methods described in this document, either as a client or as a server or plurality of servers. Computing device 1000 is intended to represent various forms of digital computers, such as laptops, desktops, workstations, personal digital assistants, servers, blade servers, mainframes, and other appropriate computers. Computing device 1000 can also represent all or parts of various forms of computerized devices, such as embedded digital controllers, media bridges, modems, network routers, network access points, network repeaters, and network interface devices including mesh network communication interfaces. Computing device 1050 is intended to represent various forms of mobile devices, such as personal digital assistants, cellular telephones, smartphones, and other similar computing devices. The components shown here, their connections and

relationships, and their functions, are meant to be exemplary only, and are not meant to limit implementations of the compositions and methods described herein.

5 Computing device 1000 includes a processor 1002, a memory 1004, a storage device 1006, a high-speed interface 1008 connecting to memory 1004 and high-speed expansion ports 1010, and a low speed interface 1012 connecting to a low speed bus 1014 and storage device 1006. Each of the components 1002, 1004, 1006, 1008, 1010, and 1012, are interconnected using various busses, and may be mounted on a common motherboard or in other manners as appropriate. The processor 1002 can process instructions for execution within the computing device 1000, including instructions stored in the memory 1004 or on 10 the storage device 1006 to display graphical information for a GUI on an external input/output device, such as display 1016 coupled to high speed interface 1008. In other implementations, multiple processors and/or multiple buses may be used, as appropriate, along with multiple memories and types of memory. Also, multiple computing devices 1000 may be connected, with each device providing portions of the necessary operations (e.g., as a 15 server bank, a group of blade servers, or a multi-processor system).

The memory 1004 stores information within the computing device 1000. In one implementation, the memory 1004 is a computer-readable medium. In one implementation, the memory 1004 is a volatile memory unit or units. In another implementation, the memory 1004 is a non-volatile memory unit or units.

20 The storage device 1006 is capable of providing mass storage for the computing device 1000. In one implementation, the storage device 1006 is a computer-readable medium. In various different implementations, the storage device 1006 may be a floppy disk device, a hard disk device, an optical disk device, or a tape device, a flash memory or other similar solid state memory device, or an array of devices, including devices in a storage area 25 network or other configurations. In one implementation, a computer program product is tangibly embodied in an information carrier. The computer program product contains instructions that, when executed, perform one or more methods, such as those described above. The information carrier is a computer- or machine-readable medium, such as the memory 1004, the storage device 1006, or memory on processor 1002.

30 The high speed controller 1008 manages bandwidth-intensive operations for the computing device 1000, while the low speed controller 1012 manages lower bandwidth-

intensive operations. Such allocation of duties is exemplary only. In one implementation, the high-speed controller 1008 is coupled to memory 1004, display 1016 (e.g., through a graphics processor or accelerator), and to high-speed expansion ports 1010, which may accept various expansion cards (not shown). In the implementation, low-speed controller 1012 is coupled to storage device 1006 and low-speed expansion port 1017 through the low-speed bus 1014. The low-speed expansion port, which may include various communication ports (e.g., Universal Serial Bus (USB), BLUETOOTH, BLUETOOTH Low Energy (BLE), Ethernet, wireless Ethernet (WiFi), High-Definition Multimedia Interface (HDMI), ZIGBEE, visible or infrared transceivers, Infrared Data Association (IrDA), fiber optic, laser, sonic, ultrasonic) may be coupled to one or more input/output devices, such as a keyboard, a pointing device, a scanner, a networking device such as a gateway, modem, switch, or router, e.g., through a network adapter 1013.

Peripheral devices can communicate with the high speed controller 1008 through one or more peripheral interfaces of the low speed controller 1012, including but not limited to a USB stack, an Ethernet stack, a WiFi radio, a BLUETOOTH Low Energy (BLE) radio, a ZIGBEE radio, an HDMI stack, and a BLUETOOTH radio, as is appropriate for the configuration of the particular sensor. For example, a sensor that outputs a reading over a USB cable can communicate through a USB stack.

The network adapter 1013 can communicate with a network 1015. Computer networks typically have one or more gateways, modems, routers, media interfaces, media bridges, repeaters, switches, hubs, Domain Name Servers (DNS), and Dynamic Host Configuration Protocol (DHCP) servers that allow communication between devices on the network and devices on other networks (e.g. the Internet). One such gateway can be a network gateway that routes network communication traffic among devices within the network and devices outside of the network. One common type of network communication traffic that is routed through a network gateway is a Domain Name Server (DNS) request, which is a request to the DNS to resolve a uniform resource locator (URL) or uniform resource indicated (URI) to an associated Internet Protocol (IP) address.

The network 1015 can include one or more networks. The network(s) may provide for communications under various modes or protocols, such as Global System for Mobile communication (GSM) voice calls, Short Message Service (SMS), Enhanced Messaging

Service (EMS), or Multimedia Messaging Service (MMS) messaging, Code Division Multiple Access (CDMA), Time Division Multiple Access (TDMA), Personal Digital Cellular (PDC), Wideband Code Division Multiple Access (WCDMA), CDMA2000, General Packet Radio System (GPRS), or one or more television or cable networks, among
5 others. For example, the communication may occur through a radio-frequency transceiver. In addition, short-range communication may occur, such as using a BLUETOOTH, BLE, ZIGBEE, WiFi, IrDA, or other such transceiver.

In some embodiments, the network 1015 can have a hub-and-spoke network configuration. A hub-and-spoke network configuration can allow for an extensible network
10 that can accommodate components being added, removed, failing, and replaced. This can allow, for example, more, fewer, or different devices on the network 1015. For example, if a device fails or is deprecated by a newer version of the device, the network 1015 can be configured such that network adapter 1013 can to be updated about the replacement device.

In some embodiments, the network 1015 can have a mesh network configuration
15 (e.g., ZIGBEE). Mesh configurations may be contrasted with conventional star/tree network configurations in which the networked devices are directly linked to only a small subset of other network devices (e.g., bridges/switches), and the links between these devices are hierarchical. A mesh network configuration can allow infrastructure nodes (e.g., bridges, switches and other infrastructure devices) to connect directly and non-hierarchically to other
20 nodes. The connections can be dynamically self-organize and self-configure to route data. By not relying on a central coordinator, multiple nodes can participate in the relay of information. In the event of a failure of one or more of the nodes or the communication links between then, the mesh network can self-configure to dynamically redistribute workloads and provide fault-tolerance and network robustness.

25 The computing device 1000 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a standard server 1020, or multiple times in a group of such servers. It may also be implemented as part of a rack server system 1024. It may also be implemented as part of network device such a modem, gateway, router, access point, repeater, mesh node, switch, hub, or security device (e.g., camera
30 server). In addition, it may be implemented in a personal computer such as a laptop computer 1022. Alternatively, components from computing device 1000 may be combined with other

components in a mobile device (not shown), such as device 1050. In some embodiments, the device 1050 can be a mobile telephone (e.g., a smartphone), a handheld computer, a tablet computer, a network appliance, a camera, an enhanced general packet radio service (EGPRS) mobile phone, a media player, a navigation device, an email device, a game console, an interactive or so-called “smart” television, a media streaming device, or a combination of any two or more of these data processing devices or other data processing devices. In some implementations, the device 1050 can be included as part of a motor vehicle (e.g., an automobile, an emergency vehicle (e.g., fire truck, ambulance), a bus). Each of such devices may contain one or more of computing device 1000, 1050, and an entire system may be made up of multiple computing devices 1000, 1050 communicating with each other through a low speed bus or a wired or wireless network.

Computing device 1050 includes a processor 1052, memory 1064, an input/output device such as a display 1054, a communication interface 1066, and a transceiver 1068, among other components. The device 1050 may also be provided with a storage device, such as a microdrive or other device, to provide additional storage. Each of the components 1050, 1052, 1064, 1054, 1066, and 1068, are interconnected using various buses, and several of the components may be mounted on a common motherboard or in other manners as appropriate.

The processor 1052 can process instructions for execution within the computing device 1050, including instructions stored in the memory 1064. The processor may also include separate analog and digital processors. The processor may provide, for example, for coordination of the other components of the device 1050, such as control of user interfaces, applications run by device 1050, and wireless communication by device 1050.

Processor 1052 may communicate with a user through control interface 1058 and display interface 1056 coupled to a display 1054. The display 1054 may be, for example, a TFT LCD display or an OLED display, or other appropriate display technology. The display interface 1056 may comprise appropriate circuitry for driving the display 1054 to present graphical and other information to a user. The control interface 1058 may receive commands from a user and convert them for submission to the processor 1052. In addition, an external interface 1062 may be provide in communication with processor 1052, so as to enable near area communication of device 1050 with other devices. External interface 1062 may provide,

for example, for wired communication (e.g., via a docking procedure) or for wireless communication (e.g., via Bluetooth or other such technologies).

5 The memory 1064 stores information within the computing device 1050. In one implementation, the memory 1064 is a computer-readable medium. In one implementation, the memory 1064 is a volatile memory unit or units. In another implementation, the memory 1064 is a non-volatile memory unit or units. Expansion memory 1074 may also be provided and connected to device 1050 through expansion interface 1072, which may include, for example, a SIMM card interface. Such expansion memory 1074 may provide extra storage space for device 1050, or may also store applications or other information for device 1050. 10 Specifically, expansion memory 1074 may include instructions to carry out or supplement the processes described above, and may include secure information also. Thus, for example, expansion memory 1074 may be provide as a security module for device 1050, and may be programmed with instructions that permit secure use of device 1050. In addition, secure applications may be provided via the SIMM cards, along with additional information, such as 15 placing identifying information on the SIMM card in a non-hackable manner.

The memory may include for example, flash memory and/or MRAM memory, as discussed below. In one implementation, a computer program product is tangibly embodied in an information carrier. The computer program product contains instructions that, when executed, perform one or more methods, such as those described above. The information 20 carrier is a computer- or machine-readable medium, such as the memory 1064, expansion memory 1074, or memory on processor 1052.

Device 1050 may communicate wirelessly through communication interface 1066, which may include digital signal processing circuitry where necessary. Communication interface 1066 may provide for communications under various modes or protocols, such as 25 GSM voice calls, Voice Over LTE (VOLTE) calls, SMS, EMS, or MMS messaging, CDMA, TDMA, PDC, WCDMA, CDMA2000, GPRS, WiMAX, LTE, 5G, among others. Such communication may occur, for example, through radio-frequency transceiver 1068. In addition, short-range communication may occur, such as using a Bluetooth, WiFi, or other such transceiver (not shown) configured to provide uplink and/or downlink portions of data 30 communication. In addition, GPS receiver module 1070 may provide additional wireless data to device 1050, which may be used as appropriate by applications running on device 1050.

Device 1050 may also communicate audibly using audio codec 1060, which may receive spoken information from a user and convert it to usable digital information. Audio codec 1060 may likewise generate audible sound for a user, such as through a speaker, e.g., in a handset of device 1050. Such sound may include sound from voice telephone calls, may include recorded sound (e.g., voice messages, music files, etc.) and may also include sound generated by applications operating on device 1050.

The computing device 1050 may be implemented in a number of different forms, as shown in the figure. For example, it may be implemented as a cellular telephone 1080. It may also be implemented as part of a smartphone 1082, personal digital assistant, or other similar mobile device.

Various implementations of the systems and techniques described here can be realized in digital electronic circuitry, integrated circuitry, specially designed ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof. These various implementations can include implementation in one or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device, and at least one output device.

These computer programs (also known as programs, software, software applications or code) include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. As used herein, the terms “machine-readable medium” “computer-readable medium” refers to any computer program product, apparatus and/or device (e.g., magnetic discs, optical disks, memory, Programmable Logic Devices (PLDs)) used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term “machine-readable signal” refers to any signal used to provide machine instructions and/or data to a programmable processor.

To provide for interaction with a user, the systems and techniques described here can be implemented on a computer having a display device (e.g., a CRT (cathode ray tube) or LCD (liquid crystal display) monitor) for displaying information to the user and a keyboard

and a pointing device (e.g., a mouse or a trackball) by which the user can provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well; for example, feedback provided to the user can be any form of sensory feedback (e.g., visual feedback, auditory feedback, or tactile feedback); and input from the user can be received in any form, including acoustic, speech, or tactile input.

The systems and techniques described here can be implemented in a computing system that includes a back end component (e.g., as a data server), or that includes a middleware component (e.g., an application server), or that includes a front end component (e.g., a client computer having a graphical user interface or a Web browser through which a user can interact with an implementation of the systems and techniques described here), or any combination of such back end, middleware, or front end components. The components of the system can be interconnected by any form or medium of digital data communication (e.g., a communication network). Examples of communication networks include a local area network ("LAN"), a wide area network ("WAN"), and the Internet.

Some communication networks can be configured to carry power as well as information on the same physical media. This allows a single cable to provide both data connection and electric power to devices. Examples of such shared media include power over network configurations in which power is provided over media that is primarily or previously used for communications. One specific embodiment of power over network is Power Over Ethernet (POE) which pass electric power along with data on twisted pair Ethernet cabling. Examples of such shared media also include network over power configurations in which communication is performed over media that is primarily or previously used for providing power. One specific embodiment of network over power is Power Line Communication (PLC) (also known as power-line carrier, power-line digital subscriber line (PDSL), mains communication, power-line telecommunications, or power-line networking (PLN), Ethernet-Over-Power (EOP)) in which data is carried on a conductor that is also used simultaneously for AC electric power transmission.

The computing system can include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

The computing system can include routers, gateways, modems, switches, hub, bridges, and repeaters. A router is a networking device that forwards data packets between computer networks and performs traffic directing functions. A network switch is a networking device that connects networked devices together by performing packet switching to receive, process, and forward data to destination devices. A gateway is a network device that allows data to flow from one discrete network to another. Some gateways can be distinct from routers or switches in that they can communicate using more than one protocol and can operate at one or more of the seven layers of the open systems interconnection model (OSI). A media bridge is a network device that converts data between transmission media so that it can be transmitted from computer to computer. A modem is a type of media bridge, typically used to connect a local area network to a wide area network such as a telecommunications network. A network repeater is a network device that receives a signal and retransmits it to extend transmissions and allow the signal can cover longer distances or overcome a communications obstruction.

It will be apparent that the present disclosure provides a skilled artisan the ability to construct a matrix in which the methylation status of one or more CpG dinucleotides and one or more genotypes (e.g., SNPs; e.g., at one or more alleles) can be evaluated as described herein, typically using a computer, to identify interactions and allow for prediction of the incidence of CVD. Although such an analysis is complex, no undue experimentation is required as all necessary information is either readily available to the skilled artisan or can be acquired by experimentation as described herein.

Methods of Treating Cardiovascular Diseases

The present disclosure provides a method for determining the likelihood that a subject will have a CVD event within, for example, one year, three years, or five years. As used herein, CVD includes, without limitation, CHD, stroke, arrhythmia, cardiac arrest, and congestive heart failure. The methods and compositions described herein provide a better ability to assess a subjects risk for cardiovascular disease, which is the first step toward more effective prevention.

Upon making a positive prognosis of a cardiac outcome (e.g., a prognosis of cardiovascular death, myocardial infarct (MI), stroke, all cause death, or a composite

thereof), a medical practitioner can advantageously use the prognostic information thereby obtained to identify the need for an intervention in the subject, such as, for example, stress testing with ECG response or myocardial perfusion imaging, coronary computed tomography angiogram, diagnostic cardiac catheterization, percutaneous coronary (e.g., balloon angioplasty with or without stent placement), coronary artery bypass graft (CABG), enrollment in a clinical trial, and administration or monitoring of effects of agents selected from, but not limited to, of agents selected from nitrates, beta blockers, ACE inhibitors, antiplatelet agents and lipid-lowering agents.

Those identified as being at higher risk (e.g., PPV of 69% for CHD incidence within three years) can be followed up promptly for further testing or more aggressive interventions. Conversely, those at lower risk can be re-tested periodically and monitored to ensure continued prevention due to the dynamic nature of DNA methylation.

Treatments for cardiovascular disease can depend on the type of cardiovascular disease and the symptoms the individual is experiencing. Treatments for cardiovascular disease can be preventative, therapeutic or palliative. Treatments for cardiovascular diseases can include, for example, lifestyle changes (e.g., diet (e.g., low fat diet), weight loss, exercise, reduction or cessation in smoking and/or drinking), pharmaceuticals (e.g., beta blockers, statins, calcium channel blocker, ACE inhibitors, vasodilator, alteplase) and/or surgical interventions (e.g., angioplasty, bypass surgery, implantable device, endarterectomy).

In accordance with the present disclosure, there may be employed conventional molecular biology, microbiology, biochemical, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. The invention will be further described in the following examples, which do not limit the scope of the methods and compositions of matter described in the claims.

EXAMPLES

Example 1—Materials and Methods

This study features data and/or biomaterial from two sources. The first set of anonymized genome-wide genetic, genome-wide DNA methylation and clinical data are

from the eighth examination cycle of the Framingham Heart Study (FHS) Offspring cohort, while the second set of anonymized clinical data and DNA are from the Intermountain Healthcare (IM) biorepository. The procedures and protocols used for the analysis of the FHS data were approved by the University of Iowa Institutional Review Board (IRB# 201503802), and the procedures and protocols used for the analyses of the IM materials were approved by the Intermountain Healthcare Institutional Review Board (IRB# 1024811).

Example 2—Framingham Heart Study Offspring Cohort

The details on the collection and preparation of clinical and biological data of the FHS cohort have been described previously (dbGAP study accession: phs000007). In brief, the demographics, risk factors and clinical information were derived from the eighth examination of the Offspring cohort, with additional clinical follow-up information used to determine incident coronary heart disease (CHD) status. Incident CHD was considered present if an individual was diagnosed with CHD within three years of the eighth examination cycle. Conversely, incident CHD was considered absent if an individual was not diagnosed with CHD within three years of the eighth examination cycle. Data from those with prevalent CHD at the eighth examination cycle were excluded from further consideration. Sources of clinical data in determining incident CHD events included subject report, review of medical records, and death certificates. The designations and dates of CHD onset used in this study are as determined by a panel of three investigators on the Framingham Endpoint Review Committee.

Genome-wide DNA methylation data profiled using the Illumina Infinium HumanMethylation450 BeadChip array (San Diego, CA, USA) was available from 2,567 subjects who were phlebotomized at the eighth examination cycle. Standard sample and probe level quality control were performed as described in previous studies, which resulted in retaining 2,560 samples and DNA methylation data from 403,192 loci (see, e.g., Dogan et al., 2018, Genes, 9:641; Pidsley et al., 2013, BMC Genomics, 14:1-10; Triche, 2014, FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes. Vol. R package version 2.2.0; Davis et al., 2018, Handle Illumina methylation data., Vol. R package version 2.22.0; and Dogan et al., 2018, PLoS One, 13:e0190549). Genome-wide genotype data obtained using the Affymetrix GeneChip HumanMapping 500K array

(Santa Clara, CA, USA) was available for 2,406 of the remaining samples. After standard sample and probe level quality control procedures were performed in PLINK on the array data as described previously, the total number of samples and SNPs remaining were 2,295 and 472,822, respectively (Dogan et al., 2018, Genes, 9:641; Dogan et al., 2018, PLoS One, 13:e0190549; and Purcell et al., 2007, Am. J. Hum. Genet., 81:559-75). A challenge in conducting biological studies of community cohorts such as the FHS is the potential for inter-relatedness of some of the subjects. Therefore, the genetic data were subjected to relatedness analysis in PLINK. Based on relatedness and incident CHD status, 1,280 subjects (18/542 males and 10/738 females diagnosed with clinical CHD within three years of the eighth examination cycle ascertainment) were part of the training set and 639 subjects (9/271 males and 5/368 females diagnosed with clinical CHD within three years of the eighth examination cycle ascertainment) were part of the test set. The demographics and conventional risk factors of these individuals are summarized in Table 1.

Table 1. Summary of demographics and conventional CHD risk factors for the individuals in the Framingham Heart Study Offspring cohort

	Training (n=1,280)		Test (n=639)	
	CHD*	No CHD†	CHD*	No CHD†
Gender (count)				
Male	18	524	9	262
Female	10	728	5	363
Age (years)				
Male	70.6±9.3	65.8±8.2	66.1±9.1	62.7±9.0
Female	71.2±10.3	66.3±8.5	66.8±9.0	64.9±9.3
Total Cholesterol (mg/dL)				
Male	171±54	177±32	161±30	182±32
Female	229±40	199±36	185±49	197±35
HDL Cholesterol (mg/dL)				
Male	50±16	50±14	48±12	51±15
Female	57±16	65±19	60±17	65±19

HbA1c (%)				
Male	5.7±0.4	5.7±0.8	5.8±0.8	5.6±0.5
Female	6.0±1.0	5.7±0.5	5.8±0.8	5.7±0.6
SBP (mmHg)				
Male	137±15	130±17	136±11	128±16
Female	140±19	129±17	132±22	127±18
DBP (mmHg)				
Male	74±11	76±11	74±8	77±10
Female	80±11	73±10	72±7	73±10
Smoker (count)				
Male	1 (6%)	35 (7%)	2 (22%)	16 (6%)
Female	2 (20%)	57 (8%)	0 (0%)	32 (9%)
Blood Pressure Treatment (count)				
Male	12 (67%)	265 (51%)	4 (44%)	112 (43%)
Female	3 (30%)	294 (40%)	4 (80%)	157 (43%)

*Those diagnosed with CHD within three years of contributing biomaterial during the Offspring Cohort eighth examination cycle.

†Those not diagnosed with CHD within three years of contributing biomaterial during the Offspring Cohort eighth examination cycle.

5 HDL: high-density lipoprotein, HbA1c: Hemoglobin A1c, SBP: systolic blood pressure, DBP: diastolic blood pressure.

Example 3—Intermountain Healthcare Cohort

10 The second de-identified cohort consisting of 159 subjects were drawn from the Intermountain Healthcare (IM) Heart Institute INSPIRE registry, where participants contributed biomaterial and have electronic medical records (EMR). These subjects were subjects who underwent coronary angiography at IM, provided consent to participate in the registry, and for whom both DNA from the time of the catheterization (i.e. index) and clinical follow up status with respect to incident CHD status were available. As documented in their

15 medical records, each of the subjects had stenosis of <50% of each of their main cardiac arteries with no other clinical evidence of an atherosclerotic heart disease event prior to or at the time of their coronary angiogram. Incident CHD status was determined based on follow-up EMR data. Incident CHD was considered present if the subject was clinically diagnosed

with CHD (>70% stenosis) on angiography, had a myocardial infarction, revascularization or death due to CHD within three years of index coronary angiography and biomaterial collection.

When available, conventional risk factor values (age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL) cholesterol level, total cholesterol level, hemoglobin A1c (HbA1c), and smoking status) also were obtained. The blood pressure values were from the admission assessment for the index coronary angiogram. For cholesterol and HbA1c, they were first available values in the 12 months prior to and 3 months after catheterization. The samples were randomly split into validation (50%) and test (50%) sets, stratified by incident CHD status, where 80 subjects (12/39 males and 11/41 females diagnosed with clinical CHD within three years of the eighth examination cycle ascertainment) were part of the validation set and 79 subjects (11/38 males and 10/41 females diagnosed with clinical CHD within three years of the eighth examination cycle ascertainment) were part of the test set. Please note that, in contrast to the FHS sample where class imbalance is evident, incident cases were intentionally selected for this cohort to ensure better balance between cases and controls. The demographics and conventional risk factors of these individuals are summarized in Table 2.

Table 2. Summary of demographics and conventional CHD risk factors for the Intermountain Healthcare validation and test sets

	Validation (n=80)		Test (n=79)	
	CHD*	No CHD†	CHD*	No CHD†
Gender (count)				
Male	12	27	11	27
Female	11	30	10	31
Age (years)				
Male	61.4±13.4	61.4±15.8	63.9±16.8	61.3±17.4
Female	62.9±16.7	68.5±11.4	66.1±11.8	63.9±15.2
Total Cholesterol (mg/dL)				
Male	144±42	178±40	169±38	172±36

Female	190±49	183±42	196±58	175±26
HDL Cholesterol (mg/dL)				
Male	37±9	39±10	41±11	38±14
Female	58±14	61±22	52±12	50±14
HbA1c (%)				
Male	6.2±1.2	5.9±0.4	6.3±1.3	6.4±1.1
Female	5.9±0.4	6.1±1.4	6.9±2.4	5.4±0.5
SBP (mmHg)				
Male	152±23	143±22	143±21	143±27
Female	141±18	152±26	153±21	147±19
DBP (mmHg)				
Male	84±11	85±12	79±11	83±16
Female	75±6	81±13	86±11	78±12
Smoker (count)				
Male	0 (0%)	2 (7%)	1 (9%)	0 (0%)
Female	1 (9%)	1 (3%)	2 (20%)	1 (3%)
Blood Pressure Treatment (count)				
Male	3 (25%)	7 (26%)	5 (45%)	10 (37%)
Female	3 (27%)	10 (33%)	3 (30%)	16 (52%)

*Those diagnosed with CHD within three years of contributing biomaterial during the Offspring Cohort eighth examination cycle.

†Those not diagnosed with CHD within three years of contributing biomaterial during the Offspring Cohort eighth examination cycle.

HDL: high-density lipoprotein, HbA1c: Hemoglobin A1c, SBP: systolic blood pressure, DBP: diastolic blood pressure.

5

Genome-wide DNA methylation and genetic assessments for each of these 159 subjects were conducted by the University of Minnesota Genome Center using the Illumina Infinium MethylationEpic Beadchip array and the Illumina Infinium Multi-Ethnic Global BeadChip array (San Diego, CA, USA), respectively. These data were then subjected to the same quality control procedure described above for the FHS samples. A total of 862,593 methylation and 818,046 SNP loci survived quality control measures. For DNA methylation,

10

loci common to both the Illumina 450K and EPIC arrays were retained, resulting in 437,242 loci for further analysis. Similarly for SNPs, those loci common to both genotyping arrays were retained, resulting in 80,371 loci for further analysis.

5 Example 4—Integrated Genetic-Epigenetic Incident Coronary Heart Disease Risk Prediction Model

Because one of the aims of this study is to translate array-based methylation loci to clinically implementable digital PCR (dPCR) assays, which has fixed constraints on precision, prior to performing data mining exclusively using data from the FHS training set, 10 the methylation variables were reduced to include loci whose delta beta ($\Delta\beta$) (absolute difference between case and controls) was at least 0.03. Covariate shift (Quionero-Candela et al, 2009, *Dataset Shift in Machine Learning*, The MIT Press) between the FHS training set and IM validation set was used to further reduce the number of methylation loci. As a result of both of these variable reduction steps, 5,571 methylation probes remained for downstream 15 analysis. All methylation loci beta values were converted into M-values and subsequently scaled to have zero mean and unit variance.

All data mining, feature selection, model development and model tuning were performed exclusively on the FHS training set. We integrated the 5,571 methylation loci with the 80,371 SNPs to mine for integrated genetic-epigenetic biomarkers that are highly 20 predictive of risk for incident CHD within three years. Our data mining approach has been outlined in previous publications (Dogan et al., 2018, *Genes*, 9:641; Dogan et al., 2018, *PLoS One*, 13:e0190549). All analyses were performed in Python. Briefly, an undersampling-based approach was implemented to account for the high class imbalance and coupled to an ensemble of machine learning algorithms (Random Forest, Support Vector Machine and 25 Logistic Regression) that incorporated cross-validation to uncover non-linear methylation-SNP interactions and highly predictive biosignatures in the FHS training set (Han et al., 2011, *Data Mining: Concepts and Techniques*, Elsevier). As a result, a marker set was selected consisting of three DNA methylation loci and five SNPs that had the best combined performance with respect to area under the receiver operating characteristic curve (AUC), 30 sensitivity and specificity. The ensemble model consisting of these eight biomarkers underwent hyperparameter tuning and was finalized for testing. The final trained integrated

genetic-epigenetic model was then applied on the FHS test, IM validation and IM test sets to determine the AUC, sensitivity and specificity in these sets.

To better understand if adding conventional CHD risk factors (age, gender, systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL) cholesterol level, total cholesterol level, hemoglobin A1c (HbA1c), and smoking status) to the integrated genetic-epigenetic model could improve performance, each risk factor was added to the final trained model and tested on the FHS test, IM validation and IM test sets.

Example 5—Polygenic Risk Score

To understand how the performance of the integrated genetic-epigenetic model described herein compared to that of Polygenic Risk Score (PRS) for incident CHD risk prediction, PRS was calculated using summary statistics from a genome-wide meta-analysis of CHD that were performed in 60,801 cases and 123,504 controls using Python Version 3.7 (Nikpay et al., 2015, Nat. Genet., 47:1121-30). Because only 80,371 SNPs overlapped between the Affymetrix array that was used to profile FHS subjects and the MultiEthnic Global BeadChip array that was used to profile IM subjects, PRS was modelled three ways.

The first was to calculate PRS based on 57,647 overlapping SNPs between both arrays that also had corresponding CHD associated log OR. For each subject, PRS was calculated by taking the product of the number of alleles associated with risk and the respective SNP's log odds ratio (log OR) for each SNP that were subsequently summed across all SNPs. Using undersampling-based logistic regression to account for class imbalance, a PRS model was fitted in the FHS training set and tested on the FHS test, IM validation and IM test sets, and the AUC, sensitivity and specificity of this model (Model 1) in each of these datasets were evaluated.

The second was to calculate PRS in the FHS cohort using 394,304 SNPs from the Affymetrix array that had corresponding CHD associated log OR. Once PRS was calculated, the same modelling approach was used to build a model in the FHS training set that was subsequently only tested on the FHS test set. The AUC, sensitivity and specificity of this model (Model 2) were evaluated.

The third was to calculate PRS in the IM cohort using 527,720 SNPs from the Illumina Multi-Ethnic Global array that had corresponding CHD associated log OR. Once

PRS was calculated, the same modelling approach was used to build a model in the IM validation set that was subsequently only tested on the IM test set. The AUC, sensitivity and specificity of this model (Model 3) were evaluated.

5 Example 6—Survival Analysis and Prognostic Scores

Using data from the FHS test, IM validation and IM test sets, a Kaplan-Meier survival curve was fitted to display the time to incident CHD event within three years as a function of risk group (high vs. low) as predicted by the integrated genetic-epigenetic model. The y-axis represents the probability of not having an incident CHD event within three years. The 95%
10 confidence interval (CI) for each of the distribution was calculated and the distributions of the high and low risk groups were compared using the log-rank test.

The two risk groups then were transformed into three clinical prognostic scores (score 1 = low risk, score 2 = intermediate risk, score 3 = high risk) using the probability of having an incident event as predicted by the integrated genetic-epigenetic model. A Kaplan-Meier
15 survival curve was fitted for these prognosis scores alongside their respective 95% CIs and compared using the log-rank test.

Example 7—Conventional Risk Factors-Based Model

To compare the performance of the integrated genetic-epigenetic model described
20 herein to two commonly used conventional risk factors-based models, FRS and PCE, these risk calculators were implemented on both cohorts to identify those at high risk for CHD incidence (>20%). The variables used in this analysis include age, gender, total cholesterol, HDL, SBP, DBP, diabetes status, smoking status, and whether individuals are undergoing blood pressure treatment. Individuals with missing values and those with values outside the
25 allowed range (e.g. for PCE, age must be between 20-79) were excluded from this analysis.

