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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 June 2016 (02.06.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/086197 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2015/062787

(22) International Filing Date:
25 November 2015 (25.11.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/084,127 25 November 2014 (25.11.2014) US

(71) Applicant: **THE BRIGHAM AND WOMEN'S HOSPITAL, INC.** [US/US]; 75 Francis Street, Boston, MA 02115 (US).

(72) Inventors: **EBERT, Benjamin, Levine**; 47 Greenough Street, Brookline, MA 02445 (US). **JAISWAL, Siddhartha**; The Brigham and Women's Hospital, Inc., 75 Francis Street, Boston, MA 02115 (US). **KATHIRESAN, Sekar**; The Broad Institute Inc., 415 Main Street, Cambridge, MA 02142 (US).

(74) Agents: **KOWALSKI, Thomas, J.** et al.; Vedder Price P.C., 1633 Broadway, New York, NY 10019 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

WO 2016/086197 A1

(54) Title: METHOD OF IDENTIFYING AND TREATING A PERSON HAVING A PREDISPOSITION TO OR AFFLICTED WITH A CARDIOMETABOLIC DISEASE

(57) Abstract: The invention relates to method for identifying and selecting a subject with increased risk of developing a cardiometabolic disease and optionally, providing a personalized medicine method, which may involve sequencing at least part of a genome of one or more cells in a blood sample of the subject and identifying from said sequencing one or more mutations in one or more somatic mutations.

METHOD OF IDENTIFYING AND TREATING A PERSON HAVING A PREDISPOSITION TO OR AFFLICTED WITH A CARDIOMETABOLIC DISEASE

RELATED APPLICATIONS AND INCORPORATION BY REFERENCE

[0001] This application claims benefit of and priority to US provisional patent application Serial No. 62/084,127 filed November 25, 2014.

[0002] The foregoing applications, and all documents cited therein or during their prosecution ("appln cited documents") and all documents cited or referenced in the appln cited documents, and all documents cited or referenced herein ("herein cited documents"), and all documents cited or referenced in herein cited documents, together with any manufacturer's instructions, descriptions, product specifications, and product sheets for any products mentioned herein or in any document incorporated by reference herein, are hereby incorporated herein by reference, and may be employed in the practice of the invention. More specifically, all referenced documents are incorporated by reference to the same extent as if each individual document was specifically and individually indicated to be incorporated by reference.

FIELD OF THE INVENTION

[0003] The present invention relates to identifying individuals with a predisposition to cardiovascular disease. In particular, the invention relates to method for identifying and selecting a subject with increased risk of developing a cardiometabolic disease and optionally a hematological cancer, and in some instances, providing a personalized medicine method.

BACKGROUND OF THE INVENTION

[0004] Cancer is thought to arise via stepwise acquisition of genetic or epigenetic changes that transform a normal cell (Nowell PC. Science 1976;194:23-8). Hence, the existence of a pre-malignant state bearing only the initiating lesions may be detectable in some individuals with no other signs of disease. For example, multiple myeloma (MM) is frequently preceded by monoclonal gammopathy of unknown significance (MGUS) (Kyle RA et al. The New England journal of medicine 2002;346:564-9), and chronic lymphocytic leukemia (CLL) is commonly preceded by monoclonal B-lymphocytosis (MBL) (Rawstron AC et al. The New England journal of medicine 2008;359:575-83).

[0005] Several lines of evidence have suggested that clonal hematopoiesis due to an expansion of cells harboring an initiating driver mutation might be an aspect of the aging hematopoietic system. Clonal hematopoiesis in the elderly was first demonstrated in studies that found that approximately 25% of healthy women over the age of 65 have a skewed pattern of X-chromosome inactivation in peripheral blood cells (Busque L et al. Blood 1996;88:59 65, Champion KM et al. British journal of haematology 1997;97:920 6) which in some cases is associated with mutations in *TET2* (Busque L et al. Nature genetics 2012;44:1179-81). Large-scale somatic events such as chromosomal insertions and deletions and loss of heterozygosity (LOH) also occur in the blood of ~2% of individuals over the age of 75 (Jacobs KB et al. Nature genetics 2012;44:651-8, Laurie CC et al. Nature genetics 2012;44:642-50). Pre-leukemic hematopoietic stem cells (HSCs) harboring only the initiating driver mutation have been found in the bone marrow of patients with AML in remission (Jan M et al. Science translational medicine 2012;4:149ra18, Shlush LI et al. Nature 2014;506:328-33). Furthermore, a substantial proportion of the population carries cells with t(14;18) translocations, although the lesion is generally present in fewer than 1 in 1000 cells (Roulland S et al. t(14;18) Journal of clinical oncology: official journal of the American Society of Clinical Oncology 2014;32:1347-55).

[0006] Recent sequencing studies have identified a set of recurrent mutations in several types of hematological malignancies (Mardis ER et al. The New England journal of medicine 2009;361:1058-66, Bejar R et al. The New England journal of medicine 2011;364:2496-506, Papaemmanuil E et al. The New England journal of medicine 2011;365:1384-95, Walter et al. Leukemia 2011;25:1153-8, Welch JS et al. Cell 2012;150:264-78, Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine 2013;368:2059-74, Papaemmanuil E et al. Blood 2013;122:3616-27; quiz 99, Walter MJ et al. Leukemia 2013;27:1275-82, Zhang J et al. Proceedings of the National Academy of Sciences of the United States of America 2013;110:1398-403, Morin RD et al. Nature 2011;476:298-303, Lenz G et al. Science 2008;319:1676-9, Lohr JG et al. Proceedings of the National Academy of Sciences of the United States of America 2012;109:3879-84, Neumann M et al. Blood 2013;121:4749-52). However, the frequency of these somatic mutations in the general population is unknown. Citation or identification of any document in this application is not an admission that such document is available as prior art to the present invention.

SUMMARY OF THE INVENTION

[0007] The present invention relates to Applicants' hypothesis that somatically acquired single nucleotide variants (SNVs) and small insertions/deletions (indels) might be increasingly detectable in the blood of otherwise healthy individuals as a function of age.

[0008] The invention relates to method for identifying and selecting a subject with increased risk of developing a cardiometabolic disease and optionally a hematological cancer, which may comprise the steps of: (a) sequencing at least part of a genome which may comprise one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject, (b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer.

[0009] The invention also relates to a method for identifying and selecting a subject with an increased risk of developing a cardiometabolic disease and optionally a hematological cancer and providing a personalized medicine method, said method which may comprise the steps of (a) sequencing at least part of a genome which may comprise one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject, (b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer, and (c) initiating a treatment or monitoring regimen to suppress said mutation(s) in the subject, thereby decreasing risk of developing a cardiometabolic disease and optionally a hematological cancer.

[0010] The presence of said mutation(s) in the above embodiments may indicate an increase in red blood cell distribution width (RDW).

[0011] The cardiometabolic disease may be atherosclerosis, coronary heart disease (CHD) or ischemic stroke (IS) or type 2 diabetes (T2D).

[0012] In embodiments wherein an increased risk for hematological cancer is also screened in addition to a cardiometabolic disease, the hematological cancer may be a leukemia, a lymphoma, a myeloma or a blood syndrome. The leukemia may be an acute myeloid leukemia (AML), chronic myelogenous leukemia (CML) or chronic lymphocytic leukemia (CLL). The

blood syndrome may be myelodysplastic syndrome (MDS) or myeloproliferative neoplasm (MPN).

[0013] The one more cells in the blood sample may be derived from hematopoietic stem cells (HSCs) , committed myeloid progenitor cells having long term self-renewal capacity or mature lymphoid cells having long term self-renewal capacity.

[0014] In some embodiments the part of the genome that is sequenced may be an exome. In other embodiments, the sequencing may be whole exome sequencing (WES) or targeted gene sequencing..

[0015] In an advantageous embodiment, the subject is a human. In other embodiments, the human may exhibit one or more risk factors of being a smoker, having a high level of total cholesterol or having high level of high-density lipoprotein (HDL).

[0016] The mutations of at least *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* may be frameshift mutations, nonsense mutations, missense mutations or splice-site variant mutations.

[0017] If the mutation is in *DNMT3A*, the mutation may advantageously be a mutation in exons 7 to 23. In a particularly advantageous embodiment, the mutation in *DNMT3A* is a mutation selected from the group consisting of P307S, P307R, R326H, R326L, R326C, R326S, R366P, R366H, R366G, A368T, F414L, F414S, F414C, C497Y, Q527H, Q527P, Y533C, G543A, G543S, G543C, L547H, L547P, L547F, M548I, M548K, G550R, W581R, W581G, W581C, G646V, G646E, L653W, L653F, V657A, V657M, R659H, Y660C, R676W, R676Q, G685R, G685E, G685A, D686Y, D686G, G699R, G699S, G699D, P700S, P700R, P700Q, D702N, D702Y, V704M, V704G, I705F, I705T, I705S, C710S, S714C, N717S, N717I, P718L, R720H, R720G, Y724C, R729Q, R729W, R729G, F731L, F732del, F732S, F732L, F734L, F734C, Y735C, Y735N, Y735S, R736H, R736C, R736P, L737H, L737V, L737F, L737R, A741V, R749C, R749L, F751L, F752del, F752C, F752L, F752I, F752V, L754R, L754H, F755S, F755I, F755L, M761I, M761V, G762C, S770W, S770P, R771Q, F772I, F772V, L773R, E774K, E774D, D781G, R792H, G796D, G796V, N797Y, N797H, P799R, P799H, R803S, P804S, P804L, S828N, K829R, Q842E, P849L, D857N, W860R, F868S, G869S, G869V, M880V, S881R, S881I, R882H, R882P, R882C, R882G, Q886R, G890D, L901R, L901H, P904L, F909C and A910P.

[0018] If the mutation is in *TET2*, the mutation is advantageously selected from the group consisting of S282F, N312S, L346P, S460F, D666G, P941S, and C1135Y.

[0019] If this mutation is in *ASXL1*, the mutation is advantageously a mutation in exon 11-12.

[0020] If the mutation is in *TP53*, the mutation is advantageously a mutation selected from the group consisting of S46F, G105C, G105R, G105D, G108S, G108C, R110L, R110C, T118A, T118R, T118I, L130V, L130F, K132Q, K132E, K132W, K132R, K132M, K132N, C135W, C135S, C135F, C135G, Q136K, Q136E, Q136P, Q136R, Q136L, Q136H, A138P, A138V, A138A, A138T, T140I, C141R, C141G, C141A, C141Y, C141S, C141F, C141W, V143M, V143A, V143E, L145Q, L145R, P151T, P151A, P151S, P151H, P152S, P152R, P152L, T155P, R158H, R158L, A159V, A159P, A159S, A159D, A161T, A161D, Y163N, Y163H, Y163D, Y163S, Y163C, K164E, K164M, K164N, K164P, H168Y, H168P, H168R, H168L, H168Q, M169I, M169T, M169V, T170M, E171K, E171Q, E171G, E171A, E171V and E171D, V172D, V173M, V173L, V173G, R174W, R175G, R175C, R175H, C176R, C176G, C176Y, C176F, C176S, P177R, P177R, P177L, H178D, H178P, H178Q, H179Y, H179R, H179Q, R181C, R181Y, D186G, G187S, P190L, P190T, H193N, H193P, H193L, H193R, L194F, L194R, I195F, I195N, I195T, V197L, G199V, Y205N, Y205C, Y205H, D208V, R213Q, R213P, R213L, R213Q, H214D, H214R, S215G, S215I, S215R, V216M, V217G, Y220N, Y220H, Y220S, Y220C, E224D, I232F, I232N, I232T, I232S, Y234N, Y234H, Y234S, Y234C, Y236N, Y236H, Y236C, M237V, M237K, M237I, C238R, C238G, C238Y, C238W, N239T, N239S, S241Y, S241C, S241F, C242G, C242Y, C242S, C242F, G244S, G244C, G244D, G245S, G245R, G245C, G245D, G245A, G245V, G245S, M246V, M246K, M246R, M246I, N247I, R248W, R248G, R248Q, R249G, R249W, R249T, R249M, P250L, I251N, L252P, I254S, I255F, I255N, I255S, L257Q, L257P, E258K, E258Q, D259Y, S261T, G262D, G262V, L265P, G266R, G266E, G266V, R267W, R267Q, R267P, E271K, V272M, V272L, R273S, R273G, R273C, R273H, R273P, R273L, V274F, V274D, V274A, V274G, V274L, C275Y, C275S, C275F, A276P, C277F, P278T, P278A, P278S, P278H, P278R, P278L, G279E, R280G, R280K, R280T, R280I, R280S, D281N, D281H, D281Y, D281G, D281E, R282G, R282W, R282Q, R282P, E285K, E285V, E286G, E286V, E286K, K320N, L330R, G334V, R337C,R337L, A347T, L348F, T377P.

[0021] If the mutation is in *JAK2*, the mutation is advantageously selected from the group consisting of N533D, N533Y, N533S, H538R, K539E, K539L, I540T, I540V, V617F, R683S,

R683G, del/ins537---539L, del/ins538---539L, del/ins540---543MK, del/ins540---544MK, del/ins541- -543K, del542---543, del543---544 and ins11546---547.

[0022] If the mutation is in *SF3B1*, the mutation is advantageously selected from the group consisting of G347V, R387W, R387Q, E592K, E622D, Y623C, R625L, R625C, H662Q, H662D, K666N, K666T, K666E, K666R, K700E, V701F, A708T, G740R, G740E, A744P, D781G and E783K.

[0023] Accordingly, it is an object of the invention to not encompass within the invention any previously known product, process of making the product, or method of using the product such that Applicants reserve the right and hereby disclose a disclaimer of any previously known product, process, or method. It is further noted that the invention does not intend to encompass within the scope of the invention any product, process, or making of the product or method of using the product, which does not meet the written description and enablement requirements of the USPTO (35 U.S.C. §112, first paragraph) or the EPO (Article 83 of the EPC), such that Applicants reserve the right and hereby disclose a disclaimer of any previously described product, process of making the product, or method of using the product.

[0024] It is noted that in this disclosure and particularly in the claims and/or paragraphs, terms such as "comprises", "comprised", "comprising" and the like can have the meaning attributed to it in U.S. Patent law; e.g., they can mean "includes", "included", "including", and the like; and that terms such as "consisting essentially of" and "consists essentially of" have the meaning ascribed to them in U.S. Patent law, e.g., they allow for elements not explicitly recited, but exclude elements that are found in the prior art or that affect a basic or novel characteristic of the invention.

[0025] These and other embodiments are disclosed or are obvious from and encompassed by, the following Detailed Description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] The following detailed description, given by way of example, but not intended to limit the invention solely to the specific embodiments described, may best be understood in conjunction with the accompanying drawings.

[0027] Figure 1. Prevalence of somatic mutation by age. Colored bands represent 50th, 75th, and 95th percentiles.

[0028] Figure 2. Characteristics of called somatic variants. A) Ten most frequently mutated genes. B) Number of subjects with 1, 2, 3, or 4 called variants. C) Percentages of type of single nucleotide base pair changes seen in the called variants. D) Allele fractions of called somatic variants. Allele fraction defined as variant reads divided by variant plus reference reads. For variants on the X chromosome in men this number was divided by 2.

[0029] Figure 3. Development of hematologic malignancies. A) Forest plot for risk of developing hematologic malignancy in those with somatic mutations overall and with VAF≥0.10, relative to those without mutations. Diamond represents the results of fixed-effects meta-analysis for the 2 cohorts, and horizontal lines are 95 percent confidence intervals. Hazard ratios were estimated by competing risks regression with death as the competing risk. The analysis includes adjudicated cancer information from MEC and unadjudicated information by annual subject interview in JHS. For interview data, leukemia, lymphoma, multiple myeloma, blood cancer, and spleen cancer were considered hematologic malignancy. All models included age groups (less than 50, 50-59, 60-69, greater than 70), diabetes status, and sex as covariates. B) Cumulative incidence plot for developing hematologic malignancy. Curves were generated from competing risks data with death as the competing risk. C) VAF in subjects that did or did not develop hematologic malignancy, p-value from Wilcoxon test.

[0030] Figure 4. Effect of somatic mutations on all-cause mortality. A) Forest plot for all-cause mortality risk associated with having a somatic clone. Diamond represents the results of fixed-effects meta-analysis of all cohorts, and horizontal lines are 95 percent confidence intervals. All models included age groups (less than 60, 60-69, 70-79, 80-89, and 90 or greater), diabetes status, and sex as covariates in a Cox proportional hazards model. Botnia includes Helsinki-sib and Diabetes-reg. B) Kaplan-Meier survival curves from the same cohorts as above, with p-values from log rank test. Left panel is those younger than 70 at time of DNA ascertainment, and right panel is those 70 or older. C) Cox proportional hazards model for all-cause mortality for those with or without mutations stratified by normal or high RDW. Diamond represents the results of fixed-effects meta-analysis of all cohorts, and horizontal lines are 95 percent confidence intervals. All models included age groups (less than 60, 60-69, 70-79, 80-89, and 90 or greater), diabetes status, and sex as covariates.

[0031] Figure 5. Association of somatic mutations with incident cardiovascular disease. A-B) Cumulative incidence plots for incident coronary heart disease (CHD) (A) and ischemic

stroke (B). Curves were generated from competing risks data with death as the competing risk. Those with prevalent events were excluded from the analyses. C-D) Forest plots for risk of developing incident CHD (C) and ischemic stroke (D) in those with somatic mutations. Diamond represents the results of fixed-effects meta-analysis using beta-coefficients from competing risks regressions for both cohorts, and horizontal lines are 95 percent confidence intervals. Age groups (less than 50, 50-59, 60-69, and 70 or greater), T2D status, sex, systolic blood pressure groups (less than 140 mm Hg, 140-160 mm Hg, and greater than 160 mm Hg), and body mass index groups (less than 25, 25-35, and greater than 35) were included as categorical covariates in the competing risks regression models, with death as the competing risk. Those with prevalent events were excluded from the analyses.

[0032] Figure 6. Characteristics of *DNMT3A* variants. A) Frequency of *DNMT3A* nonsense, frameshift, and splice-site variants called as somatic by age group. B) Frequency of R882 and non-R882 missense variants called as somatic by age group. C) Allele fraction of called *DNMT3A* variants by mutation type.

[0033] Figure 7. Co-mutations. A) Co-mutation plot, individuals are represented by columns. Black rectangles represent mutated genes, red rectangles represent 2 separate mutations in the same gene. B) Correlation plot for variant allele fraction (VAF) from the 49 subjects with 2 mutations.

[0034] Figure 8. Factors associated with clonality A) Frequency of mutation for males and females by age group. For those 60 or older, being male is associated with having a detectable clone (OR 1.3, 95% CI 1.1-1.5, p=0.005 by multivariable logistic regression using age, sex, T2D and BMI as covariates). B) Frequency of mutation for those with and without type 2 diabetes by age group. C) Frequency of mutation for non-Hispanics, Hispanics, and South Asians by age group.

[0035] Figure 9. Mutations by ethnic backgroundNumber of mutations for each gene stratified by ethnic background.

[0036] Figure 10. Blood counts for individuals with and without detectable mutations. Dots represent individuals. Box represents 25th and 75th percentiles, line in box represents median. Whiskers represent 5th and 95th percentiles. For listed genes, individuals only had mutations in that gene, and not other genes. Dashed red lines represent 11.5% and 14.5%, the normal ranges for RDW. Abbreviations: WBC-white blood cell count, PLT-platelet count, RDW-red cell

distribution width, MCV-mean corpuscular volume. Individuals were from Jackson Heart Study, Longevity Genes Project, Botnia, Helsinki-sib, or Malmo-sib.

[0037] Figure 11. Kaplan-Meier Curves for overall survival by cohorts.

[0038] Figure 12. Kaplan-Meier curves for individuals with or without clones, stratified by high (≥14.5%) or normal (<14.5%) red cell distribution width. Hash marks represent censored individuals. Individuals were from Jackson Heart Study or Longevity Genes Project.

[0039] Figure 13. Risk model for coronary heart disease and ischemic stroke. Regression parameters are same as shown for Figure 5C and 5D.Hazard ratios were estimated using competing risks regression with death as the competing risk. P-values are derived from the Fine-Gray test. Individuals with prior coronary heart disease (CHD) were excluded for CHD analysis, and individuals with prior ischemic stroke were excluded for stroke analysis. A) Coronary heart disease, B) ischemic stroke.

[0040] Figure 14. Mutation validation and serial samples. A) Eighteen variants were validated using targeted, amplicon based re-sequencing ("Rapid Heme Panel", see Methods). Comparison of VAF between the two methods is shown for the 18 variants. WES, whole exome sequencing. RHP, Rapid Heme Panel. B) Peripheral blood DNA from 4 to 8 years after the initial DNA collection was available for 13 subjects in JHS who had detectable mutations on exome sequencing. As described in the Supplementary Methods, amplicon-based targeted re-sequencing was performed for a panel of 95 genes. Graphs represent individual subjects, with mutation variant allele frequency (VAF) from the initial and later time points shown. All of the initial mutations detected on exome sequencing were still present in the later sample, and 2 subjects had acquired new mutations.

[0041] Figure 15. Risk of cancer, cardiovascular, or other death associated with clonality. Hazard ratios obtained by competing risks regression, with death by other causes as the competing risk. Results shown are risk associated with clonality. For cancer deaths, only non-hematologic cancers were included. Cardiovascular deaths included strokes (hemorrhagic or ischemic) and fatal myocardial infarction. All regressions included age groups (less than 50, 50-59, 60-69, 70 or older), diabetes status, and gender as covariates. Individuals were from Botnia, FUSION, MEC, and Jackson Heart Study. For Jackson Heart Study, only cardiovascular outcomes were adjudicated (all other deaths were considered unadjudicated).

[0042] Figure 16. (A) Forest plot for odds ratio (OR) of having a somatic mutation in those presenting with myocardial infarction (MI) compared to those without MI (referent). Results are shown for 2 cohorts (ATVB and PROMIS) and stratified by age groups. (B) Counts for number of mutations seen in cases versus controls for the most frequently mutated genes.

[0043] Figure 17. (A) Low power sections of the aortic root from recipients of Tet2-/- marrow (top row) or control marrow (bottom row) are shown at three time-points after initiation of diet. The left and center columns are oil red O stained, and the right column is Masson's trichrome. (B) Spleen (top row, H&E) and digit (bottom row, oil red O) histology are shown in mice after 14 weeks on diet. Note the large fat deposits (white arrowheads) and sheets of lipid-laden macrophages in the tissues (black arrowheads) of Tet2-/- recipients.

[0044] Figure 18. Risk of cardiovascular events from carrying a somatic clonal mutation. Risk if cardiovascular events (myocardial infarction, coronary revascularization, or stroke) in each cohort and fixed effects meta-analysis are displayed. Effect estimates are derived from a Cox proportional hazards model after accounting for age, sex, ethnicity, diabetes mellitus, smoking, hypertension, LDL cholesterol. There was no evidence of heterogeneity between the effect estimates in the two cohorts (*P* for heterogeneity = 0.79).

[0045] Figure 19. Coronary arterial calcification quantity by somatic clonal mutation carrier status. Among participants in the BioImage cohort, coronary arterial calcification (CAC) quantity was obtained as the total Agatston score. The relative difference in CAC in somatic clonal mutation carriers compared to non-carriers is estimated from linear regression with log-transformation of CAC quantity.

DETAILED DESCRIPTION OF THE INVENTION

[0046] The incidence of hematological malignancies increases with age and is associated with recurrent somatic mutations in specific genes. Applicants hypothesized that such mutations would be detectable in the blood of some individuals not known to have hematological disorders.

[0047] Applicants analyzed whole exome sequencing data from peripheral blood cell DNA of 17,182 individuals who were unselected for hematologic phenotypes. Applicants looked for somatic mutations by identifying previously characterized single nucleotide variants (SNVs) and small insertions/deletions (indels) in 160 genes recurrently mutated in hematological

malignancies. The presence of mutations was analyzed for association to hematological phenotypes, survival, and cardiovascular events.

[0048] Detectable somatic mutations were rare in individuals younger than 40, but rose appreciably with age. At ages 70-79, 80-89, and 90-108 these clonal mutations were observed in 9.6% (220 out of 2299), 11.7% (37 out of 317), and 18.4% (19 out of 103) of individuals, respectively. The majority of the variants occurred in 3 genes: DNMT3A, TET2, and ASXL1. The presence of a somatic mutation was associated with increased risk of developing hematologic malignancy (HR 11, 95% confidence interval [95% CI] 3.9-33), increased all-cause mortality (HR 1.4, 95% CI 1.1-1.8), and increased risk of incident coronary heart disease (HR 2.0, 95% CI 1.2-3.4) and ischemic stroke (HR 2.6, 95% CI 1.4-4.8).

[0049] Clonal hematopoiesis of indeterminate potential (CHIP) is a common pre-malignant condition in the elderly, and is associated with increased risk of transformation to hematologic malignancy and increased all-cause mortality, possibly due to increased cardiometabolic disease.

[0050] Cardiovascular disease is the leading cause of death worldwide. Given the association of somatic mutations with all-cause mortality beyond that explicable by hematologic malignancy and T2D, Applicants performed association analyses from two cohorts comprising 3,353 subjects with available data on coronary heart disease (CHD) and ischemic stroke (IS). After excluding those with prevalent events, Applicants found that those carrying a mutation had increased cumulative incidence of both CHD and IS (Figure 5A and 5B). In multivariable analyses that included age, sex, T2D, systolic blood pressure, and body mass index as covariates, the hazard ratio of incident CHD and IS was 2.0 (95% CI 1.2-3.5, P=0.015) and 2.6 (95% CI 1.3-4.8, P=0.003) in the individuals carrying a somatic mutation as compared to those without (Figure 5C and 5D, Figure 8).