Example 8—Digital PCR Assay Development

Array-based clinical testing can be time consuming and costly. Simple, readily available Taqman assays can be used to profile SNPs of interest from genotyping arrays.
30 However, there are limited options for profiling methylation loci of interest for clinical tests in a timely and cost effective manner. To demonstrate that the approach described herein can

be used in a clinical setting, the array-based methylation biomarkers in the integrated genetic-epigenetic model described herein were translated into dPCR assays. For each of the methylation loci, DNA from the IM cohort was bisulfite converted using the Qiagen EpiTect Bisulfite kit (Hilden, Germany). The bisulfite converted DNA was subjected to PCR amplification using custom primers. An aliquot of the amplified DNA was used to perform dPCR using custom primer and probe sets capable of distinguishing methylated and unmethylated targets. Correlation analysis was performed between the dPCR beta values and array beta values for each of the locus to demonstrate successful translation.

Example 9—Results

The clinical and demographic characteristics of the FHS and IM cohorts is outlined in Tables 1 and 2, respectively. The average age of subjects in the FHS and IM cohorts was in the mid and early 60s, respectively, with the age range in both cohorts extending from at least the lower 40s to the upper 80s. All of the subjects from the FHS cohort were of European ancestry, but at least 10 of the subjects in the IM cohort were of non-European ancestry. The most notable difference was with respect to gender. The FHS cohort had more females than males, while the IM subjects were intentionally selected to maintain gender balance in the cohort. In general, on average, total and HDL cholesterol levels were higher in FHS compared to IM and vice versa for HbA1c, SBP and DBP. Furthermore, in both the FHS and IM cohorts, on average, for incident cases and controls, total and HDL cholesterol levels were higher in females than males.

The distribution of the number of incident cases over the three year period for FHS and IM are shown in FIGs. 1 and 2, respectively. Among the 42 FHS Offspring subjects diagnosed with CHD within three years of the eighth examination cycle, the highest (29%) and lowest (7%) number of incident cases occurred between 12-18 months and 0-6 months, respectively. In contrast, among the 44 IM subjects diagnosed with CHD within three years of index coronary angiography, the highest (43%) number of subjects had their first event within 6 months of index coronary angiography, whereas the lowest (9%) number of incident cases occurred between 6-12 and 12-18 months. Still, when the entire three year incidence window was considered, the average time to event in both cohorts was similar at 1.5 ± 0.7 and 1.1 ± 1.0 years for FHS and IM, respectively.

Example 10—Integrated Genetic-Epigenetic Incident Coronary Heart Disease Risk Prediction Model

Using integrated genome-wide SNP and methylation data from the 1,919 subjects in the FHS training set, an incident CHD prediction model was built to identify those at high risk of having a heart attack or sudden death within three years. This final ensemble model consisted of a total of eight biomarkers, three of which were DNA methylation biomarkers and the remaining five were SNPs. The three methylation loci are cg00300879 (TSS200 of CNKSR1), cg09552548 (Intergenic), and cg14789911 (Body of SPATC1L), while the five SNPs are rs11716050 (LOC105376934), rs6560711 (WDR37), rs3735222 (SCIN/LOC107986769), rs6820447 (intergenic), and rs9638144 (ESYT2). The integrated genetic-epigenetic model described herein performed with an AUC, sensitivity and specificity of 0.90, 0.85, and 0.75, respectively, when evaluated with the same FHS training set.

This model was then evaluated in the FHS test, IM validation and IM test sets. The AUC sensitivity and specificity of the final model in these sets are summarized in Table 3. The ROC curves are shown in FIG. 3. Briefly, the average AUC, sensitivity and specificity in these sets are 0.79, 0.76 and 0.73, respectively. These performance metrics indicated good generalizability of the trained model to the FHS test set and the external IM cohort. The performance breakdown of the integrated genetic-epigenetic model described herein then was evaluated in each set by gender. These results are summarized in Table 4. For men, the average AUC, sensitivity and specificity of the integrated genetic-epigenetic model across the FHS test, IM validation and IM test sets were 0.79, 0.75 and 0.74, respectively. For women, the average AUC, sensitivity and specificity of the integrated genetic-epigenetic model across the FHS test, IM validation and IM test sets were 0.80, 0.77 and 0.72, respectively. The similar performance metrics for both men and women and across cohorts once again indicate good generalizability of the tool.

Table 3. Performance of our integrated genetic-epigenetic model in the Framingham Heart Study and Intermountain Healthcare cohorts

Dataset	AUC	Sensitivity	Specificity
Framingham Heart Study			
Training set	0.90	0.85	0.75

Test set	0.84	0.79	0.75
Intermountain Healthcare			
Validation set	0.78	0.78	0.74
Test set	0.74	0.71	0.71

AUC: area under the receiver operating characteristic curve.

Table 4. Performance of our integrated genetic-epigenetic model by gender in the Framingham Heart Study and Intermountain Healthcare cohorts

Dataset	AUC	Sensitivity	Specificity
FHS training			
Male	0.90	0.89	0.76
Female	0.89	0.80	0.74
FHS test			
Male	0.82	0.78	0.77
Female	0.88	0.80	0.73
IM validation			
Male	0.81	0.75	0.74
Female	0.76	0.82	0.73
IM test			
Male	0.74	0.73	0.70
Female	0.75	0.70	0.71

FHS: Framingham Heart Study cohort, IM: Intermountain Healthcare cohort, AUC: area under the receiver operating characteristic curve.

To demonstrate that the performance of the integrated genetic-epigenetic model described herein is driven by the integration of DNA methylation signatures with SNPs, models were fitted with only the five SNPs and only the three DNA methylation biomarkers. The average AUCs of these models across all four sets relative to that of the integrated genetic-epigenetic model is shown in FIG. 4. Based on these AUCs, these findings suggest that integrating both types of biomarkers allows better identification of the high and low risk groups. It was found that, on average, DNA methylation loci contributed largely to sensitivity while SNPs contribute to specificity. Similarly, to better understand if the addition of conventional CHD risk factors improves the performance of the integrated genetic-epigenetic model described herein, the average AUC of the integrated genetic-epigenetic model was compared across all four sets (baseline) to that of models that incorporated each of these risk factors. The AUCs of these models relative to the baseline also is summarized in FIG 4. None of the additions resulted in an increase in average AUC

compared to the integrated genetic-epigenetic model, suggesting that the integrated genetic-epigenetic biomarkers are capturing variance associated with conventional risk factors.

Example 11—Polygenic Risk Score

To better understand the ability of a model that only incorporates SNPs (i.e. PRS) for incident CHD risk prediction compared to the integrated genetic-epigenetic model described herein that integrates three methylation biomarkers with five SNPs, three PRS models were trained and tested. The performance of Models 1, 2 and 3 in the various sets are summarized in Table 5. The AUC, sensitivity and specificity of these models ranged from 0.41-0.54, 0.22-0.50 and 0.47-0.69, respectively.

Table 5. Performance of Polygenic Risk Score models

Model	AUC	Sensitivity	Specificity
Model 1: overlapping SNPs			
FHS test	0.54	0.50	0.59
IM validation	0.41	0.22	0.56
IM test	0.45	0.38	0.57
Model 2: FHS SNPs			
FHS test	0.41	0.43	0.47
Model 3: IM SNPs			
IM test	0.52	0.38	0.69

FHS: Framingham Heart Study cohort, IM: Intermountain Healthcare cohort
AUC: area under the receiver operating characteristic curve.

These findings indicate that all three versions, on average, had better specificity than sensitivity. Model 1, which was trained on the FHS training set with the least number of SNPs compared to Models 2 and 3, performed with the best AUC and sensitivity on the FHS test set of 0.54 and 0.50, respectively. However, the highest specificity of 0.69 was observed for Model 3, which was trained on the IM validation set with the most number of SNPs and tested on the IM test set. It was found that the models were not highly generalizable between cohorts.

The AUC, sensitivity and specificity of the integrated genetic-epigenetic model clearly outperformed that of PRS. The approach described herein also is more generalizable, consisted of far fewer SNPs and incorporated informative DNA methylation biomarkers in

addition to SNPs. However, to determine whether the addition of PRS to the integrated genetic-epigenetic model could potentially improve risk assessment, the average AUC of the integrated genetic-epigenetic model described herein (baseline) was compared with and without incorporating Model 1 PRS. The average AUC when PRS was incorporated is 0.79, which is lower than the average AUC of the baseline model as shown in FIG. 4.

Example 12—Survival Curve and Prognostic Scores

The Kaplan-Meier survival curve for the high and low risk groups is shown in FIG. 5. For those with poor prognosis (i.e. at higher risk of having an incident CHD event within three years), there is a clear rapid drop in probability of not having an incident even compared to the good prognosis group (i.e. at lower risk of having an incident CHD event within three years). The log-rank test p-value between these two groups of 2.46×10^{-16} indicate a statistically significant difference between their distributions.

These groups then were translated into clinical prognostic score of 1, 2 and 3 to indicate low, intermediate and high risk groups, respectively. The Kaplan-Meier survival score for these score are shown in FIG. 6. Once again, there is a clear rapid drop in probability of not having an incident event within three years with a high prognostic score. The log-rank test p-value between these groups is 9.39×10^{-33} , indicating a statistically significant difference between their distributions. For the high risk group with a prognostic score of 3, the positive predictive value (PPV) is 69%. For the low risk group with a prognostics score of 1, the negative predictive value (NPV) is 99%. The intermediate risk group with a prognostic score of 2 has PPV and NPV of 15% and 94%, respectively. Thus, individuals in the high and intermediate risk groups are 50 and 10 times more likely to have an incident CHD event in the next three years, respectively, compared to the low risk group.

Example 13—Conventional Risk Factors-Based Model

To compare the performance of the approach described herein to commonly used standard risk factors-based calculators, FRS and PCE, these estimators were applied to both cohorts. A majority of the risk factors considered by both of these risk calculators are the same (age, sex, smoking status, diabetes, SBP, total cholesterol and HDL cholesterol). In addition to these risk factors, the FRS calculator considers DBP, whereas the PCE calculator

considers the use of blood pressure treatment. Due to missing values and constraints in these calculators such as with respect to age, not all subjects were evaluated. The performance of these models across both cohorts are summarized in Table 6, and the breakdown by gender is summarized in Table 7. On average, FRS performed with 0.23 and 0.91 sensitivity and specificity, respectively. The PCE risk estimator, on average, performed with a sensitivity and specificity of 0.55 and 0.65, respectively. The FRS calculator had better specificity over sensitivity in both cohorts and vice versa for the PCE calculator. Similarly, for the gender breakdown, in general, FRS tended to perform with better specificity for both men and women, while PCE tended to perform better with respect to sensitivity. The integrated genetic-epigenetic approach was 52% and 51% more sensitive for men and women, respectively, compared to FRS. It was also 10% and 39% more sensitive and 10% and 6% more specific for men and women, respectively, compared to PCE.

Table 6. Performance of the Framingham Risk Score and ASCVD Pooled Cohort risk estimators in the FHS and IM cohorts.

Risk Estimator	Sensitivity	Specificity
Framingham Risk Score		
FHS cohort	0.15	0.93
IM cohort	0.31	0.89
ASCVD Pooled Cohort Equation		
FHS cohort	0.41	0.74
IM cohort	0.69	0.55

FHS: Framingham Heart Study cohort, IM: Intermountain Healthcare cohort.

Table 7. Performance of the Framingham Risk Score and ASCVD Pooled Cohort risk estimators in the FHS and IM cohorts by gender

Risk Estimator	Sensitivity	Specificity
Framingham Risk Score		
FHS cohort		
Male	0.12	0.86
Female	0.22	0.98
IM cohort		
Male	0.33	0.85
Female	0.29	0.93
ASCVD Pooled Cohort Equation		
FHS cohort		

Male	0.52	0.66
Female	0.18	0.81
IM cohort		
Male	0.78	0.61
Female	0.57	0.50

FHS: Framingham Heart Study cohort, IM: Intermountain Healthcare cohort.

Example 14—Digital PCR Assays

Because one of the goals of this study is to demonstrate the applicability of the integrated genetic-epigenetic model described herein in conventional clinical or research settings, the time consuming, labor intensive genome-wide methylation approach was translated into a simple, quick to perform methylation sensitive dPCR assays for methylation loci included in the final model. As an example, in FIG. 5, the translation of cg00300879 is shown. The Pearson correlation between methylation values as determined by dPCR to that of their corresponding array values for cg00300879 is 0.94. This high correlation suggests that dPCR is a viable alternative to array-based DNA methylation assessments.

Example 15—Additional Data

Appendix A shows a list of CpGs whose methylation is associated with CVD. Appendix B shows a list of genes whose methylation is associated with CVD. Appendix C shows a list of SNPs associated with CVD. The numerical values provided in Appendix A, B, and C are the mean of 5-fold cross validation scores, AUC ROC (Area Under The Receiver Operating Characteristic Curve), sensitivity and specificity, which were computed by the diabetes assessment / prediction algorithm described herein. Sensitivity is the true positive rate and specificity is the true negative rate.

Example 16—Materials and Methods

Study Approval:

All subjects who participated in the study provided informed written consent. All procedures used in the study were approved by the University of Iowa Institutional Review Board (IRB201706713).

Study Participants:

The 39 participants whose data are included in this study were part of a cohort of 67 subjects recruited in a series of advertisements seeking adult daily smokers, distributed to subjects and staff at the University of Iowa Hospitals and Clinics. Those subjects who were potentially interested in the study were invited to complete an online survey on their smoking habits. Those subjects who reported smoking more than 10 cigarettes a day and had at least 5 pack-years of lifetime consumption in the survey were then invited to participate in the smoking cessation protocol.

In brief, as part of the study to determine the effects of smoking cessation on pulmonary inflammation, subjects were offered USD \$400 if they successfully quit smoking. Successful quitting was defined as a self-report of quitting smoking accompanied by serum cotinine values of less than 10 ng/mL at the first-, second-, and third-monthly clinical visit. Subjects were encouraged to stop “cold turkey” and to abstain from using standard smoking cessation treatments, particularly nicotine replacement therapy, to quit smoking. Subjects were offered a brief counseling session led by a research assistant at each study visit and a weekly phone call over the first month of the study. Subjects were considered treatment failures if they had serum cotinine values above 10 ng/mL at any time point or failed to attend any of the clinical visits. Only 20 of the original 67 subjects completed all procedures, reported quitting smoking, and had serum cotinine values of <10 ng/mL at all three monthly clinic visits. Nineteen others who provided DNA for this study also completed all four visits but had serum cotinine values of >10 ng/mL at one or more visits.

Laboratory Measures:

All subjects were phlebotomized at intake and during each monthly clinic visit to provide serum and DNA for the current study. Serum cotinine levels were determined by University of Iowa Diagnostic Laboratories under standard CLIA-compliant procedures. Relative change in DNA methylation at cg05575921, a well-established epigenetic indicator of smoking intensity, and three other sites used in the Epi+ Gen CHDTM test were quantified by personnel blind to subject status. Whole blood DNA from each subject at each time point (monthly meeting) was prepared as previously described (Philibert et al., 2020, Am. J. Med. Genet. Part B Neuropsychiatr. Genet., 183:51-60).

DNA methylation at cg05575921 and the three methylation sites in the Epi+ Gen CHD™ test (cg00300879, cg09552548, and cg14789911) was performed as previously described using proprietary methylation-sensitive, nested, digital primer probe sets from Behavioral Diagnostics and Cardio Diagnostics (Coralville, IA, USA) and droplet digital PCR reagents and machinery from Bio Rad (Carlsbad, CA, USA). In brief, 1 µg of DNA from each subject at study intake (baseline) and study exit (month 3) was bisulfite-converted using a Fast 96 Epitect Kit (Qiagen, Germany), with the resulting DNA being eluted using 70 µL of 10 mM Tris buffer (pH 8.0). A 3 µL aliquot of the resulting product was pre-amplified using the assay-specific pre-amplification mix, then diluted 1:1500 for the Epi+ Gen CHD™ assay, or 1:3000 for the cg05575921 assay. After dilution, a 5 µL aliquot containing approximately 10,000 amplicons—mixed with universal droplet digital PCR reagents and fluorescent primer probe sets specific to the cg00300879, cg09552548, cg14789911, and cg05575921 loci—was partitioned into droplets and then PCR amplified using a QX-200 Droplet Digital PCR system (Bio Rad) according to manufacturer's instructions. After amplification was complete, the number of droplets containing amplicons with at least one "C" allele, one "T" allele, or neither allele was then determined using a Bio-Rad QX-200 Droplet Reader, and the absolute ratio of methylated to total CpG methylation at each was determined by the proprietary Bio Rad Quantisof™ software.

Statistical Analyses:

All data were analyzed using the JMP Version 14 (SAS Institute, Cary, SC) using standard general linear model equations (Andersen et al., 2021, Epigenetics, 1-13). Group comparisons of continuous variables were compared using T-Tests. Bivariate regression was used to analyze the relationship of changes of epigenetic-indicated smoking intensity (cg05575921) to change in cardiac methylation marker (cg00300879, cg09552548, and cg14789911) status.

Example 17—Results

The clinical and demographic characteristics of the 39 subjects who completed all four clinic visits and whose data were used in this study are given in Table 8. In brief, they tended to be in the late 30s to early 40s in age, with a slight majority being male. All but two of the subjects were White.

Table 8. Demographic and Clinical Characteristics of the Subjects

	Quitters	Non-quitters
N	20	19
Age	39.8 ± 9.9	45 ± 10.2
Gender		
Male	11	11
Female	9	8
Ethnicity		
White	20	17
African American	-	1
Other	-	1
Pack-Year Consumption	22 ± 9.6	34 ± 25
Cigarettes per day	16 ± 6	19 ± 13
Intake Cotinine (ng/mL)	206 ± 93	278 ± 135
Intake Methylation (%)		
cg00300879	52.9 ± 9.1	58.3 ± 11.7
cg09552548	30.0 ± 11.7	32.0 ± 13.4
cg14789911**	89.7 ± 9.3	81.2 ± 16.4
cg05575921	57.1 ± 22.1	47.2 ± 18.3
Delta Methylation (%) over 90 days		
cg00300879	-0.7 ± 1.6	1.0 ± 4.6
cg09552548	0.1 ± 0.8	0.2 ± 1.0
cg14789911**	-1.5 ± 3.4	-1.1 ± 5.2
cg05575921	-7.6 ± 5.8	-2.1 ± 5.5

**nominally different at $p < 0.05$

5 Only 20 of the subjects who completed all four clinic visits managed to quit smoking, as evidenced by negative cotinine values at all three monthly visits. There were no significant differences in cigarette consumption or serum cotinine values between those who quit and those who did not quit ($p > 0.05$ for both).

10 We determined methylation levels at each of the three CpG sites used in the Epi+ Gen CHD™ test in each of the subjects at study entry and study exit ninety days later (see Table 8). Please note that because the set point of these methylation sites is genetically contextual and we did not determine genotype at the five sites used in this test, a direct comparison of the methylation values for those who quit versus those who did not quit is not possible. Nevertheless, in general, lower methylation values at cg00300879 and cg09552548, but

higher methylation levels at cg14789911, are associated with increased risk for incident CHD within three years (Dogan et al., 2021, Epigenomics, 13(14):1095-112).

Over the course of the study, methylation arithmetically increased at cg00300879 and cg09552548 and decreased at cg14789911 in those who quit smoking. However, using a categorical approach to classify smoking cessation, none of the changes at these three loci were significant. The subjects who were unsuccessful in quitting smoking manifested lesser degrees of change at the three loci over the 90-day period, all of which were also not significantly associated with categorical quitting status.

However, when considering these results, it is important to realize that, just as not all cases of CHD are equally severe, not all smokers consume the same number of cigarettes. Fortunately, the use of cg05575921 as a metric for change in smoking intensity permits the stable objective measurement of smoking intensity. Recently, we have shown that changes in cg05575921 methylation in response to smoking cessation are also dose-dependent. Therefore, in order to determine whether the changes in smoking intensity were related to the changes in methylation, we analyzed the relationship of the change of methylation at each of the cardiac markers to the change in smoking intensity as measured by change in cg05575921.

As expected, even though some of the subjects showed objective evidence of decreasing the rate of smoking, there were no significant relationships between the change in smoking intensity and changes in methylation at any of the three cardiac-specific loci in the 19 subjects who did not completely quit smoking. However, in the group of 20 subjects who managed to quit smoking completely, after correction for multiple comparisons, there was a significant relationship between the smoking-cessation-induced reversion of methylation at cg05575921 to an increase in methylation at cg00300879 (Adj R² 0.26, $p < 0.04$; FIG. 11C), with the changes in methylation at cg14789911 failing to achieve statistical significance (Adj R² 0.14, $p < 0.07$; FIG. 11A).

The current findings have potential for improving CVD or CHD prevention in those with multiple risk factors. In particular, we believe that developing an epigenetic method of monitoring CVD and CHD risk may improve management of those with elevated cholesterol levels and subclinical or overt type 2 diabetes. A conundrum for clinicians is the knowledge that statin-induced decreases in serum cholesterol levels are often associated with increases

in HbA1c levels. Overall, the risk/benefit ratio for the use of statins is favorable. Whether this general benefit applies to all patients equally is not known, because current algorithms cannot simultaneously consider changes in lipid and HbA1c levels. However, because each of the methylation markers maps differently to principle components of the methylation response associated with CVD and CHD, the change in overall risk as a consequence of changes in serum cholesterol and HbA1c levels can be assessed simultaneously by the integrated assessment tools described herein. For most patients, the added information obtained by retesting methylation levels is unlikely to significantly change risk management. However, for some patients, particularly those with genetic polymorphisms that alter HbA1c levels, the added information could be valuable.

In summary, the current results of this study show changes at the CpG loci predictive of incident CHD in association with the biochemically verified treatment of a risk factor for CHD.

Example 18—Addition of Protein Biomarkers to SNP-CpG Biomarkers

At least one protein biomarker is added to a method employing a biomarker scoring system with at least one SNP and/or one methylation biomarker that offers, among other things, an improvement in the ability of the biomarker scoring system to diagnose or prognose cardiovascular disease. These experiments rely on the subjects from the Framingham Heart Study Offspring Cohort described herein.

The subjects selected for this analysis consist of those that have data on at least one protein biomarker included below in Table 9 and have information on one or more types of CVD.

Table 9. Representative List of Protein Biomarkers

Adiponectin
Alpha-1-Antitrypsin (AAT)
Alpha-2-Macroglobulin (A2Macro)
Angiopoietin-1 (ANG-1)
Angiotensin-Converting 1 Enzyme (ACE)
Apolipoprotein(a) (Lp(a))

Apolipoprotein A-I (Apo A-I)
Apolipoprotein A-II (Apo A-II)
Apolipoprotein B (Apo B)
Apolipoprotein C-I (Apo C-I)
Apolipoprotein C-III (Apo C-III)
Apolipoprotein H (Apo H)
Beta-2-Microglobulin (B2M)
Brain-Derived Neurotrophic Factor (BDNF)
C-Reactive Protein (CRP)
Carbonic anhydrase 9 (CA-9)
Carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1)
CD5 Antigen-like (CD5L)
Decorin
E-Selectin
EN-RAGE
Eotaxin-1
Factor VII
Ferritin (FRTN)
Fetuin-A
Fibrinogen
Follicle-Stimulating Hormone (FSH)
Growth Hormone (GH)
Haptoglobin
Immunoglobulin A (IgA)
Immunoglobulin M (IgM)
Insulin
Intercellular Adhesion
Molecule 1 (ICAM-1)
Interferon gamma Induced Protein 10 (IP-10)
Interleukin-1 receptor antagonist (IL-1ra)

Interleukin-6 receptor (IL-6r)
Interleukin-8 (IL-8)
Interleukin-12 Subunit p40 (IL-12p40)
Interleukin-15 (IL-15)
Interleukin-18 (IL-18)
Interleukin-18-binding protein (IL-18bp)
Interleukin-23 (IL-23)
Kidney Injury Molecule-1 (KIM-1)
Leptin
Luteinizing Hormone (LH)
Macrophage Colony-Stimulating Factor 1 (M-CSF)
Macrophage Inflammatory Protein-1 beta (MIP-1 beta)
Matrix Metalloproteinase-2 (MMP-2)
Matrix Metalloproteinase-3 (MMP-3)
Matrix Metalloproteinase-7 (MMP-7)
Matrix Metalloproteinase-9 (MMP-9)
Matrix Metalloproteinase-9, total (MMP-9, total)
Midkine
Monocyte Chemotactic Protein 1 (MCP-1)
Monocyte Chemotactic Protein 2 (MCP-2)
Monocyte Chemotactic Protein 4 (MCP-4)
Monokine Induced by Gamma Interferon (MIG)
Myeloid Progenitor Inhibitory Factor 1 (MPIF-1)
Myoglobin
N-terminal prohormone of brain natriuretic peptide (NT proBNP)
Osteopontin
Pancreatic Polypeptide (PPP)
Plasminogen Activator Inhibitor 1 (PAI-1)
Platelet endothelial cell adhesion molecule (PECAM-1)
Prolactin (PRL)

Pulmonary and Activation-Regulated Chemokine (PARC)
Pulmonary surfactant-associated protein D (SP-D)
Resistin
Serotransferrin (Transferrin)
Serum Amyloid P-Component (SAP)
Stem Cell Factor (SCF)
T-Cell-Specific Protein RANTES (RANTES)
Tamm-Horsfall Urinary Glycoprotein (THP)
Thrombomodulin (TM)
Thrombospondin-1
Thyroid-Stimulating Hormone (TSH)
Thyroxine-Binding Globulin (TBG)
Tissue Inhibitor of Metalloproteinases 1 (TIMP-1)
Transthyretin (TTR)
Troponin
Tumor necrosis factor receptor 2 (TNFR2)
Vascular Cell Adhesion Molecule-1 (VCAM-1)
Vascular Endothelial Growth (VEGF)
Vitamin D-Binding Protein (VDBP)
Vitamin K-Dependent Protein S (VKDPS)
Vitronectin von Willebrand Factor (vWF)

Machine learning methods such as ones described herein are used to identify at least one protein biomarker that, when added to the SNP and/or methylation biomarker scoring system, improve the predictive capability of the biomarker scoring system. To achieve this, subjects are split into training and test sets. The training set is used to identify the protein biomarker(s) and to quantify performance. The test set is used to verify the improvement in performance. The AUC, sensitivity, specificity and accuracy are quantified.

It is to be understood that, while the methods and compositions of matter have been described herein in conjunction with a number of different aspects, the foregoing description of the various aspects is intended to illustrate and not limit the scope of the methods and compositions of matter. Other aspects, advantages, and modifications are within the scope of
5 the following claims.

Disclosed are methods and compositions that can be used for, can be used in conjunction with, can be used in preparation for, or are products of the disclosed methods and compositions. These and other materials are disclosed herein, and it is understood that combinations, subsets, interactions, groups, etc. of these methods and compositions are
10 disclosed. That is, while specific reference to each various individual and collective combinations and permutations of these compositions and methods may not be explicitly disclosed, each is specifically contemplated and described herein. For example, if a particular composition of matter or a particular method is disclosed and discussed and a number of compositions or methods are discussed, each and every combination and
15 permutation of the compositions and the methods are specifically contemplated unless specifically indicated to the contrary. Likewise, any subset or combination of these is also specifically contemplated and disclosed.