[0051] For a subset of individuals, the traditional risk factors of smoking, total cholesterol, and high-density lipoprotein were also available; the presence of a somatic mutation remained significantly associated with incident CHD and IS even in the presence of these risk factors, and the risk was even greater in those with VAF$\geq$0.10 (Supplementary Table S12). Elevated RDW and high-sensitivity C-reactive protein (hsCRP) have also been associated with adverse cardiac outcomes (Tonelli M et al. Circulation 2008;117:163-8, Ridker PM et al. The New England journal of medicine 2002;347:1557-65), possibly reflecting an underlying inflammatory cause. In a multivariable analysis of 1,795 subjects from JHS, those with a mutation and RDW$\geq$14.5%

had a markedly increased risk of incident CHD, and this effect was independent of hsCRP (Supplementary Table S13).

[0052] Applicants find that somatic mutations leading to clonal outgrowth of hematopoietic cells are frequent in the general population. This entity, which Applicants term clonal hematopoiesis with indeterminate potential (CHIP), is present in over 10% of individuals over 70, making it one of the most common known pre-malignant lesions. The exact prevalence of CHIP is dependent on how cancer-causing mutations are defined and on the sensitivity of the technique used to detect mutations, and thus may substantially exceed this estimate. Unlike other pre-malignant lesions, CHIP appears to involve a substantial proportion of the affected tissue in most individuals; based on the proportion of alleles with the somatic mutation, Applicants find that a median of 18% of peripheral blood leukocytes are part of the abnormal clone. CHIP also persists over time; in all tested cases, the mutations were still present after 4 to 8 years.

[0053] The genes most commonly mutated in CHIP are *DNMT3A*, *TET2*, and *ASXL1*. This is consistent with previous studies that have found *DNMT3A* and *TET2* mutations to be frequent and early events in AML and MDS (Jan M et al. Science translational medicine 2012;4:149ra18, Shlush LI et al. Nature 2014;506:328-33, Papaemmanuil E et al. The New England journal of medicine 2011;365:1384-95, Welch JS et al. Cell 2012;150:264-78). Murine models of *DNMT3A* or *TET2* loss-of-function demonstrate that mutant HSCs have altered methylation patterns at pluripotency genes and a competitive advantage compared to wild-type HSCs, but mice rarely develop frank malignancy, and then only after long latency (Jeong M et al. Nature genetics 2014;46:17-23, Koh KP et al. Cell stem cell 2011;8:200-13, Challen GA et al. Nature genetics 2012;44:23-31, Moran-Crusio K et al. Cancer cell 2011;20:11-24). Similarly, Applicants' data show that humans with CHIP can live for many years without developing hematological malignancies, though they do have increased risk relative to those without mutations.

[0054] *TET2* and *DNMT3A* are frequently mutated in some lymphoid malignancies, and the initiating event for such tumors may occur in a HSC (Neumann M et al. Blood 2013;121:4749-52, Quivoron C et al. Cancer cell 2011;20:25-38, Odejide O et al. Blood 2014;123:1293-6, Asmar F et al. Haematologica 2013;98:1912-20, Couronne L et al. The New England journal of medicine 2012;366:95-6). While it is most likely that these mutations occur in a HSC, it also

possible that they occur in committed myeloid progenitors or mature lymphoid cells that have acquired long-term self-renewal capacity.

[0055] The use of somatic mutations to aid in the diagnosis of patients with clinical MDS is becoming widespread. Applicants' data demonstrate that the majority of individuals with clonal mutations in peripheral blood do not have MDS or another hematological malignancy, nor do the majority develop a clinically diagnosed malignancy in the near term. At this time, it would be premature to genetically screen healthy individuals for the presence of a somatic clone, as the positive predictive value for current or future malignancy is low.

[0056] The invention relates to method for identifying and selecting a subject with increased risk of developing a cardiometabolic disease and optionally a hematological cancer, which may comprise the steps of: (a) sequencing at least part of a genome which may comprise one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject, (b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer.

[0057] The invention also relates to a method for identifying and selecting a subject with an increased risk of developing a cardiometabolic disease and optionally a hematological cancer and providing a personalized medicine method, said method which may comprise the steps of (a) sequencing at least part of a genome which may comprise one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject, (b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer, and (c) initiating a treatment or monitoring regimen to suppress said mutation(s) in the subject, thereby decreasing risk of developing a cardiometabolic disease and optionally a hematological cancer.

[0058] The presence of said mutation(s) in the above embodiments may indicate an increase in red blood cell distribution width (RDW).

[0059] The cardiometabolic disease may be atherosclerosis, coronary heart disease (CHD) or ischemic stroke (IS).

[0060] In embodiments wherein an increased risk for hematological cancer is also screened in addition to a cardiometabolic disease, the hematological cancer may be a leukemia, a lymphoma, a myeloma or a blood syndrome. The leukemia may be an acute myeloid leukemia (AML) or chronic myelogenous leukemia (CML). The blood syndrome may be myelodysplastic syndrome (MDS).

[0061] The one more cells in the blood sample may be hematopoietic stem cells (HSCs), committed myeloid progenitor cells having long term self-renewal capacity or mature lymphoid cells having long term self-renewal capacity.

[0062] In some embodiments the part of the genome that is sequenced may be an exome. In other embodiments, the sequencing may be whole exome sequencing (WES).

[0063] In an advantageous embodiment, the subject is a human. In other embodiments, the human may exhibit one or more risk factors of being a smoker, having a high level of total cholesterol or having high level of high-density lipoprotein (HDL).

[0064] The mutations of at least *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* may be frameshift mutations, nonsense mutations, missense mutations or splice-site variant mutations.

[0065] If the mutation is in *DNMT3A*, the mutation may advantageously be a mutation in exons 7 to 23. In a particularly advantageous embodiment, the mutation in DNMT3A is a mutation selected from the group consisting of P307S, P307R, R326H, R326L, R326C, R326S, R366P, R366H, R366G, A368T, F414L, F414S, F414C, C497Y, Q527H, Q527P, Y533C, G543A, G543S, G543C, L547H, L547P, L547F, M548I, M548K, G550R, W581R, W581G, W581C, G646V, G646E, L653W, L653F, V657A, V657M, R659H, Y660C, R676W, R676Q, G685R, G685E, G685A, D686Y, D686G, G699R, G699S, G699D, P700S, P700R, P700Q, D702N, D702Y, V704M, V704G, I705F, I705T, I705S, C710S, S714C, N717S, N717I, P718L, R720H, R720G, Y724C, R729Q, R729W, R729G, F731L, F732del, F732S, F732L, F734L, F734C, Y735C, Y735N, Y735S, R736H, R736C, R736P, L737H, L737V, L737F, L737R, A741V, R749C, R749L, F751L, F752del, F752C, F752L, F752I, F752V, L754R, L754H, F755S, F755I, F755L, M761I, M761V, G762C, S770W, S770P, R771Q, F772I, F772V, L773R, E774K, E774D, D781G, R792H, G796D, G796V, N797Y, N797H, P799R, P799H, R803S, P804S, P804L, S828N, K829R, Q842E, P849L, D857N, W860R, F868S, G869S, G869V, M880V, S881R, S881I, R882H, R882P, R882C, R882G, Q886R, G890D, L901R, L901H, P904L, F909C and A910P.

[0066] If the mutation is in *TET2*, the mutation is advantageously selected from the group consisting of S282F, N312S, L346P, S460F, D666G, P941S, and C1135Y.

[0067] If this mutation is in *ASXL1*, the mutation is advantageously a mutation in exon 11-12.

[0068] If the mutation is in *TP53*, the mutation is advantageously a mutation selected from the group consisting of S46F, G105C, G105R, G105D, G108S, G108C, R110L, R110C, T118A, T118R, T118I, L130V, L130F, K132Q, K132E, K132W, K132R, K132M, K132N, C135W, C135S, C135F, C135G, Q136K, Q136E, Q136P, Q136R, Q136L, Q136H, A138P, A138V, A138A, A138T, T140I, C141R, C141G, C141A, C141Y, C141S, C141F, C141W, V143M, V143A, V143E, L145Q, L145R, P151T, P151A, P151S, P151H, P152S, P152R, P152L, T155P, R158H, R158L, A159V, A159P, A159S, A159D, A161T, A161D, Y163N, Y163H, Y163D, Y163S, Y163C, K164E, K164M, K164N, K164P, H168Y, H168P, H168R, H168L, H168Q, M169I, M169T, M169V, T170M, E171K, E171Q, E171G, E171A, E171V and E171D, V172D, V173M, V173L, V173G, R174W, R175G, R175C, R175H, C176R, C176G, C176Y, C176F, C176S, P177R, P177R, P177L, H178D, H178P, H178Q, H179Y, H179R, H179Q, R181C, R181Y, D186G, G187S, P190L, P190T, H193N, H193P, H193L, H193R, L194F, L194R, I195F, I195N, I195T, V197L, G199V, Y205N, Y205C, Y205H, D208V, R213Q, R213P, R213L, R213Q, H214D, H214R, S215G, S215I, S215R, V216M, V217G, Y220N, Y220H, Y220S, Y220C, E224D, I232F, I232N, I232T, I232S, Y234N, Y234H, Y234S, Y234C, Y236N, Y236H, Y236C, M237V, M237K, M237I, C238R, C238G, C238Y, C238W, N239T, N239S, S241Y, S241C, S241F, C242G, C242Y, C242S, C242F, G244S, G244C, G244D, G245S, G245R, G245C, G245D, G245A, G245V, G245S, M246V, M246K, M246R, M246I, N247I, R248W, R248G, R248Q, R249G, R249W, R249T, R249M, P250L, I251N, L252P, I254S, I255F, I255N, I255S, L257Q, L257P, E258K, E258Q, D259Y, S261T, G262D, G262V, L265P, G266R, G266E, G266V, R267W, R267Q, R267P, E271K, V272M, V272L, R273S, R273G, R273C, R273H, R273P, R273L, V274F, V274D, V274A, V274G, V274L, C275Y, C275S, C275F, A276P, C277F, P278T, P278A, P278S, P278H, P278R, P278L, G279E, R280G, R280K, R280T, R280I, R280S, D281N, D281H, D281Y, D281G, D281E, R282G, R282W, R282Q, R282P, E285K, E285V, E286G, E286V, E286K, K320N, L330R, G334V, R337C,R337L, A347T, L348F, T377P.

[0069] If the mutation is in *JAK2*, the mutation is advantageously selected from the group consisting of N533D, N533Y, N533S, H538R, K539E, K539L, I540T, I540V, V617F, R683S, R683G, del/ins537---539L, del/ins538---539L, del/ins540---543MK, del/ins540---544MK, del/ins541- -543K, del542---543, del543---544 and ins11546---547.

[0070] If the mutation is in *SF3B1*, the mutation is advantageously selected from the group consisting of G347V, R387W, R387Q, E592K, E622D, Y623C, R625L, R625C, H662Q, H662D, K666N, K666T, K666E, K666R, K700E, V701F, A708T, G740R, G740E, A744P, D781G and E783K.

[0071] Aging is associated with a large increase in the prevalence of atherosclerosis and cancer. Applicants recently analyzed whole exome sequencing data from over 17,000 individuals who were unselected for hematological phenotypes and found that at least 10% of humans age 70 or older harbor a mutation in a known cancer-causing gene in their blood cells (Jaiswal et al., *NEJM* 2014). Surprisingly, the presence of these mutations was associated with an increased risk of myocardial infarction (hazard ratio [HR]=2.0) and ischemic stroke (HR=2.6) in ~3,000 individuals for whom long-term follow-up information was available. It is unknown if somatic mutations that cause clonal expansion of hematopoietic stem cells also affect the function of differentiated blood cells such as macrophages, which are considered to be important mediators of atherosclerosis. Preliminary mouse data indicates that at least one of these mutations (*TET2*) does indeed directly alter blood cell function to lead to accelerated atherosclerosis. This application seeks to definitively establish whether these somatic mutations in blood cells are causally linked to atherosclerosis. Applicants expand the initial findings by assessing whether mice bearing mutations in *DNMT3A*, *TET2*, or *JAK2* in their blood cells have accelerated atherosclerosis. Applicants probe the molecular mechanism of accelerated atherosclerosis in these mice. Applicants look for alterations in expression of liver X receptor (LXR), peroxisome proliferator activated receptor gamma (PPARG), and lipopolysaccharide (LPS) target genes in mutant macrophages using RNA-seq, DNase footprinting, and bisulfite sequencing. Applicants identify and validate therapeutic targets in macrophages with these somatic mutations using drug screens in newly created cell-line models.

[0072] Atherosclerosis is the leading cause of death in the United States; however, little is known about non-lipid risk factors in humans. This application relates to a mechanism behind

the proposed causal association between these somatic mutations in blood cells and atherosclerosis.

[0073] Cancer is thought to arise via the stepwise acquisition of genetic or epigenetic mutations that transform a normal cell (P. C. Nowell, Science 194, 23-28 (1976), J. S. Welch et al., Cell 150, 264-278 (2012)). For most hematological cancers, the genes that are frequently mutated are now known (R. Bejar et al., N Engl J Med 364, 2496-2506 (2011, N. Cancer Genome Atlas Research, Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 368, 2059-2074 (2013), E. Papaemmanuil et al., Blood 122, 3616-3627; quiz 3699 (2013), T. Haferlach et al., Leukemia 28, 241-247 (2014)). Applicants hypothesized that these mutations would be detectable in the blood cells of some individuals not known to have hematological disorders, and that the presence of these mutations would represent a pre-malignant state. To test this hypothesis, Applicants examined whole exome sequencing data from large cohorts who were unselected for hematological phenotypes (S. Jaiswal et al., N Engl J Med 371, 2488-2498 (2014), G. Genovese et al., N Engl J Med 371, 2477-2487 (2014), M. Xie et al., Nat Med 20, 1472-1478 (2014)). The DNA source for these exomes was peripheral blood cells. Applicants found that ~10% of individuals over the age of 70 harbored a detectable mutation in their blood cells (Figure 1). The majority of the mutations could be accounted for by loss-of-function mutations in 3 genes, *DNMT3A*, *TET2*, and *ASXL1,* while activating mutations in *JAK2* were the fifth most common mutation. These mutations are common in myeloid malignancies. Compared to those without mutations, persons who harbored these mutations were over 10 times more likely to develop a hematologic malignancy over the next several years, confirming that this condition is a *bona fide* pre-malignant entity (D. P. Steensma et al., Blood 126, 9-16 (2015)).

[0074] Surprisingly, the presence of these mutations was also associated with an increased risk of mortality that could not be explained by hematological malignancy alone. Applicants found that those with mutations had a higher rate of having a fatal myocardial infarction (MI) or ischemic stroke. When Applicants looked at risk of developing incident coronary heart disease or ischemic stroke, Applicants found that the presence of mutations was an even stronger risk factor than smoking, hypertension, or elevated cholesterol (Figure 5).

[0075] To replicate these findings, Applicants examined whole exome sequencing data from two large case-control cohorts designed to study early-onset MI. The cases (n=4,405) were

individuals from Italy or Pakistan who had blood cells collected for DNA sequencing at the time of presentation of MI, and controls (n=3,006) were age-matched healthy individuals from the same populations. Applicants found an even greater risk associated with mutations in these cohorts; the presence of a mutation was over 8 times more likely in the MI cases as compared to age-matched controls for those 40 or younger, and 3 times more likely in those age 41-50 (Figure 16A). *DNMT3A*, *TET2*, *ASXL1*, and *JAK2* were the most frequently mutated genes, and were strongly enriched in the cases (Figure 16B).

[0076] These genetic studies have firmly established a statistical link between the presence of these mutations and atherosclerotic disease, but alone cannot establish causality. It is possible that this association is merely a correlation, and that the presence of mutations is a molecular marker of aging. However, a few observations suggest a causal relationship. First, Applicants' human genetic data shows a dose-response relationship between the size of the mutant clone and risk of atherosclerotic disease. There is a greatly increased risk of incident coronary heart disease (HR=4.4) in those with a mutant allele fraction of ≥0.10 (meaning ≥20% of the blood cells harbor the mutation, including the macrophages in vessel walls), while those with clones smaller than this size had no increased risk. Second, there is a large body of evidence to suggest that macrophages are important in the development of atherosclerosis (K. J. Moore et al., Nat Rev Immunol 13, 709-721 (2013)). They are the most prevalent cell type within vessel wall lesions, and are critical for reverse cholesterol transport (RCT), the process by which excess lipid is exported from tissues for clearance by the liver (M. Cuchel et al., Circulation 113, 2548-2555 (2006)). Third, inflammation and alterations in blood cell parameters are linked to adverse cardiovascular outcomes in epidemiological studies (R. Ross, N Engl J Med 340, 115-126 (1999), P. Libby, Nature 420, 868-874 (2002), P. Libby et al., Am J Med 116 Suppl 6A, 9S-16S (2004), P. M. Ridker et al., N Engl J Med 347, 1557-1565 (2002), M. Tonelli et al., Circulation 117, 163-168 (2008), M. Madjid et al., J Am Coll Cardiol 44, 1945-1956 (2004)).

[0077] Wild-type mice have much lower baseline cholesterol and triglyceride levels than humans in modern societies. Therefore, a standard experimental model is to use mice lacking low-density lipoprotein receptor (*Ldlr*-/-) (E. Maganto-Garcia et al., Current protocols in immunology / edited by John E. Coligan ... [et al.] Chapter 15, Unit 15 24 11-23 (2012)). These mice rapidly develop atherosclerosis when fed a high cholesterol diet. By transplanting bone

marrow into *Ldlr*-/- mice from a genetically defined mouse strain, one can test the effect of a genetic perturbation in hematopoietic cells on the development of atherosclerosis.

[0078] In general, studies of experimental atherosclerosis have revealed two relevant processes that can be influenced by macrophages: 1) reverse cholesterol transport, and 2) the local inflammatory milieu. The importance of macrophage reverse cholesterol transport is exemplified by mice in which both liver X receptor genes have been knocked out in the hematopoietic compartment (*LXRαβ*-/-) (R. K. Tangirala et al., Proceedings of the National Academy of Sciences of the United States of America 99, 11896-11901 (2002)). Recipients transplanted with marrow from these mice develop accelerated atherosclerosis and have a lipid accumulation phenotype in several tissues. LXRs are a class of nuclear receptors expressed in macrophages, hepatocytes, and intestinal epithelium that activate transcription of genes involved in cholesterol transport and lipid metabolism when bound by ligand. In the absence of LXR mediated transcription, macrophages do not efficiently upregulate expression of the cholesterol transporters ABCA1 and ABCG1. This leads to a defect in exporting cholesterol to acceptor molecules such as high-density lipoprotein (HDL) and Apo-AI, ultimately resulting in foam cell formation in tissues such as the vessel wall, skin, and spleen.

[0079] The role of inflammation in atherosclerosis has also been well studied. Mice that are deficient in toll-like receptor (TLR)-2 (A. E. Mullick et al., J Clin Invest 115, 3149-3156 (2005)), TLR4 (K. S. Michelsen et al., Proceedings of the National Academy of Sciences of the United States of America 101, 10679-10684 (2004)), or inflammasome (P. Duewell et al., Nature 464, 1357-1361 (2010)) signaling have reduced atherosclerosis, whereas mice that lack the inflammatory repressor *BCL6* in macrophages have accelerated atherosclerosis (G. D. Barish et al., Cell Metab 15, 554-562 (2012)). Macrophage mediated inflammation results in recruitment of other immune cells and intimal hyperplasia, resulting in increased lesion size and instability. Additional evidence indicates that inflammatory signaling may also directly inhibit macrophage RCT (A. Castrillo et al., Mol Cell 12, 805-816 (2003)). Finally, activation of another class of nuclear receptors, PPARs, ameliorates atherosclerosis by increasing macrophage RCT via LXR activation (A. Chawla et al., Mol Cell 7, 161-171 (2001)), as well as by attenuating inflammation via induction of the transcriptional repressors *BCL6*, *NCOR1,* and *NCOR2* (G. Pascual et al., Nature 437, 759-763 (2005), A. Chawla, Circ Res 106, 1559-1569 (2010)).

[0080] To test the hypothesis that mutations in blood cells are causally linked to atherosclerosis, Applicants turned to defined mouse models. Applicants crossed *Tet2*$^{fl/fl}$ mice to *Vav1*-Cre mice; the resultant *Tet2*$^{fl/fl}$, *Vav1*-Cre mice have exon 3 of *TET2* deleted, leading to complete loss of Tet2 function in all of their hematopoietic cells. Previous studies have demonstrated that hematopoietic stem cells (HSCs) from these mice have a differentiation defect that results in a progressive increase in HSC frequency and consequent clonal expansion of the mutant cells in a competitive transplant setting (K. Moran-Crusio et al., Cancer cell 20, 11-24 (2011)). Thus, this model system recapitulates the clonal advantage of *TET2* mutant hematopoietic cells seen in humans. Applicants then transplanted bone marrow from these mice, or control *Vav1*-Cre mice, into *Ldlr*-/- recipient mice and initiated high cholesterol diet. Tissues were then examined at several time points. The results from these preliminary experiments are clear; mice that received *Tet2*-/- bone marrow had not only increased lesion size in the aortic root (Figure 17A), but also striking numbers of lipid-laden macrophages in the spleen, skin, and peritoneal fluid (Figure 17B). This phenotype is remarkably similar to the one described for mice transplanted with *LXRαβ*-/- marrow (R. K. Tangirala et al., Proceedings of the National Academy of Sciences of the United States of America 99, 11896-11901 (2002)). In summary, these results point to a marked deficit of reverse cholesterol transport in macrophages that lack *TET2,* of which one manifestation is accelerated atherosclerosis.

[0081] The present invention relates to determining the effect of mutations in *TET2*, *DNMT3A*, and *JAK2* on atherosclerosis development *in vivo*. Applicants has already generated or obtained models for *DNMT3A* loss-of-function (*Dnmt3a*$^{fl/fl}$ mice) (S. Nguyen et al., Dev Dyn 236, 1663-1676 (2007)) and *JAK2* gain-of-function (floxed *Jak2* V617F knock-in heterozygous mice) (A. Mullally et al., Cancer cell 17, 584-596 (2010)). Unfortunately, there are no publicly available mouse models of *ASXL1* that mimic the mutations seen in humans, which are truncating mutations in exon 12. Applicants hypothesize that introducing these mutations into hematopoietic cells lead to accelerated atherosclerosis using the Ldlr-/- transplant system described above. Applicants perform the transplants described above and analyze mice after 5, 9, and 14 weeks on high-cholesterol (1.25%) diet. Lesion size and cellular composition, as well serum lipid profiles are measured. Applicants also examine lipid content of macrophages in the spleen, lung, liver, skin, intestines, and peritoneal fluid to determine if there is a global defect in macrophage RCT. It is important to note that the mice are not expected to develop a leukemic

phenotype within this time frame. Applicants also determine whether other cells within the hematopoietic compartment besides macrophages can contribute to the phenotype. Applicants cross the *TET2*, *DNMT3A*, and *JAK2* mutant mice to mice that have *CD2*-Cre (lymphoid specific) or *PF4*-Cre (platelet specific), then perform transplants into *Ldlr*-/- mice and determine the extent of atherosclerosis in each model after 9 weeks on diet. *Adgre1*-Cre mice, which express Cre in mature macrophages only, are used to test the hypothesis that the mutations do not necessarily need to occur in stem cells to exert a phenotype.

[0082] It is possible that macrophages are the cells responsible for the phenotype, but that the mutations must occur in stem cells in order to have a functional effect due to altered differentiation. If the mutant strains crossed with *Adgre1*-Cre do not have a phenotype, Applicants isolate monocyte precursors (J. Hettinger et al., Nat Immunol 14, 821-830 (2013)) from mutant Vav1-Cre mice and transplant these into recipient *Ldlr*-/- mice to test the hypothesis that monocytes/macrophages alone can recapitulate the phenotype.

[0083] The present application also relates to determining the mechanism by which mutations in *TET2*, *DNMT3A*, and *JAK2* lead to altered macrophage function and accelerated atherosclerosis. As detailed above, much experimental evidence indicates that macrophage function has a profound effect on the development of atherosclerosis. Applicants hypothesize that mutations in TET2, DNMT3A, and JAK2 lead to accelerated atherosclerosis by inhibiting macrophage reverse cholesterol transport, increasing macrophage-mediated inflammation, or both. Applicants characterize the transcriptional response of mutant macrophages to induction of RCT and inflammation by agonists of LXRs, PPARG, and TLR4.

[0084] To ensure robustness of results, Applicants utilize 2 types of primary macrophages: thioglycollate-induced peritoneal macrophages and bone-marrow derived macrophages grown *in vitro* in the presence of cytokines. The specific agonists to be used are GW3965 (LXR alpha/beta agonist), pioglitazone (PPARG agonist), and LPS (TLR4 agonist). Gene expression are assessed by RNA-sequencing. Gene set enrichment analysis are used to determine the extent of transcriptional changes for various classes of genes by comparing expression in mutant/wild-type cells that have been treated to mutant/wild-type cells that are not treated. For these experiments, gene expression are measured at two time-points after exposure to agonists to also assess the kinetics of the transcriptional response.

[0085] Applicants are uncovering the mechanistic link between the specific mutations and alterations in gene expression in response to the agonists listed above. Applicants perform DNase footprinting and bisulfite sequencing to molecularly determine the how mutant macrophages are altered in response to the specific agonists. Peritoneal macrophages are used in these experiments because they represent an *in vivo* baseline epigenetic state. DNase footprinting provides a powerful method to infer transcription factor binding in a genome-wide manner by identifying DNA sequence motifs that are protein bound. Bisulfite sequencing provides single base resolution of cytosine methylation. As *TET2* and *DNMT3A* are enzymes that alter DNA methylation, it is likely that perturbing their function result in an abnormal epigenetic state. For example, *TET2* converts 5-methylcytosine to 5-hydroxymethylcytosine, which ultimately leads to de-methylation. As methylation at promoters and enhancers anti-correlates with gene expression and transcription factor binding, Applicants hypothesize that loss of *TET2* function results in abnormal methylation of cis-regulatory elements for LXR/PPARG targets, reduced binding of transcription factors at these elements, and ultimately attenuated expression of the target genes. *DNMT3A*, a *de novo* DNA methyltransferase, has a function that opposes that of *TET2*. Applicants hypothesize that *de novo* methylation at cis-regulatory elements of pro-inflammatory genes is necessary to attenuate an inflammatory response in macrophages, and that this regulatory check is lacking in *DNMT3A* null macrophages. Unlike *TET2* and *DNMT3A*, *JAK2* is unique in that it is an activator of STAT signaling, not an epigenetic regulator. STAT transcription factors are known to activate a variety of genes involved in the immune response. Applicants hypothesize that constitutive STAT signaling in the setting of *JAK2* activating mutations results in a stronger/more protracted pro-inflammatory phenotype in mutant macrophages, which are determined by assessing the transcriptional response to LPS. Applicants are performing functional assays to assess the physiological effect of the mutations on macrophage activity. Applicants assess cholesterol efflux, cholesterol uptake, efferocytosis, LPS tolerance, and M1/M2 polarization in mutant and wild-type macrophages in response to the specific agonists. Applicants hypothesize that one or more of these processes are altered due to the mutations.

[0086] While DNase footprinting is a powerful technique, it relies upon knowing DNA motifs for specific transcription factors to accurately assign binding. Furthermore, enhancers are inferred from transcription factor binding, rather than by assessing chromatin marks. Therefore,

Applicants also consider chromatin immunoprecipitation sequencing for LXR alpha and beta, PPARG, RelA, RelB, c-REL, STAT3, and STAT5, as well as the canonical histone marks for enhancers (H3K4me1, H3K4me3, H3K27ac, H3K27me3). Finally, Applicants also consider assessment of 5-hydroxymethylcytosine as technology to detect this mark continues to improve.

[0087] The present invention also relates to identifying molecules and pathways that can reverse the pro-atherogenic phenotype of mutant macrophages. Applicants are identifying therapeutic targets for macrophages that have mutations in *TET2*, *DNMT3A*, or *JAK2*. Applicants hypothesize that specific molecules or pathways can be targeted to reverse some aspects of the pro-atherogenic phenotype of mutant macrophages. Applicants are testing whether existing agonists for LXR and PPARG can reverse the phenotype in the bone marrow transplant models. Mice are placed on drug after 8 weeks on high cholesterol diet, and plaque size are compared between treated/untreated and mutant/wild-type mice after an additional 12 weeks on diet. While these approaches may prove efficacious in mice, current LXR and PPARG agonists are not considered front-line drugs in humans because of side-effects and the potential for serious adverse events. Thus, Applicants are identifying novel targets that can reverse the phenotype specifically in mutant macrophages. Applicants utilize THP-1 cells, a monocytic leukemia cell line that retains the ability differentiate into macrophages and is responsive to LXR and PPARG agonists. Frameshift mutations in *TET2* and *DNMT3A* are created by CRISP-Cas9, while *JAK2* V617F are introduced lentivirally. Applicants also introduce artificial reporters that activate fluorescence by LXR agonists, PPARG agonists, or LPS (via NF-κB activation). Applicants are screening the cells using a highly targeted drug library of 481 molecules that affect distinct cellular pathways and processes (A. Basu et al., Cell 154, 1151-1161 (2013)). The read-out is an image-based assessment of fluorescence intensity from the reporter. The aim is to identify molecules that preferentially activate or repress the reporter in mutant, but not wild-type macrophages. Applicants are also testing individual compounds that show desired activity in the mouse bone marrow transplant model, as well as in *in vitro* assays in primary macrophages.

[0088] The small molecule approach described above may not yield any viable candidates for a variety of reasons. In this case, a second screen is performed using a CRISPR-Cas9/small-guide RNA system to inactivate all protein-coding genes. Briefly, THP-1 cells with reporter are transduced with Cas9 and a pooled library of small guide RNA molecules. After inducing macrophage differentiation, Applicants identify cells that activate or repress the reporter in

response to agonists. DNA sequencing is used to determine which guides are present, and the positive hits are further validated to ensure that they preferentially affect mutant cells.

[0089] With respect to general information on CRISPR-Cas Systems, components thereof, and delivery of such components, including methods, materials, delivery vehicles, vectors, particles, AAV, and making and using thereof, including as to amounts and formulations, all useful in the practice of the instant invention, reference is made to: US Patents Nos. 8,697,359, 8,771,945, 8,795,965, 8,865,406, 8,871,445, 8,889,356, 8,889,418, 8,895,308, 8,906,616, 8,932,814, 8,945,839, 8,993,233 and 8,999,641; US Patent Publications US 2014-0310830 (US App. Ser. No. 14/105,031), US 2014-0287938 A1 (U.S. App. Ser. No. 14/213,991), US 2014-0273234 A1 (U.S. App. Ser. No. 14/293,674), US2014-0273232 A1 (U.S. App. Ser. No. 14/290,575), US 2014-0273231 (U.S. App. Ser. No. 14/259,420), US 2014-0256046 A1 (U.S. App. Ser. No. 14/226,274), US 2014-0248702 A1 (U.S. App. Ser. No. 14/258,458), US 2014-0242700 A1 (U.S. App. Ser. No. 14/222,930), US 2014-0242699 A1 (U.S. App. Ser. No. 14/183,512), US 2014-0242664 A1 (U.S. App. Ser. No. 14/104,990), US 2014-0234972 A1 (U.S. App. Ser. No. 14/183,471), US 2014-0227787 A1 (U.S. App. Ser. No. 14/256,912), US 2014-0189896 A1 (U.S. App. Ser. No. 14/105,035), US 2014-0186958 (U.S. App. Ser. No. 14/105,017), US 2014-0186919 A1 (U.S. App. Ser. No. 14/104,977), US 2014-0186843 A1 (U.S. App. Ser. No. 14/104,900), US 2014-0179770 A1 (U.S. App. Ser. No. 14/104,837) and US 2014-0179006 A1 (U.S. App. Ser. No. 14/183,486), US 2014-0170753 (US App Ser No 14/183,429); US 2015-0184139 (U.S. App. Ser. No. 14/324,960); 14/054,414 European Patent Applications EP 2 771 468 (EP13818570.7), EP 2 764 103 (EP13824232.6), and EP 2 784 162 (EP14170383.5); and PCT Patent Publications WO 2014/093661 (PCT/US2013/074743), WO 2014/093694 (PCT/US2013/074790), WO 2014/093595 (PCT/US2013/074611), WO 2014/093718 (PCT/US2013/074825), WO 2014/093709 (PCT/US2013/074812), WO 2014/093622 (PCT/US2013/074667), WO 2014/093635 (PCT/US2013/074691), WO 2014/093655 (PCT/US2013/074736), WO 2014/093712 (PCT/US2013/074819), WO 2014/093701 (PCT/US2013/074800), WO 2014/018423 (PCT/US2013/051418), WO 2014/204723 (PCT/US2014/041790), WO 2014/204724 (PCT/US2014/041800), WO 2014/204725 (PCT/US2014/041803), WO 2014/204726 (PCT/US2014/041804), WO 2014/204727 (PCT/US2014/041806), WO 2014/204728 (PCT/US2014/041808), WO 2014/204729 (PCT/US2014/041809), WO 2015/089351 (PCT/US2014/069897), WO

2015/089354 (PCT/US2014/069902), WO 2015/089364 (PCT/US2014/069925), WO 2015/089427 (PCT/US2014/070068), WO 2015/089462 (PCT/US2014/070127), WO 2015/089419 (PCT/US2014/070057), WO 2015/089465 (PCT/US2014/070135), WO 2015/089486 (PCT/US2014/070175), PCT/US2015/051691, PCT/US2015/051830. Reference is made to PCT application designating, inter alia, the United States, application No. PCT/US14/41806, filed June 10, 2014. Reference is made to PCT application designating, inter alia, the United States, application No. PCT/US14/41806, filed June 10, 2014.

[0090] Each of these patents, patent publications, and applications, and all documents cited therein or during their prosecution ("appln cited documents") and all documents cited or referenced in the appln cited documents, together with any instructions, descriptions, product specifications, and product sheets for any products mentioned therein or in any document therein and incorporated by reference herein, are hereby incorporated herein by reference, and may be employed in the practice of the invention. All documents (e.g., these patents, patent publications and applications and the appln cited documents) are incorporated herein by reference to the same extent as if each individual document was specifically and individually indicated to be incorporated by reference.

[0091] Also with respect to general information on CRISPR-Cas Systems, mention is made of the following (also hereby incorporated herein by reference):

- *Multiplex genome engineering using CRISPR/Cas systems*. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., & Zhang, F. *Science* Feb 15;339(6121):819-23 (2013);
- *RNA-guided editing of bacterial genomes using CRISPR-Cas systems*. Jiang W., Bikard D., Cox D., Zhang F, Marraffini LA. Nat Biotechnol Mar;31(3):233-9 (2013);
- *One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering*. Wang H., Yang H., Shivalila CS., Dawlaty MM., Cheng AW., Zhang F., Jaenisch R. Cell May 9;153(4):910-8 (2013);
- *Optical control of mammalian endogenous transcription and epigenetic states*. Konermann S, Brigham MD, Trevino AE, Hsu PD, Heidenreich M, Cong L, Platt RJ, Scott DA, Church GM, Zhang F. Nature. 2013 Aug 22;500(7463):472-6. doi: 10.1038/Nature12466. Epub 2013 Aug 23;

- *Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity*. Ran, FA., Hsu, PD., Lin, CY., Gootenberg, JS., Konermann, S., Trevino, AE., Scott, DA., Inoue, A., Matoba, S., Zhang, Y., & Zhang, F. *Cell* Aug 28. pii: S0092-8674(13)01015-5. (2013*);*
- *DNA targeting specificity of RNA-guided Cas9 nucleases*. Hsu, P., Scott, D., Weinstein, J., Ran, FA., Konermann, S., Agarwala, V., Li, Y., Fine, E., Wu, X., Shalem, O., Cradick, TJ., Marraffini, LA., Bao, G., & Zhang, F. *Nat Biotechnol* doi:10.1038/nbt.2647 (2013);
- *Genome engineering using the CRISPR-Cas9 system*. Ran, FA., Hsu, PD., Wright, J., Agarwala, V., Scott, DA., Zhang, F. Nature Protocols Nov;8(11):2281-308. (2013);
- *Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells*. Shalem, O., Sanjana, NE., Hartenian, E., Shi, X., Scott, DA., Mikkelson, T., Heckl, D., Ebert, BL., Root, DE., Doench, JG., Zhang, F. *Science* Dec 12. (2013). [Epub ahead of print];
- *Crystal structure of cas9 in complex with guide RNA and target DNA*. Nishimasu, H., Ran, FA., Hsu, PD., Konermann, S., Shehata, SI., Dohmae, N., Ishitani, R., Zhang, F., Nureki, O. *Cell* Feb 27. (2014). 156(5):935-49;
- *Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells*. Wu X., Scott DA., Kriz AJ., Chiu AC., Hsu PD., Dadon DB., Cheng AW., Trevino AE., Konermann S., Chen S., Jaenisch R., Zhang F., Sharp PA. *Nat Biotechnol*. (2014) Apr 20. doi: 10.1038/nbt.2889,
- *CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling*, Platt et al., Cell 159(2): 440-455 (2014) DOI: 10.1016/j.cell.2014.09.014,
- *Development and Applications of CRISPR-Cas9 for Genome Engineering*, Hsu et al, Cell 157, 1262-1278 (June 5, 2014) (Hsu 2014),
- *Genetic screens in human cells using the CRISPR/Cas9 system*, Wang et al., Science. 2014 January 3; 343(6166): 80–84. doi:10.1126/science.1246981,
- *Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation*, Doench et al., Nature Biotechnology published online 3 September 2014; doi:10.1038/nbt.3026, and

- *In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9*, Swiech et al, Nature Biotechnology ; published online 19 October 2014; doi:10.1038/nbt.3055.
- *Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex*, Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F., Nature. Jan 29;517(7536):583-8 (2015).
- *A split-Cas9 architecture for inducible genome editing and transcription modulation*, Zetsche B, Volz SE, Zhang F., (published online 02 February 2015) Nat Biotechnol. Feb;33(2):139-42 (2015);
- *Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis*, Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, Scott DA, Song J, Pan JQ, Weissleder R, Lee H, Zhang F, Sharp PA. Cell 160, 1246–1260, March 12, 2015 (multiplex screen in mouse), and
- *In vivo* genome editing using *Staphylococcus aureus* Cas9, Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, Koonin EV, Sharp PA, Zhang F., (published online 01 April 2015), Nature. Apr 9;520(7546):186-91 (2015).
- *High-throughput functional genomics using CRISPR-Cas9,* Shalem et al., Nature Reviews Genetics 16, 299-311 (May 2015).
- *Sequence determinants of improved CRISPR sgRNA design*, Xu et al., Genome Research 25, 1147-1157 (August 2015).
- *A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks*, Parnas et al., Cell 162, 675-686 (July 30, 2015).
- *CRISPR/Cas9 cleavage of viral DNA efficiently suppresses hepatitis B virus*, Ramanan et al., Scientific Reports 5:10833. doi: 10.1038/srep10833 (June 2, 2015).
- *Crystal Structure of Staphylococcus aureus Cas9*, Nishimasu et al., Cell 162, 1113-1126 (Aug. 27, 2015).
- BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis, Canver et al., Nature 527(7577):192-7 (Nov. 12, 2015) doi: 10.1038/nature15521. Epub 2015 Sep 16.

- *Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System*, Zetsche et al., Cell 163, 759-71 (Sep 25, 2015).
- *Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems*, Shmakov et al., Molecular Cell, 60(3), 385–397 doi: 10.1016/j.molcel.2015.10.008 Epub October 22, 2015.

each of which is incorporated herein by reference, and discussed briefly below:

- Cong *et al.* engineered type II CRISPR/Cas systems for use in eukaryotic cells based on both *Streptococcus thermophilus* Cas9 and also *Streptoccocus pyogenes* Cas9 and demonstrated that Cas9 nucleases can be directed by short RNAs to induce precise cleavage of DNA in human and mouse cells. Their study further showed that Cas9 as converted into a nicking enzyme can be used to facilitate homology-directed repair in eukaryotic cells with minimal mutagenic activity. Additionally, their study demonstrated that multiple guide sequences can be encoded into a single CRISPR array to enable simultaneous editing of several at endogenous genomic loci sites within the mammalian genome, demonstrating easy programmability and wide applicability of the RNA-guided nuclease technology. This ability to use RNA to program sequence specific DNA cleavage in cells defined a new class of genome engineering tools. These studies further showed that other CRISPR loci are likely to be transplantable into mammalian cells and can also mediate mammalian genome cleavage. Importantly, it can be envisaged that several aspects of the CRISPR/Cas system can be further improved to increase its efficiency and versatility.
- Jiang *et al.* used the clustered, regularly interspaced, short palindromic repeats (CRISPR)–associated Cas9 endonuclease complexed with dual-RNAs to introduce precise mutations in the genomes of *Streptococcus pneumoniae* and *Escherichia coli.* The approach relied on dual-RNA:Cas9-directed cleavage at the targeted genomic site to kill unmutated cells and circumvents the need for selectable markers or counter-selection systems. The study reported reprogramming dual-RNA:Cas9 specificity by changing the sequence of short CRISPR RNA (crRNA) to make single- and multinucleotide changes carried on editing templates. The study showed that simultaneous use of two crRNAs enabled multiplex mutagenesis. Furthermore, when the approach was used in combination with recombineering, in *S. pneumoniae*, nearly 100% of cells that were

recovered using the described approach contained the desired mutation, and in *E. coli,* 65% that were recovered contained the mutation.

- Wang *et al.* (2013) used the CRISPR/Cas system for the one-step generation of mice carrying mutations in multiple genes which were traditionally generated in multiple steps by sequential recombination in embryonic stem cells and/or time-consuming intercrossing of mice with a single mutation. The CRISPR/Cas system will greatly accelerate the *in vivo* study of functionally redundant genes and of epistatic gene interactions.
- Konermann *et al.* addressed the need in the art for versatile and robust technologies that enable optical and chemical modulation of DNA-binding domains based CRISPR Cas9 enzyme and also Transcriptional Activator Like Effectors.
- Ran *et al.* (2013-A) described an approach that combined a Cas9 nickase mutant with paired guide RNAs to introduce targeted double-strand breaks. This addresses the issue of the Cas9 nuclease from the microbial CRISPR-Cas system being targeted to specific genomic loci by a guide sequence, which can tolerate certain mismatches to the DNA target and thereby promote undesired off-target mutagenesis. Because individual nicks in the genome are repaired with high fidelity, simultaneous nicking via appropriately offset guide RNAs is required for double-stranded breaks and extends the number of specifically recognized bases for target cleavage. The authors demonstrated that using paired nicking can reduce off-target activity by 50- to 1,500-fold in cell lines and to facilitate gene knockout in mouse zygotes without sacrificing on-target cleavage efficiency. This versatile strategy enables a wide variety of genome editing applications that require high specificity.
- Hsu *et al.* (2013) characterized SpCas9 targeting specificity in human cells to inform the selection of target sites and avoid off-target effects. The study evaluated >700 guide RNA variants and SpCas9-induced indel mutation levels at >100 predicted genomic off-target loci in 293T and 293FT cells. The authors that SpCas9 tolerates mismatches between guide RNA and target DNA at different positions in a sequence-dependent manner, sensitive to the number, position and distribution of mismatches. The authors further showed that SpCas9-mediated cleavage is unaffected by DNA methylation and that the dosage of SpCas9 and sgRNA can be titrated to minimize off-target modification.

Additionally, to facilitate mammalian genome engineering applications, the authors reported providing a web-based software tool to guide the selection and validation of target sequences as well as off-target analyses.

- Ran *et al.* (2013-B) described a set of tools for Cas9-mediated genome editing *via* non-homologous end joining (NHEJ) or homology-directed repair (HDR) in mammalian cells, as well as generation of modified cell lines for downstream functional studies. To minimize off-target cleavage, the authors further described a double-nicking strategy using the Cas9 nickase mutant with paired guide RNAs. The protocol provided by the authors experimentally derived guidelines for the selection of target sites, evaluation of cleavage efficiency and analysis of off-target activity. The studies showed that beginning with target design, gene modifications can be achieved within as little as 1–2 weeks, and modified clonal cell lines can be derived within 2–3 weeks.
- Shalem *et al.* described a new way to interrogate gene function on a genome-wide scale. Their studies showed that delivery of a genome-scale CRISPR-Cas9 knockout (GeCKO) library targeted 18,080 genes with 64,751 unique guide sequences enabled both negative and positive selection screening in human cells. First, the authors showed use of the GeCKO library to identify genes essential for cell viability in cancer and pluripotent stem cells. Next, in a melanoma model, the authors screened for genes whose loss is involved in resistance to vemurafenib, a therapeutic that inhibits mutant protein kinase BRAF. Their studies showed that the highest-ranking candidates included previously validated genes NF1 and MED12 as well as novel hits NF2, CUL3, TADA2B, and TADA1. The authors observed a high level of consistency between independent guide RNAs targeting the same gene and a high rate of hit confirmation, and thus demonstrated the promise of genome-scale screening with Cas9.
- Nishimasu *et al.* reported the crystal structure of *Streptococcus pyogenes* Cas9 in complex with sgRNA and its target DNA at 2.5 A° resolution. The structure revealed a bilobed architecture composed of target recognition and nuclease lobes, accommodating the sgRNA:DNA heteroduplex in a positively charged groove at their interface. Whereas the recognition lobe is essential for binding sgRNA and DNA, the nuclease lobe contains the HNH and RuvC nuclease domains, which are properly positioned for cleavage of the complementary and non-complementary strands of the target DNA, respectively. The

nuclease lobe also contains a carboxyl-terminal domain responsible for the interaction with the protospacer adjacent motif (PAM). This high-resolution structure and accompanying functional analyses have revealed the molecular mechanism of RNA-guided DNA targeting by Cas9, thus paving the way for the rational design of new, versatile genome-editing technologies.

- Wu *et al.* mapped genome-wide binding sites of a catalytically inactive Cas9 (dCas9) from *Streptococcus pyogenes* loaded with single guide RNAs (sgRNAs) in mouse embryonic stem cells (mESCs). The authors showed that each of the four sgRNAs tested targets dCas9 to between tens and thousands of genomic sites, frequently characterized by a 5-nucleotide seed region in the sgRNA and an NGG protospacer adjacent motif (PAM). Chromatin inaccessibility decreases dCas9 binding to other sites with matching seed sequences; thus 70% of off-target sites are associated with genes. The authors showed that targeted sequencing of 295 dCas9 binding sites in mESCs transfected with catalytically active Cas9 identified only one site mutated above background levels. The authors proposed a two-state model for Cas9 binding and cleavage, in which a seed match triggers binding but extensive pairing with target DNA is required for cleavage.
- Platt *et al.* established a Cre-dependent Cas9 knockin mouse. The authors demonstrated *in vivo* as well as *ex vivo* genome editing using adeno-associated virus (AAV)-, lentivirus-, or particle-mediated delivery of guide RNA in neurons, immune cells, and endothelial cells.
- Hsu *et al.* (2014) is a review article that discusses generally CRISPR-Cas9 history from yogurt to genome editing, including genetic screening of cells.
- Wang *et al.* (2014) relates to a pooled, loss-of-function genetic screening approach suitable for both positive and negative selection that uses a genome-scale lentiviral single guide RNA (sgRNA) library.
- Doench *et al.* created a pool of sgRNAs, tiling across all possible target sites of a panel of six endogenous mouse and three endogenous human genes and quantitatively assessed their ability to produce null alleles of their target gene by antibody staining and flow cytometry. The authors showed that optimization of the PAM improved activity and also provided an on-line tool for designing sgRNAs.

- Swiech *et al.* demonstrate that AAV-mediated SpCas9 genome editing can enable reverse genetic studies of gene function in the brain.
- Konermann *et al.* (2015) discusses the ability to attach multiple effector domains, e.g., transcriptional activator, functional and epigenomic regulators at appropriate positions on the guide such as stem or tetraloop with and without linkers.
- Zetsche *et al.* demonstrates that the Cas9 enzyme can be split into two and hence the assembly of Cas9 for activation can be controlled.
- Chen *et al.* relates to multiplex screening by demonstrating that a genome-wide *in vivo* CRISPR-Cas9 screen in mice reveals genes regulating lung metastasis.
- Ran *et al.* (2015) relates to SaCas9 and its ability to edit genomes and demonstrates that one cannot extrapolate from biochemical assays. Shalem *et al.* (2015) described ways in which catalytically inactive Cas9 (dCas9) fusions are used to synthetically repress (CRISPRi) or activate (CRISPRa) expression, showing. advances using Cas9 for genome-scale screens, including arrayed and pooled screens, knockout approaches that inactivate genomic loci and strategies that modulate transcriptional activity.
- Shalem *et al.* (2015) described ways in which catalytically inactive Cas9 (dCas9) fusions are used to synthetically repress (CRISPRi) or activate (CRISPRa) expression, showing. advances using Cas9 for genome-scale screens, including arrayed and pooled screens, knockout approaches that inactivate genomic loci and strategies that modulate transcriptional activity.
- Xu *et al.* (2015) assessed the DNA sequence features that contribute to single guide RNA (sgRNA) efficiency in CRISPR-based screens. The authors explored efficiency of CRISPR/Cas9 knockout and nucleotide preference at the cleavage site. The authors also found that the sequence preference for CRISPRi/a is substantially different from that for CRISPR/Cas9 knockout.
- Parnas *et al.* (2015) introduced genome-wide pooled CRISPR-Cas9 libraries into dendritic cells (DCs) to identify genes that control the induction of tumor necrosis factor (Tnf) by bacterial lipopolysaccharide (LPS). Known regulators of Tlr4 signaling and previously unknown candidates were identified and classified into three functional modules with distinct effects on the canonical responses to LPS.

- Ramanan *et al* (2015) demonstrated cleavage of viral episomal DNA (cccDNA) in infected cells. The HBV genome exists in the nuclei of infected hepatocytes as a 3.2kb double-stranded episomal DNA species called covalently closed circular DNA (cccDNA), which is a key component in the HBV life cycle whose replication is not inhibited by current therapies. The authors showed that sgRNAs specifically targeting highly conserved regions of HBV robustly suppresses viral replication and depleted cccDNA.
- Nishimasu *et al.* (2015) reported the crystal structures of SaCas9 in complex with a single guide RNA (sgRNA) and its double-stranded DNA targets, containing the 5'-TTGAAT-3' PAM and the 5'-TTGGGT-3' PAM. A structural comparison of SaCas9 with SpCas9 highlighted both structural conservation and divergence, explaining their distinct PAM specificities and orthologous sgRNA recognition.