APPENDIX A

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg10063260	0.842758392	XRC3	14	104181875	cg07412317	0.585431507		16	13983454
cg20904336	0.766196463	DEGS2	14	100618491	cg10311318	0.583399675	ATP8A2	13	26043578
cg18172516	0.757281748	RBM5L	2	161214856	cg16102016	0.585347915	C1orf44	17	8126456
cg17666198	0.757168153		20	52821599	cg16824643	0.585341091		18	76462312
cg25614253	0.7526974	BA1	8	143561205	cg20597767	0.585136652	RPRM1	17	45056325
cg02940070	0.748064129	PACSIN2	22	43343608	cg05255728	0.584947515	ZNF141	4	332107
cg01026744	0.743458629	NAP1L5	4	89619053	cg08019058	0.5848665		2	105735756
cg13736263	0.742711753	PDXD7	10	102778909	cg08008199	0.584826132	ZNF460	19	57791654
cg02550738	0.741934861		5	95477736	cg25859998	0.584809529		6	10424320
cg08950364	0.737449644		1	108039344	cg27613473	0.584774348		6	169240458
cg20147224	0.737266138	DDB2	11	47236405	cg08515989	0.584738186	ASPHD1	16	29912904
cg23936477	0.73559024	C18orf62	18	73139742	cg11850961	0.584715434	RCC2	1	17766227
cg03101345	0.73401513	CTU1	19	51604984	cg15923936	0.584684459		6	91318412
cg19225953	0.733172678		19	13276344	cg05482973	0.584646816	C5orf52	5	157098423
cg08210706	0.73272656	SERPINA5	14	95046686	cg24487639	0.584480974		1	149224366
cg05100739	0.732101306	RAB37	17	72733163	cg21726250	0.584364498	DNAH17	17	76472384
cg16344810	0.731884058	METAP2	12	95867376	cg05337711	0.584260731	RS11D1	16	11946064
cg27361401	0.73167565				cg20423602	0.584259817	ADARB2	10	1560592
cg04850148	0.731604753	CC14L2	17	34539744	cg10240778	0.584188704	ATAD3B	1	1407180
cg11318251	0.73115942	BAALC	8	104152588	cg10833093	0.584159384	STX2	12	131303194
cg03964373	0.730410106	CHML	1	241800323	cg02610327	0.584022476		4	44169592
cg26217827	0.729432838	ITGA11	15	68594860	cg20609752	0.583965654	EVISL	19	7894982
cg02628858	0.728091882	MTUS2	13	29910801	cg08314021	0.583937485	TSP0	22	43547679
cg06804873	0.726904775	MPSD2A	1	40419732	cg21978579	0.583937198		18	6920624
cg10712578	0.722598998		8	96619989	cg09102409	0.583869813	NDRG1	8	134308734
cg09552548	0.722085216		2	232538286	cg12810212	0.583642466	C2orf52	2	232379626
cg18052106	0.721618357		6	93354363	cg00204645	0.583539398	MMP15	16	58060579
cg06711418	0.720953013	MT2A	16	56643342	cg27181169	0.583508974	RSPRV1	16	57220014
cg13891675	0.720314943	VWA1	1	1374310	cg07054687	0.583386126		2	1788153
cg07133434	0.7184445		3	128395694	cg08351203	0.583374018	ZNF83	19	53141793
cg11170318	0.717874396		21	45246363	cg12279734	0.583376394	MORF4	4	174538350
cg04703912	0.717278487		20	62026763	cg17842821	0.583204448		5	1834682
cg00306951	0.717230733		2	221831539	cg06192336	0.583142028		17	7961272
cg14655569	0.716824882	BICD2	9	95473718	cg09593291	0.582871983		4	179734220
cg08051013	0.715966903		7	47092955	cg15683497	0.582692963	SCNN1D	1	1226929
cg14218851	0.715684488		14	103018726	cg11141497	0.582517383	TAF10	11	6633666
cg01203614	0.714925768		4	3671067	cg05905030	0.582416752		10	130540874
cg09584650	0.714176374	MYOM2	8	2002012	cg02032982	0.582348875	ESRRB	14	76905532
cg17964505	0.713888889		17	36609863	cg17016777	0.582307577		5	4207689
cg13957457	0.713825059	MUC2	11	1103876	cg11493661	0.582268387	TNX8	6	32016239

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg03226554	0.713164251		1	28573817	cg23988024	0.582178942	HCN2	19	594272
cg16148346	0.713164251		11	58826475	cg27090784	0.582138295	HTR4	5	147862681
cg16248187	0.712599054	FMNL2	2	153191727	cg19049077	0.582077093	C10orf137	10	127407606
cg15600437	0.711730915	MFAP2	1	17309539	cg02475834	0.582019135	PRPF18	10	13628849
cg25243082	0.711679139		4	40267141	cg26213155	0.581965284	SLC6A5	11	20621701
cg01247619	0.711465248	PIFC1	8	145049472	cg05524354	0.581607466	PTPRE	10	129797760
cg20162822	0.711111111	SERPINF2	17	1658265	cg01254303	0.581532171	SRRM4	12	119592035
cg15146004	0.710507246	DCP1A	3	53380803	cg15127270	0.581465743	ASAP2	2	9353077
cg27639199	0.7091006	TMC3	15	31666528	cg10592494	0.581351686	BAIAP211	7	97935764
cg04843821	0.708632026	C10orf68	10	33016632	cg09339394	0.581168558	EAPP	14	35008661
cg02999711	0.707608696	MIR135B	1	205417436	cg00160002	0.581156672	AGAP1	2	237032579
cg12304113	0.707487923	COL22A1	8	139976188	cg06889535	0.581009541	KIRREL3	11	126455712
cg08104981	0.707434515	GPR39	2	133343149	cg21144493	0.580927365	CAMTA1	1	7130061
cg01684629	0.706884058		4	121847309	cg18190417	0.580817038		12	130683439
cg05957781	0.706708274		22	48727673	cg23449010	0.580794178	CHRNA4	20	61978229
cg07796016	0.706346943	LCF1C	1	152779584	cg22507154	0.58073473		1	91185233
cg13905298	0.70605492		5	141485167	cg03830006	0.580696354	ENPP7	17	77707066
cg14928378	0.705937365		7	75372874	cg13863764	0.580666153	PTCHD2	1	11561908
cg15944457	0.705676329	WRN	8	30891341	cg09413529	0.580570074	TBX3	12	115112225
cg14326885	0.705605674	FLJ11235	5	111755158	cg13050504	0.580549742	RSPH1	21	43913111
cg00441027	0.705465161	GPC1	2	241401880	cg04206351	0.580512881	LOC440040	11	49582375
cg07525313	0.705338586	FRK	6	116262856	cg17531849	0.58046648	HRNB3	17	77311766
cg08028341	0.704909151	OBP2A	9	138437868	cg22806527	0.58014712	HRH2	5	175087093
cg03647233	0.703623188	DSCAML1	11	117387430	cg14029686	0.580043912		4	2765767
cg04247218	0.703140097	GTF3C1	16	27561328	cg01905313	0.58000189	PHLDA2	11	2951613
cg00640120	0.702901035	CYB561	17	61518241	cg11124135	0.579992004	LSM1	8	38032157
cg12471283	0.702756555		5	107120132	cg15770754	0.579989403	DPP6	7	154542235
cg23564471	0.702129079	NFASC	1	204951209	cg08847173	0.579895925	NKAIN1	1	31654792
cg11608150	0.701625532		5	135415948	cg10586042	0.57951724	SKP2	5	36152206
cg04821933	0.701383924	POMP	13	29232799	cg26668744	0.579485293		7	159127977
cg05256504	0.701328502		7	121947072	cg05348768	0.579431409	CYB5R1	1	202936122
cg22111694	0.699904366		2	91758369	cg20166714	0.579703181	CCDC42B	12	113592444
cg00837457	0.699275362	BICD1	12	32287236	cg13866263	0.5791949		6	169092843
cg08996597	0.698920438		6	130747460	cg17557366	0.579153273	SLC25A18	22	18073134
cg19344315	0.698671498	ATXN1	6	16669649	cg07104919	0.578789882	C11orf49	11	46958073
cg17516160	0.697935697	MRE11A	11	94227125	cg10983013	0.578715847		6	170816044
cg06051619	0.697705705	DIP2C	10	593275	cg11710851	0.578521031		12	130765858
cg00439630	0.697385816		15	25915520	cg23876832	0.578315464		11	62092739
cg27631593	0.697222222	SPRR2E	1	153065613	cg25620901	0.578263663	HEATR2	7	808592
cg1699880	0.6971026	TSSC1	2	3315680	cg00845883	0.578224678	ANO1	11	69923693

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg06417644	0.696859903		16	28081331	cg07972135	0.578111691	GRI4A	11	105481322
cg13903282	0.696618357		9	128819969	cg26090359	0.577880442	MUC5B	11	1268962
cg09698846	0.696376812	ZHX2	8	123964152	cg02226252	0.577836658		8	7079761
cg04099818	0.69589372		6	37318759	cg07201996	0.577612685	CDH4	20	60348277
cg10243183	0.69589372		1	88026790	cg09526886	0.577498497	SLC25A3	12	98987659
cg11314779	0.695849018		15	72567956	cg26537248	0.577475012	MEITL7B	12	56075465
cg26371327	0.695652174	EPHA2	1	16473143	cg21183536	0.577382399	CAP2	6	17393563
cg27247782	0.694500236		7	158767135	cg25753631	0.577211436		6	25732923
cg08122179	0.694202899	COMMD9	11	36311027	cg23226510	0.577158739	SRK2	19	56042469
cg27005246	0.694082126	STX4	16	31044731	cg23105471	0.577099365	CTDSP2	12	58241215
cg16928657	0.693719807	PRPF19	11	60674297	cg16483840	0.57709413	PSMB8	6	32812953
cg05379947	0.693357488	RPL10A	6	35436047	cg10364645	0.577069723		3	194650197
cg05720355	0.693115942	IQUB	7	123175289	cg14299483	0.576973545		14	101967780
cg19079513	0.692966676		17	14936230	cg24661236	0.576950283	PI16	6	36930062
cg03054255	0.692591017	BRF1	14	105685132	cg07608965	0.576862139	EPS8L1	19	5595022
cg07737560	0.691908213	ANO4	12	101470827	cg10074506	0.576713551	GSTO2	10	106035131
cg06792428	0.691897972	CDSN	6	31087240	cg13439030	0.576607578		12	132973307
cg13655986	0.691835313	DOCK1	10	129197384	cg09550909	0.576519443	TRANK1	3	36949957
cg13655169	0.691761435		14	73929570	cg11139684	0.576479231	PAM177A1	14	35515356
cg06566034	0.691637352		18	47825363	cg00746130	0.576306385	BAT5	6	31671271
cg03020684	0.69126047	THSD4	15	71532066	cg01855367	0.576252443	JAG2	14	105634387
cg22142329	0.69121844	COL6A3	2	238232841	cg05574272	0.576194641	RXFP3	5	33938213
cg22466012	0.69112525	PLD1	3	171466200	cg13388615	0.576127562		2	242878809
cg14023357	0.690942029		19	41637224	cg10445421	0.5761272		1	39284267
cg08564027	0.690760196		20	61660810	cg19169023	0.576115355	TYRO3	15	41853346
cg17474004	0.689940898	C2orf34	2	44908093	cg26869506	0.576049848		10	133775762
cg24976563	0.689884456	DCAF11	14	24587638	cg06214334	0.576045494		13	19957032
cg00433159	0.689797163	LOC652276	16	2653839	cg09895325	0.575990079	SRRM1	1	24968182
cg11345172	0.689643026	CH	12	120315133	cg17210004	0.575975097		17	63512677
cg24524099	0.689337098	PTPRN2	7	157643007	cg11403823	0.57592416		7	6856314
cg00213714	0.689084634	ASTN2	9	119449370	cg17804112	0.575839227	NCR1	19	55417496
cg25569462	0.688874196	TRIM12	4	189076860	cg11728809	0.575821611		12	34554731
cg09787311	0.68852657		17	80329175	cg15739543	0.575792068	POIS	5	6713426
cg23327070	0.688405797	IGMN	14	93214896	cg05475109	0.575675814		8	49427684
cg02190985	0.688200473		10	12582191	cg03333125	0.575645917	CCDC17	1	46088541
cg21209091	0.687355083	DKK4	8	42234975	cg24398793	0.575603865	MEOX2	7	15651770
cg18057559	0.687198068	B8X	3	107243151	cg12026479	0.575493542	ACOX3	4	8373553
cg03634833	0.686594203	ADAP1	7	9655534	cg21103385	0.575476659		7	155623788
cg20685672	0.686343564		2	181987552	cg12606933	0.575464888	SOX30	5	157078573
cg03812172	0.686234109	GCK	7	44184403	cg06706876	0.57543709	PULP	16	57317679

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg02567750	0.685941371		1	247509605	cg02631082	0.575187237	SHANK1	19	51192440
cg19909865	0.685748792	PCDH10	4	134074421	cg00546932	0.575125649		16	1947055
cg14598846	0.685628019	PLEC1	8	14508909	cg27624162	0.574833008		2	111876008
cg11502198	0.685598023	ABT1	6	26597334	cg24805360	0.574703048	LHFPL2	5	77930038
cg09086151	0.685500446	HLA-DRB1	6	32550067	cg02017486	0.57457294		17	80004854
cg13377530	0.685165485		16	88709526	cg00944166	0.57456596	SIC06A1	5	101834675
cg23630131	0.685144928		7	65973040	cg21081704	0.574522711	BRD2	6	32940306
cg26049187	0.68442029	CFHR5	1	196976319	cg05275012	0.57451955		2	96055346
cg22672060	0.684299517		4	156464550	cg16602316	0.574518834		14	93698053
cg24871132	0.684299517	PFN2	3	149688846	cg23526973	0.574485052		2	237086991
cg13147090	0.684057971	IQSEC3	12	247963	cg02611105	0.574410177	PPP1R12C	19	55628611
cg20371573	0.683937198	MVO16	13	109780516	cg09861917	0.574096283	FOXA2	20	22566877
cg09922103	0.683333333		17	80668678	cg13644262	0.574060854	PPP2R2C	4	6449564
cg22138998	0.683033511	SFRS9	12	120903935	cg01276201	0.574055174		10	134613136
cg05867245	0.68286548	ZBTB46	20	62402415	cg02844341	0.574018585	STT0T4	7	116593761
cg03369465	0.68236715	DAXX	6	33291367	cg23467747	0.573910233	AGAP1	2	237022740
cg19051015	0.682125604		17	46697414	cg08747676	0.573904438		16	1148056
cg04598251	0.682125604	NOC4L	12	132628980	cg11727653	0.573810079		1	16339496
cg03651054	0.681753584		13	50194643	cg02852327	0.573575949	SPTBN4	7	121950975
cg26490743	0.681532861		16	63650866	cg07515422	0.57355516		19	40993713
cg08967584	0.681493552		11	3535099	cg24877675	0.573520122	JAKMIP3	10	133995387
cg15029248	0.681400966		1	2379937	cg17820828	0.573465094	KCNQ1	11	2813298
cg08206881	0.681048682		2	12317357	cg22763897	0.5732557		20	44065402
cg01717830	0.680868558	DDX42	17	61851435	cg03784083	0.573204121	OPCML	11	132513895
cg04702989	0.679951691		16	31053962	cg04289036	0.572950276	C1orf210	1	43751270
cg22170801	0.679589372	NPEPL1	20	57285947	cg07722774	0.572711124		7	155915607
cg04908106	0.679264775	LMF1	16	1010488	cg03552039	0.572694661	FGFR2	10	123243495
cg19322825	0.679227053		7	2293492	cg03606898	0.572597528		19	2926882
cg27176264	0.678959115		17	81060149	cg06224587	0.572464605		4	6540449
cg04897621	0.678864734	NR4A3	9	102585321	cg17899626	0.572443425	PRPF8	17	1588142
cg00856375	0.678864734	PSMD14	2	162164676	cg10705488	0.572405041	MIB2	1	1550648
cg24137123	0.678859574		5	178404926	cg18972998	0.572333665	TTC22	1	55266296
cg18557837	0.67875652		3	39847605	cg00403616	0.572309473		8	591688
cg01844866	0.678623188	JPH3	16	87669832	cg18835078	0.572245077	PCDHGA1	5	140712209
cg17770910	0.678172226	LYNX1	8	143851427	cg22679890	0.572238418		1	244319093
cg09142166	0.677898551	PNPLA5	22	44276013	cg07940597	0.572061818	ZNF75A	16	3355079
cg11523221	0.677847754	BACH2	6	90905332	cg03879460	0.571936797		12	562272
cg15803122	0.677657005	9-Sep	17	75277207	cg01050810	0.571892577		6	168053859
cg03758011	0.677526241				cg07232577	0.571757892		6	170402031
cg12740087	0.676932367	NVL	1	224445738	cg01565320	0.571694278	SHISA3	4	42399851

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg16199098	0.676811594	TXN2	22	36878119	cg04204526	0.571675099	FLJ42875	1	2980937
cg25641745	0.676690821		13	78569349	cg03022891	0.571613304	TNNT3	11	1947791
cg07260532	0.676515366	SLC30A9	4	41992067	cg25120325	0.571581292	PNLIPR2	10	118380370
cg18376497	0.676449775	INPP4B	4	143488622	cg24168538	0.571551831		4	35527016
cg18339359	0.676312575	SLC25A37	8	23423757	cg00355678	0.571524807	PCSK4	19	1483413
cg09320367	0.676207729		1	40177469	cg08448455	0.5715247		8	49231786
cg16527629	0.676125677	SPDEF	6	34524698	cg03877418	0.571461259	NEU4	2	242757444
cg03694875	0.676086525	LOC10013436	16	435915	cg24998357	0.571443726		7	3280646
cg09543427	0.675966184	FBN2	3	13590350	cg20503657	0.57138678		10	835505
cg22264409	0.675845411	PROK2	3	71828408	cg07325827	0.571282041		17	21226936
cg10113107	0.675603865	RRM1	11	4118978	cg18835942	0.571233681	SOX12	20	306049
cg01992890	0.675483092	SHQ1	3	72897260	cg10932427	0.571111137	LNPEP	5	96271079
cg02854229	0.675362319		14	101962994	cg21895387	0.571054109	LASS3	15	101084565
cg06171406	0.675120773	ADCY9	16	4050400	cg03711791	0.570805063	KCNK1	20	49626347
cg11248869	0.674646336	FOX3	1	42801011	cg23057009	0.570772392		10	134844217
cg10719247	0.674516908	SLMAP	3	57745962	cg05385718	0.570705906	D2HGDH	2	242693323
cg14179288	0.674398867	TLE4	9	82286351	cg01516851	0.570701765		6	28945338
cg211171954	0.674396135	ROR2	9	94495731	cg06176124	0.57067619	CBLB	3	105588009
cg08120796	0.674033816	RIPK2	8	90770219	cg01352882	0.570668855	STUB1	16	730130
cg10993470	0.673854604	RP1	8	55533939	cg04243827	0.570650293		5	3103718
cg16112880	0.673819669	TMEM9	1	201123745	cg24440658	0.570562059		10	82295723
cg15823954	0.673792271	HMH1	5	143191226	cg22451782	0.570555136	KLHL29	2	23851985
cg16585380	0.673550725	C1orf67	16	31711872	cg22796353	0.57035622		7	156400281
cg11723923	0.673503577		13	112820997	cg24039603	0.57023222		14	81929044
cg21151061	0.673472813	GPD1	12	50498016	cg02083210	0.57017655	ODC1	2	10589054
cg13284574	0.67293617	ESPN	1	6519923	cg21251970	0.570149196	EXD3	9	140311919
cg03329597	0.672760793	MYH15	3	108125523	cg07476339	0.570120723		7	154706964
cg21543270	0.672567368	PACS2	14	105840286	cg18954504	0.570101461		17	34819740
cg23561752	0.672463768	ZNF828	13	115092495	cg14237930	0.570002041		9	134615126
cg08570639	0.672342995		6	113371757	cg14215586	0.569927536	HOXC9	12	54394545
cg14621900	0.672109141	KCNK7	11	65363274	cg06574716	0.569921742	ZNF283	19	44331239
cg15126201	0.672101449	TSSC1	2	3325654	cg25922808	0.569787929	MLF2	12	6862832
cg13308063	0.671618357	TNX8	6	32042344	cg02396888	0.569653384	MATN1	1	31191746
cg03718353	0.671497585	PTP4A3	8	142437137	cg24173246	0.569609082		19	14897226
cg05522042	0.671400189	KIAA0513	16	85124401	cg19664267	0.569562903	B3GALT4	6	33245163
cg21328770	0.671213239	YEATS2	3	183415364	cg04222701	0.569560524	IU8	14	23451965
cg07545317	0.671135314		12	115174598	cg18075755	0.569391974	C1QTNF7	4	15429967
cg00516051	0.671135266	ZNF1	17	4047033	cg02871313	0.569380705	SPACA1	6	88757302
cg13145293	0.671135266	DET1	15	89090195	cg14240768	0.569256919	RPS7	2	3622594
cg06915202	0.67089372	BAIAP2L1	7	98029285	cg09357276	0.56890539	ZFP42	4	188917856

Appendix A.xlsx

ILMNID	mean_cv_auc	GENE	CHR	MAPINFO	ILMNID	mean_cv_auc	GENE	CHR	MAPINFO
cg09386376	0.670696927	DRD4	11	638939	cg17099656	0.568870315	CERK	22	47135171
cg02887598	0.670515855	BIN1	2	127841945	cg23671719	0.568864273	ZNF555	19	2841288
cg16562603	0.670410628	LSM5	7	32529678	cg16739118	0.568824103	BHLHE41	12	26275661
cg06364315	0.670348507	NETO2	16	47178078	cg03119731	0.568791605	CACNG6	19	54494658
cg10727661	0.670289855	PITRM1	10	3185872	cg25418852	0.56868308	LOC643719	19	35068221
cg17217195	0.670048309	CCDC27	1	3673328	cg23087661	0.568471841		1	2940825
cg15331945	0.670048309		12	34756042	cg25578967	0.568468631	OSBPL3	7	25018440
cg19350024	0.669927536		10	63869379	cg21450137	0.568292064	C2orf79	2	25016479
cg14976442	0.669806763	SH3BP5L	1	249120571	cg04896115	0.568276012		7	47674905
cg17668163	0.669806763	DMTF1	7	86781457	cg03293206	0.568251063	ANKRD26P1	16	46603015
cg11248999	0.66968599	EPHB2	1	23067453	cg04444959	0.56822515	MGC23284	16	88744842
cg13670448	0.669444444	DNAJC9	10	75007078	cg02613370	0.568131664		10	124578366
cg15986529	0.669122931				cg22205276	0.568097297	TNFRSF11A	18	60052372
cg02873391	0.668719807		11	130470658	cg01299774	0.568092794		5	565934
cg26349672	0.668719807	KIAA0754	1	39875051	cg22786486	0.567998855	MIR1306	22	20073528
cg00356500	0.668591489	TNXB	6	32028152	cg26280578	0.567932567	ARFGAP1	20	61916036
cg19642421	0.668115942	GPR133	12	131438737	cg07395074	0.567915997		12	110283406
cg17954852	0.667753623	CFB	6	31919854	cg07101909	0.567908037		16	3202077
cg21678795	0.667753623	CSRP2	12	77272832	cg00190319	0.567892479	SCAP	3	47462780
cg07495405	0.667740154	FHOD3	18	34325827	cg23187802	0.567705334	ZCCHC24	10	81173663
cg10308629	0.667545971	BPGM	7	134354803	cg14366110	0.567589343	FIBCD1	9	133779382
cg11756073	0.667391304		10	126445291	cg02566224	0.567574115		16	89149764
cg04316537	0.667334008		2	10259889	cg13534536	0.567470339	AFAP1	4	7938304
cg21554895	0.667314894		13	20703415	cg24578428	0.567457103	CPZ	4	8621226
cg06269753	0.667028986	MSC	8	72755871	cg21537230	0.567393621	CLIC6	21	36043248
cg06979386	0.666791177		16	78027119	cg20435485	0.567365831	FAM116B	22	50753068
cg19389973	0.666750365	AOAH	7	36692197	cg17915676	0.567341186	VWA3B	2	98928898
cg07478240	0.666610585		16	87174203	cg15998773	0.567228644	RSC1A1	1	15987116
cg16472569	0.666183575		4	175357025	cg10556064	0.567145365	SMIPD3	16	68481489
cg02618555	0.66612104		8	57803480	cg10945539	0.567009693	COLEC11	2	3642447
cg02888513	0.666062802		15	39205040	cg04843568	0.566972239		5	87389026
cg15286618	0.666062802	AIP	11	67250164	cg15964187	0.56695143		12	114135623
cg12530665	0.665945626	POMP	13	29236637	cg00357671	0.566809987	NCOR2	12	124809516
cg11205696	0.665781724		8	144870535	cg17759252	0.566689246	CRTC1	19	18888081
cg19223824	0.665773879	SLC12A5	20	44682963	cg00543485	0.566667753	JUN	1	59249218
cg18656873	0.66557971	TMPRSS3	21	43809519	cg12064504	0.566560661	ZNF311	6	28972891
cg03330747	0.665462884	CDK2	12	56361238	cg05526731	0.566492116	RPH3AL	17	185263
cg07088771	0.665371158	TNXB	6	32057846	cg14301191	0.566394054	GTF2H5	6	158616065
cg24713529	0.665338164	TYMP	22	50965126	cg10724771	0.566363407	SCD5	4	83720102
cg16490805	0.665259842		16	10281150	cg09791440	0.56622636	PAX7	1	19048903

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg00999904	0.665201769	ALIC	2	3704751	cg25736002	0.566179111	REST	4	57773436
cg18873965	0.665177164	GOLGA3	12	133378609	cg25279059	0.566125334	TMEM11	17	21118261
cg04389533	0.664975845		11	34448185	cg10401356	0.565942029	KCNK9	8	140712424
cg06081609	0.664975845	SASH1	6	148663571	cg03559973	0.565925076	ZNF226	19	44669232
cg01627741	0.664855072	ZNF226	19	44568989	cg19046167	0.565866229	B3GNTL1	17	80928561
cg02068923	0.664613527		14	61547469	cg23749005	0.565798977	PTPRN2	7	158362339
cg25150572	0.664573832	PHACTR4	1	28763330	cg04936446	0.565575408	RASSF8	12	26111348
cg18501945	0.664497754	SIC20A2	8	42294924	cg04299200	0.565467261		5	534337
cg09373511	0.66430922	DHX32	10	127560294	cg22230229	0.565421765	BAHCC1	17	79372254
cg19415746	0.664300922	NRAP	10	11538944	cg09977703	0.565406718	PPP5C	19	46854563
cg09588254	0.664251208	PRR3	6	30523956	cg21724796	0.565379233	AURKB	17	8113904
cg27376941	0.664139378		22	36929349	cg18192294	0.565360811	RNLS	10	90343204
cg15950936	0.664122459		7	6914289	cg04140663	0.56531073	CPEB1	15	83315922
cg19142181	0.663565485	SIC17A9	20	61591066	cg03410223	0.565270156	C6orf89	6	36853544
cg24589459	0.663519149	BOLL	2	198651347	cg00220952	0.565254113	SCARNA16	17	75084623
cg00411300	0.663479	E2F8	11	19262578	cg09673582	0.565217376	PRPF8	17	1588270
cg23518214	0.663405797	B8X	3	107240645	cg04820853	0.565193586	HAPLN3	15	89438208
cg00156497	0.663396717	EYA4	6	133663589	cg07140289	0.565173542		8	142299684
cg10647704	0.663164251	AHNK	11	62275604	cg05253110	0.565084008		7	141130687
cg23987846	0.663092306	ERICH1	8	651301	cg12055259	0.565028576		14	104603238
cg25288140	0.662681159	BRCA1	17	41278341	cg15164702	0.565021258		10	31891625
cg25674027	0.662677278		12	103325781	cg14573833	0.564974133	LHX3	9	139096200
cg14534848	0.662660993		9	139892531	cg04161784	0.564960452	HBP1P	1	110951209
cg19705131	0.662439614	GALR1	18	74961163	cg21385821	0.56491773	CA10	17	50237844
cg00108108	0.662439614	CBFA2T3	16	88996870	cg05721645	0.564896929	CNST	1	246729578
cg02616906	0.662318841	HYDIN	16	71265770	cg05093254	0.564813617	ABCA3	16	2369796
cg26644853	0.662317258	CX3CL1	16	57406257	cg14329783	0.564776122		7	75779857
cg02137058	0.662032137		3	170451961	cg15342134	0.56473103	MVK	12	110025893
cg18693985	0.661960078	CPEB4	5	173331052	cg18982625	0.564702614	ITIL1	22	43485660
cg25215049	0.661956522	PCGF3	4	721972	cg23479905	0.564687798	MIR191	3	49059487
cg01261464	0.661851064	UBR7	14	93673175	cg00733782	0.564672569	TRIM24	7	138144684
cg10481202	0.661352657	TUBA1A	12	49581272	cg03515060	0.56465837	POFUT2	21	46705827
cg20599420	0.661244444		10	73777304	cg23370710	0.564575889	TNS3	7	47354947
cg12213037	0.661231884	SIC35E2	1	1666808	cg17296166	0.564568234	PNMAL1	19	46974567
cg242828123	0.660990338	PRMT1	19	50180204	cg11715943	0.564487441	HLA-DPA2	6	33091841
cg18171855	0.660869565		10	2543474	cg00599702	0.564464187		1	30584352
cg20253892	0.660869565	HDAC4	2	240005611	cg14108978	0.564345019	ARHGEF17	11	73021145
cg12868352	0.660507246		5	13578839	cg25450033	0.564299482		19	14444658
cg15347004	0.660416577	DHX35	20	37590907	cg14533609	0.56413392		9	38489046
cg25706012	0.6602657	ELFN1	7	1782755	cg12670061	0.564117259	HSD11B2	16	67465042

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg19590511	0.660207092	ABR	17	957402	cg26224499	0.564078266	TIMM44	19	8008578
cg27244734	0.660024155	ZNF33A	10	38299349	cg16705205	0.563992963	BRUNOL4	18	35067529
cg12534176	0.659774941	APPBP2	17	58574569	cg19593314	0.563944899	C5orf56	5	131746670
cg04158589	0.659772794		16	32937040	cg20167580	0.563749119		17	25538762
cg27180412	0.659661836		1	143277743	cg07791427	0.563637375	H0XC8	12	54402704
cg07247805	0.659661836	MXD4	4	2264197	cg03775163	0.563592966		13	114074693
cg21817284	0.659541063	MC3R	20	54823279	cg00941836	0.563570873		22	27834549
cg04415270	0.659348463	RF8	2	102091202	cg27023360	0.563453828		16	51393415
cg15264162	0.659209796	DPP6	7	154035993	cg12575883	0.563414366		7	1003710
cg25938010	0.659178744	TRIM26	6	30160080	cg13504055	0.563399335		2	102234418
cg06190807	0.659057971	LITD1	1	62660188	cg24284539	0.56331078	CCDC3	10	12999599
cg05253327	0.658937198		2	62420915	cg05129568	0.563277738	FLOT1	6	30708009
cg08102564	0.658670258	SLC47A2	17	19620263	cg23183906	0.563178614	TPM4	19	16187089
cg01267709	0.658546572	SLC7A8	14	23623756	cg25066665	0.563115749	RPRD2	1	150335507
cg16656196	0.658192102	KCNH4	17	40321577	cg10227187	0.563012715	NOXO1	16	2030509
cg09019154	0.657925474		8	19616280	cg03918377	0.562866703	PCGF3	4	713674
cg11086312	0.657608696		14	92720273	cg19252369	0.562858521	LAT52	13	21580144
cg23505252	0.657246377	C15orf48	15	45724580	cg22471075	0.562751782		22	49765229
cg24955955	0.657121587	AHRH	5	415729	cg07452164	0.562725597	TBL2	7	72993570
cg17131445	0.656763285		8	47039958	cg17800497	0.56261117	SEZ6	17	27331976
cg23384185	0.656642512	C18orf56	18	657227	cg15812020	0.56253251	LOC646405	13	25506235
cg14504512	0.656642512	IAZF1	7	27875760	cg24135583	0.56245938	PTPRN2	7	157353191
cg22080061	0.656627125	KIAA0754	1	39874802	cg00010954	0.562450869		16	54620035
cg16210973	0.656280193		5	147760896	cg07416187	0.562323197		2	193373742
cg09029571	0.656220804	YBX1	1	43147119	cg00155593	0.562319666	ANKA2	15	60690279
cg07544653	0.656038647	CY85A	18	71959407	cg08572999	0.562313336	ZNF836	19	52675203
cg13232075	0.655913102		1	204556835	cg14910495	0.56229274	CDK5R2	2	219824329
cg09563619	0.655676329	TBCD	17	80780281	cg26464185	0.562182956	NUBP2	16	1838079
cg15927357	0.655676329	C18orf32	18	47013900	cg22249752	0.561967989		12	133186738
cg17128068	0.655676329	GALNT9	12	132827523	cg08036764	0.561872181	C22orf45	22	24891220
cg02368869	0.655609456	C1orf186	1	206243998	cg23324289	0.561809671	FAM27L	17	21825321
cg13471735	0.655600473	BNC2	9	16442831	cg07280731	0.561763195	PRSS21	16	2867773
cg07699307	0.655434783	ASCC3	6	10111561	cg23593528	0.561672393	C1orf53	1	197871600
cg08481112	0.65531401	GNP7	19	2544100	cg22006640	0.561627498		4	111532035
cg18108844	0.65531401		1	2716235	cg16003378	0.561453548		14	105036533
cg06560422	0.654951691	KRTAP19-5	21	31875651	cg25132257	0.561389746	NCAM1	11	112916139
cg17616547	0.654543735		6	26688923	cg00675229	0.561272182		4	18320910
cg09669835	0.654227053	MTERF	7	91510500	cg23057721	0.561234869		17	41791209
cg15627464	0.653985507	NCOR2	12	124985655	cg17022362	0.561107304	KCNQ2	20	62084833
cg05674437	0.653985507	PIG2	3	196694153	cg07591395	0.561089442	RBM8A	1	145507220