[0092] Mention is also made of Tsai et al, "Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing," Nature Biotechnology 32(6): 569-77 (2014) which is not believed to be prior art to the instant invention or application, but which may be considered in the practice of the instant invention. Mention is also made of Konermann et al., "Genome-scale transcription activation by an engineered CRISPR-Cas9 complex," doi:10.1038/nature14136, incorporated herein by reference.

[0093] In general, the CRISPR-Cas or CRISPR system is as used in the foregoing documents, such as WO 2014/093622 (PCT/US2013/074667) and refers collectively to transcripts and other elements involved in the expression of or directing the activity of CRISPR-associated ("Cas") genes, including sequences encoding a Cas gene, a tracr (trans-activating CRISPR) sequence (e.g. tracrRNA or an active partial tracrRNA), a tracr-mate sequence (encompassing a "direct repeat" and a tracrRNA-processed partial direct repeat in the context of an endogenous CRISPR system), a guide sequence (also referred to as a "spacer" in the context of an endogenous CRISPR system), or "RNA(s)" as that term is herein used (e.g., RNA(s) to guide Cas9, e.g. CRISPR RNA and transactivating (tracr) RNA or a single guide RNA (sgRNA) (chimeric RNA)) or other sequences and transcripts from a CRISPR locus. In general, a CRISPR system is characterized by elements that promote the formation of a CRISPR complex at the site of a target sequence (also referred to as a protospacer in the context of an endogenous CRISPR system). In the context of formation of a CRISPR complex, "target sequence" refers to

a sequence to which a guide sequence is designed to have complementarity, where hybridization between a target sequence and a guide sequence promotes the formation of a CRISPR complex. A target sequence may comprise any polynucleotide, such as DNA or RNA polynucleotides. In some embodiments, a target sequence is located in the nucleus or cytoplasm of a cell. In some embodiments, direct repeats may be identified *in silico* by searching for repetitive motifs that fulfill any or all of the following criteria: 1. found in a 2Kb window of genomic sequence flanking the type II CRISPR locus; 2. span from 20 to 50 bp; and 3. interspaced by 20 to 50 bp. In some embodiments, 2 of these criteria may be used, for instance 1 and 2, 2 and 3, or 1 and 3. In some embodiments, all 3 criteria may be used. In some embodiments it may be preferred in a CRISPR complex that the tracr sequence has one or more hairpins and is 30 or more nucleotides in length, 40 or more nucleotides in length, or 50 or more nucleotides in length; the guide sequence is between 10 to 30 nucleotides in length, the CRISPR/Cas enzyme is a Type II Cas9 enzyme. In embodiments of the invention the terms guide sequence and guide RNA are used interchangeably as in foregoing cited documents such as WO 2014/093622 (PCT/US2013/074667). In general, a guide sequence is any polynucleotide sequence having sufficient complementarity with a target polynucleotide sequence to hybridize with the target sequence and direct sequence-specific binding of a CRISPR complex to the target sequence. In some embodiments, the degree of complementarity between a guide sequence and its corresponding target sequence, when optimally aligned using a suitable alignment algorithm, is about or more than about 50%, 60%, 75%, 80%, 85%, 90%, 95%, 97.5%, 99%, or more. Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences, non-limiting example of which include the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, algorithms based on the Burrows-Wheeler Transform (e.g. the Burrows Wheeler Aligner), ClustalW, Clustal X, BLAT, Novoalign (Novocraft Technologies; available at www.novocraft.com), ELAND (Illumina, San Diego, CA), SOAP (available at soap.genomics.org.cn), and Maq (available at maq.sourceforge.net). In some embodiments, a guide sequence is about or more than about 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 75, or more nucleotides in length. In some embodiments, a guide sequence is less than about 75, 50, 45, 40, 35, 30, 25, 20, 15, 12, or fewer nucleotides in length. Preferably the guide sequence is 10 - 30 nucleotides long. The ability of a guide sequence to direct sequence-specific binding of a CRISPR complex to a target sequence

may be assessed by any suitable assay. For example, the components of a CRISPR system sufficient to form a CRISPR complex, including the guide sequence to be tested, may be provided to a host cell having the corresponding target sequence, such as by transfection with vectors encoding the components of the CRISPR sequence, followed by an assessment of preferential cleavage within the target sequence, such as by Surveyor assay as described herein. Similarly, cleavage of a target polynucleotide sequence may be evaluated in a test tube by providing the target sequence, components of a CRISPR complex, including the guide sequence to be tested and a control guide sequence different from the test guide sequence, and comparing binding or rate of cleavage at the target sequence between the test and control guide sequence reactions. Other assays are possible, and will occur to those skilled in the art. A guide sequence may be selected to target any target sequence. In some embodiments, the target sequence is a sequence within a genome of a cell. Exemplary target sequences include those that are unique in the target genome. For example, for the S. pyogenes Cas9, a unique target sequence in a genome may include a Cas9 target site of the form MMMMMMMMNNNNNNNNNNNNXGG where NNNNNNNNNNNNXGG (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. A unique target sequence in a genome may include an S. pyogenes Cas9 target site of the form MMMMMMMMMNNNNNNNNNNNXGG where NNNNNNNNNNNXGG (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. For the S. thermophilus CRISPR1 Cas9, a unique target sequence in a genome may include a Cas9 target site of the form MMMMMMMMNNNNNNNNNNNNXXAGAAW where NNNNNNNNNNNNXXAGAAW (N is A, G, T, or C; X can be anything; and W is A or T) has a single occurrence in the genome. A unique target sequence in a genome may include an S. thermophilus CRISPR1 Cas9 target site of the form MMMMMMMMMNNNNNNNNNNNNXXAGAAW where NNNNNNNNNNNXXAGAAW (N is A, G, T, or C; X can be anything; and W is A or T) has a single occurrence in the genome. For the S. pyogenes Cas9, a unique target sequence in a genome may include a Cas9 target site of the form MMMMMMMMNNNNNNNNNNNNXGGXG where NNNNNNNNNNNXGGXG (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. A unique target sequence in a genome may include an S. pyogenes Cas9 target site of the form MMMMMMMMMNNNNNNNNNNNXGGXG where NNNNNNNNNNXGGXG (N is A, G, T, or C; and X can be anything) has a single occurrence in the genome. In each of these

sequences "M" may be A, G, T, or C, and need not be considered in identifying a sequence as unique. In some embodiments, a guide sequence is selected to reduce the degree secondary structure within the guide sequence. In some embodiments, about or less than about 75%, 50%, 40%, 30%, 25%, 20%, 15%, 10%, 5%, 1%, or fewer of the nucleotides of the guide sequence participate in self-complementary base pairing when optimally folded. Optimal folding may be determined by any suitable polynucleotide folding algorithm. Some programs are based on calculating the minimal Gibbs free energy. An example of one such algorithm is mFold, as described by Zuker and Stiegler (Nucleic Acids Res. 9 (1981), 133-148). Another example folding algorithm is the online webserver RNAfold, developed at Institute for Theoretical Chemistry at the University of Vienna, using the centroid structure prediction algorithm (see e.g. A.R. Gruber et al., 2008, Cell 106(1): 23-24; and PA Carr and GM Church, 2009, Nature Biotechnology 27(12): 1151-62).

[0100] In general, a tracr mate sequence includes any sequence that has sufficient complementarity with a tracr sequence to promote one or more of: (1) excision of a guide sequence flanked by tracr mate sequences in a cell containing the corresponding tracr sequence; and (2) formation of a CRISPR complex at a target sequence, wherein the CRISPR complex comprises the tracr mate sequence hybridized to the tracr sequence. In general, degree of complementarity is with reference to the optimal alignment of the tracr mate sequence and tracr sequence, along the length of the shorter of the two sequences. Optimal alignment may be determined by any suitable alignment algorithm, and may further account for secondary structures, such as self-complementarity within either the tracr sequence or tracr mate sequence. In some embodiments, the degree of complementarity between the tracr sequence and tracr mate sequence along the length of the shorter of the two when optimally aligned is about or more than about 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 97.5%, 99%, or higher. In some embodiments, the tracr sequence is about or more than about 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 40, 50, or more nucleotides in length. In some embodiments, the tracr sequence and tracr mate sequence are contained within a single transcript, such that hybridization between the two produces a transcript having a secondary structure, such as a hairpin. In an embodiment of the invention, the transcript or transcribed polynucleotide sequence has at least two or more hairpins. In preferred embodiments, the transcript has two, three, four or five hairpins. In a further embodiment of the invention, the transcript has at most five hairpins. In a

hairpin structure the portion of the sequence 5' of the final "N" and upstream of the loop corresponds to the tracr mate sequence, and the portion of the sequence 3' of the loop corresponds to the tracr sequence Further non-limiting examples of single polynucleotides comprising a guide sequence, a tracr mate sequence, and a tracr sequence are as follows (listed 5' to 3'), where "N" represents a base of a guide sequence, the first block of lower case letters represent the tracr mate sequence, and the second block of lower case letters represent the tracr sequence, and the final poly-T sequence represents the transcription terminator: (1) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataa ggcttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT; (2) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccg aaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT; (3) NNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccg aaatcaacaccctgtcattttatggcagggtgtTTTTTT; (4) NNNNNNNNNNNNNNNNNNNgttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaactt gaaaaagtggcaccgagtcggtgcTTTTTT; (5) NNNNNNNNNNNNNNNNNNNgttttagagctaGAAATAGcaagttaaaataaggctagtccgttatcaac ttgaaaaagtgTTTTTTT; and (6) NNNNNNNNNNNNNNNNNNNgttttagagctagAAATAGcaagttaaaataaggctagtccgttatcaTT TTTTTT. In some embodiments, sequences (1) to (3) are used in combination with Cas9 from *S. thermophilus* CRISPR1. In some embodiments, sequences (4) to (6) are used in combination with Cas9 from *S. pyogenes*. In some embodiments, the tracr sequence is a separate transcript from a transcript comprising the tracr mate sequence.

[0101] In some embodiments, candidate tracrRNA may be subsequently predicted by sequences that fulfill any or all of the following criteria: 1. sequence homology to direct repeats (motif search in Geneious with up to 18-bp mismatches); 2. presence of a predicted Rho-independent transcriptional terminator in direction of transcription; and 3. stable hairpin secondary structure between tracrRNA and direct repeat. In some embodiments, 2 of these criteria may be used, for instance 1 and 2, 2 and 3, or 1 and 3. In some embodiments, all 3 criteria may be used.

[0102] In some embodiments, chimeric synthetic guide RNAs (sgRNAs) designs may incorporate at least 12 bp of duplex structure between the direct repeat and tracrRNA.

[0103] For minimization of toxicity and off-target effect, it will be important to control the concentration of CRISPR enzyme mRNA and guide RNA delivered. Optimal concentrations of CRISPR enzyme mRNA and guide RNA can be determined by testing different concentrations in a cellular or non-human eukaryote animal model and using deep sequencing the analyze the extent of modification at potential off-target genomic loci. For example, for the guide sequence targeting 5'-GAGTCCGAGCAGAAGAAGAA-3' in the EMX1 gene of the human genome, deep sequencing can be used to assess the level of modification at the following two off-target loci, 1: 5'-GAGTCCTAGCAGGAGAAGAA-3' and 2: 5'-GAGTCTAAGCAGAAGAAGAA-3'. The concentration that gives the highest level of on-target modification while minimizing the level of off-target modification should be chosen for in vivo delivery. Alternatively, to minimize the level of toxicity and off-target effect, CRISPR enzyme nickase mRNA (for example S. pyogenes Cas9 with the D10A mutation) can be delivered with a pair of guide RNAs targeting a site of interest. The two guide RNAs need to be spaced as follows. Guide sequences and strategies to mimize toxicity and off-target effects can be as in WO 2014/093622 (PCT/US2013/074667).

[0104] The CRISPR system is derived advantageously from a type II CRISPR system. In some embodiments, one or more elements of a CRISPR system is derived from a particular organism comprising an endogenous CRISPR system, such as *Streptococcus pyogenes*. In preferred embodiments of the invention, the CRISPR system is a type II CRISPR system and the Cas enzyme is Cas9, which catalyzes DNA cleavage. Non-limiting examples of Cas proteins include Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (also known as Csn1 and Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4, homologues thereof, or modified versions thereof.

[0105] In some embodiments, the unmodified CRISPR enzyme has DNA cleavage activity, such as Cas9. In some embodiments, the CRISPR enzyme directs cleavage of one or both strands at the location of a target sequence, such as within the target sequence and/or within the complement of the target sequence. In some embodiments, the CRISPR enzyme directs cleavage of one or both strands within about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, 100, 200, 500, or more base pairs from the first or last nucleotide of a target sequence. In some embodiments, a

vector encodes a CRISPR enzyme that is mutated to with respect to a corresponding wild-type enzyme such that the mutated CRISPR enzyme lacks the ability to cleave one or both strands of a target polynucleotide containing a target sequence. For example, an aspartate-to-alanine substitution (D10A) in the RuvC I catalytic domain of Cas9 from S. pyogenes converts Cas9 from a nuclease that cleaves both strands to a nickase (cleaves a single strand). Other examples of mutations that render Cas9 a nickase include, without limitation, H840A, N854A, and N863A. As a further example, two or more catalytic domains of Cas9 (RuvC I, RuvC II, and RuvC III or the HNH domain) may be mutated to produce a mutated Cas9 substantially lacking all DNA cleavage activity. In some embodiments, a D10A mutation is combined with one or more of H840A, N854A, or N863A mutations to produce a Cas9 enzyme substantially lacking all DNA cleavage activity. In some embodiments, a CRISPR enzyme is considered to substantially lack all DNA cleavage activity when the DNA cleavage activity of the mutated enzyme is about no more than 25%, 10%, 5%, 1%, 0.1%, 0.01%, or less of the DNA cleavage activity of the non-mutated form of the enzyme; an example can be when the DNA cleavage activity of the mutated form is nil or negligible as compared with the non-mutated form. Where the enzyme is not SpCas9, mutations may be made at any or all residues corresponding to positions 10, 762, 840, 854, 863 and/or 986 of SpCas9 (which may be ascertained for instance by standard sequence comparison tools). In particular, any or all of the following mutations are preferred in SpCas9: D10A, E762A, H840A, N854A, N863A and/or D986A; as well as conservative substitution for any of the replacement amino acids is also envisaged. The same (or conservative substitutions of these mutations) at corresponding positions in other Cas9s are also preferred. Particularly preferred are D10 and H840 in SpCas9. However, in other Cas9s, residues corresponding to SpCas9 D10 and H840 are also preferred. Orthologs of SpCas9 can be used in the practice of the invention. A Cas enzyme may be identified Cas9 as this can refer to the general class of enzymes that share homology to the biggest nuclease with multiple nuclease domains from the type II CRISPR system. Most preferably, the Cas9 enzyme is from, or is derived from, spCas9 (*S. pyogenes* Cas9) or saCas9 (*S. aureus* Cas9). StCas9" refers to wild type Cas9 from S. thermophilus, the protein sequence of which is given in the SwissProt database under accession number G3ECR1. Similarly, S pyogenes Cas9 or spCas9 is included in SwissProt under accession number Q99ZW2. By derived, Applicants mean that the derived enzyme is largely based, in the sense of having a high degree of sequence homology with, a wildtype enzyme, but

that it has been mutated (modified) in some way as described herein. It will be appreciated that the terms Cas and CRISPR enzyme are generally used herein interchangeably, unless otherwise apparent. As mentioned above, many of the residue numberings used herein refer to the Cas9 enzyme from the type II CRISPR locus in *Streptococcus pyogenes*. However, it will be appreciated that this invention includes many more Cas9s from other species of microbes, such as SpCas9, SaCa9, St1Cas9 and so forth. Enzymatic action by Cas9 derived from *Streptococcus pyogenes* or any closely related Cas9 generates double stranded breaks at target site sequences which hybridize to 20 nucleotides of the guide sequence and that have a protospacer-adjacent motif (PAM) sequence (examples include NGG/NRG or a PAM that can be determined as described herein) following the 20 nucleotides of the target sequence. CRISPR activity through Cas9 for site-specific DNA recognition and cleavage is defined by the guide sequence, the tracr sequence that hybridizes in part to the guide sequence and the PAM sequence. More aspects of the CRISPR system are described in Karginov and Hannon, The CRISPR system: small RNA-guided defence in bacteria and archaea, Mole Cell 2010, January 15; 37(1): 7. The type II CRISPR locus from *Streptococcus pyogenes* SF370, which contains a cluster of four genes Cas9, Cas1, Cas2, and Csn1, as well as two non-coding RNA elements, tracrRNA and a characteristic array of repetitive sequences (direct repeats) interspaced by short stretches of non-repetitive sequences (spacers, about 30bp each). In this system, targeted DNA double-strand break (DSB) is generated in four sequential steps. First, two non-coding RNAs, the pre-crRNA array and tracrRNA, are transcribed from the CRISPR locus. Second, tracrRNA hybridizes to the direct repeats of pre-crRNA, which is then processed into mature crRNAs containing individual spacer sequences. Third, the mature crRNA:tracrRNA complex directs Cas9 to the DNA target consisting of the protospacer and the corresponding PAM via heteroduplex formation between the spacer region of the crRNA and the protospacer DNA. Finally, Cas9 mediates cleavage of target DNA upstream of PAM to create a DSB within the protospacer. A pre-crRNA array consisting of a single spacer flanked by two direct repeats (DRs) is also encompassed by the term "tracr-mate sequences"). In certain embodiments, Cas9 may be constitutively present or inducibly present or conditionally present or administered or delivered. Cas9 optimization may be used to enhance function or to develop new functions, one can generate chimeric Cas9 proteins. And Cas9 may be used as a generic DNA binding protein.

[0106] Typically, in the context of an endogenous CRISPR system, formation of a CRISPR complex (comprising a guide sequence hybridized to a target sequence and complexed with one or more Cas proteins) results in cleavage of one or both strands in or near (e.g. within 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, or more base pairs from) the target sequence. Without wishing to be bound by theory, the tracr sequence, which may comprise or consist of all or a portion of a wild-type tracr sequence (e.g. about or more than about 20, 26, 32, 45, 48, 54, 63, 67, 85, or more nucleotides of a wild-type tracr sequence), may also form part of a CRISPR complex, such as by hybridization along at least a portion of the tracr sequence to all or a portion of a tracr mate sequence that is operably linked to the guide sequence.

[0107] An example of a codon optimized sequence, is in this instance a sequence optimized for expression in a eukaryote, e.g., humans (i.e. being optimized for expression in humans), or for another eukaryote, animal or mammal as herein discussed; see, e.g., SaCas9 human codon optimized sequence in WO 2014/093622 (PCT/US2013/074667). Whilst this is preferred, it will be appreciated that other examples are possible and codon optimization for a host species other than human, or for codon optimization for specific organs is known. In some embodiments, an enzyme coding sequence encoding a CRISPR enzyme is codon optimized for expression in particular cells, such as eukaryotic cells. The eukaryotic cells may be those of or derived from a particular organism, such as a mammal, including but not limited to human, or non-human eukaryote or animal or mammal as herein discussed, e.g., mouse, rat, rabbit, dog, livestock, or non-human mammal or primate. In some embodiments, processes for modifying the germ line genetic identity of human beings and/or processes for modifying the genetic identity of animals which are likely to cause them suffering without any substantial medical benefit to man or animal, and also animals resulting from such processes, may be excluded. In general, codon optimization refers to a process of modifying a nucleic acid sequence for enhanced expression in the host cells of interest by replacing at least one codon (e.g. about or more than about 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more codons) of the native sequence with codons that are more frequently or most frequently used in the genes of that host cell while maintaining the native amino acid sequence. Various species exhibit particular bias for certain codons of a particular amino acid. Codon bias (differences in codon usage between organisms) often correlates with the efficiency of translation of messenger RNA (mRNA), which is in turn believed to be dependent on, among other things, the properties of the codons being translated and the

availability of particular transfer RNA (tRNA) molecules. The predominance of selected tRNAs in a cell is generally a reflection of the codons used most frequently in peptide synthesis. Accordingly, genes can be tailored for optimal gene expression in a given organism based on codon optimization. Codon usage tables are readily available, for example, at the "Codon Usage Database" available at www.kazusa.orjp/codon/ and these tables can be adapted in a number of ways. See Nakamura, Y., et al. "Codon usage tabulated from the international DNA sequence databases: status for the year 2000" Nucl. Acids Res. 28:292 (2000). Computer algorithms for codon optimizing a particular sequence for expression in a particular host cell are also available, such as Gene Forge (Aptagen; Jacobus, PA), are also available. In some embodiments, one or more codons (e.g. 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more, or all codons) in a sequence encoding a CRISPR enzyme correspond to the most frequently used codon for a particular amino acid.

[0108] In some embodiments, a vector encodes a CRISPR enzyme comprising one or more nuclear localization sequences (NLSs), such as about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs. In some embodiments, the CRISPR enzyme comprises about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs at or near the amino-terminus, about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more NLSs at or near the carboxy-terminus, or a combination of these (e.g. zero or at least one or more NLS at the amino-terminus and zero or at one or more NLS at the carboxy terminus). When more than one NLS is present, each may be selected independently of the others, such that a single NLS may be present in more than one copy and/or in combination with one or more other NLSs present in one or more copies. In a preferred embodiment of the invention, the CRISPR enzyme comprises at most 6 NLSs. In some embodiments, an NLS is considered near the N- or C-terminus when the nearest amino acid of the NLS is within about 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, or more amino acids along the polypeptide chain from the N- or C-terminus. Non-limiting examples of NLSs include an NLS sequence derived from: the NLS of the SV40 virus large T-antigen, having the amino acid sequence PKKKRKV; the NLS from nucleoplasmin (e.g. the nucleoplasmin bipartite NLS with the sequence KRPAATKKAGQAKKKK); the c-myc NLS having the amino acid sequence PAAKRVKLD or RQRRNELKRSP; the hRNPA1 M9 NLS having the sequence NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY; the sequence RMRIZFKNKGKDTAELRRRRVEVSVELRKAKKDEQILKRRNV of the IBB domain from importin-alpha; the sequences VSRKRPRP and PPKKARED of the myoma T protein; the

sequence POPKKKPL of human p53; the sequence SALIKKKKKMAP of mouse c-abl IV; the sequences DRLRR and PKQKKRK of the influenza virus NS1; the sequence RKLKKKIKKL of the Hepatitis virus delta antigen; the sequence REKKKFLKRR of the mouse Mx1 protein; the sequence KRKGDEVDGVDEVAKKKSKK of the human poly(ADP-ribose) polymerase; and the sequence RKCLQAGMNLEARKTKK of the steroid hormone receptors (human) glucocorticoid. In general, the one or more NLSs are of sufficient strength to drive accumulation of the CRISPR enzyme in a detectable amount in the nucleus of a eukaryotic cell. In general, strength of nuclear localization activity may derive from the number of NLSs in the CRISPR enzyme, the particular NLS(s) used, or a combination of these factors. Detection of accumulation in the nucleus may be performed by any suitable technique. For example, a detectable marker may be fused to the CRISPR enzyme, such that location within a cell may be visualized, such as in combination with a means for detecting the location of the nucleus (e.g. a stain specific for the nucleus such as DAPI). Cell nuclei may also be isolated from cells, the contents of which may then be analyzed by any suitable process for detecting protein, such as immunohistochemistry, Western blot, or enzyme activity assay. Accumulation in the nucleus may also be determined indirectly, such as by an assay for the effect of CRISPR complex formation (e.g. assay for DNA cleavage or mutation at the target sequence, or assay for altered gene expression activity affected by CRISPR complex formation and/or CRISPR enzyme activity), as compared to a control no exposed to the CRISPR enzyme or complex, or exposed to a CRISPR enzyme lacking the one or more NLSs.

[0109] Aspects of the invention relate to the expression of the gene product being decreased or a template polynucleotide being further introduced into the DNA molecule encoding the gene product or an intervening sequence being excised precisely by allowing the two 5' overhangs to reanneal and ligate or the activity or function of the gene product being altered or the expression of the gene product being increased. In an embodiment of the invention, the gene product is a protein. Only sgRNA pairs creating 5' overhangs with less than 8bp overlap between the guide sequences (offset greater than -8 bp) were able to mediate detectable indel formation. Importantly, each guide used in these assays is able to efficiently induce indels when paired with wildtype Cas9, indicating that the relative positions of the guide pairs are the most important parameters in predicting double nicking activity. Since Cas9n and Cas9H840A nick opposite strands of DNA, substitution of Cas9n with Cas9H840A with a given sgRNA pair should have

resulted in the inversion of the overhang type; but no indel formation is observed as with Cas9H840A indicating that Cas9H840A is a CRISPR enzyme substantially lacking all DNA cleavage activity (which is when the DNA cleavage activity of the mutated enzyme is about no more than 25%, 10%, 5%, 1%, 0.1%, 0.01%, or less of the DNA cleavage activity of the non-mutated form of the enzyme; whereby an example can be when the DNA cleavage activity of the mutated form is nil or negligible as compared with the non-mutated form, e.g., when no indel formation is observed as with Cas9H840A in the eukaryotic system in contrast to the biochemical or prokaryotic systems). Nonetheless, a pair of sgRNAs that will generate a 5' overhang with Cas9n should in principle generate the corresponding 3' overhang instead, and double nicking. Therefore, sgRNA pairs that lead to the generation of a 3' overhang with Cas9n can be used with another mutated Cas9 to generate a 5' overhang, and double nicking. Accordingly, in some embodiments, a recombination template is also provided. A recombination template may be a component of another vector as described herein, contained in a separate vector, or provided as a separate polynucleotide. In some embodiments, a recombination template is designed to serve as a template in homologous recombination, such as within or near a target sequence nicked or cleaved by a CRISPR enzyme as a part of a CRISPR complex. A template polynucleotide may be of any suitable length, such as about or more than about 10, 15, 20, 25, 50, 75, 100, 150, 200, 500, 1000, or more nucleotides in length. In some embodiments, the template polynucleotide is complementary to a portion of a polynucleotide comprising the target sequence. When optimally aligned, a template polynucleotide might overlap with one or more nucleotides of a target sequences (e.g. about or more than about 1, 5, 10, 15, 20, or more nucleotides). In some embodiments, when a template sequence and a polynucleotide comprising a target sequence are optimally aligned, the nearest nucleotide of the template polynucleotide is within about 1, 5, 10, 15, 20, 25, 50, 75, 100, 200, 300, 400, 500, 1000, 5000, 10000, or more nucleotides from the target sequence.

[0110] In some embodiments, one or more vectors driving expression of one or more elements of a CRISPR system are introduced into a host cell such that expression of the elements of the CRISPR system direct formation of a CRISPR complex at one or more target sites. For example, a Cas enzyme, a guide sequence linked to a tracr-mate sequence, and a tracr sequence could each be operably linked to separate regulatory elements on separate vectors. Or, RNA(s) of the CRISPR System can be delivered to a transgenic Cas9 animal or mammal, e.g., an animal or

mammal that constitutively or inducibly or conditionally expresses Cas9; or an animal or mammal that is otherwise expressing Cas9 or has cells containing Cas9, such as by way of prior administration thereto of a vector or vectors that code for and express in vivo Cas9. Alternatively, two or more of the elements expressed from the same or different regulatory elements, may be combined in a single vector, with one or more additional vectors providing any components of the CRISPR system not included in the first vector. CRISPR system elements that are combined in a single vector may be arranged in any suitable orientation, such as one element located 5' with respect to ("upstream" of) or 3' with respect to ("downstream" of) a second element. The coding sequence of one element may be located on the same or opposite strand of the coding sequence of a second element, and oriented in the same or opposite direction. In some embodiments, a single promoter drives expression of a transcript encoding a CRISPR enzyme and one or more of the guide sequence, tracr mate sequence (optionally operably linked to the guide sequence), and a tracr sequence embedded within one or more intron sequences (e.g. each in a different intron, two or more in at least one intron, or all in a single intron). In some embodiments, the CRISPR enzyme, guide sequence, tracr mate sequence, and tracr sequence are operably linked to and expressed from the same promoter. Delivery vehicles, vectors, particles, nanoparticles, formulations and components thereof for expression of one or more elements of a CRISPR system are as used in the foregoing documents, such as WO 2014/093622 (PCT/US2013/074667). In some embodiments, a vector comprises one or more insertion sites, such as a restriction endonuclease recognition sequence (also referred to as a "cloning site"). In some embodiments, one or more insertion sites (e.g. about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more insertion sites) are located upstream and/or downstream of one or more sequence elements of one or more vectors. In some embodiments, a vector comprises an insertion site upstream of a tracr mate sequence, and optionally downstream of a regulatory element operably linked to the tracr mate sequence, such that following insertion of a guide sequence into the insertion site and upon expression the guide sequence directs sequence-specific binding of a CRISPR complex to a target sequence in a eukaryotic cell. In some embodiments, a vector comprises two or more insertion sites, each insertion site being located between two tracr mate sequences so as to allow insertion of a guide sequence at each site. In such an arrangement, the two or more guide sequences may comprise two or more copies of a single guide sequence, two or more different guide sequences, or combinations of these. When multiple different guide

sequences are used, a single expression construct may be used to target CRISPR activity to multiple different, corresponding target sequences within a cell. For example, a single vector may comprise about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, or more guide sequences. In some embodiments, about or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more such guide-sequence-containing vectors may be provided, and optionally delivered to a cell. In some embodiments, a vector comprises a regulatory element operably linked to an enzyme-coding sequence encoding a CRISPR enzyme, such as a Cas protein. CRISPR enzyme or CRISPR enzyme mRNA or CRISPR guide RNA or RNA(s) can be delivered separately; and advantageously at least one of these is delivered via a nanoparticle complex. CRISPR enzyme mRNA can be delivered prior to the guide RNA to give time for CRISPR enzyme to be expressed. CRISPR enzyme mRNA might be administered 1-12 hours (preferably around 2-6 hours) prior to the administration of guide RNA. Alternatively, CRISPR enzyme mRNA and guide RNA can be administered together. Advantageously, a second booster dose of guide RNA can be administered 1-12 hours (preferably around 2-6 hours) after the initial administration of CRISPR enzyme mRNA + guide RNA. Additional administrations of CRISPR enzyme mRNA and/or guide RNA might be useful to achieve the most efficient levels of genome modification.

[0111] In one aspect, the invention provides methods for using one or more elements of a CRISPR system. The CRISPR complex of the invention provides an effective means for modifying a target polynucleotide. The CRISPR complex of the invention has a wide variety of utility including modifying (e.g., deleting, inserting, translocating, inactivating, activating) a target polynucleotide in a multiplicity of cell types. As such the CRISPR complex of the invention has a broad spectrum of applications in, e.g., gene therapy, drug screening, disease diagnosis, and prognosis. An exemplary CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within the target polynucleotide. The guide sequence is linked to a tracr mate sequence, which in turn hybridizes to a tracr sequence. In one embodiment, this invention provides a method of cleaving a target polynucleotide. The method comprises modifying a target polynucleotide using a CRISPR complex that binds to the target polynucleotide and effect cleavage of said target polynucleotide. Typically, the CRISPR complex of the invention, when introduced into a cell, creates a break (e.g., a single or a double strand break) in the genome sequence. For example, the method can be used to cleave a disease gene in a cell. The break created by the CRISPR complex can be

repaired by a repair processes such as the error prone non-homologous end joining (NHEJ) pathway or the high fidelity homology-directed repair (HDR). During these repair process, an exogenous polynucleotide template can be introduced into the genome sequence. In some methods, the HDR process is used modify genome sequence. For example, an exogenous polynucleotide template comprising a sequence to be integrated flanked by an upstream sequence and a downstream sequence is introduced into a cell. The upstream and downstream sequences share sequence similarity with either side of the site of integration in the chromosome. Where desired, a donor polynucleotide can be DNA, e.g., a DNA plasmid, a bacterial artificial chromosome (BAC), a yeast artificial chromosome (YAC), a viral vector, a linear piece of DNA, a PCR fragment, a naked nucleic acid, or a nucleic acid complexed with a delivery vehicle such as a liposome or poloxamer. The exogenous polynucleotide template comprises a sequence to be integrated (e.g., a mutated gene). The sequence for integration may be a sequence endogenous or exogenous to the cell. Examples of a sequence to be integrated include polynucleotides encoding a protein or a non-coding RNA (e.g., a microRNA). Thus, the sequence for integration may be operably linked to an appropriate control sequence or sequences. Alternatively, the sequence to be integrated may provide a regulatory function. The upstream and downstream sequences in the exogenous polynucleotide template are selected to promote recombination between the chromosomal sequence of interest and the donor polynucleotide. The upstream sequence is a nucleic acid sequence that shares sequence similarity with the genome sequence upstream of the targeted site for integration. Similarly, the downstream sequence is a nucleic acid sequence that shares sequence similarity with the chromosomal sequence downstream of the targeted site of integration. The upstream and downstream sequences in the exogenous polynucleotide template can have 75%, 80%, 85%, 90%, 95%, or 100% sequence identity with the targeted genome sequence. Preferably, the upstream and downstream sequences in the exogenous polynucleotide template have about 95%, 96%, 97%, 98%, 99%, or 100% sequence identity with the targeted genome sequence. In some methods, the upstream and downstream sequences in the exogenous polynucleotide template have about 99% or 100% sequence identity with the targeted genome sequence. An upstream or downstream sequence may comprise from about 20 bp to about 2500 bp, for example, about 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, or 2500 bp. In some methods, the exemplary upstream or downstream sequence have

about 200 bp to about 2000 bp, about 600 bp to about 1000 bp, or more particularly about 700 bp to about 1000 bp. In some methods, the exogenous polynucleotide template may further comprise a marker. Such a marker may make it easy to screen for targeted integrations. Examples of suitable markers include restriction sites, fluorescent proteins, or selectable markers. The exogenous polynucleotide template of the invention can be constructed using recombinant techniques (see, for example, Sambrook et al., 2001 and Ausubel et al., 1996). In a method for modifying a target polynucleotide by integrating an exogenous polynucleotide template, a double stranded break is introduced into the genome sequence by the CRISPR complex, the break is repaired via homologous recombination an exogenous polynucleotide template such that the template is integrated into the genome. The presence of a double-stranded break facilitates integration of the template. In other embodiments, this invention provides a method of modifying expression of a polynucleotide in a eukaryotic cell. The method comprises increasing or decreasing expression of a target polynucleotide by using a CRISPR complex that binds to the polynucleotide. In some methods, a target polynucleotide can be inactivated to effect the modification of the expression in a cell. For example, upon the binding of a CRISPR complex to a target sequence in a cell, the target polynucleotide is inactivated such that the sequence is not transcribed, the coded protein is not produced, or the sequence does not function as the wild-type sequence does. For example, a protein or microRNA coding sequence may be inactivated such that the protein or microRNA or pre-microRNA transcript is not produced. In some methods, a control sequence can be inactivated such that it no longer functions as a control sequence. As used herein, "control sequence" refers to any nucleic acid sequence that effects the transcription, translation, or accessibility of a nucleic acid sequence. Examples of a control sequence include, a promoter, a transcription terminator, and an enhancer are control sequences. The target polynucleotide of a CRISPR complex can be any polynucleotide endogenous or exogenous to the eukaryotic cell. For example, the target polynucleotide can be a polynucleotide residing in the nucleus of the eukaryotic cell. The target polynucleotide can be a sequence coding a gene product (e.g., a protein) or a non-coding sequence (e.g., a regulatory polynucleotide or a junk DNA). Examples of target polynucleotides include a sequence associated with a signaling biochemical pathway, e.g., a signaling biochemical pathway-associated gene or polynucleotide. Examples of target polynucleotides include a disease associated gene or polynucleotide. A "disease-associated" gene or polynucleotide refers to any

gene or polynucleotide which is yielding transcription or translation products at an abnormal level or in an abnormal form in cells derived from a disease-affected tissues compared with tissues or cells of a non disease control. It may be a gene that becomes expressed at an abnormally high level; it may be a gene that becomes expressed at an abnormally low level, where the altered expression correlates with the occurrence and/or progression of the disease. A disease-associated gene also refers to a gene possessing mutation(s) or genetic variation that is directly responsible or is in linkage disequilibrium with a gene(s) that is responsible for the etiology of a disease. The transcribed or translated products may be known or unknown, and may be at a normal or abnormal level. The target polynucleotide of a CRISPR complex can be any polynucleotide endogenous or exogenous to the eukaryotic cell. For example, the target polynucleotide can be a polynucleotide residing in the nucleus of the eukaryotic cell. The target polynucleotide can be a sequence coding a gene product (e.g., a protein) or a non-coding sequence (e.g., a regulatory polynucleotide or a junk DNA). Without wishing to be bound by theory, it is believed that the target sequence should be associated with a PAM (protospacer adjacent motif); that is, a short sequence recognized by the CRISPR complex. The precise sequence and length requirements for the PAM differ depending on the CRISPR enzyme used, but PAMs are typically 2-5 base pair sequences adjacent the protospacer (that is, the target sequence) Examples of PAM sequences are given in the examples section below, and the skilled person will be able to identify further PAM sequences for use with a given CRISPR enzyme. In some embodiments, the method comprises allowing a CRISPR complex to bind to the target polynucleotide to effect cleavage of said target polynucleotide thereby modifying the target polynucleotide, wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said target polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence. In one aspect, the invention provides a method of modifying expression of a polynucleotide in a eukaryotic cell. In some embodiments, the method comprises allowing a CRISPR complex to bind to the polynucleotide such that said binding results in increased or decreased expression of said polynucleotide; wherein the CRISPR complex comprises a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence within said polynucleotide, wherein said guide sequence is linked to a tracr mate sequence which in turn hybridizes to a tracr sequence. Similar considerations and conditions apply as above for methods of modifying a target

polynucleotide. In fact, these sampling, culturing and re-introduction options apply across the aspects of the present invention. In one aspect, the invention provides for methods of modifying a target polynucleotide in a eukaryotic cell, which may be *in vivo, ex vivo* or *in vitro*. In some embodiments, the method comprises sampling a cell or population of cells from a human or non-human animal, and modifying the cell or cells. Culturing may occur at any stage *ex vivo*. The cell or cells may even be re-introduced into the non-human animal or plant. For re-introduced cells it is particularly preferred that the cells are stem cells.

[0094] Indeed, in any aspect of the invention, the CRISPR complex may comprise a CRISPR enzyme complexed with a guide sequence hybridized to a target sequence, wherein said guide sequence may be linked to a tracr mate sequence which in turn may hybridize to a tracr sequence.

[0095] As used herein "diagnosis" or "identifying a patient having" refers to a process of determining if an individual is afflicted with, or has a genetic predisposition to develop, a cardiometabolic disease.

[0096] As used herein, a "companion diagnostic" refers to a diagnostic method and or reagent that is used to identify subjects susceptible to treatment with a particular treatment or to monitor treatment and/or to identify an effective dosage for a subject or sub-group or other group of subjects. For purposes herein, a companion diagnostic refers to reagents, such as DNA isolation and sequencing reagents, that are used to detect somatic mutations in a sample. The companion diagnostic refers to the reagents and also to the test(s) that is/are performed with the reagent.

[0097] The present invention may be applied to other diseases in addition to cardiometabolic diseases and diseases affiliated with cardiometabolic diseases. In the present application, atherosclerosis is considered a cardiometabolic disease. The methods of the present invention may also be utilized for the diagnosis, prediction and treatment of other cancers in addition to hematological cancer. Other diseases which may be diagnosed, predicted or treated by methods of the present invention include, but are not limited to, autoimmune diseases (such as arthritis), other diseases involving decreased immunity (such as severe infection), dementia (such as Alzheimer's disease), diabetes, hypertension, Progeroid syndromes and other diseases involving in premature aging (such as Alzheimer's disease and Parkinson's disease).

[0098] The terms "treat," treating," "treatment," and the like refer to reducing or ameliorating a cardiovascular disease or symptoms associated therewith. It will be appreciated that, although not precluded, treating a cardiovascular disease or the risk of developing a cardiometabolic disease does not require that the disease or the risk be completely eliminated.

[0099] In the context of the present invention, a "treatment" is a procedure which alleviates or reduces the negative consequences of a cardiometabolic disease. Many cardiometabolic disease treatments are known in the art, and some are set forth herein. Any treatments or potential treatments can be used in the context of the present invention.

[00100] A treatment is not necessarily curative, and may reduce the effect of a cardiovascular disease by a certain percentage over an untreated a cardiovascular disease. The percentage reduction or diminution can be from 10% up to 20, 30, 40, 50, 60, 70, 80, 90, 95, 99 or 100%.

[00101] Methods of treatment may be personalized medicine procedures, in which the DNA of an individual is analyzed to provide guidance on the appropriate therapy for that specific individual. The methods of the invention may provide guidance as to whether treatment is necessary, as well as revealing progress of the treatment and guiding the requirement for further treatment of the individual.

[00102] As used herein, "inhibiting the development of," "reducing the risk of," "prevent," "preventing," and the like refer to reducing the probability of developing a cardiometabolic disease in a patient who may not have a cardiometabolic disease, but may have a genetic predisposition to developing a cardiometabolic disease. As used herein, "at risk," "susceptible to," or "having a genetic predisposition to," refers to having a propensity to develop a cardiometabolic disease. For example, a patient having a genetic mutation in a gene associated with a cardiometabolic disease has increased risk (e.g., "higher predisposition") of developing the disease relative to a control subject having a "lower predisposition" (e.g., a patient without a genetic mutation in a gene associated with a cardiometabolic disease).

[00103] As used herein, "reduces," "reducing," "inhibit," or "inhibiting," may mean a negative alteration of at least 10%, 15%, 25%, 50%, 75%, or 100%.

[00104] As used herein, "increases" or "increasing" may mean a positive alteration of at least 10%, 15%, 25%, 50%, 75%, or 100%.

[00105] A "therapeutically effective amount" refers to the amount of a compound required to improve, inhibit, or ameliorate a condition of a patient, or a symptom of a disease, in a clinically

relevant manner. Any improvement in the patient is considered sufficient to achieve treatment. A sufficient amount of an active compound used to practice the present invention for the treatment of cardiovascular disease varies depending upon the manner of administration, the age, body weight, genotype, and general health of the patient. Ultimately, the prescribers or researchers will decide the appropriate amount and dosage regimen. Such determinations are routine to one of ordinary skill in the art.

[00106] From a therapeutic perspective, antihypertensives (such as diuretic medicines, beta-blocking agents, calcium-channel blockers, renin-angiotensin system agents), lipid-modifying medicines, anti-inflammatory agents, nitrates and antiarrhythmic medicines are considered strong candidates for a cardiometabolic disease treatment. Aspects of the invention relate to the administration of antihypertensives (such as diuretic medicines, beta-blocking agents, calcium-channel blockers, renin-angiotensin system agents), lipid-modifying medicines, nitrates and antiarrhythmic medicines separately to individuals in need thereof that may also possess different gene variants associated with a favorable response to each type of administration.

[00107] In other embodiments, treatment and/or prevention of cardiometabolic disease may involve aspirin, statins, steroidal or non-steroidal anti-inflammatory drugs, and/or epigenetic modifiers. The epigenetic modifiers may be non-specific DNA synthesis inhibitors, such as DNA methyltransferase inhibitors (such as, but not limited to 5-aza-2'-deoxycytidine or 5-azacytidine) or histone deacetylase inhibitors (such as varinostat, romidepsin, panobinostat, belinostat and entinostat).

[00108] Proprotein convertase subtilisin kexin 9 (PCSK9) is a member of the subtilisin serine protease family. PCSK9 is primarily expressed by the liver and is critical for the down regulation of hepatocyte LDL receptor expression. LDL-C levels in plasma are highly elevated in humans with gain of function mutations in PCSK9, classifying them as having severe hypercholesterolemia. When PCSK9 binds to the LDL receptor, the receptor is broken down and can no longer remove LDL cholesterol from the blood. If PCSK9 is blocked, more LDL receptors will be present on the surface of the liver and will remove more LDL cholesterol from the blood. Therefore, PCSK9 is an attractive target for CRISPR. PCS9K-targeted CRISPR may be formulated in a lipid particle and for example administered at about 15, 45, 90, 150, 250 and 400 μg/kg intraveneously (see, e.g., http://www.alnylam.com/capella/wp-content/uploads/2013/08/ALN-PCS02-001-Protocol-Lancet.pdf).

[00109] Bailey et al. (J Mol Med (Berl). 1999 Jan;77(1):244-9) discloses insulin delivery by ex-vivo somatic cell gene therapy involves the removal of non-B-cell somatic cells (e.g. fibroblasts) from a diabetic patient, and genetically altering them in vitro to produce and secrete insulin. The cells can be grown in culture and returned to the donor as a source of insulin replacement. Cells modified in this way could be evaluated before implantation, and reserve stocks could be cryopreserved. By using the patient's own cells, the procedure should obviate the need for immunosuppression and overcome the problem of tissue supply, while avoiding a recurrence of cell destruction. Ex-vivo somatic cell gene therapy requires an accessible and robust cell type that is amenable to multiple transfections and subject to controlled proliferation. Special problems associated with the use of non-B-cell somatic cells include the processing of proinsulin to insulin, and the conferment of sensitivity to glucose-stimulated proinsulin biosynthesis and regulated insulin release. Preliminary studies using fibroblasts, pituitary cells, kidney (COS) cells and ovarian (CHO) cells suggest that these challenges could be met, and that ex-vivo somatic cell gene therapy offers a feasible approach to insulin replacement therapy. The system of Bailey et al. may be used/and or adapted to the CRISPR Cas system of the present invention for delivery to the liver.

[00110] The methods of Sato et al. (Nature Biotechnology Volume 26 Number 4 April 2008, Pages 431-442) may be applied to the CRISPR Cas system of the present invention for delivery to the liver. Sato et al. found that treatments with the siRNA-bearing vitamin A–coupled liposomes almost completely resolved liver fibrosis and prolonged survival in rats with otherwise lethal dimethylnitrosamine-induced liver cirrhosis in a dose- and duration-dependent manner. Cationic liposomes (Lipotrust) containing O,O'-ditetradecanoyl-N-(a-trimethylammonioacetyl) diethanolamine chloride (DC-6-14) as a cationic lipid, cholesterol and dioleoylphosphatidylethanolamine at a molar ratio of 4:3:3 (which has shown high transfection efficiency under serumcontaining conditions for in vitro and in vivo gene delivery) were purchased from Hokkaido System Science. The liposomes were manufactured using a freeze-dried empty liposomes method and prepared at a concentration of 1 mM (DC-16-4) by addition of double-distilled water (DDW) to the lyophilized lipid mixture under vortexing before use. To prepare VA-coupled liposomes, 200 nmol of vitamin A (retinol, Sigma) dissolved in DMSO was mixed with the liposome suspensions (100 nmol as DC-16-4) by vortexing in a 1.5 ml tube at 25 1C. To prepare VA-coupled liposomes carrying siRNAgp46 (VA-lip-siRNAgp46), a solution of

siRNAgp46 (580 pmol/ml in DDW) was added to the retinol-coupled liposome solution with stirring at 25 C. The ratio of siRNA to DC-16-4 was 1:11.5 (mol/mol) and the siRNA to liposome ratio (wt/wt) was 1:1. Any free vitamin A or siRNA that was not taken up by liposomes were separated from liposomal preparations using a micropartition system (VIVASPIN 2 concentrator 30,000 MWCO PES, VIVASCIENCE). The liposomal suspension was added to the filters and centrifuged at 1,500g for 5 min 3 times at 25 1C. Fractions were collected and the material trapped in the filter was reconstituted with PBS to achieve the desired dose for in vitro or in vivo use. Three injections of 0.75 mg/kg siRNA were given every other day to rats. The system of Sato et al. may be used/and or adapted to the CRISPR Cas system of the present invention for delivery to the liver by delivering about 0.5 to 1 mg/kg of CRISPR Cas RNA in the liposomes as described by Sato et al. to humans.

[00111] The methods of Rozema et al. (PNAS, August 7, 2007, vol. 104, no. 32) for a vehicle for the delivery of siRNA to hepatocytes both in vitro and in vivo, which Rozema et al. have named siRNA Dynamic PolyConjugates may also be applied to the present invention. Key features of the Dynamic Poly-Conjugate technology include a membrane-active polymer, the ability to reversibly mask the activity of this polymer until it reaches the acidic environment of endosomes, and the ability to target this modified polymer and its siRNA cargo specifically to hepatocytes in vivo after simple, low-pressure i.v. injection. SATA-modified siRNAs are synthesized by reaction of 5' aminemodified siRNA with 1 weight equivalents (wt eq) of Nsuccinimidyl-S-acetylthioacetate (SATA) reagent (Pierce) and 0.36 wt eq of $NaHCO_3$ in water at 4°C for 16 h. The modified siRNAs are then precipitated by the addition of 9 vol of ethanol and incubation at 80°C for 2 h. The precipitate is resuspended in 1X siRNA buffer (Dharmacon) and quantified by measuring absorbance at the 260-nm wavelength. PBAVE (30 mg/ml in 5mMTAPS, pH 9) is modified by addition of 1.5 wt % SMPT (Pierce). After a 1-h incubation, 0.8 mg of SMPT-PBAVE was added to 400 μl of isotonic glucose solution containing 5 mM TAPS (pH 9). To this solution was added 50 μg of SATA-modified siRNA. For the dose–response experiments where [PBAVE] was constant, different amounts of siRNA are added. The mixture is then incubated for 16 h. To the solution is then added 5.6 mg of Hepes free base followed by a mixture of 3.7 mg ofCDM-NAGand 1.9mg of CDM-PEG. The solution is then incubated for at least 1 h at room temperature before injection. CDM-PEG and CDM-NAG are synthesized from the acid chloride generated by using oxalyl chloride. To the acid chloride is

added 1.1 molar equivalents polyethylene glycol monomethyl ether (molecular weight average of 450) to generate CDM-PEG or (aminoethoxy)ethoxy-2-(acetylamino)-2-deoxy-β-D-glucopyranoside to generate CDM-NAG. The final product is purified by using reverse-phase HPLC with a 0.1% TFA water/acetonitrile gradient. About 25 to 50 μg of siRNA was delivered to mice. The system of Rozema et al. may be applied to the CRISPR Cas system of the present invention for delivery to the liver, for example by envisioning a dosage of about 50 to about 200 mg of CRISPR Cas for delivery to a human.

[00112] In an aspect the invention provides methods, regents and companion diagnostics for identifying and treating subjects at risk for, or having a PCSK9 mediated disorder. The invention can be used to select therapeutic compositions and dosages, to predict and monitor responses and outcomes. Subjects may be treated to prevent, delay the onset of, or ameliorate the PCSK9 mediated disorder by promoting or inhibiting PCSK9 activity.