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg18144065	0.653864734	CAPZA1	1	113161296	cg13686104	0.560972364	ADIPOR1	1	202927632
cg22708112	0.653864734	GCOM1	15	57883392	cg07221298	0.560971886		16	1170843
cg25687358	0.653825115	DUOX1	15	45409942	cg12162424	0.560950048	ABCC13	21	15646312
cg04558424	0.653623188	C16orf74	16	85743497	cg24769849	0.560874877	JAKMIP1	4	6072170
cg23972860	0.653140097	CAMTA1	1	7395024	cg02776867	0.560868523	DOT1L	19	2210488
cg23287902	0.652997192	LMAN2L	2	97405364	cg24550683	0.560731418	TIMP2	17	76905821
cg02010478	0.652898551	TMEM151A	11	66059798	cg12576557	0.560708641	TRPM5	11	2441080
cg18550846	0.652898551	TMEM189	20	48767148	cg09848508	0.560684407	BAIAP3	16	1393584
cg12472351	0.652777778	COL11A2	6	33158526	cg11213690	0.560655621		7	149112402
cg13600364	0.652346141		6	362095	cg14899357	0.560625165		2	85153373
cg11227987	0.652190079	PTPRN2	7	158250469	cg08911368	0.560568729		8	11471085
cg00135841	0.652049426		22	18737348	cg01875838	0.560557287	TMED1	19	10947446
cg19707653	0.651960229	KIAA1671	22	25571929	cg05799507	0.560550068	MTDP	16	56677223
cg19592003	0.651823725	DEFB114	6	49928115	cg20456258	0.560420219	DGKQ	4	962124
cg08969950	0.651749882	NR2F1	5	92929679	cg22898362	0.560396541	KLR9	19	51506358
cg00613177	0.65173948	PPP1R10	6	30570163	cg19389001	0.560197131	SEMA5B	3	122640778
cg17341969	0.651696465	SNPH	20	1287050	cg22163463	0.560069937	PIFNM3	17	6458640
cg09595202	0.651462861		7	156812104	cg17806623	0.560066192	KI	13	33590002
cg16899088	0.651371631	STYK1	12	10827583	cg06521960	0.559821772	CACNA1H	16	1220382
cg27003122	0.651329435	OCA2	15	28339563	cg03977071	0.559821772	ZADH2	18	72917163
cg07790733	0.651328502		19	1868952	cg20571761	0.559797431		10	44805674
cg04904561	0.651231718	TDRD9	14	104394831	cg14845385	0.559797143	MRPL55	1	228297092
cg25178900	0.651086957		10	27546458	cg24049348	0.55979492	PDGFRL	8	17433926
cg04784163	0.650966184	GPSM3	6	32164210	cg06213635	0.559777955		11	129488336
cg11757144	0.650966184	MTERF	7	91510093	cg03891843	0.559767871	KNDC1	10	135030439
cg17386240	0.650780505	TGFB1	5	135384080	cg01992512	0.559716408	ADARB1	21	46493156
cg02501715	0.650603865	KIAA1530	4	1373985	cg12249227	0.559570692	IRIG1	3	66550918
cg09788123	0.650542317		6	26533540	cg04125223	0.559442568	CCDC127	5	218200
cg24122049	0.650483092		12	30032232	cg248606475	0.559399377	PHILA2	11	2951281
cg10436333	0.650429787	NKAIN4	20	61883856	cg15605097	0.559361117	DGAT2	11	75479463
cg03371778	0.65027565		10	132365775	cg24671760	0.559722397	C6orf186	6	110678078
cg22098375	0.650120773	GA8R1	6	29590966	cg12568043	0.559727057	C7orf33	7	148287720
cg15625558	0.649879227	C17orf63	17	27171324	cg09685500	0.559165419	KCNH4	17	40333380
cg11716681	0.649879227	CXADR	21	18884682	cg16488953	0.559165288	KCNK2	18	77638446
cg25858232	0.64972104				cg26886268	0.559157904	PTPRN2	7	157387156
cg03193308	0.649680586	RBM28	7	127950724	cg16738915	0.559028611	SIRT2	19	39390806
cg02196731	0.649668558	AFF1	4	87855982	cg23928920	0.559021038	STIS/A2	15	92944756
cg11358777	0.649275362	MOG	6	29638498	cg23103993	0.55866615	PDE4A	19	10531434
cg23332689	0.649275362	IDUA	4	987652	cg06479755	0.558611575	JAM2	21	27012313
cg00542760	0.649212819	ARFGAP1	20	61917750	cg16101543	0.558585736		12	131162896

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg23627354	0.649023168	CCBL2	1	89459658	cg23312375	0.558554456	RASA3	13	114814024
cg12252547	0.649013097	MALZ	8	120220032	cg18709881	0.558485607		18	72837627
cg07864327	0.648913043	KBTBD12	3	127690108	cg26170569	0.558448186		2	130986931
cg09144424	0.648578251	PDLIM1	10	97050675	cg16928293	0.55830285		13	107027718
cg15709435	0.648548334		5	180612719	cg08923033	0.558233309	HBA1	16	226002
cg15244965	0.648374673	CSMD1	8	3429480	cg01146980	0.558067517	DNAJC21	5	34930244
cg15717617	0.648350439	PLEKHA2	8	38782090	cg19385090	0.558038542		7	27292614
cg23188000	0.648309179	IGMN	14	93215136	cg08422181	0.558005995	UHRF1	19	4909468
cg16288101	0.648228101		14	88621538	cg24336702	0.557966516	PRKAR1B	7	762576
cg01648609	0.648125717		4	2049057	cg08594606	0.557913521	DNAJA4	15	78556512
cg21901577	0.647826087		4	118736876	cg09664314	0.557827814	NEGR1	1	72190606
cg06906965	0.647644444		19	58450175	cg12287936	0.557499473	NDUFS6	5	1800606
cg00834796	0.647463768	JAKMIP2	5	147161924	cg17258335	0.557462933		17	1474778
cg11806703	0.647342995	CTDSP1	2	219264807	cg08202754	0.557453426		7	152621979
cg23252336	0.647342995		2	132548153	cg13022905	0.557423455		5	147699892
cg16060486	0.647222222	CYP11B1	2	38303472	cg00369351	0.557320906	BTN2A1	6	26458121
cg07453773	0.647222222	ST5	11	8779319	cg02187348	0.557275237	SPG7	16	89574699
cg13740834	0.647211182		7	93926617	cg06849719	0.557078272		1	205399945
cg13329852	0.647174592	EBPL	13	50265193	cg16330146	0.556830979		14	102422355
cg04522003	0.647152246	LOC10019093	17	40913702	cg07647179	0.556811783	MUC2	11	1078956
cg10349685	0.646859903	DLC1	8	13372491	cg16983817	0.556623816		21	15399683
cg24844769	0.646739113	INSR	19	7269698	cg10493162	0.556620671		16	88204283
cg18391209	0.6466165	CAPN8	1	223747670	cg13322722	0.556483401		12	131714809
cg12578563	0.646474232		11	1159210	cg06633637	0.556462534	SCARNA2	1	109643190
cg10012722	0.646458775		2	237085763	cg13005428	0.556450829	RPTOR	17	78916378
cg02788857	0.646455474	PIWIL2	8	22132959	cg23530758	0.556402691	RPTOR	17	78857712
cg08031982	0.646395227		16	87577539	cg25390635	0.556388179	GRAMD4	22	47022618
cg07959978	0.646014493	LEPR	1	65994091	cg14743534	0.556338962	FLJ35390	7	44079036
cg27008363	0.645926714	HOXC4	12	54449761	cg00984540	0.556295384		7	155589293
cg02272814	0.645652174	TDRD6	6	46655782	cg21998512	0.556229161	GATAD1	7	92077031
cg10242372	0.645652174		11	45248652	cg03041029	0.556213562		2	240405158
cg09682190	0.645410628	FAM54A	6	136571455	cg19754622	0.555952449	STK32C	10	134045578
cg17491304	0.645339007				cg11218091	0.555890457	LOC399815	10	124638976
cg19734370	0.645169082	NPTX1	17	78444348	cg02575675	0.555872619	LOC440356	16	29875187
cg23613051	0.645169082	SH3BP2	4	2820428	cg14271908	0.555716335	TFAM	10	60144732
cg25252482	0.644927536		7	119893751	cg07152177	0.555640053	C1orf198	1	231004507
cg13049471	0.644927536	ST6GALNAC1	17	74641167	cg11685316	0.555639855	MFS6L	17	8702564
cg13526469	0.644890974		3	13245693	cg08373528	0.555619482	PRPH2	6	42672105
cg18750833	0.644806763	MCEE	2	71358451	cg05495984	0.555529132	EHMT2	6	31865909
cg24744500	0.644806763	PHK3CG	7	106508732	cg09688773	0.555413734		3	172469100

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg22286978	0.64468599	A1BG	19	58858806	cg01252672	0.555404654	NPR3	5	32711881
cg05243705	0.64468599	PHF10	6	170124933	cg05115862	0.55516016		17	79439144
cg20118431	0.644604255	B3GAT1	11	134278896	cg22330492	0.555114138	TTL16	17	46894465
cg09817641	0.644558392				cg13523818	0.555058138	HAPLN4	19	19371900
cg08334527	0.644525986		1	229252042	cg22470218	0.555056089	YIPF3	15	43484939
cg163338321	0.644499761		17	48994958	cg04875709	0.555044136	CDKN3B	12	12870463
cg11417092	0.644444444	C7orf50	7	1057285	cg20909380	0.555009712		15	99987288
cg09300980	0.644323671	FAM8A1	6	17601309	cg24539779	0.554926005	HES3	1	6302896
cg22088529	0.644323671	FAM81A	15	59730587	cg15988970	0.554836931	DNAJC15	13	43597736
cg18988170	0.644252863	CABYR	18	21719263	cg15360451	0.554705267		12	131057032
cg00421735	0.644082126	ZNF862	7	149559471	cg04258219	0.554693702		7	155142415
cg041158046	0.64384058	ALG6	1	63833953	cg16418105	0.554607493		16	876216
cg12614687	0.643719807	OTF	5	76926752	cg15739997	0.554598562	UBE2V1	20	48729771
cg19399653	0.643719807	CEACAM8	19	43085416	cg04482075	0.554551033	SEPI1	16	1991308
cg00939347	0.643714421	DIP2C	10	652260	cg25830048	0.554548818	HES7	17	8027460
cg05896902	0.643599034	LOC145663	15	45671018	cg26500033	0.554464354		6	166585954
cg15828711	0.643537795		12	114235931	cg14001750	0.55434763	PRKCA	17	64302651
cg20250343	0.643478261		2	92160107	cg20418529	0.554281737		6	166260012
cg11197418	0.643357488	SUC1G2	3	67705285	cg13750902	0.554252358	PIP5K1B	9	17393017
cg08806109	0.64333617		10	51489223	cg18322589	0.554189354	TACC2	10	123909456
cg27122958	0.643236715	NF1	17	29425209	cg16841133	0.554181509	CTDP1	18	77496360
cg19550904	0.643088944	C14orf73	14	103569127	cg13457396	0.554057064	C1orf25	1	185127093
cg16357661	0.642995169	EXO1	1	24201898	cg04717802	0.553978741	WBP2NL	22	42394638
cg22516566	0.642995169	LYSDMD2	15	52031467	cg07747299	0.553959421	C21orf56	21	47604052
cg08345189	0.642874396		5	65437789	cg07110217	0.553918609	PCBP4	3	52001995
cg24049841	0.642823138	RTDR1	22	23484598	cg14337129	0.553759946		9	131594230
cg01017349	0.642780647	COL1C11	2	3648229	cg06150148	0.553612189	LONRF2	2	100902129
cg02725269	0.64263285	KLK1	19	51327177	cg01540869	0.553430049		13	25116150
cg01147665	0.642391304		3	32146075	cg27624178	0.553377853	C15orf24	15	34393959
cg16423096	0.642276694		14	22279816	cg21158664	0.553351808		7	159870
cg02519772	0.642186822	LRP1	12	57578672	cg25744700	0.553321615	CTSH	15	79237217
cg07671603	0.642149758	C7orf13	7	156433350	cg09841842	0.553284795	FRMD6	14	51974848
cg25968664	0.642028986		8	113049519	cg17085352	0.553221116	HOXC13	12	54332026
cg11996093	0.641958865	SNHG2	2	1234555	cg22946460	0.553218086	NAV2	11	19372282
cg18535437	0.641922931	PPM1A	14	60715218	cg12397584	0.553202883		12	58068688
cg22946066	0.64178744	TRMT11	6	126308385	cg09662920	0.553201607	GIPC2	1	78511433
cg01814565	0.641666667	C8orf34	8	69490348	cg24585418	0.553109125		5	179059426
cg13830619	0.641425121		12	9555480	cg26455175	0.553107997		16	853533
cg09684079	0.641002966		10	134787625	cg17050616	0.55296541	MOV10L1	22	50585229
cg23104954	0.640640875		13	50701501	cg08500872	0.552956298	BANP	16	87990909

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg21138405	0.640532169	IRF1	5	131827807	cg09000779	0.552866302		3	136867925
cg16864139	0.640496255	PLEKHH2	2	43864212	cg22198853	0.552692344		6	1594411
cg02765031	0.640458937		10	85462980	cg14260521	0.552655302	ZNF22	10	45496672
cg23903129	0.640217391	GPR133	12	131617086	cg07671805	0.55244757		5	74907694
cg21955801	0.640217391		15	258339600	cg26012716	0.552187371	TMEM135	11	86748913
cg14126608	0.640040189		11	122846540	cg14561063	0.552183979		10	44202513
cg04940582	0.639975845	CYP26B1	2	72356976	cg02574731	0.552171615		1	161510041
cg08317263	0.639899211	CCDC69	5	150603169	cg16836675	0.552149909	TOM1L1	17	52981853
cg02321381	0.639855072	OXA1L	14	23233613	cg12709669	0.552138253		18	12287173
cg10189882	0.639855072	ERIC1	8	665934	cg26545162	0.551986454	ZACN	17	74075092
cg03594819	0.6397343	TIMM8B	11	111956829	cg10290504	0.551962838		11	116578271
cg23885108	0.639639093		17	37755750	cg00042144	0.551948433	C1orf74	1	209958035
cg18024113	0.639613527		1	16862202	cg01243586	0.55194542	SAA2	11	18270548
cg13450266	0.639492754	MCF2L	13	113625160	cg01885783	0.551941345		22	20288286
cg16576255	0.639492754	LOC400891	22	21400640	cg00084338	0.55182682	DLL1	6	170595920
cg00075608	0.638647343	ADAMTS6	5	64770553	cg12078775	0.551799713		6	30419543
cg09091657	0.638570718		8	41004167	cg18011273	0.551732107	SORCS2	4	7203326
cg26208159	0.63852657	OR4X2	11	48266856	cg01422370	0.551666609	NANS	9	100819542
cg01070148	0.638461922	BCL3	19	45250754	cg01422370	0.551655537		2	73384389
cg15743799	0.638405797	RCVRN	17	9805578	cg16490124	0.551512308	SNORA14B	1	235292369
cg09554300	0.638334085	TMEM9	1	201123894	cg19631264	0.551435507	MIR496	14	101526802
cg21864942	0.638326282	PARP4	13	25080095	cg12796015	0.551343701	E2F2	1	23857532
cg06686857	0.638285024	C8orf39	8	94751454	cg49219655	0.551318512	KLF11	2	10187086
cg07589192	0.638164251	IGFALS	16	1845627	cg08849813	0.551163026	SERGEF	11	17825098
cg22972858	0.638129251	GPR81	12	123215248	cg01056242	0.551147723	GPR133	12	131622739
cg02275014	0.638043478		1	109647126	cg23804620	0.551000774		4	6690894
cg11331837	0.638029122		17	35161825	cg22439381	0.550987883		5	179061795
cg02139034	0.637560386	COPG2	7	130146351	cg08783793	0.550908251	ERAP1	5	96143578
cg22304519	0.637536529		2	227560785	cg12618699	0.550637473		12	133183756
cg23481419	0.637360757		11	71282370	cg06394109	0.550510131		16	1152511
cg24371114	0.637351244	CCDC149	4	24913973	cg16276982	0.550411311		15	29968032
cg09226051	0.637309403	NLRP3	1	247611502	cg06904356	0.550343656		5	1849983
cg05194618	0.637274798	KHLI28	14	45432309	cg10362742	0.550254993	HSPH1	13	31736223
cg15160709	0.636995048	CYP26B1	2	72357937	cg02156870	0.550253421	SRCIN1	17	36717733
cg14204060	0.636898818				cg06962620	0.550249445	C21orf67	21	46360149
cg00744433	0.63688747	CXADR	21	18884067	cg10194536	0.550231286	NXN	17	807457
cg19513321	0.63667237	VAMP5	2	85811432	cg03987199	0.550155112	OPRD1	1	29189655
cg04896959	0.636594203		15	78267971	cg10907148	0.550094181	C17orf28	17	72948349
cg27092704	0.636352657		3	153066830	cg08888354	0.550065141	ZNF570	19	37958707
cg23336241	0.636352657		2	134699952	cg08419873	0.55006142		13	27296010

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg22648182	0.636231884		5	1663755	cg07218516	0.550051863	FAM134C	17	40761703
cg14189381	0.636231884		9	125106354	cg26707709	0.549876053	SNED1	2	241975756
cg04196064	0.635990338	TRIM48	11	55029786	cg03854913	0.549798355	TBCD	17	80798068
cg06963711	0.635869565	GNG12	1	68170392	cg03936135	0.549796882	LOC732275	16	86371248
cg21874832	0.635736643	FAM178B	2	97652249	cg10127463	0.549736322	RPS16	19	39926709
cg21381779	0.635628019	APPL2	12	105567515	cg07698580	0.549718131	RGR	10	86016643
cg26356683	0.635535508		13	60841985	cg21461564	0.549482206	JTPA	8	63999030
cg09655329	0.635507246		12	68975128	cg16889990	0.549444505	USP29	19	57631478
cg27554954	0.635250795	ANXA2	15	60691595	cg11541601	0.549437055	PDLIM2	8	22437870
cg20290983	0.635223178	MRPS18A	6	43655470	cg07099640	0.549171826		16	88292973
cg00117018	0.635010203	ZNF251	8	145955983	cg19952015	0.549150399	ATXN7L1	7	105432653
cg27102995	0.634903382	JAZF1	7	282221194	cg20648847	0.549125523	ACTN3	11	66326767
cg06375652	0.634665465		16	86100423	cg11827998	0.549110377	TMEM121	14	105995685
cg25900813	0.634541063		8	140717297	cg09320190	0.549091395	WNT3	17	44847220
cg00359010	0.634541063	MDG	6	29635692	cg17658885	0.54909007	ZNF33A	10	38299503
cg02440562	0.63442029	SPATA21	1	16763949	cg13047869	0.549081015	C3orf24	3	10149882
cg10418598	0.634299517	SNX33	15	75940305	cg00016223	0.54902849	STK38	6	36515007
cg00035529	0.634179283	FBXO47	17	371233835	cg22734086	0.549008207	C2orf70	2	26785367
cg070354208	0.634085806	DCDC2	6	24358566	cg27638458	0.54900433	ODF3L2	19	467740
cg10115987	0.634057971		2	220375548	cg08906194	0.548930567	CNTFR	9	34577972
cg20132862	0.634048945		11	119970504	cg09866143	0.548902925	SLC12A8	3	124861521
cg16786868	0.633937198	HEIZ	17	65242413	cg18319818	0.548872436	MMAA	4	146540545
cg21110795	0.633937198	INTS1	7	1535471	cg15964611	0.548871209	OPCML	11	133402545
cg01751181	0.633816425	OPCML	11	133032650	cg11832804	0.548835756	TERT	5	1279449
cg06407043	0.633729527		12	8403616	cg23513183	0.548745423		19	2858944
cg100808058	0.633596101		6	1400797	cg23197280	0.548729289	TSTA3	8	144699837
cg04017513	0.633454106	NRM	6	30658894	cg19927510	0.548683282	DNM2	19	10829071
cg18397073	0.63321256	POU2F2	19	42600278	cg16454495	0.548609297		16	1199966
cg05072413	0.633091787	FAM19A5	22	49015998	cg12449366	0.548589212	C6orf174	6	127840755
cg18424635	0.633030453	HLA-DRB5	6	32490421	cg15426006	0.548586199	UBAC2	13	9869897
cg17841765	0.632841958		1	22560699	cg21137501	0.548496195		14	108892303
cg07023327	0.632526689	MRPS17	7	56019384	cg09894276	0.548448288	WDR27	6	169977394
cg19383689	0.632487923	WSB1	17	25621092	cg11554525	0.548327448	MAK16	8	33342681
cg11645318	0.632487923		7	37491346	cg11401866	0.548278528	HSP81	7	75932851
cg03312205	0.632487923	COPS7A	12	6832737	cg07477090	0.5482383	PHGDH	1	120255318
cg22508145	0.632487227	CPAMD8	19	17015427	cg16914890	0.548220891		2	242839615
cg06193597	0.632325954		2	241896910	cg01434559	0.54821191		11	8290939
cg19181528	0.63216222		20	59542589	cg00601042	0.548173732	GALNT9	12	132900274
cg05654340	0.632125604	AMID1	6	111197834	cg25823419	0.548123881	GNB4	3	179169252
cg17922326	0.632125604	ADAR2	10	1558808	cg20141509	0.547953957	CASZ1	1	10781431

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg11857238	0.63184081	TRIM10	6	30122593	cg06616976	0.547866299		7	1361231
cg17873612	0.631673286		1	23279531	cg13808071	0.547829117	GREB1	2	11679872
cg19701213	0.631521739	TMEV145	19	42817154	cg10318066	0.54770217		2	74045551
cg10348972	0.631521739		8	57505432	cg21202759	0.547678175		2	241085043
cg23197939	0.631400966	CDC40	17	78039368	cg25432232	0.547600321	AURKC	19	57742423
cg16575248	0.630797101	DYRK3	1	206808599	cg10981736	0.547513775	C4orf42	4	1242757
cg09404428	0.630561262		7	121078303	cg19818764	0.54750737		14	21439700
cg10339911	0.630555556	EIF1	17	39847455	cg11466815	0.547338508	SLC38A9	5	55008477
cg02075087	0.630247376	ING2	4	184436383	cg09702881	0.547326712	MRPS188	6	30584746
cg16087940	0.630217967		10	129947807	cg05857996	0.547249524	EBF4	20	2675418
cg04225510	0.630193337	CDC123	10	12237758	cg24954453	0.547216326		5	780549
cg22915373	0.630091783		13	110337639	cg15967501	0.547150992	SEC1	19	49143404
cg23163279	0.629876848	FLJ35220	17	78389065	cg06263372	0.547053926	BRUNOL4	18	34850864
cg24390932	0.629710145	SLC18A1	8	20040785	cg14182974	0.546846243	RCL1	9	4791918
cg10756475	0.629500188		4	1757242	cg08277196	0.546823417	BMPIR1A	10	88516091
cg06938878	0.629468599	CAICB	11	15094364	cg11392858	0.546592448	PLD6	17	17109651
cg03502446	0.629227053	NUP37	12	102513777	cg04124626	0.546517409	PLCXD3	5	41509934
cg26772847	0.62910628		15	62589487	cg02675646	0.546350826	SPTBN4	19	40996118
cg08409553	0.628953941	KIAA0664	17	2601321	cg09693358	0.546342908	ERGIC3	20	34129717
cg08814536	0.62883744		13	106567330	cg25610515	0.546332042	LUZP1	1	23495577
cg24748746	0.628743961		7	155738916	cg00093478	0.546286036		13	112996967
cg04325591	0.628694235	FA2H	16	74806009	cg16789863	0.546282331		2	132589606
cg02614045	0.628623188	DOCK2	5	169129494	cg14138549	0.546053807	1-Dec	9	117932648
cg15248935	0.628518845	DNAI86	7	157180055	cg20141398	0.54602456		1	2928616
cg02849924	0.628502415	POLD1	19	50887530	cg21672705	0.545932251	AGPAT5	8	6565867
cg21725754	0.628502415	NAF1	4	164087380	cg07217030	0.54569468	SLC45A4	8	142233642
cg05423392	0.628488416		16	34586837	cg14463292	0.545383518	FGF12	3	192289252
cg19311470	0.628434974	RPL9	4	39460490	cg11282895	0.545119178	SON	21	34914652
cg01485790	0.628413695	NCOR2	12	124911358	cg00488734	0.545077049	MMP15	16	58076165
cg16976520	0.628406619	ESYT2	7	158588852	cg17945560	0.545038089	TINAGL1	1	32052651
cg04257152	0.628381643	KRT27	17	38933754	cg10983623	0.545019864	DDX6	11	118662299
cg09848405	0.628367544	ZNF700	16	3285198	cg02414650	0.54498693	IFT140	16	1561036
cg20675251	0.62826087	ATAD3A	1	1464803	cg06651273	0.544941837	ASNSD1	2	190526628
cg09276445	0.627657005	ITGBL1	13	102106423	cg23323827	0.544901716	INHBB	2	121108755
cg18528696	0.627593381		12	64925934	cg27080194	0.544789248		12	131198873
cg23361708	0.627448873	MOGS	2	74697613	cg15028160	0.544638553	PPFIA3	19	49622717
cg17913877	0.627294686		2	60602483	cg08357651	0.544617826	COQ10B	2	198318030
cg21209485	0.62722099	MMEL1	1	2529359	cg00321703	0.544493011		2	129079801
cg15727583	0.627080502	MF12	3	196757701	cg20255370	0.544203827	EIF2AK4	15	40268687
cg13496998	0.62705314	C14orf165	14	24404827	cg17431280	0.544181462	GRLF1	19	47507461

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg22447508	0.626932367	MACROD2	20	13975439	cg13996962	0.544171864	LRRCA	9	131643933
cg18947110	0.626932367	NADK	1	1688383	cg06202585	0.544154035		2	38325802
cg13205771	0.626811594	KIF1	6	33376588	cg05597766	0.543936944	NAP1L4	11	3013541
cg23531285	0.626703801	POT1	22	16288696	cg19916659	0.543786034	MIR548N	2	179387931
cg16971664	0.626328502	PIRMA1	10	3208440	cg09322003	0.543721104		7	15584248
cg04137418	0.626302512	LOC785733	6	131148736	cg24467349	0.543695841	GSTM5	1	110754835
cg27237671	0.626230609	TMEM18	2	676223	cg21091547	0.543603326	CDKN1A	6	36645500
cg20913114	0.626224896		7	1315546	cg06889571	0.543570787	ADAR2	10	1416791
cg20185145	0.626207729	RREB1	6	7230113	cg00256231	0.543463423	TBC1D16	17	77916733
cg06713229	0.626190966		10	134624679	cg07482372	0.543440037	PHLDA2	11	2951201
cg17793819	0.626101966	ZHX2	8	123793727	cg11235602	0.543392565	MOBP	3	39543776
cg21703068	0.626086957	ITSN1	21	35015322	cg24364535	0.543365854	OPCML	11	133098499
cg01718065	0.626030936		19	51774429	cg25628461	0.543320794		4	687023
cg19987356	0.625902342	PLEKHH3	17	40823930	cg11421509	0.543303417		8	105379602
cg07315018	0.625899996	MEST	7	130131916	cg17372223	0.543264177	NT5DC2	3	52568218
cg06234051	0.625817607	SOX9	17	70120541	cg09792204	0.543216047	OCA2	15	28342183
cg04156077	0.625773823	VWTR1	3	149421196	cg03124680	0.543179326	NAP1L4	11	3013791
cg04581753	0.625673217	ZNF295	21	43429161	cg12308909	0.543039209	RPS6KA2	6	167042108
cg14197071	0.625571063	PRDM8	4	81110459	cg18183042	0.542843248	TBCD	17	80733038
cg02797871	0.625451664	COL20A1	20	61953935	cg11541678	0.54281582	MLC1	22	50506001
cg07240846	0.625391129	CAMK1D	10	12438782	cg13069441	0.542664203	STK19	6	31939322
cg21044572	0.624966891		2	219766752	cg13552710	0.542614098	ALK	2	30144547
cg01275006	0.624927914	GGN	19	38876250	cg04240373	0.542583452	PSMA5	1	109969611
cg27114706	0.624918258		12	92527244	cg20982735	0.542520702	PSKH2	8	87082023
cg19041132	0.624913805	SPHK1	17	74380824	cg14861803	0.542322066	NUDT16	3	131100698
cg22488717	0.624761229	C2orf70	2	26785946	cg27402782	0.542306352	SCYL1	11	65296794
cg14839656	0.624637681	ACAD10	12	112124200	cg13700197	0.542230852	PTCH1	9	98221975
cg25435686	0.624575949		6	32952826	cg19120251	0.54219763	SH3PXD2A	10	105598517
cg18333405	0.62456086		7	148764182	cg24677002	0.542103032	CLQA	1	22965435
cg05676441	0.624479582	CDH18	5	19988800	cg27622722	0.541803412	SGK3	8	67624757
cg05401945	0.624395974	CCDC66	3	56590734	cg10636020	0.541765151	KCNH2	7	150656979
cg20949223	0.624269693	MUC1	1	155161679	cg22674497	0.541727533	USH1G	17	72916000
cg07630858	0.623842969	DBN1	5	176900819	cg25753693	0.541539763	KIK8	19	51504234
cg17568255	0.623636093	IAH1	2	9614530	cg05753675	0.541503964	RNF216	7	5821294
cg11761615	0.623607092	TNRC18	7	5348924	cg20333904	0.54147866		2	240724165
cg18909131	0.623482953	KIAA0427	18	46289505	cg24258529	0.541325536		4	1570685
cg19872923	0.623429952	SYT1	12	79457822	cg26243551	0.541267833	SFRS18	6	99873348
cg00673651	0.623309179		6	5007264	cg07051728	0.541246101	SULF1	8	70378515
cg15341866	0.623188406		14	101158095	cg22496723	0.541009139	CACNG2	22	36960779
cg25708755	0.623155781	PTPRN2	7	157951411	cg09141300	0.540946349		5	30346020

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg20670274	0.622848118	ERIC11	8	681577	cg13295614	0.540841141		2	239330383
cg08977311	0.622826087	C3orf50	3	168308798	cg23908794	0.540823698		6	32383465
cg22411472	0.622826087	SH3RF3	2	109950128	cg00624150	0.540809351	PCBD2	5	134240771
cg15288329	0.622705314		6	170268691	cg21436520	0.540723249		6	53413189
cg06942649	0.622584541		8	435813	cg16010433	0.54051145	HGN2	19	588603
cg07615383	0.622485106	IGF1R	15	99409194	cg09011167	0.540175602		11	62423268
cg11064039	0.622463768	PRKAR1B	7	766100	cg052227549	0.539924778		12	1770782
cg05731801	0.622377782	EMIP2	16	10647200	cg18074403	0.539796341	DIP2C	10	403367
cg03003073	0.622342995		1	9489434	cg13608616	0.539791377	ERCC1	19	45926999
cg14705524	0.622222222	TUSC4	3	50388322	cg10343367	0.539502696	LOC732275	16	86375065
cg26064535	0.622222222	OPRD1	1	29138043	cg27048098	0.539451422	MIA3	1	222791638
cg14325184	0.622033143		22	21025079	cg16135989	0.539378202		5	176166629
cg16132351	0.621818913	COL9A3	20	61447823	cg02222324	0.539268286		2	10231621
cg08386165	0.621618357	TMEM126B	11	85345554	cg04522671	0.539169104		5	139144055
cg02386599	0.621497585	DAAM2	6	39769542	cg09473826	0.538991574	TNR	1	175568216
cg27193377	0.621301324	TMEM170A	16	75498489	cg14278501	0.538958828		17	39804486
cg00459289	0.621256039	ACTR3B	7	152456746	cg22262702	0.538938867	TAF4	20	60638679
cg08546707	0.621256039		1	846195	cg21474955	0.53868382	GCHFR	15	41056252
cg02869364	0.621211641	C7orf50	7	1081709	cg13553473	0.538639065	ZBTB9	6	33422300
cg07392829	0.621081324	KRTAP6-3	21	31964616	cg07140595	0.538600813	CENPF	1	214776313
cg14789911	0.621021911	C21orf56	21	47582049	cg24112000	0.538539027		20	60950667
cg04876606	0.620772947	RBM6	3	49977596	cg24176037	0.538287832	TBPL1	6	134274008
cg18579447	0.620652174	MAPKSP1	4	100815442	cg09508938	0.538258363		6	13770042
cg12466610	0.620582897	MOSC2	1	220950205	cg20112079	0.538250988	VENTX	10	135050142
cg13747681	0.620348569		11	66179783	cg16726374	0.538193959		16	21567052
cg21303655	0.620301048	KCND3	1	112438797	cg22261895	0.538107363	PLBD1	12	14720100
cg00744351	0.620289855	CHRNE	17	4804374	cg06015422	0.537895853		8	70907139
cg01263716	0.62023764	PHLDA2	11	2951590	cg19542630	0.537848464	ACOX3	4	8442666
cg01802772	0.620217251	ACOF1	1	55014160	cg10433376	0.537838273		21	46766039
cg02665463	0.620185864	ARL17A	17	44657154	cg24774960	0.537739988	MAPK8IP1	11	45906625
cg05573109	0.620157325	ALDH1L1	3	125876246	cg13733394	0.537667534	CLDN10	13	96085504
cg16655778	0.620023834		12	132981736	cg13878677	0.537511307	IQSEC1	3	13060915
cg04231541	0.619874052	TESSP1	16	2849110	cg23350274	0.53746884	RDH12	14	68188663
cg17611045	0.61968599	FGF12	3	192289245	cg04206699	0.537432959		8	70855057
cg27447006	0.619552395	ASAP3	1	23763279	cg12586386	0.537412314	AQP11	11	77299805
cg07909498	0.619509328		4	79627477	cg09241929	0.537364884	GNAS	20	57465560
cg00057593	0.619286998	GML	8	143916959	cg27469738	0.537292347	ANKRD2	10	99338074
cg18865445	0.619279683		13	110522265	cg18335504	0.537291886	TEP1	14	20881748
cg12097325	0.619031065		1	1076431	cg03277550	0.537285416	AFG3L1	16	90038083
cg13149387	0.61884058	LOC10013087	4	1204205	cg01218619	0.537187428		4	25090298

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg17181941	0.618754188	NCAPH	2	97004840	cg11946719	0.537078516		2	233216220
cg19105362	0.618642856	SAR1A	10	71929497	cg25407557	0.537029777		12	132168981
cg04649852	0.618620574		16	29229650	cg24898738	0.537024566	GPR123	10	134938144
cg26748794	0.618478261	FAM38A	16	88804051	cg03727754	0.536811912	CACNA1C	12	2564100
cg25594486	0.618413681	SHANK1	19	51165441	cg04368832	0.536624998		16	5529642
cg13707567	0.618357488		14	84720664	cg10629682	0.536597751	SLC30A3	2	27486061
cg20360416	0.61811434	SORCS2	4	7246127	cg19807241	0.536591693		11	14402761
cg22327976	0.61805993	GPR137B	1	236318493	cg14882966	0.536551235		2	3699353
cg11176853	0.617820564	PGBD5	1	230488694	cg15247269	0.536478937	HSPA1B	6	31795125
cg09639108	0.617764902		3	127006893	cg26266427	0.536309062	TNXB	6	32063838
cg03167812	0.617504965	AGAP1	2	236970809	cg03940643	0.536245267		16	11343701
cg14158712	0.617281149	CL7orf103	17	21156871	cg01084566	0.536172558	HPGD	4	175444482
cg23829584	0.617270531		6	20032755	cg09668030	0.536093617		1	27852173
cg18152887	0.617270531		2	920593	cg13176198	0.536065514		15	30336915
cg07365741	0.617171859		6	170478434	cg17133982	0.535890684	VMA5B1	1	20672105
cg08560373	0.617152572	SNRPN	15	25123381	cg23831021	0.535851485	ATP5E	20	57607762
cg03034934	0.616917223	CBFA2T3	16	88965044	cg16433211	0.535821572	EPN2AIP1	3	37034693
cg03261155	0.616908213	S100A712	1	153413940	cg01808810	0.535810435	SDHA	5	218483
cg08777812	0.616806122		11	70289549	cg11198895	0.535706527	SYT7	11	61347003
cg10426939	0.616731873	LYSMD4	15	100273706	cg19551541	0.535591146		4	190935955
cg19926776	0.6165349	UBA52	19	18682709	cg07510611	0.535466199	JARID2	6	15513503
cg05825120	0.616425121		16	30682503	cg10023862	0.535319866	CUEDC1	17	55962841
cg23201527	0.616325752		20	61783225	cg17052170	0.535300332	LY6E	8	144099482
cg16696923	0.616183575	CDH17	8	95229872	cg08734618	0.535125906	TSLP	5	110408916
cg10701252	0.616181107	WNT5A	3	55508585	cg16065768	0.534931945		12	114107712
cg14921757	0.616036009	PRC1	15	91517551	cg18337575	0.534857227	CTDP1	18	77482837
cg18584561	0.615700483	GREB1	2	11682017	cg04034967	0.534847069		7	1309108
cg15935291	0.615458937		14	29130634	cg12109260	0.534739359		8	144850435
cg12514963	0.615346311	C7orf27	7	2580905	cg19108952	0.534685986	HAUS2	15	42840882
cg15798862	0.615251434	DTX2	7	76129360	cg10940203	0.534647841	FAM153C	5	177434438
cg20594982	0.6152488	AGR1	1	976707	cg26115216	0.534592459	PRR23B	3	138739916
cg22855876	0.615232575	TMEM214	2	27256034	cg10467557	0.534584161		13	21893662
cg04140932	0.615131915		20	61674143	cg08375775	0.534484568	DDX60	4	169239733
cg04371780	0.614820692	ELFN1	7	1785297	cg13813355	0.534400293		2	175210231
cg05695995	0.614691218		6	7468848	cg19994023	0.534121671		8	896972
cg13184736	0.614441645	GNG12	1	58290409	cg06794034	0.53402341	SORCS2	4	7301854
cg09988062	0.614273466	SAMD11	1	877489	cg11871345	0.534011351	SOX12	20	305874
cg26658509	0.61425641		11	3225550	cg07205796	0.533818074		16	1187722
cg15255455	0.614228735	PLIN5	19	4534986	cg19983815	0.533650545	ATP6VOC	16	2563934
cg26429499	0.614195496	SHANK2	11	70563792	cg22532774	0.533493874	ESPN1	2	239037049

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg02337902	0.614169485	C7orf50	7	1103177	cg09184378	0.533419139		6	126698055
cg234746885	0.613774073	MCCC1	3	182782313	cg01463540	0.533383159	DLGAP4	20	35064637
cg19388625	0.613647343	TPO	2	1452665	cg26228351	0.533308087	KIF21B	1	200992656
cg03421195	0.613425059		10	7517125	cg04895288	0.533257257		7	98999949
cg20546778	0.613405797	GAP1	8	75263599	cg07534843	0.533039886	MATN4	20	43935283
cg14068721	0.613346435	SBR2	19	56047224	cg06951626	0.533029976		12	22095330
cg22953237	0.613258701		7	31425682	cg18223453	0.532930563	TBX4	17	59554746
cg17005319	0.613208004	DEAF1	11	655579	cg11401682	0.532882737	NEU1	6	31830616
cg08072848	0.613032038		8	1245672	cg15205435	0.532852799	CHD5	1	6187920
cg23845936	0.612922705	INSR	19	7140630	cg22248382	0.532802348	RAB11FIP4	17	29758931
cg10619318	0.612913767	SMAD5	5	135470751	cg20276402	0.532647117	LOC400657	18	72263688
cg09117448	0.612747959	COX6A2	16	31439393	cg13707894	0.532605185		3	27674461
cg23278040	0.612715839		19	266712	cg04556542	0.532522985	SLC24A4	14	92790580
cg25786640	0.612692552	MAD1L1	7	1946447	cg26555463	0.532240733	HIST1H2AB	6	26034022
cg10510478	0.612617406	CLDN18	3	137728810	cg03996735	0.532207576	H19	11	2020104
cg20586124	0.612604071	CTAGE1	18	19998018	cg08385906	0.532001556	BAT4	6	31634065
cg00947859	0.612560386	KRTAP10-11	21	46067411	cg02472801	0.531939457		8	2480483
cg19369022	0.612465525		19	51111388	cg00595030	0.531785461	ICAM4	19	10398582
cg16520539	0.612407106	TRIM27	6	28890872	cg12699647	0.531605379	C17orf62	17	80401857
cg02478118	0.612245767		6	13773541	cg15175351	0.531579696	LRRC27	10	134165822
cg08141172	0.612198068	KIRREL2	19	36351774	cg16807101	0.531501422	UHRF18P1L	12	100517446
cg06389098	0.612185405	CAUB2	16	71424252	cg06792262	0.531396692	ZNF167	3	44622596
cg08837037	0.611918529	C22orf27	22	31318179	cg02761523	0.531275351	NFIX	19	13106643
cg23285459	0.611846314	GNAI2	7	2802560	cg21221161	0.531256574	PRKCZ	1	1999183
cg05519582	0.611602091	GRAMD4	22	47027906	cg20784591	0.531096898	PILRA	7	99972461
cg16203863	0.611507189	WDR73	15	85197975	cg10899547	0.530962806	RASGEF1C	5	179554111
cg23270808	0.611382703	OR7E91P	2	71251169	cg02931642	0.530891241		3	184320734
cg24408057	0.610869565	SHD	19	4277333	cg26438705	0.530873757	IFT140	16	1661097
cg09973791	0.610755861	OCA2	15	28015719	cg03306486	0.530872843	APC2	19	1467952
cg15725440	0.610748792	PRDM13	6	100053161	cg20303441	0.530536222	DNMT3A	2	25496390
cg11559162	0.610592991		13	113000664	cg17171962	0.530508605	HIC1	17	1958164
cg13969265	0.610531307	KCNK9	8	140716673	cg10320410	0.530462162		7	64343315
cg26058778	0.610516142		10	133376617	cg27538751	0.530414394		16	24586390
cg02229097	0.610259136	SNN	16	11761951	cg22838050	0.530384086	MED16	19	872690
cg13502474	0.610258033		2	88583557	cg08980987	0.530291514	BCAT1	12	25054905
cg03850936	0.609973829		6	168216524	cg24042517	0.530000684	HOXB9	17	46704410
cg16397968	0.609957037	PFKP	10	3164740	cg03929570	0.529878832		16	70484978
cg14793596	0.609853723	MXRA8	1	1290092	cg02020616	0.529852318	ZNF362	1	33744259
cg01363474	0.609541063	PPP2R2C	4	6344484	cg05710090	0.52982573		12	34442705
cg06349450	0.609293861		16	85867170	cg20165381	0.529431372	CD164L2	1	27709912

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg01285435	0.609721415	VP528	8	145654565	cg13348530	0.529403278	SORCS2	4	7283184
cg18810947	0.609208894	TNPO2	19	12831659	cg01678701	0.529381876		17	8250623
cg15835500	0.609192703		6	32952707	cg03294164	0.529355667	DPY19L1	7	35077725
cg19518104	0.608955461	PRDM14	8	70981345	cg22186263	0.529306812	CREB3L2	7	137570950
cg06073139	0.608890588	SIC38A7	16	58718950	cg16782885	0.529277203		8	21541421
cg21487265	0.608816425	PAX8	2	113992762	cg25796986	0.529218562	F10	13	113783990
cg11721177	0.608737119	EPHA8	1	22919873	cg14175932	0.529143251		14	23018807
cg03844381	0.608661166	ARHGFB7	13	111805922	cg01699727	0.529006623	FUJ36000	17	21903947
cg18890615	0.608466423		13	112161629	cg16480841	0.528925337		12	52430851
cg07037412	0.60844489	CACNA1H	16	1234809	cg10354134	0.528891999	MYBPC3	11	47372960
cg05706517	0.608333333	ARSK	5	94893759	cg04524933	0.528486578	KCNG1	20	49621299
cg14196011	0.608333333	CLTC	17	57759682	cg02534163	0.528463768	ENPP2	8	120650994
cg19666555	0.608319712		4	76865528	cg00257786	0.528409489		1	173991688
cg08841829	0.608118607	ABCG1	21	43638893	cg19113686	0.528247887	RPL24	3	101406347
cg16705744	0.608084174		3	27756039	cg12496800	0.528225793	PCBD1	10	72648987
cg07037750	0.60801665	TMEM121	14	105992115	cg20402552	0.528139115		13	28397359
cg24805396	0.607608696	KLHL3	5	137043431	cg18291331	0.528112584	PC	11	66628290
cg21070081	0.60736715		5	110105162	cg05664039	0.528093849	TRIM15	6	30131755
cg26387458	0.607317478		11	28642652	cg11467141	0.528091248		2	132348705
cg25873514	0.607246377		17	81014413	cg00086113	0.528088771	PTPN9	15	75760823
cg27154418	0.607095711	GP6	19	55526208	cg04987465	0.527881475	RPRML	17	45056797
cg15279616	0.60692777	FRMD1	6	168467854	cg25191041	0.527774958		21	44782470
cg11864326	0.606870281	ABR	17	942320	cg18586983	0.527664483		7	100993034
cg26986928	0.60683854		5	3059243	cg09392615	0.52763624	GREM1	15	33023237
cg03152602	0.606508841	THRA	17	38244572	cg19706320	0.527599931		14	100204528
cg13615998	0.606418434	HOXC10	12	54380956	cg12890750	0.527506173	VTRNA1-2	5	140097950
cg12039197	0.606400966	ACTG2	2	74120092	cg15773886	0.527252032		19	50651024
cg15308271	0.606249695	GCKR	2	27720440	cg14625731	0.527131719		9	38672608
cg02845274	0.606044153	ABCA13	7	48494362	cg10076071	0.527127965	XRN2	20	21284343
cg09724798	0.605549427	LOC144742	12	119741507	cg03494017	0.527081194		5	1724892
cg23036947	0.605463951	MXRA8	1	1294018	cg10750464	0.527014296	CCDC151	19	11534062
cg07844442	0.605301426	DOCK1	10	129144269	cg00525874	0.526985699	SIC6A3	5	1409985
cg11800794	0.605193237	MAPRE1	20	31407338	cg13013644	0.526951077	SIC9A3	5	502571
cg00792783	0.60517182	PLCL1	2	198669748	cg08967106	0.526860242		17	35014412
cg23845206	0.605122941	FAHD2B	2	97760726	cg14786686	0.526454738	ADARB2	10	1713335
cg13993183	0.604559021		9	134609541	cg00158539	0.526385313	MVT1L	2	1859035
cg26113972	0.604384537	RPH3AL	17	104350	cg11310341	0.526357178		12	38532129
cg15720089	0.604191128	CACNB2	10	18550223	cg10948795	0.526340723	RIMBP2	12	130918994
cg03838806	0.604073062		4	6477277	cg00232160	0.526231235		9	129468157
cg01854228	0.604008568	LOC646762	7	29690023	cg23081604	0.52614375	PWILL1	12	130824224

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg24690094	0.603632665		11	67383802	cg08757611	0.526010061		17	76250500
cg05969021	0.603425872	ID3	1	23886005	cg14366742	0.529963742		4	8546989
cg06809298	0.603254812	C0X7A2	6	75953853	cg15814923	0.525751557	PRACA	19	14228610
cg11967675	0.603987028	MYK	12	110011641	cg11738976	0.52595206	JPH2	20	42744590
cg13997975	0.60298466	ZNF503	10	77158502	cg20982046	0.525362804		8	74282931
cg20737388	0.602877929	DNAJB13	11	73668626	cg18755531	0.525154974		5	110729987
cg22234479	0.602486576	TUBA3C	13	19756050	cg11594731	0.524873841	PTH1R	3	46940212
cg17608082	0.602294686	SEMASA	5	9119047	cg07673807	0.524836651		17	12526662
cg14837792	0.60215509		14	106091981	cg09368670	0.524772132	PRP388	1	109234826
cg06950634	0.602126741		11	304351	cg07707498	0.524726274	CL6orf87	16	46865555
cg07587706	0.601959056	C19orf43	19	12845012	cg00853714	0.524533354	BAT1	6	31510077
cg03736247	0.601932367		11	115044192	cg16474696	0.524410124	MRI1	19	13875014
cg14853771	0.60184132	PSMD1	2	23191295	cg08625990	0.524181311	LOC154449	6	170571794
cg11347582	0.60178464		6	168502341	cg25367084	0.524062784	SLC2B1	11	74870452
cg18377044	0.601692312		4	6568049	cg19870512	0.523863753	KCNA6	12	4919081
cg09841001	0.601489659		17	79455736	cg13494933	0.523666418		1	148864629
cg19726408	0.601366226	SH3PX2A	10	105428651	cg02354839	0.523516793	BT3	5	72794086
cg19639530	0.601242415	ZNF777	7	149158187	cg21028319	0.523513558	TM2D2	8	38847958
cg23634348	0.60103186	DGKQ	4	967822	cg06483699	0.523366229	LOC256880	4	100871150
cg03085224	0.600823445	DAXX	5	33290949	cg05320642	0.523211385		2	3063448
cg08515845	0.600818366	CDK5	7	150755180	cg13697193	0.52311706	GPR19	12	12849138
cg13720737	0.600443632		13	113611494	cg09556994	0.523036324		2	216072673
cg07089161	0.600303144		7	153287529	cg18762806	0.522857568		8	48676000
cg01923252	0.600230714		6	108436849	cg07350977	0.522757356	CDH9	5	27038836
cg02841482	0.600940466	VEPH1	3	157217803	cg17215843	0.522478899	HERC3	4	89513558
cg06146665	0.599917561	USP36	17	76799964	cg12510614	0.522474553	PTPRN2	7	157782916
cg17682441	0.59958168	CGREF1	2	27342146	cg00300879	0.522366291	CNKSR1	1	26503847
cg02974499	0.599508847	FAU	11	64889668	cg00773700	0.522345358	ZC3H3	8	144557035
cg07936109	0.599033816	NXN1	19	17566626	cg00677217	0.522312251		4	3747401
cg09127314	0.598879996		1	152161683	cg01415527	0.522219489	SBNQ2	19	1110591
cg19623012	0.5988696		11	116371330	cg24804544	0.522207043	GRH2IP	7	6544107
cg09464728	0.598847333	CYGB	17	74533239	cg11415705	0.52208743	CLorf122	1	38273375
cg05146307	0.598802578		16	29227937	cg12062537	0.521994018	SSU72	1	1496856
cg05419622	0.598800677	NRN1	6	6001380	cg03114558	0.521981844	CRYM	16	21289539
cg03998338	0.5987698	CES4	16	55794910	cg07908755	0.521976757	COB1	7	51385729
cg03586390	0.59859548	NRXN2	11	64428252	cg22447799	0.521935928		1	1855324
cg13002506	0.598594665	GDPD4	11	76998821	cg11447335	0.521817188	BRD9	5	865350
cg19282259	0.598462895	NCRNA00200	10	1205611	cg13902645	0.521734658		11	5959945
cg12869334	0.598433203	GPR124	8	37699360	cg18558455	0.521245448	CUX1	7	101559167
cg10724529	0.598429952		6	16959014	cg26932586	0.521156736	DYNC1H1	14	102514382

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg03550075	0.598361803	SFRS8	12	132278448	cg00806481	0.521126806	PRDM16	1	2996650
cg12537405	0.59803305		10	134898881	cg14955916	0.521105758	DIP2C	10	459968
cg10817497	0.597873457		2	234261386	cg18165381	0.521010165		3	44552316
cg11757417	0.597769566		7	98424289	cg26504835	0.52098631	PSORS1C3	6	31143710
cg07911905	0.59774618	CHST11	12	105096644	cg12853648	0.520618499	BA11	8	243609917
cg19821361	0.597690592	KRTAP10-6	21	46012520	cg13135459	0.52044836	LOC387646	10	27541298
cg06746318	0.597584541	CUX1	7	101755186	cg02289741	0.520141957	PAK4	19	39616340
cg20873718	0.597383917	LFNG	7	2559423	cg16059665	0.519572686		5	81576371
cg11008123	0.597369989	LOC283267	11	33097335	cg17849733	0.519443426	PSORS1C1	6	31082200
cg15576576	0.597292711	ZNF331	19	54040818	cg16419441	0.519338111	DPYS	8	105478683
cg16645815	0.596980676	INPP5A	10	134556992	cg26754472	0.519169578	EMR1	19	6926394
cg12280692	0.596977302		13	113107959	cg02937313	0.519096375	THAP1	8	42698971
cg27561954	0.596649439	RBP7	1	10057312	cg08886301	0.519076759	KIF26B	1	245710401
cg18944383	0.596573898	ENPEP	4	111397179	cg08043345	0.518945679	REEP6	19	1496580
cg26162794	0.596473901	CCDC144B	17	18528568	cg19731340	0.518737931	SPRY4	5	141704709
cg18480974	0.596323008		14	104668827	cg15034757	0.518702198		3	186617939
cg11615395	0.596276606	MAML3	4	140925869	cg01989857	0.518661649	DYNC2L1	2	44001018
cg14383174	0.596132686		17	77896174	cg17179051	0.518486226	NR1H4	12	100953546
cg22171626	0.595805374	ACLY	17	40073284	cg03372334	0.518475295	MBNL2	13	98002933
cg02139338	0.595727388	STK32C	10	134108333	cg17537073	0.518456872		13	24902500
cg19630376	0.595713608		4	26065653	cg20153737	0.51808846		18	77835867
cg13458561	0.595683012	INPP5A	10	134556389	cg06710672	0.518069519		16	1133378
cg26490671	0.595681867	PTPRN2	7	157670732	cg22988655	0.518057518	HLA-H	6	29856070
cg05005586	0.595630181	MYO18A	17	27413207	cg26014401	0.517926274		22	25082111
cg04195581	0.595619349	COX7C	5	85914155	cg17871792	0.517899011	LOC400657	18	72265233
cg13712197	0.5953349682	NCOR2	12	124844874	cg14659662	0.51770203	GLIS1	1	54151053
cg14568768	0.595285748	LCORL	4	18024356	cg15278109	0.517748431		10	29214117
cg10429608	0.595224867		2	238864632	cg09402367	0.517711998		16	1153273
cg11713274	0.595204733		12	54609870	cg12956567	0.517507517	HLA-I	6	30228003
cg07086592	0.595057025	DCAF4L2	8	88886306	cg11452571	0.517368379	PHLDB3	19	43979464
cg17611936	0.595049839	PRKAG2	7	151411526	cg23821359	0.517265246		13	112903015
cg05127899	0.594800324		6	28911474	cg22207602	0.517174336	PAK1	11	77185127
cg05184456	0.594779794	SCRT1	8	145557639	cg12660364	0.516990682	ZMIZ2	7	44788915
cg09373148	0.594647991	ESYT3	3	138152837	cg05616010	0.516880568	ZSCAN18	19	58630089
cg02831900	0.594571305		4	184320640	cg25632672	0.516634461	GTPBP1	22	39101984
cg21161649	0.594503765	AGPAT3	21	45284362	cg09486166	0.516550285	ASCC2	22	30234222
cg08926642	0.594056297	PER3	1	7887455	cg12598340	0.516364004	TMEM54	1	33367416
cg00840412	0.594045703		12	127359199	cg22540324	0.516317142	TRMT61B	2	29092552
cg00675156	0.593986156	TTL6	17	46871867	cg19001789	0.51630989		13	111464554
cg01642579	0.59398553	ANKS3	16	4746990	cg11791078	0.51628275	RANBP3L	5	36273196

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg11735305	0.593624173	DIRC2	3	122512541	cg27611584	0.516157672	ILG12	17	73552146
cg22646995	0.593538795		5	180541779	cg19293748	0.516012278	ZNF611	19	53237645
cg12472483	0.593437687	PTPRN2	7	157996945	cg14612544	0.515787671	ADAMTS14	10	72432573
cg20996561	0.593380197	CACNA11	22	40082316	cg10812016	0.515320126		7	155385272
cg16891165	0.593371554		12	105676993	cg26115312	0.515286961		1	111747236
cg25094735	0.593291733	NAPS8	19	50848024	cg03444077	0.515258496		10	45359721
cg11115431	0.593268149		6	168661137	cg01963754	0.515141723	C13orf16	13	111977248
cg03042666	0.593183677	HDAC4	2	240014781	cg22462726	0.515140373		3	184209261
cg16989719	0.593134222		2	238392110	cg07894004	0.514947113	NTSR2	2	1810683
cg11759446	0.593071341	BAG4	8	38034186	cg10894690	0.514812312	LOC728875	1	144521462
cg03704061	0.592854003	CAPN9	1	230883300	cg0067824	0.514801718		10	23492914
cg00473416	0.59277079		3	127176006	cg03508928	0.514448477	KIF9	3	47324552
cg02814135	0.592512746		11	119887950	cg00510330	0.514406175		6	39916111
cg02883229	0.592302397		7	155616337	cg09826364	0.514201448		7	158789723
cg17062305	0.591994849	TMEM132C	12	129127997	cg24942922	0.514153137		3	120003547
cg00599530	0.591837354		6	170589809	cg21740826	0.514110621	PIAS4	19	4037881
cg00169964	0.591833749	PCDH21	10	85974172	cg21994712	0.514035071		19	21861136
cg23662927	0.591819171	TNFAIP8L1	19	4639377	cg11631644	0.514030396	C21orf29	21	46131923
cg16820615	0.591815319	RASA3	13	114884918	cg15904664	0.513994267	ASPG	14	104569653
cg23034788	0.591810701	MYBPC3	11	47364745	cg12984948	0.51386653		12	76283091
cg25588969	0.591696981	GINS1	20	25384199	cg06343669	0.513649373	GALNT9	12	132690782
cg13354250	0.5913632		6	169558175	cg05517610	0.513431442		8	142626593
cg04482943	0.591290737	TMEM199	17	26684713	cg02128655	0.51342894	MYO9B	19	17186722
cg00172803	0.591232211	SLC38A8	16	84076941	cg00343414	0.513366446	C11orf87	11	109292465
cg09925166	0.591193695	CDC34	19	540049	cg00108944	0.513205518	GLIPR112	12	75784855
cg01788773	0.59116244		8	1113662	cg24515054	0.513199551		13	19240493
cg15862680	0.590996417	CASKIN2	17	73512560	cg11921736	0.513057202	WDR52	3	113160437
cg06033559	0.590931273	CPPED1	16	12897810	cg24632646	0.512925999	SLC9A3	5	506343
cg11471805	0.590809585	SHISA6	17	11461316	cg07450214	0.512900605	ANKRD26	10	27389879
cg12549600	0.590776075		1	113287034	cg25996001	0.512834662	CYFIP1	15	23003087
cg24040576	0.590664899	SMYD3	1	246670459	cg14538238	0.512549397	GLDC	9	6646386
cg13617192	0.590464617		12	34534365	cg25830305	0.512471493	TNNI2	11	1859381
cg13337350	0.590457315	CD44	11	35239907	cg03326715	0.512461895	MVT11	2	2328391
cg15742848	0.59045641		2	169769501	cg08543278	0.512279689		7	1295706
cg27363327	0.590280329	TTRK1	6	43211208	cg04039119	0.512249586	GRIP1	13	113987285
cg26576206	0.590224123	ABCA7	19	1064038	cg20467929	0.511926185	KIF19	17	72321958
cg22334962	0.590128135	DPH2	1	44435615	cg22986662	0.511923147	C3orf21	3	194790434
cg22700015	0.589955988		1	228743131	cg17459023	0.51180104	PCDHA2	5	140188392
cg13175861	0.589941855	SP3	2	174829296	cg02016328	0.511327889	RAB18	10	27793124
cg26654790	0.589931549	SIFN13	17	33775475	cg21974350	0.511299351		13	112616931

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg03456771	0.58991879		7	56242072	cg09463917	0.51116916	SCARF2	22	20784958
cg04817271	0.589918666	LOC401431	7	149571094	cg11045331	0.510888192	AGAP1	2	236412034
cg16712549	0.589820752	LOC349114	7	39773156	cg04253276	0.510830036	LTBP3	11	65314021
cg07888205	0.589773346		16	34257432	cg25679870	0.510707237	INSR	19	7293776
cg17932096	0.589754661		4	58060773	cg08242633	0.510660556		8	2591411
cg23440882	0.589724624	SAMD11	1	875880	cg18815765	0.510514076	SEC14L1	17	75136729
cg06609049	0.589613527	THOP1	19	2785107	cg00871381	0.510470333	ATCAY	19	3879827
cg05301470	0.589586547		6	28558006	cg00598429	0.51041989	SP6	17	45928750
cg21284119	0.589462415	WRNIP1	6	2765000	cg14952449	0.510392067		5	1154998
cg03687084	0.589235356		2	162364203	cg00893875	0.510315448		10	91597593
cg11950805	0.589089138	CDC20	1	43824379	cg06042504	0.510230259		8	55087323
cg25800500	0.589049009	ANKRD23	2	97505275	cg19090533	0.510149236		12	133536292
cg20959174	0.588984531	TRIP12	2	230785180	cg13906646	0.509995487	RPS6KA2	6	166892921
cg10711039	0.588780751	CEP68	2	65283143	cg22044566	0.509953357	FAM84A	2	14772312
cg13414629	0.588755867	FREQ	9	132935646	cg20070631	0.509813199	CHN1	2	175870490
cg26685375	0.588705679		11	120039029	cg05319515	0.509730735	CALHM2	10	105212620
cg11173002	0.588623898	SLIT3	5	168302634	cg03731646	0.509718214		18	74499372
cg21724239	0.588583498		8	58056113	cg20211711	0.509701076	C17orf61	17	7307397
cg23289581	0.588579588	RTN4RL1	17	1840546	cg07970752	0.509574376	MSLN	16	817463
cg20826526	0.588532841	SSR3	3	156266748	cg00027499	0.509370628	PHKG2	16	30759507
cg00489213	0.58835806	FAM120B	6	170686956	cg05516272	0.509319864	ZDHHC24	11	66311334
cg16716449	0.587997846	A2BP1	16	6976709	cg18107355	0.509283966		10	37940899
cg14557787	0.587937147	AP2A1	19	50305342	cg13007784	0.509105975	SEC14L4	22	30901249
cg22862357	0.587922705		6	32774788	cg03827835	0.509034591		5	134388224
cg21067780	0.587901603		6	887772	cg22226438	0.509032624	FLJ39609	1	854766
cg04483506	0.587769667	ID3	1	23886151	cg15511186	0.508826818	KCNK7	11	65363744
cg18133957	0.58761968	APC2	19	1450493	cg15782228	0.508752684	LAMA5	20	60932415
cg19279342	0.587594944	CFD	19	863054	cg25792518	0.508749202	ACS2	17	48545950
cg24550644	0.587463387	MYO1D	17	30846204	cg15833099	0.50872422	GPR39	2	133402999
cg10450421	0.587431819		16	68564110	cg26224173	0.508543216		13	29329153
cg16523141	0.586829433		16	25271956	cg05877129	0.508428575		22	39784769
cg13462219	0.58670428	C8orf73	8	144649510	cg00431813	0.50805966	C7orf50	7	1051703
cg12417704	0.586633151	KIAA0495	1	3659656	cg17414900	0.508055092		20	56293634
cg10730497	0.586329233	RAB18	11	66036017	cg03175771	0.508022981	PRR14	16	30661534
cg02548132	0.586322064	ANGPT2	8	6420242	cg10073723	0.507686515	BMPER	7	33944213
cg11383165	0.586252836	SSC5D	19	56002283	cg22535103	0.507625947	C8orf71	8	58192502
cg10435123	0.586178999	HVCN1	12	111127010	cg16021939	0.507601697	LOC284379	19	54106876
cg06582782	0.586175303		12	5621261	cg22286906	0.507562441		12	132340356
cg03976326	0.586169878	SDK1	7	4231054	cg13172906	0.507402979	ARC	8	143694295
cg18821144	0.586120316		12	68881398	cg22722822	0.507107128	CHIC2	4	54931066