[00113] PCSK9 inhibitors can be used to treat patients with familial hypercholesterolemia (FH), clinical atherosclerotic cardiovascular disease (CVD), and other disorders requiring lowering of LDL cholesterol (LDL-C). PCSK9 inhibitors include evolocumab (Repatha™), Praulent, and Alnylam. Evolocumab is a human monoclonal IgG2 antibody which binds to PCSK9 and inhibits circulating PCSK9 from binding to the low density lipoprotein (LDL) receptor (LDLR), preventing PCSK9-mediated LDLR degradation and permitting LDLR to recycle back to the liver cell surface. By inhibiting the binding of PCSK9 to LDLR, evolocumab increases the number of LDLRs available to clear LDL from the blood, thereby lowering LDL-C levels. Evolocumab can be administered alone or in combination with other agents. Evolocumab is used as an adjunct to diet and maximally tolerated statin therapy.

[00114] The liver X receptor (LXR) is a member of the nuclear receptor family of transcription factors and is related to nuclear receptors such as the PPARs, FXR and RXR. Liver X receptors (LXRs) are nuclear receptors involved in regulation of lipid metabolism, cholesterol homeostasis, and inflammatory responses in the central nervous system. Two isoforms of LXR have been identified and are referred to as LXRα and LXRβ. LXRα expression is restricted to liver, kidney, intestine, fat tissue, macrophages, lung, and spleen and is highest in liver, whereas LXRβ is expressed in almost all tissues and organs. Defects contribute to the pathogenesis of neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, multiple sclerosis, and Huntington's disease.

[00115] In an aspect the invention provides methods, regents and companion diagnostics for identifying and treating subjects at risk for, or having an LXR mediated disorder. Subjects are identified and can be treated to prevent, delay the onset of, or ameliorate the LXR mediated disorder by activating or inhibiting LXR activity. LXR agonists (e.g., hypocholamide, T0901317, GW3965, or N,N-dimethyl-3beta-hydroxy-cholenamide (DMHCA)) are useful to reduce the cholesterol level in serum and liver and inhibits the development of atherosclerosis in murine disease models. Certain LXR agonists (e.g. GW3965) improve glucose tolerance in a model of diet-induced obesity and insulin resistance by regulating genes involved in glucose metabolism in liver and adipose tissue.

[00116] Peroxisome proliferator-activated receptors (PPARs) are a group of nuclear receptor proteins that function as transcription factors regulating the expression of genes and play essential roles in the regulation of cellular differentiation, development, and metabolism. There are three main forms: PPARα; PPARβ/δ , and PPARγ. The PPARγ1 isoform is expressed in the spleen, intestine, and white adipose tissue, the PPARγ2 is preferentially expressed in white and brown fat, and PPARγ2 is most abundantly in fat cells, and plays a pivotal role in fat cell differentiation and lipid storage. Hereditary disorders of all PPARs have been described, leading to, for example, lipodystrophy, and insulin resistance.

[00117] In an aspect, the invention provides methods, reagents and diagnostics for identifying, monitoring, and treating inherited disorders. In certain embodiments, the disorders may have single genetic components. Advantageously, the invention is useful for complex multifactorial disorders that do not have a single genetic cause, such as, but not limited to, heart disease, diabetes, and obesity. Thus, multiple gene loci can be evaluated and treatments designed or adjusted accordingly. Accordingly, a subject's disease or disorder can be diagnosed as to relative contributions of variations in multiple genes. Accordingly, patient-specific treatments cam be selected involving multiple drugs in specific combinations and/or dosages. For example, depending on the presence of gene variants at genetic loci that determine propensity to develop a metabolic disease, patient-specific combinations and dosages of drugs to treat diabetes can be selected. According to the invention, in one embodiment, a patient suffering from a particular disease is diagnosed to determine the presence or absence of genetic variants associated with a disease or condition, and a treatment regime assigned accordingly. In particular, treatments that are likely to be effective or to which a subject is likely to be most sensitive are selected. In

another embodiment, treatments which are predicted to result in unwanted patient-specific side effects are avoided.

[00118] Advantageously, the invention provides methods, reagents and diagnostics for identifying, monitoring, and treating diseases or disorders that arise spontaneously or develop over time. In certain embodiments, such disorders are clonal. By clonal it is meant that there is an aspect to the disease that results from appearance or enlargement (or disappearance) of a population of cells. One non-limiting example is a clonal population of cells that arises through a spontaneous mutation. Another non-limiting example is a clonal population of cells that arises from an external stimulation. For example, an autoimmune disorder may arise or be exacerbated by expansion or activation of a population of immune cells. Yet another example is loss of a highly diverse population of cells leaving a population arising from a small number of clones. For example, it has been observed that clonal diversity of immune cells capable of mounting an immune response diminishes with age. In certain embodiments, there is activation or inhibition of expression of one or more genes that develops over time. For example, expression levels can be monitored by measuring transcription or evaluating DNA methylation.

[00119] In an embodiment of the invention, a cholesterol level is monitored and a treatment to reduce cholesterol is initiated. One or more of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* is tested for sequence and a cholesterol medication is selected accordingly, taking into account the alleles of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* present in the subject and their correlation with effectiveness of treatment and minimization of undesired side effects such as liver function and muscle pain.

[00120] Provided herein are sensitive and specific methods to detect and closely monitor somatic mutations associated with disease, particularly a cardiometabolic disease and a hematological cancer. The companion diagnostic methods provided herein are based on the finding that clonal hematopoiesis due to somatic mutation is a common finding in the elderly, and most frequently involves *DNMT3A*, *TET2*, and *ASXL1*. This clinical entity, CHIP, is associated with increased risk of developing hematological malignancy, minimal changes in blood counts, increased overall mortality, and increased risk of cardiometabolic disease. In addition, the companion diagnostic method provided herein also is based on the finding that detecting mutations in *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* specifically provides superior prognostic and treatment selection information as compared to other markers involved

in cardiometabolic disease and a hematological cancer. In exemplary methods provided herein, the diagnostic and prognostic methods are companion methods to therapy with any treatment described herein.

[00121] In one embodiment, the companion diagnostic is used to monitor clonality of somatic mutations in a patient. In another embodiment, clonality is determined. Not being bound by a theory, the size of a clone harboring a somatic mutation, as described herein, can be used to monitor the effectiveness of a treatment.

[00122] In one embodiment, a patient is treated with a cholesterol lowering drug if a mutation in *TET2, DNMT3A,* and/or *JAK2* is observed. Not being bound by a theory, *TET2, DNMT3A,* and/or *JAK2* mutations result in a deficit of reverse cholesterol transport in macrophages and treatment with a cholesterol lowering drug described herein can ameliorate this deficit. In another embodiment, a patient is treated with an anti-inflammatory drug described herein, if a mutation in *TET2, DNMT3A,* and/or *JAK2* is observed. Not being bound by a theory, *TET2, DNMT3A,* and/or *JAK2* mutations result in a protracted inflammatory phenotype in macrophages and treatment with an anti-inflammatory drug described herein can ameliorate this phenotype.

[00123] The present invention also relates to identifying molecules, advantageously small molecules or biologics, that may be involved in inhibiting one or more of the mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*. The invention contemplates screening libraries of small molecules or biologics to identify compounds involved in suppressing or inhibiting expression of somatic mutations or alter the cells phenotypically so that the cells with mutations behave more normally in one or more of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*.

[00124] High-throughput screening (HTS) is contemplated for identifying small molecules or biologics involved in suppressing or inhibiting expression of somatic mutations in one or more of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*. The flexibility of the process has allowed numerous and disparate areas of biology to engage with an equally diverse palate of chemistry (see, e.g., Inglese et al., Nature Chemical Biology 3, 438 - 441 (2007)). Diverse sets of chemical libraries, containing more than 200,000 unique small molecules, as well as natural product libraries, can be screened. This includes, for example, the Prestwick library (1,120 chemicals) of off-patent compounds selected for structural diversity, collective coverage of multiple therapeutic areas, and known safety and bioavailability in humans, as well as the NINDS Custom

Collection 2 consisting of a 1,040 compound-library of mostly FDA-approved drugs (see, e.g., US Patent No. 8,557,746) are also contemplated.

[00125] The NIH's Molecular Libraries Probe Production Centers Network (MLPCN) offers access to thousands of small molecules – chemical compounds that can be used as tools to probe basic biology and advance our understanding of disease. Small molecules can help researchers understand the intricacies of a biological pathway or be starting points for novel therapeutics. The Broad Institute's Probe Development Center (BIPDeC) is part of the MLPCN and offers access to a growing library of over 330,000 compounds for large scale screening and medicinal chemistry. Any of these compounds may be utilized for screening compounds involved in suppressing or inhibiting expression of somatic mutations in one or more of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*.

[00126] The phrase "therapeutically effective amount" as used herein refers to a nontoxic but sufficient amount of a drug, agent, or compound to provide a desired therapeutic effect.

[00127] As used herein "patient" refers to any human being receiving or who may receive medical treatment.

[00128] A "polymorphic site" refers to a polynucleotide that differs from another polynucleotide by one or more single nucleotide changes.

[00129] A "somatic mutation" refers to a change in the genetic structure that is not inherited from a parent, and also not passed to offspring.

[00130] Therapy or treatment according to the invention may be performed alone or in conjunction with another therapy, and may be provided at home, the doctor's office, a clinic, a hospital's outpatient department, or a hospital. Treatment generally begins at a hospital so that the doctor can observe the therapy's effects closely and make any adjustments that are needed. The duration of the therapy depends on the age and condition of the patient, the stage of the a cardiovascular disease, and how the patient responds to the treatment. Additionally, a person having a greater risk of developing a cardiovascular disease (e.g., a person who is genetically predisposed) may receive prophylactic treatment to inhibit or delay symptoms of the disease.

[00131] The medicaments of the invention are prepared in a manner known to those skilled in the art, for example, by means of conventional dissolving, lyophilizing, mixing, granulating or confectioning processes. Methods well known in the art for making formulations are found, for example, in Remington: The Science and Practice of Pharmacy, 20th ed., ed. A. R. Gennaro,

2000, Lippincott Williams & Wilkins, Philadelphia, and Encyclopedia of Pharmaceutical Technology, eds. J. Swarbrick and J. C. Boylan, 1988-1999, Marcel Dekker, New York.

[00132] Administration of medicaments of the invention may be by any suitable means that results in a compound concentration that is effective for treating or inhibiting (e.g., by delaying) the development of a cardiovascular disease. The compound is admixed with a suitable carrier substance, e.g., a pharmaceutically acceptable excipient that preserves the therapeutic properties of the compound with which it is administered. One exemplary pharmaceutically acceptable excipient is physiological saline. The suitable carrier substance is generally present in an amount of 1-95% by weight of the total weight of the medicament. The medicament may be provided in a dosage form that is suitable for oral, rectal, intravenous, intramuscular, subcutaneous, inhalation, nasal, topical or transdermal, vaginal, or ophthalmic administration. Thus, the medicament may be in form of, e.g., tablets, capsules, pills, powders, granulates, suspensions, emulsions, solutions, gels including hydrogels, pastes, ointments, creams, plasters, drenches, delivery devices, suppositories, enemas, injectables, implants, sprays, or aerosols.

[00133] In order to determine the genotype of a patient according to the methods of the present invention, it may be necessary to obtain a sample of genomic DNA from that patient. That sample of genomic DNA may be obtained from a sample of tissue or cells taken from that patient.

[00134] The tissue sample may comprise but is not limited to hair (including roots), skin, buccal swabs, blood, or saliva. The tissue sample may be marked with an identifying number or other indicia that relates the sample to the individual patient from which the sample was taken. The identity of the sample advantageously remains constant throughout the methods of the invention thereby guaranteeing the integrity and continuity of the sample during extraction and analysis. Alternatively, the indicia may be changed in a regular fashion that ensures that the data, and any other associated data, can be related back to the patient from whom the data was obtained. The amount/size of sample required is known to those skilled in the art.

[00135] Generally, the tissue sample may be placed in a container that is labeled using a numbering system bearing a code corresponding to the patient. Accordingly, the genotype of a particular patient is easily traceable.

[00136] In one embodiment of the invention, a sampling device and/or container may be supplied to the physician. The sampling device advantageously takes a consistent and

reproducible sample from individual patients while simultaneously avoiding any cross-contamination of tissue. Accordingly, the size and volume of sample tissues derived from individual patients would be consistent.

[00137] According to the present invention, a sample of DNA is obtained from the tissue sample of the patient of interest. Whatever source of cells or tissue is used, a sufficient amount of cells must be obtained to provide a sufficient amount of DNA for analysis. This amount will be known or readily determinable by those skilled in the art.

[00138] DNA is isolated from the tissue/cells by techniques known to those skilled in the art (see, e.g., U.S. Pat. Nos. 6,548,256 and 5,989,431, Hirota et al., Jinrui Idengaku Zasshi. September 1989; 34(3):217-23 and John et al., Nucleic Acids Res. Jan. 25. 1991 ;19(2):408; the disclosures of which are incorporated by reference in their entireties). For example, high molecular weight DNA may be purified from cells or tissue using proteinase K extraction and ethanol precipitation. DNA may be extracted from a patient specimen using any other suitable methods known in the art.

[00139] It is an object of the present invention to determine the genotype of a given patient of interest by analyzing the DNA from the patent, in order to identify a patient carrying specific somatic mutations of the invention that are associated with developing a cardiovascular disease. In particular, the kit may have primers or other DNA markers for identifying particular mutations such as, but not limited to, one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*.

[00140] There are many methods known in the art for determining the genotype of a patient and for identifying or analyzing whether a given DNA sample contains a particular somatic mutation. Any method for determining genotype can be used for determining genotypes in the present invention. Such methods include, but are not limited to, amplimer sequencing, DNA sequencing, fluorescence spectroscopy, fluorescence resonance energy transfer (or "FRET")-based hybridization analysis, high throughput screening, mass spectroscopy, nucleic acid hybridization, polymerase chain reaction (PCR), RFLP analysis and size chromatography (e.g., capillary or gel chromatography), all of which are well known to one of skill in the art.

[00141] The methods of the present invention, such as whole exome sequencing and targeted amplicon sequencing, have commercial applications in diagnostic kits for the detection of the somatic mutations in patients. A test kit according to the invention may comprise any of the

materials necessary for whole exome sequencing and targeted amplicon sequencing, for example, according to the invention. In a particular advantageous embodiment, a companion diagnostic for the present invention may comprise testing for any of the genes in disclosed herein. The kit further comprises additional means, such as reagents, for detecting or measuring the sequences of the present invention, and also ideally a positive and negative control.

[00142] The present invention further encompasses probes according to the present invention that are immobilized on a solid or flexible support, such as paper, nylon or other type of membrane, filter, chip, glass slide, microchips, microbeads, or any other such matrix, all of which are within the scope of this invention. The probe of this form is now called a "DNA chip". These DNA chips can be used for analyzing the somatic mutations of the present invention. The present invention further encompasses arrays or microarrays of nucleic acid molecules that are based on one or more of the sequences described herein. As used herein "arrays" or "microarrays" refers to an array of distinct polynucleotides or oligonucleotides synthesized on a solid or flexible support, such as paper, nylon or other type of membrane, filter, chip, glass slide, or any other suitable solid support. In one embodiment, the microarray is prepared and used according to the methods and devices described in U.S. Pat. Nos. 5,446,603; 5,545,531; 5,807,522; 5,837,832; 5,874,219; 6,114,122; 6,238,910; 6,365,418; 6,410,229; 6,420,114; 6,432,696; 6,475,808 and 6,489,159 and PCT Publication No. WO 01/45843 A2, the disclosures of which are incorporated by reference in their entireties.

[00143] For the purposes of the present invention, sequence identity or homology is determined by comparing the sequences when aligned so as to maximize overlap and identity while minimizing sequence gaps. In particular, sequence identity may be determined using any of a number of mathematical algorithms. A nonlimiting example of a mathematical algorithm used for comparison of two sequences is the algorithm of Karlin & Altschul, Proc. Natl. Acad. Sci. USA 1990;87: 2264-2268, modified as in Karlin & Altschul, Proc. Natl. Acad. Sci. USA 1993;90: 5873-5877.

[00144] Another example of a mathematical algorithm used for comparison of sequences is the algorithm of Myers & Miller, CABIOS 1988;4: 11-17. Such an algorithm is incorporated into the ALIGN program (version 2.0) which is part of the GCG sequence alignment software package. When utilizing the ALIGN program for comparing amino acid sequences, a PAM120 weight residue table, a gap length penalty of 12, and a gap penalty of 4 can be used. Yet another

useful algorithm for identifying regions of local sequence similarity and alignment is the FASTA algorithm as described in Pearson & Lipman, Proc. Natl. Acad. Sci. USA 1988;85: 2444-2448.

[00145] Advantageous for use according to the present invention is the WU-BLAST (Washington University BLAST) version 2.0 software. WU-BLAST version 2.0 executable programs for several UNIX platforms can be downloaded from the FTP site for Blast at the Washington University in St. Louis website . This program is based on WU-BLAST version 1.4, which in turn is based on the public domain NCBI-BLAST version 1.4 (Altschul & Gish, 1996, Local alignment statistics, Doolittle ed., Methods in Enzymology 266: 460-480; Altschul et al., Journal of Molecular Biology 1990;215: 403-410; Gish & States, 1993;Nature Genetics 3: 266-272; Karlin & Altschul, 1993;Proc. Natl. Acad. Sci. USA 90: 5873-5877; all of which are incorporated by reference herein).

[00146] In all search programs in the suite the gapped alignment routines are integral to the database search itself. Gapping can be turned off if desired. The default penalty (Q) for a gap of length one is Q=9 for proteins and BLASTP, and Q=10 for BLASTN, but may be changed to any integer. The default per-residue penalty for extending a gap (R) is R=2 for proteins and BLASTP, and R=10 for BLASTN, but may be changed to any integer. Any combination of values for Q and R can be used in order to align sequences so as to maximize overlap and identity while minimizing sequence gaps. The default amino acid comparison matrix is BLOSUM62, but other amino acid comparison matrices such as PAM can be utilized.

[00147] Alternatively or additionally, the term "homology" or "identity", for instance, with respect to a nucleotide or amino acid sequence, can indicate a quantitative measure of homology between two sequences. The percent sequence homology can be calculated as $(N_{ref}-N_{dif})* 100/-N_{ref}$, wherein N_{dif} is the total number of non-identical residues in the two sequences when aligned and wherein N_{ref} is the number of residues in one of the sequences. Hence, the DNA sequence AGTCAGTC will have a sequence identity of 75% with the sequence AATCAATC (N N_{ref} =8; N N_{dif} =2). "Homology" or "identity" can refer to the number of positions with identical nucleotides or amino acids divided by the number of nucleotides or amino acids in the shorter of the two sequences wherein alignment of the two sequences can be determined in accordance with the Wilbur and Lipman algorithm (Wilbur & Lipman, Proc Natl Acad Sci USA 1983;80:726, incorporated herein by reference), for instance, using a window size of 20 nucleotides, a word length of 4 nucleotides, and a gap penalty of 4, and computer-assisted analysis and interpretation

of the sequence data including alignment can be conveniently performed using commercially available programs (e.g., Intelligenetics.TM. Suite, Intelligenetics Inc. CA). When RNA sequences are said to be similar, or have a degree of sequence identity or homology with DNA sequences, thymidine (T) in the DNA sequence is considered equal to uracil (U) in the RNA sequence. Thus, RNA sequences are within the scope of the invention and can be derived from DNA sequences, by thymidine (T) in the DNA sequence being considered equal to uracil (U) in RNA sequences. Without undue experimentation, the skilled artisan can consult with many other programs or references for determining percent homology.

[00148] The invention further encompasses kits useful for screening nucleic acids isolated from one or more patients for any of the somatic mutations described herein and instructions for using the oligonucleotide to detect variation in the nucleotide corresponding to one or more of the somatic mutations, such as but not limited to, one or more genes selected from the group consisting of DNMT3A, TET2, ASXL1, TP53, JAK2 and SF3B1, of the isolated nucleic acid.

[00149] Another aspect of the invention is a method of screening patients to determine those patients more likely to develop a cardiovascular disease comprising the steps of obtaining a sample of genetic material from a patient; and assaying for the presence of a genotype in the patient which is associated with developing cardiovascular diseases, any of the herein disclosed somatic mutations.

[00150] In other embodiments of this invention, the step of assaying is selected from the group consisting of: restriction fragment length polymorphism (RFLP) analysis, minisequencing, MALD-TOF, SINE, heteroduplex analysis, single strand conformational polymorphism (SSCP), denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE).

[00151] The present invention also encompasses a transgenic mouse which may express one or more of the herein disclosed somatic mutations. Methods for making a transgenic mouse are well known to one of skill in the art, see e.g., U.S. Patent Nos. 7,709,695; 7,667,090; 7,655,700; 7,626,076; 7,566,812; 7,544,855; 7,538,258; 7,495,147; 7,479,579; 7,449,615; 7,432,414; 7,393,994; 7,371,920; 7,358,416; 7,276,644; 7,265,259; 7,220,892; 7,214,850; 7,186,882; 7,119,249; 7,112,715; 7,098,376; 7,045,678; 7,038,105; 6,750,375; 6,717,031; 6,710,226; 6,689,937; 6,657,104; 6,649,811; 6,613,958; 6,610,905; 6,593,512; 6,576,812; 6,531,645; 6,515,197; 6,452,065; 6,372,958; 6,372,957; 6,369,295; 6,323,391; 6,323,390; 6,316,693;

6,313,373; 6,300,540; 6,255,555; 6,245,963; 6,215,040; 6,211,428; 6,201,166; 6,187,992; 6,184,435; 6,175,057; 6,156,727; 6,137,029; 6,127,598; 6,037,521; 6,025,539; 6,002,067; 5,981,829; 5,936,138; 5,917,124; 5,907,078; 5,894,078; 5,850,004; 5,850,001; 5,847,257; 5,837,875; 5,824,840; 5,824,838; 5,814,716; 5,811,633; 5,723,719; 5,720,936; 5,688,692; 5,631,407; 5,620,881; 5,574,206 and 5,569,827. The transgenic mouse may be utilized to mimic cardiosvascular disease conditions and may be useful to test novel treatments of cardiovascular disease in a mouse model.

[00152] Although the present invention and its advantages have been described in detail, it should be understood that various changes, substitutions and alterations can be made herein without departing from the spirit and scope of the invention as defined in the appended claims.

[00153] The present invention will be further illustrated in the following Examples which are given for illustration purposes only and are not intended to limit the invention in any way.

Examples

Example 1: Methods

[00154] Sample ascertainment. Subjects were ascertained from 22 population-based cohorts in 3 consortia (see Supplementary Table S1). These studies were performed using protocols approved by the ethics committees of all involved institutions, as well as with informed consent from all participants. Samples with missing age (116 subjects) or cell lines as the source of DNA (492 subjects) were excluded.

[00155] Whole Exome Sequencing and Targeted Amplicon Sequencing . DNA was obtained from individual cohorts and further processed at the Broad Institute of MIT and Harvard. Briefly, genomic DNA was subject to hybrid capture, sequencing, and alignment using the Broad Genomics Platform and Picard pipeline. BAM files were analyzed for SNVs using MuTect with OxoG filtering and indels using Indelocator (Cibulskis K et al. Nature biotechnology 2013;31:213-9, Costello M et al. Nucleic acids research 2013;41:e67). A clinically validated, targeted amplicon assay was used for sequencing of 95 genes in select samples.

Variant calling

[00156] Applicants defined a list of pathogenic variants reported in the literature and/or the Catalog of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) in human hematologic malignancies

from 160 genes (see Supplementary Table S3). As a negative control, Applicants also searched for variants that were recurrently seen in non-hematologic malignancies (see Supplementary Table S5) (Lawrence MS et al. Nature 2014;505:495-501).

Statistics

[00157] All statistical analysis was performed using R (http://www.r-project.org/).

Example 2: Results

[00158] Identification of candidate somatic mutations. To investigate the extent of clonal hematopoiesis with somatic mutation, Applicants analyzed whole exome sequencing data from peripheral blood cell DNA of 17,182 individuals who were selected without regard to hematological characteristics. Of these, 15,801 were cases and controls ascertained from 22 cohorts for type 2 diabetes (T2D) association studies, and the remaining 1,381 were additional, previously un-sequenced individuals from the Jackson Heart Study, a population-based cohort (Supplementary Table S1). The median age of individuals was 58 (range 19-108), 8,741 were women, and 7,860 had T2D.

[00159] The identification of somatic driver mutations in cancer has come largely from studies that have compared differences in DNA sequence between tumor and normal tissue from the same individual. Once mutations are identified, investigators may genotype samples for these somatic variants without relying on a matched normal tissue. Because Applicants had DNA from only one source (blood), Applicants limited Applicants' examination to variants previously described in the literature for 160 recurrently mutated candidate genes in myeloid and lymphoid malignancies (Supplementary Table S2). Potential false-positives were removed by utilizing variant-calling algorithms with filters for known artifacts such as strand-bias and clustered reads, as well as additional filtering for rare error modes using a panel of normal (Cibulskis K et al. Nature biotechnology 2013;31:213-9). The lower limit of detection for variants depended on the depth of coverage. The median average sequencing depth over exons from the 160 genes was 84X, and ranged from 13 to 144. At a sequencing depth of 84X, the limit of detection for SNVs was at a variant allele fraction (VAF) of 0.035; for indels, the limit was 0.07.

[00160] With this approach, Applicants identified a total of 805 candidate somatic variants (hereafter referred to as mutations) from 746 individuals in 73 genes (Supplementary Table S3).

As a negative control, Applicants searched for previously described, cancer-associated variants in 40 non-hematologic genes (Supplementary Table S4) (Lawrence MS et al. Nature 2014;505:495-501) and found only 10 such variants in these genes. Below, Applicants show that the frequency of apparent mutations is exceedingly low in young people, and rises with age. These internal controls indicate that the rate of false discovery due to technical artifacts is low. Applicants also verified a subset of the variants using amplicon-based, targeted sequencing; 18/18 variants were confirmed with a correlation coefficient of 0.97 for the VAF between the two methods (Figure 14A).