Appendix A.xlsx

ILMNID	mean cv auc	GENE	CHR	MAPINFO	ILMNID	mean cv auc	GENE	CHR	MAPINFO
cg22213386	0.586065741	NIPAL4	5	156887005	cg25752163	0.507041489	CUX2	12	111470657
cg02925162	0.585996917	SOX8	16	1031205	cg15123035	0.506824785	ZNF642	1	40943211
cg07831351	0.585807451		20	3229402	cg25442600	0.506626223		5	1888033
cg21952685	0.585739046	FAM63B	15	59063274	cg13798431	0.506605834	NOTCH1	9	139402542
cg05890377	0.585602666		2	74357713	cg00744229	0.506450389	RNF25	2	219536566
cg06348651	0.585487358	IFI181	12	110562118	cg03899598	0.506192691	IMID8	16	734426
cg12064069	0.585465939	D4S234E	4	4389509	cg04682135	0.505796934	C3orf55	3	157260764
					cg06426293	0.505570813	SYTL3	6	159151854
					cg15064681	0.505316424	CASQ2	1	116312839
					cg02949992	0.505239426	CDCA4	14	105487248
					cg05269678	0.505007946	FSCN1	7	5643158

APPENDIX B

GENE	GENE	GENE	GENE
XRCC3	MCF2L	SOX8	WNT3
DEGS2	LOC400891	FAM63B	C3orf24
RBMS1	ADAMTS6	IFT81	STK38
BAI1	OR4X2	D45234E	ODF3L2
PACSIN2	BCL3	ATP8A2	CNTFR
NAP1L5	RCVRN	C17orf44	SLC12A8
PDZD7	PARP4	RPRML	MMAA
DDB2	C8orf39	ZNF141	TERT
C18orf62	IGFALS	ZNF460	TSTA3
CTU1	GPR81	ASPHD1	DNM2
SERPINA5	COPG2	RCC2	C6orf174
RAB37	CCDC149	C5orf52	UBAC2
METAP2	NLRP3	DNAH17	WDR27
CCL4L2	KLHL28	RSL1D1	MAK16
BAALC	VAMP5	ATAD3B	HSPB1
CHML	TRIM48	STX2	PHGDH
ITGA11	GNG12	EVISL	GN84
MTUS2	FAM178B	TSPO	CASZ1
MFSD2A	APPL2	NDRG1	AURKC
MT2A	ANXA2	C2orf52	C4orf42
VWA1	MRPS18A	MMP15	SLC38A9
BICD2	ZNF251	RSPRY1	MRPS18B
MYOM2	SPATA21	ZNF83	EBF4
MUC2	SNX33	MORF4	SEC1
FMNL2	FBXO47	SCNN1D	RCL1
MFAP2	DCDC2	TAF10	BMPRI1A
PLEC1	HELZ	ESRRB	PLD6
SERPINF2	INTS1	HCN2	PLCXD3
DCP1A	OPCML	HTR4	ERGIC3
TMC3	NRM	C10orf137	LUZP1
C10orf68	POU2F2	PRPF18	AGPAT5
MIR135B	FAM19A5	SLC6A5	SLC45A4
COL22A1	HLA-DRB5	PTPRE	SON
GPR39	MRPS17	SRRM4	TINAGL1
LCE1C	WSB1	ASAP2	DDX6
WRN	COPS7A	EAPP	IFT140
FLJ11235	CPAMD8	KIRREL3	ASNSD1
GPC1	AMD1	CHRNA4	INHBB
FRK	ADARB2	ENPP7	PPFIA3
OBP2A	TRIM10	PTCHD2	COQ10B
DSCAML1	TMEM145	TBX3	EIF2AK4
GTF3C1	CCDC40	RSPH1	GRLF1
CYB561	DYRK3	LOC440040	LRRC8A
NFASC	EIF1	HRNBP3	NAP1L4
POMP	ING2	HRH2	MIR548N
BICD1	CDC123	LSM1	GSTM5

GENE	GENE	GENE	GENE
ATXN1	FLJ35220	NKAIN1	CDKN1A
MRE11A	SLC18A1	SKP2	TBC1D16
DIP2C	CALCB	CYB5R1	MOBP
SPRR2E	NUP37	CCDC42B	NT5DC2
TSSC1	KIAA0664	SLC25A18	RPS6KA2
ZHX2	FA2H	C11orf49	MLC1
EPHA2	DOCK2	HEATR2	STK19
COMMD9	DNAJB6	ANO1	ALK
STX4	POLD1	GRIA4	PSMA5
PRPF19	NAF1	MUC5B	PSKH2
RPL10A	RPL9	CDH4	NUDT16
IQUB	ESYT2	SLC25A3	SCYL1
BRF1	KRT27	METTL7B	PTCH1
ANO4	ZNF200	CAP2	C1QA
CDSN	ATAD3A	CTDSP2	SGK3
DOCK1	ITGBL1	PSMB8	KCNH2
THSD4	MOGS	PI16	USH1G
COL6A3	MMEL1	EPS8L1	KLK8
PLD1	MFI2	GSTO2	RNF216
C2orf34	C14orf165	TRANK1	SFRS18
DCAF11	MACROD2	FAM177A1	SULF1
LOC652276	NADK	BAT5	CACNG2
CIT	KIFC1	JAG2	PCBD2
PTPRN2	POTEH	RXFP3	ERCC1
ASTN2	LOC285733	TYRO3	MIA3
TRIML2	TMEM18	SRRM1	TNR
LGMN	RREB1	NCR1	TAF4
DKK4	ITSN1	POLS	GCHFR
BBX	PLEKHH3	CCDC17	ZBTB9
ADAP1	MEST	MEOX2	CENPF
GCK	SOX9	ACOX3	TBPL1
PCDH10	WWTR1	SOX30	VENTX
ABT1	ZNF295	PLL2	PLBD1
HLA-DRB1	PRDM8	LHFPL2	MAPK8IP1
CFHR5	COL20A1	SLC6A1	CLDN10
PFN2	CAMK1D	BRD2	IQSEC1
IQSEC3	GGN	PPP1R12C	RDH12
MYO16	SPHK1	FOXA2	AQP11
SFRS9	C2orf70	ST7OT4	GNAS
ZBTB46	ACAD10	SPTBN4	ANKRD2
DAXX	CDH18	JAKMIP3	TEP1
NOC4L	CCDC66	KCNQ1	AFG3L1
DDX42	MUC1	C1orf210	GPR123
NPEPL1	DBN1	FGFR2	CACNA1C
LMF1	IAH1	PRPF8	SLC30A3
NR4A3	TNRC18	MIB2	HSPA18

GENE	GENE	GENE	GENE
PSMD14	KIAA0427	TTC22	HPGD
JPH3	SYT1	PCDHGA1	VWA5B1
LYNX1	C3orf50	ZNF75A	ATP5E
PNPLA5	SH3RF3	SHISA3	EPM2AIP1
BACH2	IGF1R	FLJ42875	SDHA
NVL	PRKAR1B	TNNT3	SYT7
TXN2	EMP2	PNLIPRP2	JARID2
SLC30A9	TUSC4	PCSK4	CUEDC1
INPP4B	OPRD1	NEU4	LY6E
SLC25A37	COL9A3	SOX12	TSLP
SPDEF	TMEM126B	LNPEP	HAUS2
LOC100134368	DAAM2	LASS3	FAM153C
FBLN2	TMEM170A	KCNG1	PRR23B
PROK2	ACTR3B	D2HGDH	DDX60
RRM1	KRTAP6-3	CBLB	ATP6VOC
SHQ1	C21orf56	STUB1	ESPNL
ADCY9	RBM6	KLHL29	DLGAP4
FOXJ3	MAPKSP1	ODC1	KIF21B
SLMAP	MOSC2	EXD3	MATN4
TLE4	KCND3	HOXC9	TBX4
ROR2	CHRNE	ZNF283	NEU1
RIPK2	PHLDA2	MLF2	CHD5
RP1	ACOT11	MATN1	RAB11FIP4
TMEM9	ARL17A	B3GALT4	LOC400657
HMHB1	ALDH1L1	JUB	SLC24A4
C16orf67	TESSP1	C1QTNF7	HIST1H2A8
GPD1	FGF12	SPACA1	H19
ESPN	ASAP3	RP57	BAT4
MYH15	GML	ZFP42	ICAM4
PAC52	LOC100130872	CERK	C17orf62
ZNF828	NCAPH	ZNF555	LRRC27
KCNK7	SAR1A	BHLHE41	UHRF1BP1L
TNX8	FAM38A	CACNG6	ZNF167
PTP4A3	SHANK1	LOC643719	NFIX
KIAA0513	SORCS2	OSBPL3	PRKCZ
YEATS2	GPR137B	C2orf79	PILRA
ZZEF1	PGBD5	ANKRD26P1	RASGEF1C
DET1	AGAP1	MGC23284	DNMT3A
BAIAP2L1	C17orf103	TNFR5F11A	HIC1
DRD4	SNRPN	MIR1306	MED16
BIN1	S100A7L2	SCAP	BCAT1
LSM5	LYSMD4	ZCCHC24	HOXB9
NETO2	UBA52	FIBCD1	ZNF362
PITRM1	CDH17	AFAP1	CD164L2
CCDC27	WNT5A	CPZ	DPY19L1
SH3BP5L	PRC1	CLIC6	CREB3L2

GENE	GENE	GENE	GENE
DMTF1	GREB1	FAM116B	F10
EPHB2	C7orf27	VWA38	FLJ36000
DNAJC9	DTX2	RSC1A1	ENPP2
KIAA0754	AGRN	SMPD3	RPL24
GPR133	TMEM214	CRTC1	PCBD1
CFB	SAMD11	JUN	PC
CSRP2	PLIN5	ZNF311	TRIM15
FHOD3	SHANK2	GTF2H5	PTPN9
BPGM	MCCC1	SCD5	GREM1
MSC	TPO	PAX7	VTRNA1-2
AOAH	GDAP1	REST	XRN2
AIP	SBK2	TMEM11	CCDC151
SLC12A5	DEAF1	B3GNTL1	SLC6A3
TMPRSS3	SMAD5	RASSF8	SLC9A3
CDK2	COX6A2	BAHCC1	MYT1L
TYMP	MAD1L1	PPP5C	RIMBP2
ALLC	CLDN18	AURKB	PIWIL1
GOLGA3	CTAGE1	RNLS	PRKACA
SASH1	KRTAP10-11	CPEB1	JPH2
ZNF226	TRIM27	C6orf89	PTH1R
PHACTR4	KIRREL2	SCARNA16	PRPF38B
SLC20A2	CALB2	HAPLN3	C16orf87
DHX32	C22orf27	LHX3	BAT1
NRAP	GNA12	HBXIP	MRI1
PRR3	GRAMD4	CA10	LOC154449
SLC17A9	WDR73	CNST	SLCO2B1
BOLL	OR7E91P	ABCA3	KCNA6
E2F8	SHD	TTLL1	BTF3
EYA4	PRDM13	MIR191	TM2D2
AHNAK	KCNK9	TRIM24	LOC256880
ERICH1	SNN	POFUT2	GPR19
BRCA1	PFKP	TNS3	CDH9
GALR1	MXRA8	PNMAL1	HERC3
CBFA2T3	PPP2R2C	HLA-DPB2	CNKSR1
HYDIN	VPS28	ARHGEF17	ZC3H3
CX3CL1	TNPO2	HSD11B2	SBNO2
CPEB4	PRDM14	TIMM44	GRID2IP
PCGF3	SLC38A7	BRUNOL4	C1orf122
UBR7	PAX8	C5orf56	SSU72
TUBA1A	EPHA8	HOXC8	CRYM
SLC35E2	ARHGEF7	CCDC3	COBL
PRMT1	CACNA1H	FLOT1	BRD9
HDAC4	ARSK	TPM4	DYNC1H1
DHX35	CLTC	RPRD2	PRDM16
ELFN1	ABCG1	NOXO1	PSORS1C3
ABR	TMEM121	LATS2	LOC387646

GENE	GENE	GENE	GENE
ZNF33A	KLHL3	TBL2	PAK4
APPBP2	GP6	SEZ6	PSORS1C1
MXD4	FRMD1	LOC646405	DPYS
MC3R	THRA	ZNF836	EMR1
RFX8	HOXC10	CDK5R2	THAP1
DPP6	ACTG2	NUBP2	KIF26B
TRIM26	GCKR	C22orf45	REEP6
L1TD1	ABCA13	FAM27L	SPRY4
SLC47A2	LOC144742	PRSS21	DYNC2LI1
SLC7A8	MAPRE1	C1orf53	NR1H4
KCNH4	PLCL1	NCAM1	MBNL2
C15orf48	FAHD2B	KCNQ2	HLA-H
AHRR	RPH3A1	RBM8A	GLIS1
C18orf56	CACNB2	ADIPOR1	HLA-L
JAZF1	LOC646762	ABCC13	PHLDB3
YBX1	ID3	JAKMIP1	PAK1
CYB5A	COX7A2	DOT1L	ZMIZ2
TBCD	MVK	TIMP2	ZSCAN18
C18orf32	ZNF503	TRPM5	GTPBP1
GALNT9	DNAJB13	BAIAP3	ASCC2
C1orf186	TUBA3C	TMED1	TMEM54
BNC2	SEMA5A	MT1DP	TRMT61B
ASCC3	C19orf43	KLK9	RANBP3L
GNG7	PSMD1	SEMA5B	LLGL2
KRTAP19-5	SH3PXD2A	PITPNM3	ZNF611
MTERF	ZNF777	KL	ADAMTS14
NCOR2	DGKQ	ZADH2	C13orf16
PIGZ	CDK5	MRPL55	NTSR2
CAPZA1	VEPH1	PDGFRL	LOC728875
GCOM1	USP36	KNDC1	KIF9
DUOXA1	CGREF1	ADARB1	PIAS4
C16orf74	FAU	LRIG1	C21orf29
CAMTA1	NXNL1	CCDC127	ASPG
LMAN2L	CYGB	DGAT2	MYO9B
TMEM151A	NRN1	C6orf186	C11orf87
TMEM189	CES4	C7orf33	GLIPR1L2
COL11A2	NRXN2	KCNG2	WDR52
KIAA1671	GDPD4	SIRT2	ANKRD26
DEFB114	NCRNA00200	ST8SIA2	CYFIP1
NR2F1	GPR124	PDE4A	GLDC
PPP1R10	SFRS8	JAM2	TNNI2
SNPH	CHST11	HBA1	GRTP1
STYK1	KRTAP10-6	DNAJC21	KIF19
OCA2	CUX1	UHRF1	C3orf21
TDRD9	LFNG	DNAJA4	PCDHA2
GP5M3	LOC283267	NEGR1	RAB18

GENE	GENE	GENE	GENE
TGFB1	ZNF331	NDUFS6	SCARF2
KIAA1530	INPP5A	BTN2A1	LTBP3
NKAIN4	RBP7	SPG7	SEC14L1
GABBR1	ENPEP	SCARNA2	ATCAY
C17orf63	CCDC144B	RPTOR	SP6
CXADR	MAML3	FLJ35390	FAM84A
RBM28	ACLY	GATAD1	CHN1
AFF1	STK32C	LOC399815	CALHM2
MOG	MYO18A	LOC440356	C17orf61
IDUA	COX7C	TFAM	MSLN
ARFGAP1	LCORL	C1orf198	PHKG2
CCBL2	DCAF4L2	MFSD6L	ZDHHC24
MAL2	PRKAG2	PRPH2	SEC14L4
KBTBD12	SCRT1	EHMT2	FLJ39609
PDLIM1	ESYT3	NPR3	LAMA5
CSMD1	AGPAT3	HAPLN4	ACSF2
PLEKHA2	PER3	YIPF3	PRR14
JAKMIP2	TTLL6	CDKN1B	BMPER
CTDSP1	ANKS3	HES3	C8orf71
CYP11B1	DIRC2	DNAJC15	LOC284379
ST5	CACNA1I	UBE2V1	ARC
EBPL	NAP5B	SEPX1	CHIC2
LOC100190938	BAG4	HES7	CUX2
DLC1	CAPN9	PRKCA	ZNF642
INSR	TMEM132C	PIPSK1B	NOTCH1
CAPN8	PCDH21	TACC2	RNF25
PIWIL2	TNFAIP8L1	CTDP1	JMJD8
LEPR	RASA3	C1orf25	C3orf55
HOXC4	MYBPC3	WBP2NL	SYTL3
TDRD6	GIN51	PCBP4	CASQ2
FAM54A	TMEM199	LONRF2	CDCA4
NPTX1	SLC38A8	C15orf24	FSCN1
SH3BP2	CDC34	CTSH	
ST6GALNAC1	CASKIN2	FRMD6	
MCEE	CPPED1	HOXC13	
PIK3CG	SHISA6	NAV2	
A1BG	SMYD3	GIPC2	
PHF10	CD44	MOV10L1	
B3GAT1	TTBK1	BANP	
C7orf50	ABCA7	ZNF22	
FAM8A1	DPH2	TMEM135	
FAM81A	SP3	TOM1L1	
CABYR	SLFN13	ZACN	
ZNF862	LOC401431	C1orf74	
ALG6	LOC349114	SAA2	
OTP	THOP1	DLL1	

GENE		GENE		GENE		GENE
CEACAM8		WRNIP1		NANS		
LOC145663		CDC20		SNORA14B		
SUCLG2		ANKRD23		MIR496		
NF1		TRIP12		E2F2		
C14orf73		CEP68		KLF11		
EXO1		FREQ		SERGEF		
LYSMD2		SLIT3		ERAP1		
RTDR1		RTN4RL1		HSPH1		
COLEC11		SSR3		SRCIN1		
KLK1		FAM120B		C21orf67		
LRP1		A2BP1		NXN		
C7orf13		AP2A1		C17orf28		
SNTG2		APC2		ZNF570		
PPM1A		CFD		FAM134C		
TRMT11		MYO1D		SNED1		
C8orf34		C8orf73		LOC732275		
IRF1		KIAA0495		RPS16		
PLEKHH2		RAB1B		RGR		
CYP26B1		ANGPT2		TTPA		
CCDC69		SSC5D		USP29		
OXA1L		HVCN1		PDLIM2		
TIMM88		SDK1		ATXN7L1		
		NIPAL4		ACTN3		

APPENDIX C

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
JHU_2.144695321	0.709178744	rs4356466	0.571407606	rs10742717	0.546358054
rs34556593	0.694806763	rs2281597	0.57137398	rs2355988	0.546347608
JHU_12.3475542	0.694565217	rs13027801	0.571295385	rs1961707	0.546325395
rs863002	0.69365942	rs1781653	0.571283622	rs2569824	0.546271704
rs6839051	0.693478261	rs17146982	0.571215608	rs7088954	0.546213565
rs13291305	0.6897343	rs9842777	0.5711694	rs7980248	0.546164114
rs10163171	0.686292271	rs977977	0.571167577	rs2925702	0.546078933
exm159581	0.68321256	rs4748177	0.571167557	rs2766551	0.546077736
2:213836206-T-C	0.682608696	rs17409304	0.571144344	rs1911409	0.546076238
JHU_15.42389006	0.680917874	rs1332340	0.571136589	rs10519806	0.54606166
rs339567	0.679468599	rs1332260	0.571089789	rs1446046	0.546037899
exm2258729	0.674939614	rs1920379	0.571080104	rs1481779	0.546031524
rs4671658	0.674818841	rs4427133	0.571043626	rs17765618	0.546031479
exm2269420	0.674516908	rs2894604	0.571036053	rs7963488	0.545945294
rs939291	0.673913043	rs3922692	0.571017269	rs7163730	0.545935482
JHU_10.2079926	0.673429952	rs2899395	0.571006029	rs10897794	0.545916866
rs131861	0.673128019	rs2084347	0.571001036	rs1429238	0.545879355
rs7783747	0.673067633	rs10192834	0.570868426	rs2624087	0.545877396
rs12193236	0.673067633	rs8075406	0.57081562	rs7946983	0.545872844
rs35125498	0.672826087	rs1166940	0.570720286	rs11054692	0.545858764
rs11689032	0.672705314	rs4632084	0.570591853	rs9879195	0.545856175
rs1168889	0.672644928	rs17635255	0.570497034	rs9964424	0.54582722
rs7900896	0.672342995	rs2985697	0.570487201	rs4750059	0.545774921
seq-tid-19-59546984-C-A	0.671678744	rs12345187	0.570415936	rs2963672	0.545774859
rs6460763	0.671376812	rs236387	0.570408199	rs11956139	0.545713542
rs7045243	0.671256039	rs4559194	0.570386632	rs7230070	0.545689794
rs12201777	0.67071256	rs7863063	0.570274711	rs6073714	0.545605437
rs11090409	0.669806763	rs2737191	0.570252503	rs2694777	0.545577042
rs17798302	0.669202899	rs292557	0.570339801	rs6708139	0.545565572
rs827540	0.668297101	rs7039618	0.570234969	rs1027548	0.545516734
rs258401	0.66763285	rs4304686	0.570145245	rs41492948	0.545499962
rs20692220	0.666908213	rs4782578	0.570017701	rs7300366	0.545491735

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs9649963	0.666666667	rs1696937	0.569986182	rs6998613	0.545479293
rs11264045	0.665536643	rs704102	0.569861417	rs301854	0.545424849
rs11940537	0.665603546	rs6509554	0.569847402	rs720528	0.545375406
rs160632	0.66557971	rs10780460	0.569767885	rs8006730	0.545371566
rs10487320	0.664613527	rs11177183	0.569761551	rs2249152	0.545366298
rs10734819	0.66455314	rs2793783	0.569686924	rs9892725	0.545359852
rs4906939	0.664492754	rs9275523	0.569635917	rs6443219	0.545326387
rs13281284	0.664009662	rs1868419	0.569603007	rs3818100	0.545322078
rs2546275	0.664009662	rs4677766	0.569423127	rs11952802	0.545320366
rs11775949	0.663949275	rs7960156	0.569358143	rs1497427	0.545310772
JHU_1.213096113	0.663949275	rs3793735	0.569348475	rs4795067	0.545307895
JHU_2.156588547	0.663586957	rs4851752	0.56917036	rs4383207	0.54530373
rs11658592	0.663586957	rs10833102	0.5690886	rs4743859	0.54527347
rs2390409	0.663586957	rs6781111	0.569072322	rs6991762	0.545231592
rs34104622	0.662922705	rs149391	0.568989595	rs497471	0.545230646
rs1924381	0.662831678	rs7072877	0.568936168	rs4238859	0.545218451
rs3733051	0.662801932	rs41387252	0.568851156	rs11767371	0.545205696
JHU_13.87733477	0.662258454	rs11672085	0.568822609	rs11184169	0.545202033
rs6925860	0.662077295	rs9345943	0.56876421	rs11638902	0.545171185
rs17247253	0.661896135	rs6547082	0.568751101	rs741192	0.545162233
rs13116304	0.661835749	rs2167061	0.568679779	rs4619255	0.545144062
JHU_9.112406610	0.661835749	rs7152947	0.568662575	rs11222916	0.545140078
rs6478581	0.661835749	rs4758330	0.568418435	rs17073281	0.545137526
JHU_10.44932835	0.661714976	rs4842266	0.568415904	rs17081444	0.545125545
rs2142569	0.661231884	rs3761893	0.568399354	rs2012517	0.545105522
rs10784245	0.661058629	rs885883	0.568371984	rs10050439	0.54509915
JHU_6.46111558	0.660990338	rs4791824	0.568361666	rs256930	0.545074596
rs1015287	0.660990338	rs10512080	0.568359205	rs7852123	0.545039064
exm2257330	0.660929952	rs11240341	0.568354031	rs1749296	0.545038537
rs6021886	0.660869565	rs1885894	0.568328633	rs7652331	0.544967803
JHU_9.16734455	0.660748792	rs6480547	0.568242543	rs6128156	0.544951139
rs469353	0.660088416	rs11038178	0.568214313	rs7012551	0.544928617
rs6089568	0.659903382	rs6961889	0.56818735	rs6461517	0.544875058

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
JHU_X_2608134	0.659541063	rs1462027	0.56809181	rs11036390	0.544839573
JHU_3_41299515	0.65942029	rs2065945	0.567973317	rs2037719	0.5448333
rs25206	0.65942029	rs1005100	0.567882045	rs3981102	0.544791389
rs1397199	0.659359903	rs17561584	0.567877884	rs2051930	0.544753696
rs12100690	0.659299517	rs3803761	0.567849415	rs1163860	0.544747663
JHU_8_24865116	0.659178744	rs10512187	0.56772253	rs10503290	0.544742584
rs34569523	0.658574879	rs3846608	0.567600468	rs7198575	0.544691404
rs1869094	0.657971014	rs4767592	0.567554865	rs10894900	0.544679966
rs1558085	0.657732151	rs10894798	0.567523334	rs4732686	0.544667676
rs9908590	0.657729469	rs12357	0.567444739	rs1895941	0.544641179
rs13155975	0.657487923	rs9315310	0.567352558	rs7046236	0.544635549
JHU_13.93702009	0.657487923	rs12198460	0.567347529	rs894911	0.544634762
rs12967485	0.657427536	rs945748	0.567301214	rs1606237	0.544570474
JHU_3_150146401	0.657065217	rs163503	0.567294291	rs7794252	0.54457028
JHU_7_9003738	0.656884058	rs7540760	0.567270551	rs2183004	0.544550138
JHU_3_61554354	0.656763285	rs7799	0.567203916	rs12268030	0.544529798
rs9596775	0.656642512	rs2367119	0.567189589	rs10166909	0.544527661
rs584329	0.656582126	rs4900200	0.567106792	rs11691352	0.5444976
rs4758263	0.656311584	rs10005962	0.567031626	rs2247593	0.544478824
rs2382089	0.656280193	rs2954018	0.566946594	rs7723508	0.544475477
rs4686667	0.656219807	rs1583366	0.566901283	rs16936133	0.544429751
exm1095171	0.656219807	rs1479923	0.56688896	rs918425	0.54440418
rs3815550	0.65615942	rs11731479	0.566840468	rs11638981	0.544400747
rs12299455	0.655857488	rs2042586	0.56679761	rs10456045	0.544355605
JHU_3_118307706	0.655857488	rs10434594	0.566738215	rs11570190	0.544311431
rs2840795	0.655374396	rs4746989	0.566545444	rs2299546	0.544206688
rs941904	0.65531401	rs4752824	0.566517441	rs9868373	0.544201474
rs17087253	0.65531401	rs7147119	0.566466533	rs11071118	0.544193007
rs2073425	0.655012077	rs5473525	0.566457404	rs7554095	0.544171156
rs677011	0.654589372	rs2617101	0.566424403	rs10751506	0.544146647
JHU_6.867058	0.653985507	rs1890112	0.566319484	rs269957	0.544127068
JHU_10.18505522	0.653743961	rs17779700	0.566317586	rs3731843	0.544125986
rs4436895	0.653502415	rs153226	0.566270337	rs10793294	0.54402782

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs1401935	0.653502415	rs16970500	0.566259793	rs9614255	0.54402454
rs6427327	0.653381643	rs735956	0.566182515	rs978979	0.544008793
rs7662632	0.65326087	rs646610	0.566182078	rs876347	0.54393966
rs1144430	0.653019324	rs4418897	0.566163158	rs10040260	0.54385564
rs7209678	0.652958937	rs889162	0.566027934	rs1465514	0.543834539
rs2345408	0.652898551	rs4978125	0.565964349	rs4593008	0.543792566
JHU_14.95570691	0.652475845	rs667335	0.565947228	rs2025804	0.543778095
rs7806365	0.651932367	rs8031751	0.565790548	rs934778	0.543689972
1:33906788-G-T	0.651932367	rs16874525	0.565784218	rs9397338	0.543687128
JHU_17.43803188	0.651690821	rs12437164	0.565555368	rs7593050	0.543644204
rs695317	0.651680851	rs2011170	0.565555216	rs10502069	0.543571071
rs9651740	0.651559338	rs11221510	0.565548564	rs836472	0.543568544
JHU_5.146827430	0.650845411	rs12427533	0.565545976	rs111013565	0.543545829
rs1448547	0.650724638	rs10819808	0.565523915	rs11666543	0.54347947
rs782750	0.650362319	rs11005606	0.565476143	rs2176377	0.543459002
rs4872450	0.650301932	rs10865499	0.565451218	rs1352516	0.54340844
JHU_17.37037770	0.650301932	rs1041566	0.565430791	rs1895385	0.543398019
JHU_5.114409639	0.650301932	rs1899734	0.565430787	rs826057	0.54333187
rs2416451	0.650120773	rs4743294	0.5654062	rs7162520	0.543266688
rs1992599	0.650120773	rs10488719	0.565335025	rs2166840	0.543197189
rs1889486	0.650068322	rs17502818	0.565329621	rs1473160	0.543193892
JHU_13.65299557	0.650060386	rs1680726	0.565272222	rs6440072	0.543183726
rs449114	0.65	rs8132937	0.565268971	rs1512130	0.543179775
rs336206	0.649879227	rs12708823	0.565219396	rs983741	0.543145001
rs12418581	0.649558629	rs7327483	0.565202513	rs1441951	0.543131406
rs62293654	0.649335749	rs8001719	0.565199455	rs669116	0.543119055
rs1973655	0.649275362	rs990395	0.565167747	rs441463	0.543007871
rs4889107	0.649214976	rs4745072	0.565079941	rs2459020	0.542990552
rs4710024	0.649214976	rs35146	0.565030556	rs372543	0.542979522
rs7624691	0.649154589	rs989430	0.564952928	rs10736313	0.542950497
JHU_10.19572206	0.648913043	rs1401404	0.564912322	rs4885005	0.542874758
rs10512177	0.648728369	rs8083533	0.564896855	rs12712638	0.542860518
rs2636683	0.648550725	rs4906974	0.564894599	rs1808529	0.542848319

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs10513799	0.648490338	rs2620441	0.564884302	rs1792137	0.542839066
rs32261764	0.648429952	rs4517810	0.564871148	rs702279	0.542800905
rs4777110	0.648369565	rs2181522	0.564851552	rs2046510	0.54276846
JHU_1.208529470	0.648369565	rs2836757	0.564849844	rs7085788	0.54275992
JHU_11.93578004	0.648188406	rs7518010	0.564786445	rs10946988	0.542747634
rs11935705	0.648067633	rs4328484	0.564709282	rs966878	0.54274065
rs12790256	0.648067633	rs320947	0.564700108	rs1837253	0.542692907
JHU_13.86954690	0.647705314	rs2151037	0.564697182	rs17767419	0.542673357
rs2842854	0.647612766	rs17154652	0.564654974	rs10516462	0.542658894
rs2357486	0.647524155	rs17315253	0.564592143	rs816850	0.542623383
JHU_16.59222039	0.647463768	rs778809	0.564573552	rs6803827	0.54261204
rs2503044	0.647342995	rs17477583	0.56456574	rs4977400	0.542591111
rs748919	0.647161836	rs3005774	0.564559826	rs5750781	0.542586423
rs10496992	0.647101449	rs2162019	0.564542959	rs10741246	0.542585118
14:104179267-T-C	0.647041063	rs11137568	0.564507518	rs10604	0.542550896
JHU_12.130668362	0.64692029	rs12457970	0.564432697	rs16957962	0.542527851
rs10517155	0.64673913	rs1990877	0.56438758	rs444556	0.542519595
JHU_2.239478852	0.646618357	rs11534043	0.564350279	rs2868371	0.542515068
rs17172610	0.646617967	rs11255142	0.564346381	rs7489433	0.542471457
rs323593	0.646557971	rs1532365	0.564335861	rs1851062	0.542439959
rs7725091	0.646533097	rs34104149	0.564286307	rs10508561	0.54241151
rs4810056	0.646497585	rs5992962	0.564178856	rs11251508	0.54240544
rs2978608	0.646497585	rs2042726	0.564157866	rs761676	0.542390853
JHU_14.44919016	0.646437198	rs274717	0.56415324	rs12143872	0.542357227
rs12077113	0.646432861	rs7217216	0.564089614	rs1317057	0.542354927
rs3771598	0.64643026	rs1353581	0.564068845	rs1786566	0.542347531
rs1048608	0.646316425	rs4391483	0.564049312	rs10884651	0.542300072
rs756928	0.646305201	rs216855	0.56399365	rs4851262	0.542272841
rs13192959	0.646135266	rs12510308	0.563980142	rs2839734	0.542258535
rs4866126	0.646014493	rs7313354	0.563959798	rs10853670	0.542255473
rs11021976	0.645833333	rs1463143	0.563927612	rs4974496	0.542153204
rs2702600	0.645728369	rs741194	0.5639101	rs2221480	0.542066999
rs10928469	0.645728132	rs7767991	0.563872699	rs1031287	0.5420163

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs2277547	0.64571256	rs13287948	0.563865987	rs542610	0.541993968
JHU_2.116844732	0.64571256	rs1006737	0.563830167	rs17767298	0.541970463
rs10903713	0.645591787	rs4235576	0.563817021	rs1968285	0.541969924
rs13096777	0.645471014	rs6959554	0.56373722	rs6811329	0.541957597
rs17790352	0.645363357	rs13134673	0.563727264	rs160380	0.541942348
rs17816780	0.645169082	rs1558322	0.563718728	rs3757340	0.541875355
rs2400000	0.645139007	rs331603	0.563716551	rs6074514	0.541851797
rs7653853	0.645048309	rs1159148	0.563715514	rs2026739	0.54181098
rs2874792	0.644927536	rs9646114	0.563648957	rs4293757	0.541808318
rs1560233	0.644874941	rs1163249	0.563646932	rs9807802	0.541788019
rs471722	0.64486715	rs11596076	0.56363684	rs980160	0.541779141
rs4059236	0.644806763	rs1243188	0.563615438	rs4662667	0.541724759
rs26784	0.644746377	rs9644902	0.563526233	rs9569549	0.541655449
rs4976606	0.644739007	rs7559750	0.563522203	rs6464644	0.541647078
rs11104658	0.644639953	rs7106757	0.563501361	rs2113545	0.541644769
rs11136250	0.64465957	rs1291828	0.563496225	rs10840269	0.541615604
rs7924795	0.644323671	rs4619182	0.563495829	rs1859716	0.541583983
rs7275012	0.644323671	rs9898388	0.563484523	rs10125991	0.54152446
rs7152071	0.644202899	rs10839663	0.563468011	rs587527	0.54151797
rs2793700	0.644202899	rs8023364	0.563456273	rs4661310	0.541452817
JHU_14.101807547	0.644202899	rs4507535	0.563401002	rs6775519	0.541388252
JHU_13.31564483	0.644202899	rs2028355	0.563382493	rs1673667	0.541374884
JHU_11.77906385	0.644142512	rs3809618	0.563365729	rs6931355	0.541358849
rs12154311	0.644082126	rs10070001	0.563326144	rs2336384	0.541331479
rs7610221	0.644082126	rs2117383	0.563204428	rs17804441	0.54131786
rs246330	0.644082126	rs17210088	0.563200353	rs10968202	0.541302348
rs9538998	0.644070922	rs11859365	0.563196159	rs13210587	0.541295153
rs3018058	0.643780193	rs10152504	0.56315431	rs1867523	0.541216694
rs9810211	0.643780193	rs16889280	0.563112033	rs1475793	0.541210105
JHU_6.32900615	0.643599034	rs7958163	0.563106118	rs8009711	0.541125707
rs2189517	0.643599034	rs4469949	0.563049123	rs1546918	0.541119167
rs2079674	0.643538647	rs826969	0.562979797	rs9394575	0.541093208
rs4747808	0.643478261	rs6590700	0.562846182	rs9377361	0.541068065

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs1033394	0.643357488	rs9680422	0.562832542	rs6482683	0.541040139
rs7720449	0.642885106	rs2113404	0.562826862	rs11732212	0.541016807
rs2963327	0.642874396	rs13118105	0.562816186	rs4704484	0.540999327
rs8127751	0.64281401	rs7143791	0.562783461	rs12162232	0.540927552
rs2209651	0.642753623	rs8038746	0.562719066	rs11680933	0.540902919
rs10774959	0.642753623	rs1483179	0.562715946	rs7748981	0.540818698
rs355466	0.642753623	rs6075996	0.562608643	rs284386	0.540794867
rs9948487	0.642589125	rs10929194	0.562541885	rs9293683	0.540762098
rs11228984	0.642391304	rs16869089	0.562519257	rs2989426	0.540735468
JHU_18.1934018	0.642391304	rs7979054	0.562476609	rs2677879	0.540635801
rs17009311	0.642378251	rs33967759	0.562440555	rs10742265	0.540622893
rs2591605	0.642270531	rs1866732	0.562412448	rs745696	0.54062163
rs10791311	0.642270531	rs11121382	0.562373879	rs12822139	0.540612299
rs13163149	0.642149758	rs6565965	0.562363989	rs1811302	0.540583077
rs4698186	0.642028986	rs509342	0.562335598	rs11194423	0.540539158
rs10503852	0.642021513	rs6007211	0.562330042	rs9806973	0.540520052
rs2838055	0.642007329	rs585124	0.562271845	rs17089525	0.540509784
JHU_5.101138144	0.641908213	rs1015020	0.562196674	rs737605	0.540494884
rs4786738	0.641908213	rs10512395	0.562185018	rs11710743	0.54048671
rs735358	0.641847826	rs624613	0.562178112	rs7991300	0.540475343
rs10811222	0.641828369	rs7631577	0.562127537	rs7929370	0.540464358
rs12425616	0.64178744	rs7937849	0.562119486	rs1367819	0.540462403
rs79008	0.641765721	rs7244595	0.562058894	rs10873925	0.540431839
rs4961625	0.641727053	rs237894	0.562013489	rs4898851	0.540428888
rs726863	0.641666667	rs878554	0.561994959	rs10510829	0.540394393
rs17050272	0.64160628	rs1433394	0.561968869	rs4905614	0.540347622
JHU_4.187085944	0.641485507	rs9390201	0.561966054	rs6793582	0.540332434
JHU_20.59746811	0.641304348	rs4986223	0.56196561	rs523329	0.540125711
rs4913487	0.64130331	rs8041060	0.561886521	rs6473667	0.540115603
rs2721764	0.641282033	rs1095290	0.561879182	rs963666	0.540107252
rs394307	0.641267849	rs1560061	0.561871757	rs4261154	0.540034666
rs2785079	0.641249409	rs562918	0.561732092	rs7786287	0.540029813
rs7131784	0.641062802	rs6105944	0.561720782	rs36453	0.540027628

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs7338029	0.641002415	rs2968845	0.561613476	rs27295475	0.539961763
rs12492560	0.640994563	rs7964182	0.561573087	rs3865014	0.539907961
JHU_9.112985385	0.640942029	rs6434355	0.561566366	rs8089366	0.539897054
rs12158564	0.640942029	rs10007790	0.561493529	rs941684	0.539890823
rs2211332	0.640904255	rs844395	0.561426117	rs1229469	0.539792843
rs12916632	0.640821256	rs12692519	0.561420458	rs328094	0.539769811
rs2888659	0.640821256	rs719787	0.561305257	rs630075	0.539745256
JHU_2.225182399	0.64076087	rs4910755	0.561287011	rs1202592	0.5397417
rs7152431	0.640458937	rs1532121	0.561277348	rs17512962	0.53967374
JHU_4.176510511	0.640338164	rs12701937	0.561266494	rs944868	0.539656956
rs1931054	0.64033617	rs4739148	0.561228127	rs7269546	0.539654256
rs1158393	0.640217391	rs1655627	0.561197345	rs4360905	0.539604035
JHU_13.40401276	0.640157005	rs8019537	0.561167921	rs16972457	0.539589173
rs12593066	0.640096618	rs7318879	0.561152499	rs7701376	0.53955002
JHU_13.106438798	0.640096618	rs9417900	0.561115992	rs155387	0.53951343
rs976451	0.640096618	rs880953	0.561081568	rs10937329	0.539445697
rs688244	0.640036232	rs6804425	0.561068694	rs4243226	0.539420245
rs9459108	0.639975845	rs6876647	0.560973117	rs41331946	0.539394599
JHU_2.79226943	0.639975845	rs3943387	0.560952604	rs38111	0.539389438
rs12405870	0.639794686	rs1507599	0.56095225	rs17379007	0.539373996
rs1584260	0.6397343	rs1164270	0.560925218	rs7169485	0.539359002
rs7101071	0.639697636	rs6934462	0.560897415	rs1753458	0.539356586
rs1550559	0.639613527	rs10420793	0.560885056	rs10003627	0.539249975
rs6829805	0.639613527	rs4335778	0.560864074	rs4405487	0.539234862
JHU_17.1398269	0.63955314	rs11228258	0.560811202	rs7669431	0.539230618
rs17391190	0.639434752	rs2822577	0.560758146	rs2703403	0.53921134
rs4302458	0.639426714	rs4881109	0.56074877	rs3794989	0.539209858
rs10180220	0.639371981	rs6070529	0.560721383	rs11962613	0.539167182
rs11631886	0.639371981	rs11700764	0.56071095	rs1603530	0.539165733
rs3729617	0.639311594	rs11772635	0.560663396	rs1009170	0.539158082
rs16861196	0.639253664	rs7912500	0.560633561	rs3845265	0.539113672
rs608358	0.639251208	rs4670300	0.560619333	rs2570665	0.539097205
rs7790846	0.639251208	rs797208	0.560597877	rs11212569	0.538982597

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs26698	0.639190821	rs7090752	0.560549105	rs2449837	0.538930985
JHU_18.57715607	0.639190821	rs728340	0.56051785	rs7283473	0.538926791
seq_rs40728	0.639130435	rs7445966	0.560512203	rs293257	0.538898129
rs4389718	0.639009662	rs10447366	0.560476181	rs10060763	0.538896429
rs13100749	0.638949275	rs7260635	0.560467596	rs9686689	0.538870664
rs619662	0.638899054	rs6930903	0.56034246	rs480039	0.538856535
rs8066210	0.638888889	rs1075705	0.560317485	rs238627	0.538854329
JHU_16.86360890	0.638888889	rs17300302	0.560253983	rs11054800	0.538845953
rs6851949	0.638888889	rs10751692	0.560253888	rs7161192	0.538843636
rs4665203	0.638828502	rs12452504	0.560132839	rs315515	0.538716145
JHU_11.120399725	0.638707729	rs2282035	0.560039028	rs1465511	0.538691661
rs1436214	0.638707729	rs12994268	0.560001076	rs2173479	0.538691586
rs12919132	0.638707729	rs11605489	0.559986448	rs747400	0.538672028
rs12957838	0.638647343	rs10106296	0.559979225	rs929713	0.538566339
rs11862419	0.638647343	rs17616128	0.559974439	rs10489777	0.538553741
rs2981309	0.638606619	rs3812389	0.559968405	rs518590	0.538512109
rs2027742	0.638586957	rs11582843	0.55995458	rs12548089	0.53850778
rs17528240	0.63852657	rs582435	0.559949723	rs7560385	0.538441622
rs2422334	0.63852657	rs2078326	0.559887217	rs2024269	0.538362015
rs9423399	0.638405797	rs3750552	0.559865852	rs2146615	0.538293602
rs2502521	0.638405797	rs2530709	0.559858044	rs2939739	0.538259042
rs362836	0.638405797	rs2047375	0.559836724	rs10461909	0.538222543
JHU_18.70535900	0.638345411	rs6888551	0.559820097	rs7599327	0.538211493
rs62300891	0.638345411	rs6730320	0.559806943	rs11062486	0.538197273
rs6997533	0.638345411	rs1472411	0.559804794	rs3129877	0.538143072
rs1248480	0.638224638	rs7044488	0.559749429	rs8061801	0.538132272
rs4745969	0.638043478	rs10090760	0.559749367	rs1017002	0.538132231
rs10146843	0.637983092	rs6023640	0.559713169	rs9287423	0.538128959
rs74291645	0.637862319	rs5564859	0.559697463	rs10148671	0.538071992
rs17726944	0.637807565	rs1268565	0.559695022	rs11234221	0.538041943
JHU_6.5597845	0.637741546	rs7324699	0.559669027	rs4669573	0.537979527
rs1438538	0.637711584	rs2492779	0.559659124	rs1859785	0.537930335
exm2265606	0.637681159	rs6075750	0.559633738	rs7323349	0.537869689

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs7048053	0.637681159	rs1887752	0.559612212	rs11968978	0.537829758
rs3370480	0.637620773	rs409346	0.559552151	rs2424582	0.537828029
rs2184513	0.637560386	rs4734305	0.559414733	rs1196845	0.537814579
JHU_2.56239896	0.6375	rs1333656	0.559390376	rs10485664	0.537765613
rs11044078	0.637439614	rs2877056	0.559377148	rs285014	0.53767763
rs11131163	0.637439614	rs10829661	0.559363026	rs11664246	0.537589309
rs7801660	0.637439614	rs4140979	0.559358244	rs1677703	0.53756545
rs4727625	0.637382979	rs2066711	0.55934779	rs1345390	0.537565142
rs640887	0.637379227	rs11192037	0.559344789	rs583641	0.537552391
rs2436219	0.637318841	rs16939937	0.559314164	rs1246116	0.537531339
rs3202930	0.637207565	rs2298028	0.55919644	rs11060202	0.537528956
rs17210043	0.637077295	rs7080882	0.559177384	rs6595989	0.537503224
rs4910994	0.636956522	rs16862146	0.559131291	rs2263664	0.537498993
rs196548	0.636844917	rs6975557	0.559123397	rs1892110	0.537467132
rs7620170	0.636777305	rs2291183	0.55911403	rs8005917	0.537447648
JHU_22.44292370	0.636714976	rs776771	0.559071654	rs12684544	0.537400753
rs3118242	0.636714976	rs7762756	0.559065295	rs17777132	0.537390818
exm87609	0.636714976	rs2060000	0.559054544	rs1888207	0.537353767
rs9892788	0.636714976	rs12635000	0.559040756	rs599486	0.537328957
rs10964831	0.636554846	rs4842173	0.559002566	rs7820792	0.537311144
rs7527726	0.63647343	rs9858228	0.558968438	rs7972536	0.537288688
rs11688662	0.63647343	rs7192816	0.558966108	rs10044480	0.537269797
rs73073429	0.636352657	rs10991043	0.558937854	rs7986	0.53725475
rs3125902	0.636352657	rs1327242	0.558935265	rs4665041	0.537245502
rs11134720	0.636352657	rs2726959	0.558825579	rs2034039	0.537244839
rs560248	0.636171498	rs4238562	0.558806408	rs233928	0.537206982
rs10915421	0.636143262	rs11201618	0.558784919	rs2283004	0.537184341
rs10837346	0.636111111	rs2419549	0.558754755	rs1412839	0.537168207
rs282134	0.636111111	rs2808365	0.558751742	rs885704	0.537139788
rs7794935	0.636077305	rs17223208	0.558668846	rs2296050	0.537079907
rs9507401	0.635943262	rs13010824	0.558626445	rs1350927	0.53706579
rs9290889	0.635869565	rs1463388	0.558623441	rs224013	0.537029843
JHU_3.69879758	0.635869565	rs579861	0.558590823	rs6919280	0.536984771

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs139957	0.635798582	rs10740329	0.558578476	rs11658329	0.536953972
JHU_3.181664733	0.635748792	rs17745923	0.558568017	rs9358642	0.536930924
rs1026453	0.635640898	rs7903263	0.55859251	rs11063610	0.536902858
rs1123804	0.635507246	rs239865	0.558556534	rs6709176	0.536901587
JHU_20.60339579	0.635507246	rs2306597	0.558546158	rs2172963	0.536868813
rs2427507	0.635386473	rs11654616	0.558545253	rs2965375	0.536866825
rs1385951	0.635326087	rs9303397	0.558524003	rs4942198	0.536862326
rs2426956	0.6352657	rs2389594	0.55851201	rs11063699	0.536674889
rs6794755	0.6352657	rs4433972	0.558506005	rs17394450	0.53661186
rs12476772	0.6352657	rs936672	0.558488179	rs10490750	0.536599603
rs6950264	0.635220095	rs41487149	0.558481664	rs2257172	0.536593866
JHU_6.168451969	0.635144928	rs6450136	0.558476207	rs9655382	0.536581103
rs17290929	0.635121513	rs12138550	0.558473869	rs7681895	0.53650701
rs491920	0.635024155	rs2491393	0.55843437	rs1451355	0.536482736
rs10092702	0.635008511	rs10814496	0.558410605	rs12534073	0.536475558
rs11988761	0.634965012	rs2165666	0.558407819	rs2169820	0.536431408
rs476484	0.634903382	rs778771	0.558376983	rs854140	0.536418373
JHU_5.177409018	0.634842995	rs17073748	0.558310102	rs6073285	0.536397942
rs982094	0.634722222	rs321452	0.55829121	rs12367493	0.536390945
rs11658574	0.634541063	rs2283822	0.558272035	rs6837598	0.536296735
rs3751757	0.634541063	rs2500406	0.55824408	rs2736017	0.536293628
JHU_2.46015634	0.634541063	rs1667746	0.558198	rs297324	0.536289055
rs636407	0.634541063	rs1389177	0.558035209	rs11922609	0.536202809
rs886933	0.634541063	rs7413797	0.558025187	rs6607343	0.536146509
rs29996	0.63442029	rs3773010	0.558022869	rs3820515	0.53613503
rs7192609	0.63442029	rs7251848	0.557982889	rs1669084	0.536133157
JHU_14.83059877	0.634359903	rs3407	0.557949139	rs16858314	0.536129058
rs2297203	0.634299517	rs9390757	0.557946094	rs7019268	0.536037181
exm692523	0.634299517	rs5889533	0.55791383	rs10965099	0.536036672
rs1897140	0.63423913	rs7177662	0.557864881	rs298028	0.536026612
rs11171357	0.634224586	rs7705033	0.557819385	rs7273527	0.535977165
rs3858691	0.634208983	rs12714811	0.557818735	rs12031275	0.535950869
rs1936960	0.634178744	rs2138559	0.557739424	rs6115202	0.535790123

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs59432775	0.634178744	rs9962099	0.557724701	rs7262725	0.535788728
rs2343545	0.634178744	rs11201271	0.557617551	rs918702	0.53576938
rs41518	0.634150355	rs2025878	0.557612448	rs9506266	0.53576128
rs291472	0.634118913	rs10974007	0.557591897	rs827628	0.535755925
rs10977619	0.634057971	rs1053733	0.557578258	rs1258732	0.535694225
rs6731963	0.634057971	rs9638144	0.557570063	rs431103	0.535671193
rs4371172	0.633997872	rs104669	0.557556477	rs7214958	0.535667814
rs2341310	0.633937198	rs1163929	0.557550698	rs2740354	0.535665752
rs17810229	0.633929078	rs13268364	0.55750547	rs10207132	0.535658002
JHU_11.128750053	0.633816425	rs4921227	0.557493287	rs473295	0.535609844
rs7161829	0.633756039	rs2730078	0.557473877	rs1535658	0.535589075
rs4955455	0.633635266	rs168899	0.55745597	rs4788423	0.53549989
rs14616	0.633629314	rs11579242	0.557451026	rs41527748	0.535412689
rs26305	0.633574879	rs11697967	0.557402736	rs10485587	0.535173286
JHU_2.154539954	0.633574879	rs4480736	0.557379992	rs11024460	0.535167417
rs13337719	0.633574879	rs10773557	0.557339995	rs4416006	0.535143854
JHU_17.54889735	0.633574879	rs7320670	0.557276291	rs2808783	0.535125136
rs599774	0.633574879	rs1176280	0.557206656	rs7558386	0.535088197
rs927501	0.633548227	rs40687	0.557181998	rs9568494	0.535049908
rs164369	0.633454106	rs972449	0.557135663	rs10194900	0.5349656
rs17136076	0.633454106	rs12486603	0.557114718	rs13153654	0.534962435
rs12270171	0.63329669	rs17119662	0.557101259	rs10499563	0.534735997
JHU_17.14262666	0.633272947	rs355297	0.55706255	rs663486	0.534721629
JHU_1.238134417	0.633272947	rs10008376	0.556986276	rs2514668	0.534702445
rs10754479	0.63321256	rs1035730	0.556955589	rs7872727	0.534624768
1:201360281-AG	0.63321256	rs4534818	0.556926066	rs17260153	0.534583951
rs1536620	0.633091787	rs7911917	0.556895453	rs17590583	0.534391209
rs13250607	0.633091787	rs1028879	0.556886892	rs818821	0.534358518
rs4363051	0.633091787	rs1415701	0.556863379	rs10233260	0.534335818
rs1007468	0.633091787	rs7149001	0.556845932	rs2828013	0.534299378
rs9883783	0.633036407	rs1428491	0.556841384	rs1756673	0.534134162
rs2384598	0.633031401	rs17174714	0.556815187	rs17833849	0.534098046
rs7192514	0.633031401	rs297804	0.556799856	rs2059271	0.534075479

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs10988819	0.632971014	rs17747544	0.556798411	rs11858195	0.534058082
rs12022725	0.632971014	rs890319	0.556775227	rs7916855	0.534034848
rs19994	0.632850242	rs773618	0.556756871	rs2253981	0.534034227
rs2697852	0.632789855	rs17100627	0.556723142	rs2374219	0.534023242
rs17817727	0.632789855	rs7802120	0.556645654	rs6847630	0.534019484
6:130642109-C-T	0.632729469	rs2415611	0.556591836	rs4801368	0.534017628
rs9814702	0.632669082	rs11135930	0.556536071	rs882924	0.534007688
rs2263115	0.632608696	rs2555576	0.556458567	rs1655127	0.533992007
exm207717	0.632427536	rs11088200	0.556414137	rs2378373	0.533908852
JHU_5.157199546	0.632246377	rs17186115	0.556404481	rs668816	0.533863492
JHU_4.27194070	0.632246377	rs1875108	0.556398275	rs9610510	0.533825359
rs11618140	0.63218599	rs10746026	0.556337637	rs1833868	0.533809176
rs10865518	0.63218599	rs1888557	0.556328372	rs7309163	0.533806583
rs1150010	0.632125604	rs7922034	0.556275373	rs742132	0.533735985
rs7669852	0.632125604	rs4760514	0.556268023	rs4770603	0.533579416
rs12890235	0.632125604	rs999428	0.556256527	rs1108967	0.533572773
rs12187443	0.632065217	rs6667450	0.556230787	rs2617841	0.533572078
rs10928979	0.632065217	rs2521624	0.556204121	rs11119314	0.53356592
exm-rs6017342	0.632004831	rs7011450	0.556185546	rs17032441	0.533562945
rs10816666	0.632004831	rs16978198	0.556174022	rs4287691	0.533500875
rs2297109	0.631989835	rs10841392	0.556165095	rs11056659	0.5334452
rs4808046	0.631763285	rs7581524	0.556152962	rs2015461	0.533434211
JHU_8.141135654	0.631763285	rs10501664	0.55614189	rs65264	0.533403437
rs2403303	0.631763285	rs4752530	0.556130144	rs2476188	0.533335416
rs635857	0.631698345	rs6123727	0.556103613	rs6924957	0.533324542
rs11055643	0.631642512	rs6897761	0.556062204	rs174605	0.533293291
rs162713	0.631642512	rs111161249	0.556028311	rs10103353	0.533287372
rs17227021	0.631642512	rs4847000	0.555986877	rs7148455	0.533286911
rs7012323	0.631582126	rs2372321	0.555981757	rs1372786	0.533201328
rs6733550	0.631582126	rs251884	0.555973254	rs2366777	0.533196307
JHU_2.207423644	0.631521739	rs9819844	0.555954585	rs12434022	0.533181123
rs10868372	0.631280193	rs6506633	0.555929787	rs978671	0.533066624
exm-rs6497618	0.631219807	rs171512	0.555923017	rs760894	0.533035853

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs12937209	0.63115942	rs7898455	0.555912518	rs10108270	0.533023262
rs2788407	0.63115942	rs2021935	0.555908142	rs12718295	0.532948775
rs2872899	0.631038647	rs8068318	0.553881143	rs638908	0.532877959
rs6428597	0.631038647	rs11642659	0.553853863	rs17247190	0.532871489
rs10514019	0.631014184	rs1443993	0.555786739	rs11223274	0.532779382
rs1567426	0.63101182	rs7656326	0.555766173	rs11632280	0.532774649
rs7942991	0.631007565	rs1387754	0.555727649	rs2785264	0.532669536
rs4641492	0.630939243	rs9544383	0.553674461	rs17433780	0.532653764
JHU_9.18847817	0.630917874	rs4721036	0.555643094	rs2115063	0.532623764
rs12734153	0.630917874	rs385872	0.555613061	rs4764478	0.532462623
rs1433265	0.630917874	rs11011346	0.555608118	rs1481507	0.532459231
rs971626	0.630797101	rs6746082	0.55557292	rs17799849	0.532393037
rs10037523	0.630736715	rs10918602	0.555570397	rs1114978	0.532383962
rs384901	0.630736715	rs13250078	0.555546431	rs399211	0.53238013
rs1485142	0.630728605	rs6746541	0.555540685	rs2907799	0.532379702
rs17141724	0.630702364	rs163171	0.555484324	rs10244108	0.532379348
JHU_11.30424807	0.630676329	rs13243874	0.555468075	rs1373641	0.532318698
rs1731986	0.630676329	rs8074097	0.555466951	rs17816598	0.532267489
rs10478479	0.630674232	rs3813402	0.555428715	rs4683139	0.532251409
rs8075555	0.630648463	rs8082973	0.555378816	rs279109	0.532242066
rs2717871	0.63051844	rs10756949	0.555342473	rs4905798	0.532208378
rs175707	0.630495169	rs12151188	0.55532806	rs17063802	0.532159911
rs10510930	0.630482033	rs10491238	0.555291865	rs6908425	0.532153186
rs9911562	0.630434783	rs11845763	0.555274196	rs11903301	0.532152749
rs2101178	0.630361229	rs2499762	0.55526685	rs11252694	0.532120272
rs17158861	0.630340189	rs17804174	0.555211986	rs2015843	0.532085864
rs58452037	0.630253623	rs6015693	0.555211698	rs17171090	0.531971844
rs1532560	0.630178723	rs7010302	0.555176348	rs1929409	0.531958373
rs176910	0.63013285	rs2201331	0.555154193	rs11690621	0.531944063
rs9926029	0.630072464	rs666371	0.555101144	rs1024488	0.531829389
rs1405377	0.630019622	rs1479835	0.555063859	rs13854	0.531685073
rs12750195	0.630012077	rs9898576	0.555045071	rs9923227	0.531663477
rs7500962	0.629951691	rs4953911	0.554993529	rs17168967	0.531635984

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs1746250	0.629891304	rs11975146	0.554966616	rs12474831	0.53162967
rs15222594	0.629830918	rs1482571	0.554963113	rs12641050	0.531594229
rs12490375	0.629814184	rs209677	0.554901319	rs959214	0.53157093
rs887269	0.629770531	rs4436756	0.554867137	rs7104875	0.531559924
rs140690408	0.629710145	rs3919602	0.554853827	rs1043424	0.531494162
rs2013296	0.629710145	rs11917356	0.554830334	rs7150406	0.53148933
JHU_16.6312999	0.629710145	rs9318052	0.554706321	rs2009579	0.531453387
rs524770	0.629710145	rs1461282	0.554645313	rs16826658	0.531444205
rs17792757	0.629649758	rs10503800	0.554614868	rs9303915	0.531430993
rs1763382	0.629597872	rs4470164	0.554589507	rs11679555	0.531422947
rs4835860	0.629589372	rs7310431	0.554588573	rs6657073	0.531388094
rs10878557	0.629528986	rs4834142	0.554537644	rs4780476	0.5313573
rs1500619	0.629408213	rs2144834	0.554497684	rs17779442	0.531324987
JHU_15.38816832	0.629347826	rs10847460	0.554483176	rs7645545	0.531130915
rs13162753	0.629227053	rs17712825	0.554475487	rs7979636	0.531126828
rs6061832	0.62910628	rs9287852	0.554460049	rs8222559	0.531043311
JHU_4.146216160	0.62910628	rs12474952	0.554436215	rs6549009	0.531008878
JHU_8.3417188	0.62910628	rs9301668	0.554407869	rs6728378	0.5309559
rs4909996	0.629065957	rs16953563	0.554367695	rs12519981	0.530922566
rs368269	0.628985507	rs345224	0.554295076	rs6806687	0.530849964
rs11023121	0.628743961	rs1896634	0.554286544	rs10496481	0.530849363
rs12359343	0.628730969	rs3017906	0.554230882	rs4901047	0.530739179
rs6594659	0.628683575	rs4621031	0.554177912	rs10492925	0.530707105
rs6806752	0.628683575	rs939815	0.554176101	rs17258448	0.530632906
rs4823813	0.628616548	rs12641893	0.554144768	rs7221489	0.530625164
rs7002098	0.628562802	rs10506525	0.554099585	rs10891169	0.530580121
rs1122517	0.62853948	rs4732066	0.554095308	rs11076193	0.530564341
JHU_19.31384487	0.628442029	rs6112243	0.55408974	rs6432085	0.530539469
rs12333630	0.628440662	rs2973319	0.554021553	rs10013995	0.530440167
rs2583903	0.628381643	rs2034472	0.553997982	rs2169990	0.530356358
rs1826243	0.628339243	rs1265140	0.553986639	rs151402	0.530354604
rs10484652	0.628296927	rs7130004	0.553981322	rs2833435	0.530302836
rs9914458	0.62826087	rs1953228	0.553975955	rs2830675	0.530188105