[00161] The frequency of clonal somatic mutation increases with age. Hematological malignancies, as well as other cancers and pre-malignant states, increase in frequency with age. Mutations were very rare in samples collected before the age of 40, but rose in frequency with each decade of life thereafter (Figure 1). Mutations in genes implicated in hematological malignancies were found in 5.6% (95% CI 5.0-6.3%) of individuals age 60-69, 9.5% (95% CI 8.4-10.8%) of individuals age 70-79, 11.7% (95% CI 8.6-15.7%) of individuals age 80-89, and 18.4 % (95% CI 12.1-27.0%) of individuals older than 90. These rates greatly exceed the incidence of clinically diagnosed hematologic malignancy in the general population (Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2012) (http://www.seer.cancer.gov/popdata). National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch. 2014).

[00162] Though Applicants searched for mutations in genes implicated in multiple hematologic malignancies, Applicants primarily identified genes that were most frequently mutated in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS). The most commonly mutated gene was *DNMT3A* (403 variants, Figure 2A, Figure 6), followed by *TET2* (72 variants) and *ASXL1* (62 variants). Of note, only exon 3 of *TET2* was baited by exon capture (corresponding to ~50% of the coding region), and the portion of exon 12 of *ASXL1* that accounts for ~50% of the mutations in this gene had poor coverage depth. Thus, mutations in *TET2* and *ASXL1* are likely underrepresented in this study. Other frequently mutated genes included *TP53* (33 variants), *JAK2* (31 variants), and *SF3B1* (27 variants).

[00163] In sequencing studies of MDS and AML, most patients have mutations in 2 or more driver genes (the median number of recurrently mutated genes in *de novo* AML patients is five (Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute

myeloid leukemia. The New England journal of medicine 2013;368:2059-74)). In this study, Applicants found that 693 of 746 individuals with a detectable mutation had only 1 mutation in the set of genes Applicants examined, consistent with the hypothesis that these subjects had pre-malignant clones harboring only the initiating lesion (Figure 2B, Figure 7A-B).

[00164] The most common base pair change in the somatic variants was cytosine to thymine (C-T) transition (Figure 2C), which is considered to be a somatic mutational signature of aging (Welch JS et al. Cell 2012;150:264-78, Alexandrov LB et al. Nature 2013;500:415-21). The median VAF for the identified mutations was 0.09, (Figure 2D), suggesting that they are present in only a subset of blood cells, and supports their somatic, rather than germline, origin.

[00165] Somatic mutations persist over time. Blood cell DNA obtained 4 to 8 years after the initial DNA collection was available for targeted sequencing in 13 subjects with 17 somatic mutations (4 subjects had 2 mutations). In all cases, the mutations detected at the earlier time point were still present at the later time point. For 10 mutations, the VAF stayed the same or slightly decreased, and for 7 mutations, the VAF clearly increased. New mutations were detected in 2 subjects, but no subjects were diagnosed with hematological malignancies (Figure 14B).

[00166] Risk factors associated with somatic mutation. To understand risk factors that contributed to having a detectable mutation, Applicants performed a multivariable logistic regression that included age, sex, T2D status, and ancestry as covariates (Supplementary Tables S6-7 and Figure 8). As expected, age was the largest contributor to risk of having a mutation. MDS is characterized by a slight male preponderance. In those age 60 or older, men had an increased likelihood of having a detectable mutation compared to women (OR 1.3, 95% CI 1.1-1.5, p=0.005 by logistic regression). Hispanics and Native Americans are reported to have significantly lower incidence of MDS than other groups in the United States (Rollison DE et al. Blood 2008;112:45-52). Applicants found that Hispanics had a lower risk of having a mutation relative to those of European ancestry, whereas other groups were not significantly different (Supplementary Table S6 and Figure 8). Of the genes Applicants queried, the spectrum of mutations did not differ between ancestry groups (Figure 9).

[00167] Association of somatic mutation with risk of hematologic malignancy. Pre-malignant states such as MGUS, MBL, and others are associated with an increased risk of subsequently developing a malignancy. Of the cohorts that contributed data to the study, two had longitudinal

follow-up information on cancer that arose subsequent to DNA collection (JHS and MEC). Together, these comprised 3,342 subjects, including 134 (4%) in whom Applicants detected somatic mutations in the blood. In a median follow-up period of 95 months, 16 hematological malignancies were reported, of which 5 (31%) were in the group that had detectable mutations (Supplementary Table S11).

[00168] In a fixed-effects meta-analysis of the 2 cohorts adjusted for age, sex, and T2D, hematological malignancies were 11-fold more common in individuals with a detectable clone (95% CI 3.9-33), a difference that was highly significant ($p<0.001$). In individuals with a VAF greater than 0.10 (indicating a higher proportion of cells in the blood carrying the mutation), the risk of developing a diagnosed hematological malignancy was nearly 50-fold (HR 49, 95% CI 21-120, $p<0.001$) (Figure 3A). Consistent with this finding, individuals with a mutation who went on to be diagnosed with a hematological malignancy had a significantly higher mean VAF at the time of blood collection than those who did not (25.2% vs 12.0%, $p=0.003$ by Wilcoxon rank sum test, Figure 3C).

[00169] While individuals with detectable mutations had a markedly increased risk of developing hematological malignancy, the absolute risk remained small; overall, approximately 4% of individuals with a clone developed a hematological malignancy during the study period (Figure 3B). This translates to a risk of developing hematological malignancy of ~0.5% per year overall, and ~1% per year in those with VAF>0.10.

[00170] Blood cell indices of individuals with somatic mutations. Somatic mutations found in MDS and AML lead to abnormal differentiation, ineffective hematopoiesis, and cytopenias. Blood count data was available on 3,107 individuals from 5 cohorts (JHS, UA non-diabetic controls, Botnia, Malmö-sib, Helsinki-sib), including 139 subjects with a detectable mutation. When looking at individuals with single mutations (*ASXL1*, *DNMT3A*, *JAK2*, *SF3B1*, and *TET2*) or mutations in more than 1 gene versus those with no mutations, Applicants found no significant differences in mean white blood cell count, hemoglobin, platelet count, or white blood cell differential after accounting for age and sex (Figure 10). The only statistically significant difference in blood cell indices was an increase in red blood cell distribution width (RDW), a parameter describing the variation in size of red blood cells (13.8% vs 13.4%, $p=0.002$ by Wilcoxon rank-sum test, Supplementary Table S8). This finding suggests that those carrying

somatic mutations might have perturbations in hematopoiesis similar to those seen in MDS, even in the absence of cytopenias.

[00171] Applicants also asked whether the presence of a mutation was associated with increased likelihood of having abnormally low blood counts (Supplementary Table S9). Most of those with a mutation had no cytopenias, nor did they have a higher rate of any single cytopenia than those without mutations. A small fraction of subjects had multiple cytopenias, and these were enriched in the fraction with mutations (OR 3.0, p=0.037 by Fisher's exact test). Furthermore, among anemic subjects, those with mutations had a higher fraction of unexplained anemias than those without mutations (Supplementary Table S10).

[00172] Association of somatic clones with overall survival. Applicants next assessed whether the presence of a somatic mutation had an effect on overall survival based on available data from 5,132 individuals in 7 cohorts (Figure 4) with a median follow-up period of 96 months. In a model adjusted for age, sex, and T2D, carrying a mutation was associated with increased all-cause mortality (HR 1.4, p=0.018 by fixed-effects meta-analysis with beta-coefficients derived from Cox proportional hazards models for individual cohorts, Figure 4A). Kaplan-Meier survival analysis within subjects 70 or older showed an increased risk of death in those with a mutation (p=0.002 by rank-sum test, Figure 4B and Figure 11). Death from hematologic neoplasms alone could not account for the observed increase in mortality, as only 1 individual with a mutation died from hematological malignancy. When Applicants performed a cause-specific mortality analysis, Applicants found that those with mutations had a higher risk of death due to cardiovascular causes, but not cancer (Figure 15).

[00173] Because Applicants found that the presence of a somatic mutation was significantly associated with higher RDW, Applicants also examined whether harboring mutations was synergistic with elevated RDW for risk of death. High RDW has been associated with increased all-cause mortality in the aging and critically ill population (Patel KV et al. Archives of internal medicine 2009;169:515-23, Perlstein TS et al. Archives of internal medicine 2009;169:588-94, Bazick HS et al. Critical care medicine 2011;39:1913-21), but the mechanism behind this association is uncertain. Information on RDW was available on 2,409 subjects in 2 cohorts. In an analysis adjusted for age, sex, and T2D status, Applicants found that having a mutation in conjunction with an RDW≥14.5% (the upper limit of normal) was associated with a marked increase in the risk of death compared to those without mutations and normal RDW (HR 3.7,

$p<0.001$, by fixed-effects meta-analysis with beta-coefficients estimated from Cox models for the two cohorts). In contrast, those with no mutation and high RDW had a more modest increase in mortality (Figure 4C, Figure 12).

[00174] Association of somatic clones with cardiometabolic disease. A recent paper reported that large, somatic chromosomal alterations in peripheral blood cells were associated with having T2D (Bonnefond A et al. Nature genetics 2013;45:1040-3). Applicants also found that somatic mutations in genes known to cause hematologic malignancies were significantly associated with increased risk of T2D, even after adjustment for potential confounding variables (OR 1.3, $p<0.001$, Figures 11, 12). Those with T2D were slightly more likely to have mutations than those without T2D at each age group (Figure 8).

[00175] Cardiovascular disease is the leading cause of death worldwide. Given the association of somatic mutations with all-cause mortality beyond that explicable by hematologic malignancy and T2D, Applicants performed association analyses from two cohorts comprising 3,353 subjects with available data on coronary heart disease (CHD) and ischemic stroke (IS). After excluding those with prevalent events, Applicants found that those carrying a mutation had increased cumulative incidence of both CHD and IS (Figure 5A and 5B). In multivariable analyses that included age, sex, T2D, systolic blood pressure, and body mass index as covariates, the hazard ratio of incident CHD and IS was 2.0 (95% CI 1.2-3.5, P=0.015) and 2.6 (95% CI 1.3-4.8, P=0.003) in the individuals carrying a somatic mutation as compared to those without (Figure 5C and 5D, Figure 8).

[00176] For a subset of individuals, the traditional risk factors of smoking, total cholesterol, and high-density lipoprotein were also available; the presence of a somatic mutation remained significantly associated with incident CHD and IS even in the presence of these risk factors, and the risk was even greater in those with VAF$\geq$0.10 (Supplementary Table S12). Elevated RDW and high-sensitivity C-reactive protein (hsCRP) have also been associated with adverse cardiac outcomes (Tonelli M et al. Circulation 2008;117:163-8, Ridker PM et al. The New England journal of medicine 2002;347:1557-65), possibly reflecting an underlying inflammatory cause. In a multivariable analysis of 1,795 subjects from JHS, those with a mutation and RDW$\geq$14.5% had a markedly increased risk of incident CHD, and this effect was independent of hsCRP (Supplementary Table S13).

[00177] Further validation was performed showing the relationship between clonal hematopoesis and risk for cardiovascular disease by analyzing two additional cohort studies. The BioImage Study is a study of the characteristics of subclinical cardiovascular disease, as measured by imaging modalities, unsupervised circulating biomarker measurements, and risk factors that predict progression to overt clinical cardiovascular disease, in a diverse, population-based sample of 7,300 men (aged 55-80) and women (aged 60-80). The socio-demographics of the study population aims to mirror the US population as a whole with approximately 69% of the cohort will be white, 12% African-American, 13% Hispanic, 4% Asian, predominantly of Chinese descent and 2% other (U.S. Census Bureau: 2000). The Malmö Diet and Cancer study is a 10-year prospective case-control study in 45-64-year-old men and women (n = 53,000) living in a city with 230,000 inhabitants.

[00178] Participants from the nested case-control studies that passed sample quality control (contamination, prevalent cardiovascular disease, germline Ti/Tv, germline total variants / depth, germline F inbreeding coefficient) are displayed (Table 1). Participants were matched by age, sex, diabetes status, and smoking status. LDL cholesterol and total cholesterol are adjusted accounting for statin medications as previously described. Individuals that carried at least one mutation in a gene conferring risk of clonal expansion with a variant allele fraction > 0.10 are defined as carrying a somatic clone.

[00179] Table 1. Baseline characteristics of study participants

	BioImage		**Malmo Diet & Cancer**	
	Cases N = 150	**Controls N = 344**	**Cases N = 536**	**Controls N = 531**
Age, y	70.2 (5.8)	70.3 (5.9)	59.9 (5.4)	59.9 (5.4)
Female	60 (42.9%)	136 (39.5%)	233 (43.5%)	233 (43.9%)
Total cholesterol, mg/dL	208.8 (38.1)	214.0 (37.5)	271.2 (173.1)	252.7 (133.8)
LDL cholesterol, mg/dL	127.4 (31.4)	122.9 (32.8)	186.9 (117.1)	172.0 (92.5)
HDL cholesterol, mg/dL	51.0 (14.7)	53.1 (15.1)	49.1 (13.0)	51.8 (13.7)
Triglycerides, mg/dL	181.6 (94.1)	169.6 (93.6)	127.4 (57.4)	123.7 (58.9)
Diabetes mellitus type 2	36 (24.0%)	89 (25.9%)	82 (15.3%)	80 (15.1%)
Hypertension	130 (86.7%)	256 (74.4%)	421 (78.5%)	354 (66.7%)
Smoker	24 (16.0%)	55 (16.0%)	167 (31.2%)	166 (31.3%)
BMI	27.8 (4.7)	27.4 (4.7)	26.5 (4.0)	26.2 (4.0)
Somatic clone carrier	**24 (16.0%)**	**34 (9.9%)**	**31 (5.8%)**	**22 (4.1%)**

[00180] Applicants show an increased risk of cardiovascular events from carrying a somatic clonal mutation in both studies (Figure 18). Furthermore, Applicants show that an increase in

coronary arterial calcification quantity is associated with somatic clonal mutation carrier status (Figure 19).

Example 3: Discussion

[00181] Applicants find that somatic mutations leading to clonal outgrowth of hematopoietic cells are frequent in the general population. This entity, which Applicants term clonal hematopoiesis with indeterminate potential (CHIP), is present in over 10% of individuals over 70, making it one of the most common known pre-malignant lesions. The exact prevalence of CHIP is dependent on how cancer-causing mutations are defined and on the sensitivity of the technique used to detect mutations, and thus may substantially exceed this estimate. Unlike other pre-malignant lesions, CHIP appears to involve a substantial proportion of the affected tissue in most individuals; based on the proportion of alleles with the somatic mutation, Applicants find that a median of 18% of peripheral blood leukocytes are part of the abnormal clone. CHIP also persists over time; in all tested cases, the mutations were still present after 4 to 8 years.

[00182] The genes most commonly mutated in CHIP are *DNMT3A*, *TET2*, and *ASXL1*. This is consistent with previous studies that have found *DNMT3A* and *TET2* mutations to be frequent and early events in AML and MDS (Jan M et al. Science translational medicine 2012;4:149ra18, Shlush LI et al. Nature 2014;506:328-33, Papaemmanuil E et al. The New England journal of medicine 2011;365:1384-95, Welch JS et al. Cell 2012;150:264-78). Murine models of *DNMT3A* or *TET2* loss-of-function demonstrate that mutant HSCs have altered methylation patterns at pluripotency genes and a competitive advantage compared to wild-type HSCs, but mice rarely develop frank malignancy, and then only after long latency (Jeong M et al. Nature genetics 2014;46:17-23, Koh KP et al. Cell stem cell 2011;8:200-13, Challen GA et al. Nature genetics 2012;44:23-31, Moran-Crusio K et al. Cancer cell 2011;20:11-24). Similarly, Applicants' data show that humans with CHIP can live for many years without developing hematological malignancies, though they do have increased risk relative to those without mutations.

[00183] Certain genes commonly mutated in AML and MDS are absent or very rare in this study. Their rarity likely indicates that they are cooperating rather than initiating mutations. While mutations in genes specific for lymphoid malignancies were rarely detected, it is

important to note that *TET2* and *DNMT3A* are frequently mutated in some lymphoid malignancies, and the initiating event for such tumors may occur in a HSC (Neumann M et al. Blood 2013;121:4749-52, Quivoron C et al. Cancer cell 2011;20:25-38, Odejide O et al. Blood 2014;123:1293-6, Asmar F et al. Haematologica 2013;98:1912-20, Couronne L et al. The New England journal of medicine 2012;366:95-6). While it is most likely that these mutations occur in a HSC, it also possible that they occur in committed myeloid progenitors or mature lymphoid cells that have acquired long-term self-renewal capacity.

[00184] The use of somatic mutations to aid in the diagnosis of patients with clinical MDS is becoming widespread. Applicants' data demonstrate that the majority of individuals with clonal mutations in peripheral blood do not have MDS or another hematological malignancy, nor do the majority develop a clinically diagnosed malignancy in the near term. At this time, it would be premature to genetically screen healthy individuals for the presence of a somatic clone, as the positive predictive value for current or future malignancy is low. Further studies are needed to definitively assess whether the detection of a mutation in conjunction with blood count abnormalities is sufficient to make a presumptive diagnosis of MDS or another hematological malignancy.

[00185] Perhaps the most surprising finding in Applicants' study is the lower overall rate of survival in those with clones as compared to those without. This effect is much larger than can be explained by hematological malignancies alone, is synergistic with high RDW (which could be a marker of perturbation of hematopoiesis due to the clone), and may be related to the increased risk of incident CHD and IS in those with clones. The association of somatic mutations with non-hematological disease may be due to confounding by variables that are currently unknown, or may simply represent a shared consequence of the underlying process of aging. Alternatively, it may represent an underlying shared pathophysiology of seemingly unrelated disorders. For example, cells of the monocyte/macrophage lineage are considered important mediators of atherosclerosis and type 2 diabetes (Libby P. 2002;420:868-74, Olefsky JM, Glass CK. Annual review of physiology 2010;72:219-46). Applicants propose that one possible explanation for these findings is that somatic mutations that lead to clonal hematopoiesis cause functional abnormalities in differentiated blood cells that modulate the risk of cardiometabolic disease.

[00186] In summary, Applicants find that clonal hematopoiesis due to somatic mutation is a common finding in the elderly, and most frequently involves *DNMT3A*, *TET2*, and *ASXL1*. This clinical entity, CHIP, is associated with increased risk of developing hematological malignancy, minimal changes in blood counts, increased overall mortality, and increased risk of cardiometabolic disease.

Example 4: Supplemental Methods

[00187] Sample ascertainment and cohort descriptions. Subjects were ascertained as cases and controls for T2D from 22 cohorts in 3 consortia (Supplementary Table 1). Details on sample ascertainment for cases and controls from the most of the GoT2D (8 cohorts, 2,376 subjects) and SIGMA (4 cohorts, 3,435 subjects) consortia have been previously described (Flannick J et al. Nature genetics 2013;45:1380-5, Consortium STD et al. Nature 2014;506:97-101). The other ascertained individuals were part of T2D-GENES (9,990 subjects), a consortium comprised of 10 population-based cohorts. Details on sample ascertainment can be found in Supplementary Table 1. The remaining 1,381 individuals were additional subjects in the Jackson Heart Study (JHS), a large population-based cohort of African-Americans in Jackson, Mississippi, who had given consent for genetic testing but were not in any previously sequenced cohorts (Sempos CT et al. The American journal of the medical sciences 1999;317:142-6.). Including the 1,027 subjects from JHS enrolled through T2D-GENES (513 with T2D), a total of 2,408 subjects from JHS were in this study. Since 3,400 subjects in JHS were consented for genetic studies, ~70% of consented subjects from JHS are represented in this study, with a modest overall enrichment for T2D.

[00188] Subjects for which age was not available (116 subjects) or with cell lines as the source DNA (492 subjects) were excluded, including all subjects from the Wellcome Trust Case Control Consortium (WTCCC).

[00189] Vital status for Finland-United States Investigation of NIDDM Genetics Study (FUSION), The Botnia Study (Botnia), Helsinki Siblings with Diabetes cohort (Helsinki_sib), and Scania Diabetes Register (Diabetes_reg) was ascertained from the Finnish or Swedish Hospital Death Registries. Subjects from Botnia, Helsinki_sib, and Diabetes_reg were pooled for survival analysis. Vital status for subjects from the Multiethnic Cohort (MEC) was ascertained from Center for Medicare Services (CMS) data. Vital status for subjects from the JHS was

ascertained from vital records and annual follow-up interviews. For individuals lost to follow-up, if there was no death certificate, the individual was assumed to be alive. Vital status for non-diabetic Ashkenazis in the Longevity Genes Project (LGP) was ascertained from hospital death records and annual follow-up interviews.

[00190] Malignancy information for MEC was ascertained through linkage of the MEC with cancer registries of California and Hawaii. Malignancy information for JHS was ascertained from annual interview. Malignancy information for some subjects from FUSION and Botnia was available from the Finnish Hospital Discharge Register and Death Register, but not included because it was deemed to have significant ascertainment bias.

[00191] Standard blood cell indices (white blood cell count, hemoglobin, hematocrit, platelet count, and white blood cell differential) were available for most (but not all) subjects from JHS, LGP, Botnia, Malmo-sib, and Helsinki-sib. Information on red blood cell distribution width (RDW) was available on most subjects in JHS and LGP.

[00192] Data on cardiovascular outcomes for JHS was obtained from annual patient interview and adjudicated from hospital records. Data on cardiovascular outcomes for FUSION was obtained from Finnish Hospital Discharge and Death Registries. Coronary heart disease (CHD) included fatal and non-fatal myocardial infarctions as well as coronary revascularization procedures. For CHD analysis, those with prior CHD events were excluded. For ischemic stroke analysis, those with prior ischemic stroke were excluded. Lab data (blood pressure, body mass index, serum high density lipoprotein, serum total cholesterol, and high-sensitivity C-reactive protein) was obtained at the same time as blood collection for DNA.

[00193] Exome sequencing. DNA was obtained from individual cohorts and further processed at the Broad Institute of MIT and Harvard. DNA libraries were bar coded using the Illumina index read strategy, exon capture was performed using Agilent Sure-Select Human All Exon v2.0, and sequencing was performed by Illumina HiSeq2000. Sequence data were aligned by the Picard (http://picard.sourceforge.net) pipeline using reference genome hg19 with the BWA algorithm (Li H, Durbin R. Bioinformatics 2009;25:1754-60) and processed with the Genome Analysis Toolkit (GATK) to recalibrate base-quality scores and perform local realignment around known insertions and deletions (indels) (DePristo MA et al. Nature genetics 2011;43:491-8.). BAM files were then analyzed for single nucleotide variants using MuTect (http://www.broadinstitute.org/cancer/cga/mutect) with Oxo-G filtering

(http://www.broadinstitute.org/cancer/cga/dtoxog) and for indels using Indelocator (http://www.broadinstitute.org/cancer/cga/indelocator), followed by annotation using Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator/) (Cibulskis K et al. Nature biotechnology 2013;31:213-9). All MuTect and Indelocator analyses were performed using the Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose) at the Broad Institute.

[00194] Variant calling. Cancer genome studies typically compare sequence from tumor and germline DNA, and define somatic mutations as the sequence variants present in tumor but not germline DNA. To circumvent the lack of matched tissue in this study, Applicants defined a list of pathogenic variants reported in the literature and/or the Catalog of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) in human hematologic malignancies from 160 genes (see Supplementary Table S2). Applicants specifically excluded genes known to be involved in hematologic malignancy that had a relatively high frequency of heterozygous loss-of-function germline mutations in the population (*NF1, SH2B3, BRCA1/2*). Identical frameshift variants that were seen 3 or more times from the same ancestry group were also excluded, unless such variants were previously reported as somatic. Frameshift and nonsense mutations were further excluded if they occurred in the first or last 10% of the gene open reading frame (Ng PC et al. PLoS genetics 2008;4:e1000160), unless mutations in those regions had been previously reported, (e.g. DNMT3A (Walter MJ et al. Leukemia 2011;25:1153-8)). Applicants used minimum variant read counts of 3 for MuTect and 6 for Indelocator. Python (http://www.python.org) scripts were used to parse mutation annotation format (MAF) files produced by MuTect and Indelocator for variants of interest.

[00195] To further confirm that Applicants were detecting bona fide somatic mutations, Applicants examined variants in 40 driver genes involved in non-hematologic malignancies (Lawrence MS et al. Nature 2014;505:495-501). Applicants hypothesized that very few variants would be detected from these genes if Applicants' methodology had high specificity for real mutations. Using the same calling approach as with hematologic genes (Supplementary Table S4), Applicants detected only 10 variants that were the same as those mutated in non-hematologic cancers, and most of these appeared to be rare germline polymorphisms as evidenced by allele fraction (Supplementary Table S5).

[00196] Targeted re-sequencing. Validation of variants discovered by whole exome sequencing was done with "Rapid Heme Panel" (RHP), a Laboratory Developed Test designed

and validated at a CLIA-certified lab (Center for Advanced Molecular Diagnostics, Brigham and Women's Hospital). RHP uses TruSeq Custom Amplicon Kit (Illumina, Inc. San Diego, CA, USA) and contains 95 genes (50 for AML/MDS, 8 for MPN, 27 for ALL, and 10 others). For oncogenes, known mutation hotspots are targeted; and for tumor suppressor genes the entire coding sequence is analyzed. The average amplicon size is 250-bp and about 50% of the regions are covered on both strands. Library preparation was according to manufacturer's instruction and sequencing was 150 bp paired-end reads with MiSeq v2.2 chemistry. Raw data was analyzed with Illumina on-board Real-Time-Analysis (RTA v.2.4.60.8) software and MiSeq Reporter. The VCF files were filtered with a cutoff for any nucleotide position with 10 or more variant reads or with 5-9 variant reads (if allele frequency >33%) as well as a Q score greater than 30 and germline single nucleotide polymorphisms were removed by comparison to dbSNP database (NCBI Human Build 141). The filtered variant lists were manually reviewed and BAM file examined in Integrated Genome Viewer (IGV, Broad Institute).