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs4495279	0.62826087	rs12435660	0.553951202	rs11692121	0.530156014
rs10333379	0.628200483	rs12929499	0.553922013	rs2824407	0.530106098
rs625578	0.628140097	rs1006378	0.55388721	rs3131313	0.530042558
rs6946852	0.628140097	rs111129517	0.553858128	rs1475114	0.529967787
rs4266037	0.628048463	rs4489652	0.553803577	rs1227087	0.529955789
rs539713	0.628019324	rs2866422	0.55379507	rs1248671	0.52994666
rs17591211	0.627986288	rs1864500	0.553764436	rs16879224	0.529830324
JHU_16.3117353	0.627958937	rs6441647	0.553729057	rs2606167	0.529806789
rs958091	0.627951064	rs407931	0.553728658	rs882061	0.529648867
rs1010797	0.627898551	rs1040155	0.553678437	rs7876037	0.529537605
rs4696925	0.627898551	rs4929918	0.553629331	rs470747	0.52952839
JHU_2.221730130	0.627898551	rs543736	0.553621404	rs3217916	0.52944786
rs62491007	0.627838164	rs4978627	0.553604427	rs4742875	0.529370632
rs4243829	0.627777778	rs6531119	0.553584618	rs7948698	0.529361671
JHU_18.47054856	0.627717391	rs11989122	0.553554589	rs4919129	0.529353325
rs7202909	0.627717391	rs1686166	0.553492062	rs4958109	0.529259086
JHU_1.203100297	0.627717391	rs4404633	0.553410219	rs10245531	0.529240857
JHU_17.78974410	0.627657005	rs1553497	0.553392406	rs4831945	0.529122672
rs8026295	0.627647281	rs13249753	0.553367308	rs1010544	0.529079432
rs7321466	0.627596618	rs10773591	0.553343132	rs1413057	0.52890035
rs2834606	0.627596618	rs177698	0.553320833	rs4781419	0.528882763
rs870092	0.627596618	rs136067	0.55326029	rs1948853	0.528824558
rs2517580	0.627596618	rs848386	0.553259907	rs1996651	0.528783013
rs13031421	0.627536232	rs1760349	0.55318702	rs10519435	0.528688029
rs10840868	0.627536232	rs16908163	0.553185716	rs3912027	0.528648826
rs1864213	0.627536232	rs1571227	0.553183876	rs10813957	0.528571099
exm975441	0.627355072	rs7037905	0.553099494	rs1905045	0.52854254
rs1391812	0.627294686	rs973309	0.553050376	rs2373452	0.528513552
rs1028802	0.627294686	rs1423056	0.552985409	rs2704302	0.528457067
rs7084874	0.627294686	rs201768	0.55296229	rs4944092	0.528391812
rs1397933	0.6272343	rs7922854	0.552943597	rs17655898	0.52838579
rs7152595	0.6272343	rs443673	0.552930755	rs4888764	0.52837036
JHU_15.70049423	0.627173913	rs2459602	0.552895014	rs9379274	0.528357642

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs73022977	0.627173913	rs7627367	0.5528796	rs10502353	0.528334668
rs7219847	0.627165248	rs2299219	0.552855679	rs2119235	0.528312426
JHU_18.7589317	0.627113527	rs8004280	0.552843229	rs17551494	0.528301758
1:162048621-C-A	0.627113527	rs11177589	0.552809628	rs16978566	0.528280488
rs567379	0.627113527	rs1330714	0.552798256	rs1866389	0.528230358
rs2086673	0.62705314	rs412147	0.552765696	rs10431158	0.528128994
rs16941388	0.627032861	rs595107	0.552728716	rs6419158	0.528116816
rs7686623	0.626999764	rs10886170	0.552677252	rs4710998	0.528027931
rs10976830	0.626992754	rs817510	0.55267384	rs1897023	0.527948299
rs4711976	0.626932367	rs297012	0.552633196	rs9860870	0.527911981
rs9301025	0.626932367	rs2641698	0.55262834	rs7558848	0.527863983
JHU_12.31811879	0.626811594	rs10181248	0.552626619	rs1395103	0.527834872
rs1605245	0.626811594	rs6127987	0.552579193	rs7337170	0.527820759
15:45620385-C-T	0.626690821	rs8041529	0.552530018	rs9392162	0.52761639
rs8056759	0.626683452	rs6499165	0.552475669	rs613444	0.52740718
rs7855214	0.626647281	rs1351950	0.552425987	rs1534043	0.527388861
rs11207605	0.626630435	rs1040806	0.552425436	rs149658	0.527384301
rs9847915	0.626619149	rs1510737	0.552424078	rs10869160	0.527320135
rs6981187	0.626589598	rs840616	0.552368049	rs4624587	0.527240059
rs2358513	0.626586998	rs1735884	0.5523596	rs10819136	0.527223225
rs10803270	0.626570048	rs7539399	0.552343054	rs10242225	0.527112408
rs1598022	0.626509662	rs2748241	0.552315227	rs6765153	0.527108173
rs6690747	0.626509662	rs7616565	0.552285861	rs7711608	0.527100814
rs7895833	0.626478251	rs1172417	0.552280523	rs6474884	0.527079342
rs6475800	0.626449275	rs10462717	0.552275642	rs41319446	0.527073098
JHU_11.33957781	0.626449275	rs399760	0.552268476	rs4564573	0.526980781
rs931913	0.626328502	rs4761903	0.552245243	rs9324943	0.526965923
rs8080754	0.626268116	rs16913804	0.552216407	rs12949531	0.526965063
JHU_2.45480249	0.626268116	rs2046616	0.552180353	rs302739	0.526953559
rs17535443	0.626268116	rs1401072	0.55216217	rs4783192	0.526853352
rs10079187	0.626268116	rs7035163	0.552100285	rs1324073	0.526833181
rs10899855	0.626267849	rs1997754	0.552097795	rs713812	0.526679835
JHU_8.105779942	0.626207729	rs1559862	0.55207154	rs2148895	0.526635755

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs3884187	0.626147343	rs13033902	0.552063444	rs4985124	0.526594959
rs6856130	0.626112057	rs9316232	0.552013927	rs7177529	0.52637467
rs2897454	0.62602435	rs2883367	0.551979301	rs941886	0.526303195
rs1426389	0.626018676	rs2598404	0.551927928	rs1859572	0.526140383
rs6835720	0.625968558	rs1043178	0.551924837	rs17011692	0.526018416
rs12955131	0.62587234	rs11162351	0.551874859	rs1527878	0.526017198
rs4823776	0.62573617	rs733392	0.551864319	rs7310929	0.525967734
rs39162230-A-C	0.625724638	rs12727814	0.551840245	rs1001064	0.525790508
rs16874044	0.625621513	rs6128386	0.55183387	rs12879285	0.525761068
rs150613	0.625572813	rs1104696	0.551833446	rs11126157	0.525721713
rs206184	0.625543478	rs17515	0.55180808	rs527458	0.525718857
rs11858834	0.625543478	rs7866342	0.551804915	rs1172479	0.525671846
exm292415	0.625483092	rs2392055	0.55177825	rs10771158	0.525458784
exm2267263	0.625483092	rs12605064	0.551669387	rs4909917	0.525449894
JHU_7.33683840	0.625483092	rs2104286	0.551668679	rs10926188	0.525421006
rs6886412	0.625422705	rs10965529	0.551570546	rs817771	0.525401472
rs13748	0.625422705	rs2672587	0.551545502	rs1601422	0.525344514
rs2081	0.625382979	rs1446929	0.551516345	rs1437565	0.525336747
rs1514828	0.625362319	rs10275417	0.551514802	rs12755273	0.525328898
rs11036246	0.625362319	rs7644123	0.551501096	rs11074352	0.525250151
JHU_3.158197077	0.625241546	rs6074148	0.551487025	rs10507887	0.525220155
rs4304404	0.625241546	rs153060	0.551463626	rs4732297	0.525191485
rs9318420	0.625181159	rs11197045	0.551426749	rs10935064	0.5251188
rs11177441	0.625181159	rs17094882	0.551425296	rs2361150	0.525054672
rs4709103	0.625120773	rs6502656	0.551416357	rs10485136	0.524951296
rs17354120	0.625119622	rs9290245	0.551403355	rs10502983	0.524764357
rs3777958	0.625060386	rs4757744	0.55139172	rs4661747	0.52474733
rs164697	0.625	rs7406119	0.551359637	rs7760502	0.524745626
rs1716486	0.623307164	rs594490	0.55133244	rs271174	0.524566293
rs6961566	0.613467649	rs7636107	0.551324221	rs7660174	0.524559811
rs955196	0.611068463	rs7788668	0.551317347	rs436282	0.524449409
rs12602961	0.610725009	rs11159460	0.551271777	rs8088894	0.524449335
rs4986122	0.608052981	rs7951657	0.551247276	rs9508	0.524211887

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs10752336	0.605110364	rs1870130	0.551206773	rs2027858	0.524191086
rs1857501	0.604721899	rs10493634	0.551122646	rs17040558	0.52418078
rs6056752	0.600923277	rs170880	0.551060049	rs898845	0.52411736
rs11096623	0.600613655	rs9547461	0.551019537	rs4899366	0.524021083
rs3774275	0.600391164	rs7946913	0.55100924	rs1980751	0.523953202
rs4869280	0.599200278	rs1321742	0.550976157	rs7914572	0.523894013
rs7742625	0.598356897	rs7681323	0.550898155	rs4790850	0.523806141
rs10739280	0.598328317	rs17800095	0.550869221	rs2052496	0.523781215
rs17488624	0.598188545	rs2111890	0.550847613	rs4878077	0.52368426
rs11716050	0.597713036	rs1157774	0.550764877	rs9905551	0.523367439
rs4880716	0.596889665	rs7902673	0.550760371	rs7336995	0.523255943
rs12373390	0.596803052	rs535457	0.550738516	rs1342606	0.523153904
rs12975389	0.596675331	rs9598099	0.55071166	rs7675393	0.52302671
rs9273363	0.596252381	rs9599058	0.550700799	rs11148711	0.522932932
rs6983566	0.595990522	rs7776725	0.550673186	rs9322655	0.52290464
rs10491999	0.59554054	rs1511036	0.550656509	rs13424270	0.522900166
rs56566	0.595159784	rs4938621	0.550642964	rs9656462	0.522771276
rs1338176	0.595143416	rs2816662	0.550612602	rs8018803	0.522585642
rs9819550	0.594941187	rs9398172	0.550567867	rs970689	0.522513509
rs2537872	0.594692346	rs10744391	0.550562595	rs13332500	0.522460119
rs9647441	0.594550446	rs2159700	0.550557911	rs7753862	0.522422636
rs11257103	0.59454715	rs2149074	0.550555149	rs2207418	0.522231919
rs2618114	0.59423035	rs2172802	0.550489634	rs10494952	0.522136318
rs4766096	0.594209404	rs9542547	0.550485901	rs3171297	0.522120669
rs6778486	0.59374914	rs7828391	0.550476209	rs266274	0.522009132
rs11711241	0.593160637	rs2532001	0.550465446	rs2765912	0.521867232
rs13386062	0.592630174	rs10743827	0.550458988	rs2813427	0.521840668
rs6493620	0.592190682	rs42188	0.55043915	rs10896300	0.521825613
rs4689093	0.592068472	rs153464	0.550298077	rs4580353	0.521782792
rs9428225	0.591519599	rs1332408	0.550237798	rs647080	0.52162084
rs7194009	0.591508627	rs10493578	0.550223162	rs7719129	0.521553852
rs2293370	0.590309666	rs4812779	0.550187565	rs403904	0.521286305
rs2523	0.590209401	rs1004553	0.550142077	rs17086628	0.521127531

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs1861777	0.590039053	rs809375	0.550127396	rs2219863	0.520844237
rs1155742	0.589908871	rs17824620	0.550126877	rs10774840	0.520791695
rs2290207	0.589519028	rs1107785	0.550042895	rs2246436	0.520690323
rs3749078	0.589352557	rs2515000	0.549987397	rs13061996	0.520658837
rs6429214	0.589246061	rs651568	0.549948166	rs12538361	0.520628903
rs1001290	0.588907291	rs2284280	0.549921792	rs11777456	0.520626039
rs8037418	0.588426411	rs10955844	0.549912877	rs17318350	0.520336642
rs4303700	0.588233397	rs2330522	0.549906333	rs12611038	0.520303711
rs1381894	0.588095806	rs461863	0.549869649	rs4954599	0.520243287
rs9844484	0.587706889	rs11242704	0.549852445	rs10176316	0.520089785
rs6105044	0.587691245	rs13194498	0.549825289	rs4720128	0.519955483
rs12061042	0.587387821	rs17086212	0.549825186	rs10975641	0.519936069
rs2921446	0.587311156	rs9837834	0.549801187	rs16921571	0.519931801
rs4891895	0.587076392	rs7109838	0.549774282	rs2437896	0.51985863
rs4764937	0.586830038	rs6677615	0.549701606	rs1439495	0.51978569
rs561102	0.586489547	rs1180939	0.549689682	rs1888665	0.51966718
rs11133935	0.586471718	rs231005	0.549632432	rs6871440	0.519639403
rs6096889	0.586327772	rs436760	0.549617384	rs6481165	0.519622767
rs1430706	0.586037721	rs3750695	0.549608778	rs9922767	0.519560845
rs4789580	0.586035646	rs7833751	0.549601699	rs158955	0.519508039
rs4723801	0.585924137	rs3801778	0.549534641	rs1350924	0.519378989
rs1077224	0.585815003	rs11885902	0.549529043	rs4796808	0.519236579
rs2942194	0.585563669	rs6458307	0.549526191	rs244005	0.519203727
rs508378	0.585468096	rs6679531	0.549523384	rs940155	0.519182004
rs7102454	0.585423613	rs11152931	0.549499566	rs1598492	0.519165034
rs2384550	0.584962612	rs7899611	0.549384349	rs1011814	0.519103067
rs3810040	0.584791086	rs219002	0.549308038	rs11616892	0.518895207
rs2079685	0.584512905	rs6465353	0.549290352	rs588629	0.518694847
rs7929583	0.584400766	rs11099629	0.549279516	rs1630958	0.518674795
rs9388399	0.584262237	rs6667720	0.549230978	rs11135570	0.518656418
rs294244	0.58420213	rs4667005	0.549216067	rs2277080	0.518617507
rs1013264	0.584184918	rs7192266	0.549206674	rs17826560	0.518502619
rs7983347	0.58406193	rs2055831	0.549164162	rs8091851	0.518311709

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs8110245	0.58394952	rs10950917	0.54912463	rs6035795	0.517910132
rs2467864	0.583910107	rs13264395	0.549102759	rs6777242	0.517260262
rs6121785	0.583670815	rs10486663	0.549062276	rs7712888	0.517230155
rs12481420	0.582797392	rs780240	0.549001387	rs9992247	0.517141706
rs3748863	0.582429865	rs10269378	0.548994255	rs12067454	0.517012578
rs7237444	0.582406656	rs4432842	0.548965572	rs7388381	0.516934979
rs4868776	0.582325707	rs12553631	0.548965131	rs1508733	0.516920047
rs12446319	0.582317484	rs4353251	0.548962349	rs12503223	0.516878872
rs373533	0.581790165	rs221035	0.548937346	rs7901695	0.516779019
rs1351205	0.581535213	rs9527514	0.548932102	rs3135093	0.516727099
rs12717991	0.581349879	rs1150438	0.548920743	rs909302	0.516520951
rs7666340	0.581098298	rs11185115	0.54890146	rs4939008	0.516507435
rs2833145	0.580655435	rs1558738	0.548876239	rs7708153	0.516447201
rs10020857	0.580403916	rs481377	0.548869736	rs17098973	0.516249618
rs16867655	0.58036775	rs4374782	0.548842806	rs11122213	0.51613143
rs1869890	0.580262839	rs34337	0.548842539	rs4577845	0.51609556
rs9419604	0.580045616	rs6022230	0.54883492	rs17119018	0.515978549
rs546343	0.580013344	rs2248353	0.548801467	rs6575944	0.51595357
rs7843510	0.579677677	rs7261610	0.548796919	rs1499306	0.515933962
rs12469420	0.579441407	rs5770661	0.548722876	rs2468147	0.515921134
rs1881516	0.579423803	rs7905804	0.548702145	rs12465448	0.515891109
rs6868704	0.579368109	rs403608	0.548678541	rs4873135	0.515883906
rs6603832	0.579308751	rs17300655	0.548669943	rs309543	0.5157324
rs2282498	0.579037821	rs8132470	0.548641655	rs2278556	0.515730136
rs6820447	0.578929625	rs395119	0.548612968	rs10182071	0.51570077
rs9815265	0.57882764	rs10878938	0.548595867	rs2296634	0.515645532
rs6871296	0.578781893	rs162057	0.548589735	rs2378956	0.515590089
rs7993686	0.578767533	rs7422221	0.548577268	rs1565153	0.515548243
rs2256270	0.578747382	rs5030780	0.548572621	rs2045308	0.515356608
rs6677662	0.578709595	rs7021746	0.548536439	rs2180233	0.515268325
rs2678666	0.578672109	rs2245747	0.548514227	rs9321263	0.515208584
rs1010776	0.578432537	rs2803439	0.548499743	rs13164082	0.515138789
rs9948473	0.578197769	rs7081687	0.548461244	rs2445271	0.514884396

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs2262994	0.578053758	rs790742	0.548425404	rs12245332	0.514739158
rs3735222	0.577943792	rs2824318	0.548423498	rs8652210	0.514625147
rs10491984	0.577817886	rs11186188	0.548376554	rs2293348	0.514187014
rs10805632	0.577765542	rs9473045	0.548351818	rs7793306	0.513816741
rs842301	0.577530481	rs954645	0.548339717	rs9357271	0.513788071
rs6560711	0.577494513	rs3845976	0.548295246	rs12054271	0.513672401
rs6539463	0.577473695	rs10847718	0.548275318	rs5771040	0.513660876
rs6846226	0.57746506	rs9291428	0.54826232	rs7048811	0.513654542
rs9955849	0.577426105	rs9915136	0.548251269	rs1117361	0.513552393
rs2216774	0.577415284	rs7243349	0.54823083	rs9921236	0.51354158
rs2031609	0.577273825	rs158342	0.548209831	rs11593681	0.513373138
rs913165	0.577115305	rs16826701	0.548129339	rs981900	0.513255081
rs9533034	0.577074621	rs2075469	0.548074554	rs948969	0.5131313184
rs244043	0.577053852	rs10760078	0.548048674	rs12516484	0.5131313071
rs7836754	0.576981353	rs7137605	0.548038841	rs4619008	0.51281407
rs4739074	0.57674309	rs691336	0.54800007	rs3934815	0.512694149
rs10833748	0.576670171	rs17023520	0.547996568	rs1164768	0.512567926
rs10004030	0.576668874	rs1781931	0.547970585	rs683266	0.512399825
rs12464286	0.576638092	rs12478492	0.547947532	rs11163192	0.512110308
rs986126	0.576604541	rs2069962	0.547905831	rs2962442	0.511761187
rs3774581	0.576602758	rs6134639	0.547853116	rs10498555	0.511355832
rs1020088	0.576531769	rs1539519	0.547813844	rs2456470	0.511204544
rs8121916	0.576435472	rs3784484	0.54779503	rs4900337	0.511183294
rs7843177	0.576100464	rs7092910	0.547750263	rs17591495	0.511109963
rs10761739	0.576036715	rs12521931	0.547740644	rs10732827	0.510872437
rs7080842	0.575839462	rs9589960	0.547724037	rs10814993	0.510757808
rs1060939	0.575675402	rs4689126	0.547705747	rs12374409	0.510599046
rs11071136	0.575670858	rs9508795	0.54768965	rs10778637	0.510302364
rs6956956	0.57560927	rs11691608	0.547598341	rs1969815	0.510160414
rs1488935	0.575559493	rs9317527	0.547518709	rs9303642	0.510153862
rs4804217	0.575570079	rs10956511	0.547499089	rs947095	0.509981407
rs7428879	0.575288696	rs6939307	0.547431129	rs776385	0.509840424
rs1534362	0.575278814	rs235360	0.547405039	rs2963794	0.50980151

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs41473644	0.575126163	rs4719344	0.547374093	rs12709959	0.509598318
rs4798896	0.575082622	rs11610	0.547317291	rs1295348	0.509570927
rs7784828	0.575032097	rs4781935	0.547249121	rs6534418	0.509515759
rs14976	0.57496003	rs6793348	0.54724228	rs12246110	0.509035715
rs8060979	0.574926218	rs1543416	0.547229674	rs9903786	0.50897774
rs180619	0.574524201	rs2030834	0.547190273	rs13115900	0.508868017
rs12205802	0.5743413	rs543686	0.547171859	rs7100118	0.50886479
rs17791802	0.574289074	rs1867971	0.547144934	rs41481747	0.508801049
rs10746573	0.574286004	rs3816963	0.547102459	rs4736691	0.508223551
rs4053955	0.574184599	rs10756653	0.547100179	rs362794	0.508156184
rs984802	0.574143907	rs1977057	0.547079547	rs2074932	0.508130588
rs2600057	0.574103317	rs4719155	0.547079489	rs3765096	0.508101617
rs7688017	0.574046313	rs7752164	0.547053201	rs849162	0.508085635
rs11886398	0.573943122	rs1197671	0.547033244	rs11712379	0.508008559
rs220749	0.573892189	rs2641409	0.547021539	rs13193063	0.50784736
rs11630094	0.573891095	rs566337	0.547013715	rs12650571	0.50768363
rs13121806	0.573828724	rs6855567	0.546994951	rs10513196	0.507585386
rs9525817	0.573685433	rs10824749	0.546984888	rs2342468	0.507395912
rs9320772	0.57357192	rs113080	0.546970025	rs132663	0.507374806
rs135264	0.57347617	rs360816	0.546947483	rs12447438	0.507189242
rs11036247	0.57338424	rs728329	0.546929362	rs9939674	0.507022973
rs17438276	0.572971653	rs7142050	0.54692681	rs437548	0.506861885
rs1385237	0.572483941	rs10103791	0.546921105	rs6441990	0.506614913
rs2147104	0.572343148	rs6989495	0.546879635	rs1462372	0.506604237
rs6982811	0.572313424	rs220860	0.546869876	rs1143901	0.506545077
rs17593746	0.572180114	rs469999	0.546811753	rs5993445	0.506523049
rs1852458	0.57216954	rs10018816	0.546745168	rs2469480	0.506470001
rs1574317	0.572133671	rs1538589	0.546721844	rs12233670	0.506467091
rs11129180	0.572091102	rs9652942	0.546689395	rs4694548	0.50635317
rs1717275	0.572010556	rs10757392	0.54659794	rs2164560	0.506208451
rs11641868	0.571879904	rs725925	0.546654654	rs17753780	0.50613214
rs11683907	0.571806767	rs6557618	0.546593217	rs7937934	0.506129629
rs7829784	0.571782751	rs1293405	0.546584751	rs1464906	0.505856493

Appendix C.xlsx

ILMNID	mean_cv_auc	ILMNID	mean_cv_auc	ILMNID	mean_cv_auc
rs1876314	0.571764432	rs4241857	0.54656977	rs9526823	0.505812903
rs3792252	0.571751665	rs11123232	0.546551619	rs6565994	0.505810257
rs11957402	0.571714153	rs36318	0.546528587	rs289754	0.505673349
rs12233446	0.571650618	rs6533940	0.546467957	rs10850378	0.505536582
rs2987532	0.571636962	rs2222973	0.546447304	rs8468	0.505442157
rs7708319	0.571463498	rs532644	0.546381798	rs11199483	0.505376803
rs10025805	0.505108544	rs12713517	0.505179957	rs4709854	0.505369378
				rs980238	0.50529338

WHAT IS CLAIMED IS:

1. A kit for determining methylation status of at least one CpG dinucleotide and
5 a genotype of at least one single-nucleotide polymorphism (SNP), the kit comprising:
at least one first nucleic acid primer at least 8 nucleotides in length that is
complementary to a bisulfite-converted nucleic acid sequence comprising a first CpG
dinucleotide at a GC locus selected from the group consisting of cg00300879, cg09552548,
and cg14789911 or at a second CpG dinucleotide in linkage disequilibrium with the first
10 CpG dinucleotide at a GC locus selected from the group consisting of cg00300879,
cg09552548, and cg14789911, wherein the linkage disequilibrium has a value of $R > 0.3$,
wherein the at least one first nucleic acid primer detects a methylated or unmethylated CpG
dinucleotide, and
at least one second nucleic acid primer at least 8 nucleotides in length that is
15 complementary to a DNA sequence or a bisulfite-converted DNA sequence of a first SNP
selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and
rs9638144 or a second SNP in linkage disequilibrium with the first SNP selected from the
group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, wherein
the linkage disequilibrium has a value of $R > 0.3$.
20
2. The kit of claim 1, wherein the at least one first nucleic acid primer detects the
unmethylated CpG dinucleotide.
3. The kit of claim 1, wherein the at least one first nucleic acid primer detects the
25 methylated CpG dinucleotide.
4. The kit of any of claims 1-3, wherein the at least one first nucleic acid primer
comprises one or more nucleotide analogs.
5. The kit of any of claims 1-4, wherein the at least one first nucleic acid primer
30 comprises one or more synthetic or non-natural nucleotides.

6. The kit of any one of claims 1-5, further comprising a solid substrate to which the at least one first nucleic acid primer is bound.

5 7. The kit of claim 6, wherein the substrate is a polymer, glass, semiconductor, paper, metal, gel or hydrogel.

8. The kit of claim 6, wherein the solid substrate is a microarray or microfluidics card.

10

9. The kit of any one of claims 1-8, further comprising a detectable label.

10. The kit of any one of claims 1-9, further comprising at least a third nucleic acid primer at least 8 nucleotides in length that is complementary to a nucleic acid sequence upstream of the CpG dinucleotide.

15

11. The kit of any one of claims 1-9, further comprising at least a third nucleic acid primer at least 8 nucleotides in length that is complementary to a nucleic acid sequence downstream of the CpG dinucleotide.

20

12. A method of determining the presence of biomarkers associated with predicting a three-year incidence of CVD using a biological sample from a subject, comprising:

25 (a) providing a first portion of the biological sample and a second portion of the biological sample, wherein the nucleic acid from at least the first portion is bisulfite converted;

(b) contacting the first portion of the biological sample with a first oligonucleotide primer at least 8 nucleotides in length that is complementary to a sequence that comprises a first CpG dinucleotide at a GC locus selected from the group consisting of cg00300879, cg09552548, and cg14789911, or a second CpG dinucleotide in linkage disequilibrium with the first CpG dinucleotide at a GC locus selected from the group

30

consisting of cg00300879, cg09552548, and cg14789911, wherein the linkage disequilibrium has a value of $R > 0.3$, wherein the first nucleic acid primer detects a methylated or unmethylated CpG dinucleotide; and

5 (c) contacting the second portion of the biological sample with a nucleic acid primer at least 8 nucleotides in length that is complementary to a DNA sequence or a bisulfite-converted DNA sequence of a first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 or a second SNP in linkage disequilibrium with a first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144, wherein the linkage disequilibrium has a
10 value of $R > 0.3$,

wherein the percentage of methylation of the CpG dinucleotide at the GC locus selected from the group consisting of cg00300879, cg09552548, and cg14789911, and the identity of the nucleotide at the first SNP selected from the group consisting of rs11716050, rs6560711, rs3735222, rs6820447, and rs9638144 or the second SNP in linkage
15 disequilibrium with the first SNP are biomarkers associated with the three-year incidence of CVD.

13. The method of claim 12, wherein the biological sample is selected from the group consisting of blood and saliva.
20

14. The method of claim 12, wherein the at least one first nucleic acid primer detects the unmethylated CpG dinucleotide.

15. The method of claim 12, wherein the at least one first nucleic acid primer
25 detects the methylated CpG dinucleotide.

16. The method of claim 12, wherein the at least one first nucleic acid primer comprises one or more nucleotide analogs.

17. The method of claim 12, wherein the at least one first nucleic acid primer
30 comprises one or more synthetic or non-natural nucleotides.

18. The method of claim 12, wherein the window of incidence is three years.

19. A method of determining the presence of a biomarker associated with CVD in
5 a subject sample, the method comprising:

(a) isolating nucleic acid sample from the subject sample;

(b) performing a genotyping assay on a first portion of the nucleic acid sample
to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from
Appendix C and/or is a second SNP in linkage disequilibrium ($R > 0.3$) with a first SNP from
10 Appendix C to obtain genotype data; and/or

(c) bisulfite converting the nucleic acid in a second portion of the nucleic acid
and performing methylation assessment on a second portion of the nucleic acid sample to
detect methylation status of at least one CpG site from Appendix A and/or a CpG site
collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data; and

15 (d) entering the genotype data from step (b) and/or methylation data from step
(c) into an algorithm that accounts for at least one SNP main effect and/or at least one CpG
main effect and/or at least one interaction effect, wherein the algorithm is a machine learning
algorithm capable of accounting for linear and non-linear effects.

20 20. The method of claim 19, wherein the at least one interaction effect is selected
from the group consisting of a gene-environment interaction (SNPxCpG) effect, a gene-gene
interaction (SNPxSNP) effect, and an environment-environment interaction (CpGxCpG)
effect.

25 21. The method of claim 19, wherein the at least one interaction effect is a gene-
environment interaction effect (SNPxCpG) between a CpG site from Appendix A or a CpG
site that is collinear ($R > 0.3$) with a CpG site from Appendix A and a SNP from Appendix C
or a SNP within moderate linkage disequilibrium ($R > 0.3$) from a SNP from Appendix C.

22. The method of claim 19, wherein the at least one interaction effect is an environment-environment interaction effect (CpGxCpG) between at least two CpG sites from Appendix A.

5 23. The method of claim 22, wherein one or both of the at least two CpG sites are collinear ($R > 0.3$) with one or both of the at least two CpG sites from Appendix A.

24. The method of claim 19, wherein the at least one interaction effect is a gene-gene interaction effect (SNP x SNP) between at least two SNPs from Appendix C.

10

25. The method of claim 24, wherein one or both of the at least two SNPs are collinear ($R > 0.3$) with one or both of the at least two SNPs from Appendix C.

15 26. The method of any of claims 19-25, wherein the biological sample is a saliva sample.

27. A system for determining methylation status of at least one CpG dinucleotide and a genotype of at least one single-nucleotide polymorphism (SNP), the system comprising:

20

a nucleic acid isolation module configured to isolate a nucleic acid sample from a subject sample;

a genotyping assay module configured to perform a genotyping assay on a first portion of the nucleic acid sample to detect the presence of at least one SNP, wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage

25

disequilibrium ($R > 0.3$) with a first SNP from Appendix C to obtain genotype data;

a methylation assay module configured to bisulfite convert the nucleic acid in a second portion of the nucleic acid and perform a methylation assessment on a second portion of the nucleic acid sample to detect methylation status of at least one CpG site from Appendix A and/or a CpG site collinear ($R > 0.3$) with a CpG from Appendix A to obtain methylation data; and

30

an identification system configured to account for at least one SNP main

effect and/or at least one CpG main effect and/or at least one interaction effect based on the genotype data from step (b) and/or methylation data from step (c).

5 28. The system of claim 27, wherein the algorithm is a machine learning algorithm capable of accounting for linear and non-linear effects.

 29. The system of claim 27 or 28, further comprising an output module configured to provide an output based on an identification by the identification system, wherein the identification accounts for at least one SNP main effect and/or at least one CpG
10 main effect and/or at least one interaction effect based on the genotype data from step (b) and/or methylation data from step (c).

 30. A non-transitory computer-readable medium storing instructions executable by a processing device to perform operations comprising:
15 accounting for at least one SNP main effect and/or at least one CpG main effect and/or at least one interaction effect based on genotype data and/or methylation data, wherein:
 the genotype data is based on a genotyping assay on a first portion of a nucleic acid sample isolated from a subject sample to detect the presence of at least one SNP,
20 wherein the at least one SNP is a first SNP from Appendix C and/or is a second SNP in linkage disequilibrium ($R>0.3$) with a first SNP from Appendix C to obtain the genotype data; and

 the methylation data is based on a methylation assay on a bisulfite converted nucleic acid in a second portion of the nucleic acid sample to detect methylation
25 status of at least one CpG site from Appendix A and/or a CpG site collinear ($R>0.3$) with a CpG from Appendix A to obtain methylation data.

 31. The non-transitory computer-readable medium of claim 30, wherein the operations further comprise providing an output based on the accounting.

30

32. The non-transitory computer-readable medium of claim 31, wherein the output comprises one or more of storing a report based on the accounting to another non-transitory computer-readable medium, modifying a display based on the accounting, triggering an audible alert based on the accounting, triggering a haptic or vibratory alert based on the accounting, triggering the printing of a report based on the accounting, or triggering the delivery of a therapeutic based on the accounting.
- 5