[00197] For 13 subjects from JHS, DNA obtained from a peripheral blood sample collected 4 to 8 years after the original DNA was available for analysis. RHP was used as described above to assess VAF of the previously detected mutations at the second time point, and to assess for the acquisition of new mutations.

[00198] Genes: ABL1, ASXL1, ATM, BCL11B, BCOR, BCORL1, BRAF, BRCC3, CALR, CBL, CBLB, CD79B, CEBPA, CNOT3, CREBBP, CRLF2, CSF1R, CSF3R, CTCF, CTNNB1, CUX1, CXCR4, DMNT3A, DNMT3B, EED, EGFR, EP300, ETV6, EZH2, FANCL, FBXW7, FLT3, GATA1, GATA2, GATA3, GNAS, GNB1, IDH1, IDH2, IKZF1, IKZF2, IKZF3, IL7R, JAK1, JAK2, JAK3, KIT, KRAS, LUC7L2, MAP2K1, MEF2B, MPL, MYD88, NOTCH1, NOTCH2, NOTCH3, NPM1, NRAS, NT5C2, PAX5, PDGFRA, PDS5B, PHF6, PIGA, PIK3CA, PIM1, PRPF40B, PRPF8, PTEN, PTPN11, RAD21, RET, RIT1, RPL10, RUNX1, SETBP1, SETD2, SF1, SF3A1, SF3B1, SH2B3, SMC1A, SMC3, SRSF2, STAG2, STAT3, TET2, TLR2, TP53, U2AF1, U2AF2, WHSC1, WT1, XPO1, ZRSR2.

[00199] Statistics and analysis plan. All statistical analyses were performed using R. Cox proportional hazards and Kaplan-Meier analysis was performed using the **survival** package (http://cran.r-project.org/web/packages/survival/index.html). Competing risks regression (CRR) was used to estimate hazard ratios for developing hematologic malignancy with death as the competing risk (Fine JP and Gray, RJ. Journal of the American Statistical Association

1999;94:496-509). CRR and cumulative incidence analysis was performed using the **cmprsk** package (http://cran.r-project.org/web/packages/cmprsk/index.html). Fixed-effects meta-analysis using beta-coefficients for risk estimates from individual cohorts was used to provide summary hazard ratios across heterogenous cohorts. Meta-analysis was performed using the **meta** package (http://cran.r-project.org/web/packages/meta/index.html).

[00200] Analyses for factors associated with clonal hematopoiesis were performed using logistic regression with the pre-determined variables age, sex, type 2 diabetes status, and ancestry.

[00201] Primary outcomes for clinical associations with clonal hematopoiesis were pre-determined to be all-cause mortality and hematologic malignancy, using Cox proportional hazards models and CRR, respectively. Association of mutations with blood counts was also a primary analysis. For hematologic malignancy outcomes, only events incident to the time of DNA collection were considered. Red cell distribution width was also included as a variable in the survival analysis because of its association with mutations and previous reports of its association with increased mortality.

[00202] An analysis of cause specific mortality revealed an increase in cardiovascular deaths. For this reason, Applicants examined incident events of coronary heart disease and ischemic stroke using CRR and the pre-determined variables age, sex, type 2 diabetes status, body mass index, and systolic blood pressure. For some subjects, data was also available on smoking status, total cholesterol, and high-density lipoprotein. This was examined as a subgroup analysis in a separate regression (see Supplementary Tables S12-13).

[00203] For some analyses, a cutoff of VAF at 0.10 was used. This value was chosen because it was close to the median VAF in the dataset and is roughly the lower limit of detection using Sanger DNA sequencing, and was thus used to designate large versus small clone size.

Example 5: Supplemental Tables

Supplementary Table S1
Cohort descriptions and baseline characteristics

TOTAL	No T2D	T2D	T2D unknown	Clinical information available	Cohort descriptions and sample ascertainment	References
Overall (number with clone)	8903(5867)	7861(5378)	18(1)			
Female (number with clone)	4589(2184)	3845(1573)	3(0)			
Mean age (SD)	57(15)	59(10)				
Mean BMI (SD)	28(5.8)	25(5.6)				
Population-based	**No T2D**	**T2D**				
JHS (Jackson Heart Study)						
Overall (number with clone)	1801(57)	589(25)	18(1)	Survival, malignancy, blood counts, cardiovascular events	The Jackson Heart Study is a population-based cohort of 5,306 African-Americans living in the Jackson, Mississippi metropolitan area. Subjects were enrolled as random members of community, volunteers, as part of the Atherosclerosis Risk in Communities (ARIC), or secondary family members. At 12-month intervals after the baseline clinic visit (Exam 1), participants were contacted by telephone to update information; another clinic examination, documented interim medical events, hospitalizations, and functional status, and obtain additional socioeconomic information. Questions about medical events, symptoms of cardiovascular disease and functional status were repeated annually. Ongoing cohort surveillance involves abstraction of medical records and death certificates for relevant International Classification of Diseases (ICD) codes and adjudication of nonfatal events and deaths.	Wilson, J.G., et al. (2005) "Study design for genetic analysis in the Jackson Heart Study." Ethnicity and Disease 15 30-37; Taylor, H.A. (2005). "The Jackson Heart Study: An Overview." Ethnicity and Disease 15:1-3; Fuqua, S.R. et al., (2005). "Recruiting African-American Research Participation in the Jackson Heart Study: Methods, Response Rates, and Sample Description." Ethnicity and Disease 15 43-48; Fuqua, S.R. et al., (2005).
Female (number with clone)	1003(30)	401(18)	3(0)			
Mean age (SD)	53(13)	58(11)				
Mean BMI (SD)	31(6.8)	34(6.4)				
GoT2D	**No T2D**	**T2D**				
Botnia (The Botnia Study), Diabetes_reg (Scania Diabetes Registry), Helsinki-sib (Helsinki siblings with diabetes cohort), Malmö-sib (siblings in Malmö, Sweden)				Survival, blood counts, cardiovascular events	The Botnia Project was started in 1990 to study risk factors for T2D in western Finland and southern Sweden and includes ~11,000 people from ~3400 families. The Botnia Diabetes Registry contains over 7000 diabetes patients recruited at hospitals in Scania, Sweden as from 1996. The majority of the patients come from the city of Malmö, and they account for about 25% of all diabetic patients in the region. Death information was obtained from the Finnish or Swedish Death and Hospital Discharge Registers. Individuals were ranked according to a liability model that measured risk for T2D. Briefly, liability scores were computed as the difference between diabetes status and the predicted risk based on age, BMI and gender. Extreme cases were selected to have the highest liability scores (with diabetes but with low predicted risk for diabetes), and extreme controls were selected to have the lowest liability scores (without diabetes but with high predicted risk for diabetes).	Franks, J., et al. (2013). "Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes." Nat Genet 45(11): 1380-1385; Group, L., et al. (1996). "Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects." Diabetes 45(11): 1585-1593; Lindholm, E., et al. (2001). "Classifying diabetes according to the new WHO clinical stages." Eur J Epidemiol 17(11): 553-559.
Overall (number with clone)	324(12)	369(14)				
Female (number with clone)	108(8)	211(5)				
Mean age (SD)	65(10)	57(10)				
Mean BMI (SD)	30(3.7)	25(2.7)				
FUSION (Finland-United States Investigation of NIDDM Genetics Study)				Survival, blood counts, cardiovascular events	The Finland-United States Investigation of NIDDM Genetics (FUSION) study is a long-term effort to identify genetic variants that predispose to type 2 diabetes (T2D) or that impact the variability of T2D-related quantitative traits. Unrelated T2D cases were selected from FUSION affected-sibpair families and from stage 2 replication. NGT controls with higher age and BMI were prioritized and were frequency matched to cases by birth province.	Valle, T., et al. (1998). "Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study." Diabetes Care 21(6): 949-958; Scott, L. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829): 1341-1345 (2007)
Overall (number with clone)	474(17)	470(21)				
Female (number with clone)	213(11)	201(10)				
Mean age (SD)	53(7.2)	58(8.6)				
Mean BMI (SD)	28(3.9)	31(5.5)				
KORA					KORA is a regional research platform for population-based studies, subsequent follow-up studies and family studies, established in 1999. Cases and controls were all of German ancestry. Cases were identified by self report of T2D in a personal interview which was validated by a questionnaire mailed to the treating physician and/or by medical chart review. Controls were non-diabetic as defined by self-report.	Wichmann, H. E., et al. (2005). "MONICA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes." Gesundheitswesen 67 Suppl 1: S26-30.
Overall (number with clone)	91(11)	97(6)				
Female (number with clone)	58(8)	43(3)				
Mean age (SD)	70(5.6)	61(8.1)				
Mean BMI (SD)	35(6.5)	28(2.8)				
MPP (Malmo Preventive Project)					The MPP was started in the early 1970s as a screening survey in the middle-aged population of Malmö, the third largest city of Sweden. Subjects born in Malmö and residents of the city were invited for a clinical examination, questionnaire and blood sampling. In all 33,444 men and 20,900 women participated during the period 1974-1992. Cases and controls were selected as for Botnia.	Berglund, G., et al. (2003). "Long-term outcome of the Malmo preventive project: mortality and cardiovascular morbidity." J Intern Med 247(1): 19-29
Overall (number with clone)	212(14)	110(8)				
Female (number with clone)	87(7)	51(5)				
Mean age (SD)	68(6.2)	47(6.1)				
Mean BMI (SD)	36(1.8)	23(1.4)				
STT (UKT2D Consortium, Controls)					Non-diabetic controls selected from the Twins UK Study. A twin pair was considered for selection if there was no recorded family history of diabetes, neither twin was ever recorded as impaired glucose tolerant, there were available quantitative trait and genetic data, and no evidence of admixture. From set of qualifying twin pairs, the best control twin was selected from each pair with the lowest ratio of fasting glucose level to BMI across all readings	Moayyeri, A., et al. (2013). "The UK Adult Twin Registry (TwinsUK Resource)." Twin Res Hum Genet 16(1): 144-149.
Overall (number with clone)	819(14)	*				
Female (number with clone)	285(11)	*				
Mean age (SD)	61(10)	*				
Mean BMI (SD)	31(5.9)	*				
SIGMA	**No T2D**	**T2D**				
MEC (Multiethnic cohort, Hispanics in Los Angeles)				Survival, malignancy, cardiovascular events	The MEC consists of 215,251 men and women in Hawaii and Los Angeles. Those self-identified as Latinos were used for sequencing in a prior study. Between 1993 and 1996, adults between 45 and 75 years old were enrolled. Potential cohort members were identified through Department of Motor Vehicles drivers' license files, voter registration files and Health Care Financing Administration data files. Between 1995 and 2008, blood specimens were collected from ~67,000 MEC participants. Controls were frequency matched to cases on sex, ethnicity and age at entry into the cohort (5-year age groups) and place of birth (U.S. vs. Mexico, South or Central America). Persons in the cohort who develop cancer are identified through Surveillance, Epidemiology, and End Results (SEER) Program registries that have been established by state statute in Hawaii and California.	Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, et al. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 151: 346-357.
Overall (number with clone)	445(24)	486(27)				
Female (number with clone)	215(13)	263(14)				
Mean age (SD)	68(7.2)	68(7.6)				
Mean BMI (SD)	27(4.5)	30(5.7)				
MexB1 (UNAM/INCMNSZ Diabetes Study)					Cases were recruited at the outpatient diabetes clinic of the Department of Endocrinology and Metabolism of the Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran (INCMNSZ). All Mexican-mestizo individuals were invited to participate in the study. Control subjects were recruited from a cohort of adults aged 45 years or older among participants	SIGMA T2D Consortium, et al. (2014). "Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico." Nature 505(7484): 97-101
Overall (number with clone)	543(7)	550(20)				

	Female (number with clone)	338(8)	328(15)	
	Mean age (SD)	55(9.4)	55(13)	
	Mean BMI (SD)	28(3.9)	28(4.4)	
MexB2 (Diabetes in Mexico Study)			Individuals were recruited from two tertiary level institutions (INSS and SSSTE) located in Mexico City. Unrelated healthy subjects older than 45 years and with fasting glucose levels below 100 mg/dl were classified as controls. Unrelated individuals, older than 18 years, with either previous T2D diagnosis or fasting glucose levels above 125 mg/dl were included as T2D cases	SIGMA T2D Consortium, et al. (2014). "Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico." Nature 506(7486): 97-101
	Overall (number with clone)	177(3)	393(11)	
	Female (number with clone)	134(3)	275(9)	
	Mean age (SD)	56(9.3)	57(12)	
	Mean BMI (SD)	28(4.6)	29(5.5)	
MexB3 (Mexico City Diabetes Study)			The Mexico City Diabetes Study is a population based prospective investigation. All 35-64 years of age men and non-pregnant women residing in the study site (low income neighborhoods equivalent to 6 census tracks with a total population of 35,000 inhabitants) were interviewed and invited to participate in the study. There was a response rate of 67% for the initial exam. Diagnostic criteria for type 2 diabetes were recommended by the ADA.	SIGMA T2D Consortium, et al. (2014). "Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico." Nature 506(7486): 97-101.
	Overall (number with clone)	550(25)	392(9)	
	Female (number with clone)	334(14)	184(2)	
	Mean age (SD)	62(7.5)	64(7.5)	
	Mean BMI (SD)	29(4.9)	30(5.5)	

T2D-GENES	No T2D	T2D		
AW (Wake Forest School of Medicine Study)			Cases are self-reported diabetics with diabetic nephropathy, recruited from diabetic clinics with age of onset ≥25. Controls were recruited from community and internal medicine clinics and had a current diagnosis of diabetes or renal disease.	Palmer, N. D. et al. A genome-wide association search for type 2 diabetes genes in African Americans. PLoS One 7, e29202. (2012)
	Overall (number with clone)	531(18)	529(48)	
	Female (number with clone)	308(9)	317(28)	
	Mean age (SD)	52(12)	64(9.1)	
	Mean BMI (SD)	30(7.0)	29(6.6)	
EK (Korea Association Research Project)			Cases selected for age of onset of T2D ≥40 years. Participants with early onset T2D and family history were prioritized. Controls were selected for not having current or past diabetes and no diabetic medications. Older subjects with normal glucose were prioritized.	Cho, Y. S. et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat. Genet. 41, 527–534 (2009)
	Overall (number with clone)	554(32)	521(20)	
	Female (number with clone)	324(18)	287(4)	
	Mean age (SD)	68(8.6)	54(7.5)	
	Mean BMI (SD)	24(3.1)	26(3.3)	
Study and Singapore Prospective Study Program			Cases were clinically ascertained T2D from primary care clinics. Individuals with early age of diagnosis and with at least one first degree relative with T2D were preferentially selected. Controls were defined as "fasting blood glucose <6 mmol/l", no personal history of diabetes, and no anti-diabetic medication. Older controls preferentially selected	Sim, X. et al. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. PLoS Genet. 7(4), e1001363 (2011)
	Overall (number with clone)	592(25)	478(24)	
	Female (number with clone)	383(11)	250(11)	
	Mean age (SD)	58(7.0)	58(8.3)	
	Mean BMI (SD)	23(3.4)	26(3.5)	
HA (Hispanics from San Antonio Family Heart Study, San Antonio Family Diabetes/ Gallbladder Study, Veterans Administration Genetic Epidemiology Study, and the Investigation of Nephropathy and Diabetes Study family component)			Cases were drawn from four separate family studies and met the following criteria: ADA 2002 criterion, WHO 1999 criteria, or physician reported diagnosis with current medical therapy. Controls defined by not having fasting glucose <126 mg/dL at each visit and no history of prior diabetic medication.	Mitchell, B. D. et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. Circulation 94, 2159–2170 (1996); Hunt, K. J. et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. Diabetes 54, 3655–3661 (2005); Colletta, D. K. et al. Genome-wide linkage scan for genes influencing plasma triglyceride levels in the Veterans Administration Genetic Epidemiology Study. Diabetes 58, 275–284 (2009); Knowler, W. C. et al. The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. J. Diabetes Complicat. 19, 1–9 (2005)
	Overall (number with clone)	111(6)	142(3)	
	Female (number with clone)	62(4)	59(1)	
	Mean age (SD)	44(14)	51(12)	
	Mean BMI (SD)	30(5.5)	33(6.3)	
HS (Hispanics in Starr County, Texas)			Cases defined by fasting glucose ≥ 140 mg/dl on more than 1 occasion or self-reported provider-diagnosed diabetes with current medical therapy. In instances where cases were drawn from families, the individual with youngest age at onset was chosen. Controls ascertained from epidemiologically represented cohorts of individuals in Starr County, TX with individuals with known diagnosis of diabetes excluded. Controls are significantly younger on average.	Hanis, C. L. et al. Diabetes among Mexican Americans in Starr County, Texas. Am. J. Epidemiol. 118, 659–672 (1983); Below JE, et al. Genome-wide association and meta-analysis in populations from Starr County, Texas and Mexico City identify type 2 diabetes susceptibility loci and enrichment for eQTLs in top signals. Diabetologia 54, 2047–2055 (2011)
	Overall (number with clone)	704(5)	751(32)	
	Female (number with clone)	506(3)	449(15)	
	Mean age (SD)	55(8.9)	56(12)	
	Mean BMI (SD)	30(6.2)	32(6.4)	

SL (London Life Sciences Population Study (UK Indian Asians))					
Overall (number with clone)	538(24)	581(18)			
Female (number with clone)	85(4)	75(2)			
Mean age (SD)	68(9.2)	58(5.6)			
Mean BMI (SD)	27(3.5)	27(2.9)		A population-based cohort study of Indian Asians living in West London, UK with all 4 grandparents born on the Indian subcontinent. Prevalent T2D defined as previous physician diagnosis of diabetes on treatment, with onset of diabetes after the age of 15 years and without insulin use in the first year after diagnosis, or fasting plasma glucose ≥7.0 mmol/L. Controls defined as no previous history of diabetes, no anti-diabetic medication, and fasting plasma glucose <5.6 mmol/L.	Chambers, J.C. et al. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nat. Genet. 41, 1170-1172 (2009); Chambers, J.C. et al. Common genetic variation near melatonin receptor MTNR1B contributes to raised plasma glucose and increased risk of type 2 diabetes among Indian Asians and European Caucasians. Diabetes 58, 2703-2708 (2009); van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. Nature 492, 369-375 (2012)
SS (Singapore Indian Eye Study (Singapore Indians))					
Overall (number with clone)	585(15)	565(31)			
Female (number with clone)	288(5)	250(14)			
Mean age (SD)	56(10)	61(9.7)			
Mean BMI (SD)	25(4.8)	27(5.1)		Cases selected for HbA1c>6.5% or personal history of diabetes with age at diagnosis available. Preferentially selected cases with at least one first degree relative with T2D. Controls selected for HbA1c <6%, no personal history of diabetes, and not taking antidiabetes medication. Older controls preferentially selected.	Tan, S. et al. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. PLoS Genet. 7(4), e1001363 (2011)
UA (Ashkenazis)					
Overall (number with clone)	342(42)	506(49)	Survival, blood counts	Subjects in this cohort are of Ashkenazi Jewish origin, defined as having all four grandparents born in Northern or Eastern Europe, subjects with known or suspected Sephardic Jewish or non-Jewish ancestry excluded. T2D cases were selected from two separate DNA collections: 1. Genome-wide, affected-sibling-pair linkage study (Permutt et al. Diabetes 2001) or 2. Study to determine genetic risk for diabetic complications (Bleich et al. PLoS One 2013). Controls were selected for fasting blood glucose <7 mmol/L, no personal history of diabetes, and no anti-diabetic medications. Controls included elderly (>80 years) individuals who were part of the Longevity Genes Project	Atzmon, G. et al. Lipoprotein genotype and conserved pathway for exceptional longevity in humans. PLoS Biol. 4(4), e113 (2006); Atzmon, G. et al. Evolution in health and medicine Sackler colloquium: Genetic variation in human telomerase is associated with telomere length in Ashkenazi centenarians. Proc Natl Acad Sci U S A. 107 Suppl 1, 1710-1717 (2010); Permutt, M.A. et al. A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. Diabetes 50(3), 681-685 (2001); Bleich et al. Predicting diabetic nephropathy using a multifactorial genetic model. PLoS One 8(4), e58743 (2013)
UM (Metabolic Syndrome in Men Study)					
Overall (number with clone)	503(9)	487(24)			
Female (number with clone)	*	*			
Mean age (SD)	55(4.6)	60(6.7)			
Mean BMI (SD)	26(3.2)	31(5.1)		The METSIM Study includes 10,197 men, aged from 45 to 73 years, randomly selected from the population register of the town of Kuopio, Eastern Finland, and examined in 2005-2010. The aim of the study is to investigate genetic and non-genetic factors associated with the risk of type 2 diabetes (T2D), cardiovascular disease (CVD), and insulin resistance–related traits in a cross-sectional and longitudinal setting. Unrelated T2D cases with family history of diabetes were selected. Unrelated NGT controls were selected, prioritizing older individuals with no family history of diabetes.	Stancakova, A. et al. Changes in insulin sensitivity and insulin release in relation to glycemia and glucose tolerance in 6,414 Finnish men. Diabetes 58, 1212–1221 (2009)

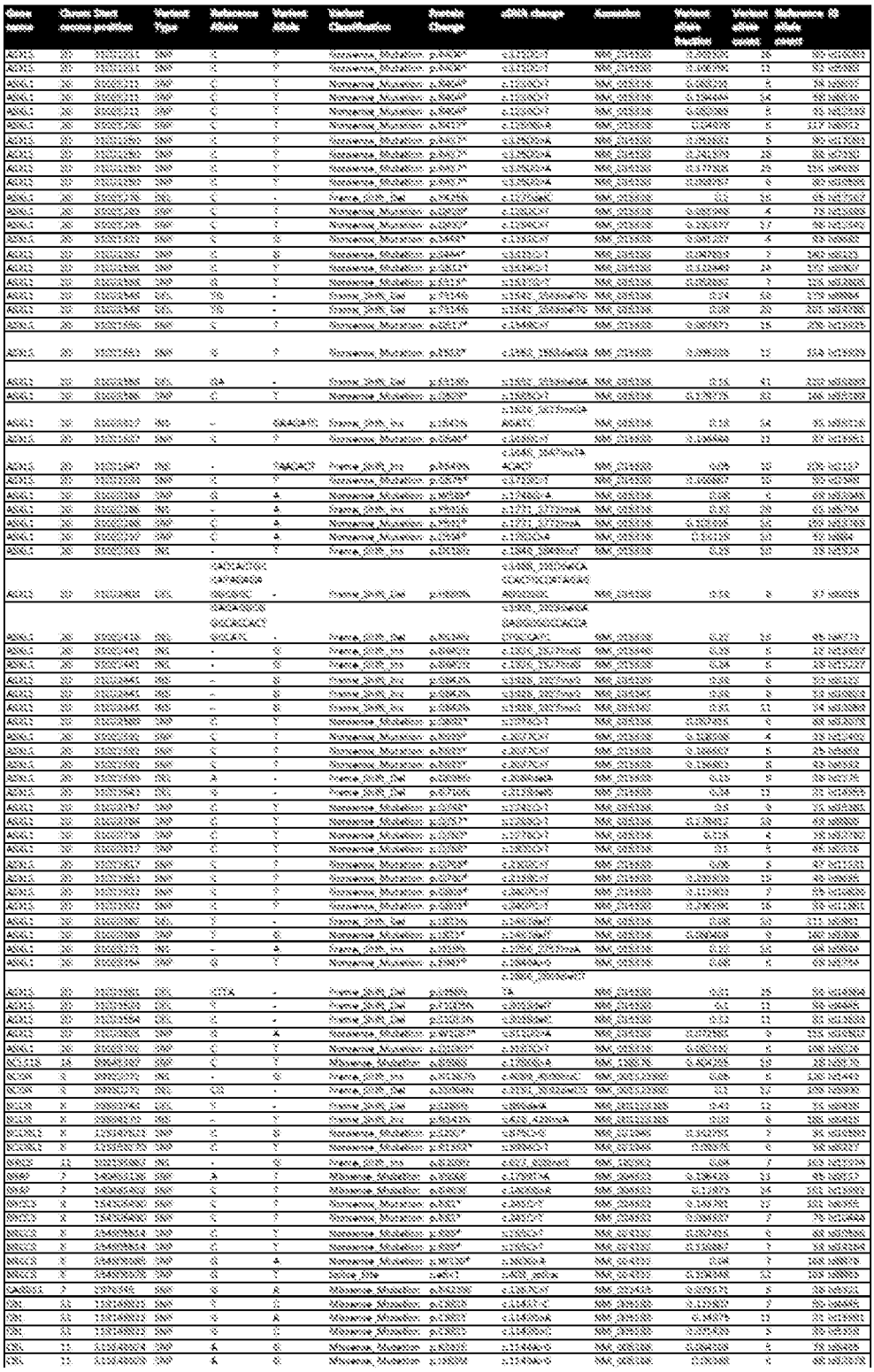
Supplementary Table S2
List of hematopoletic genes and variants queried

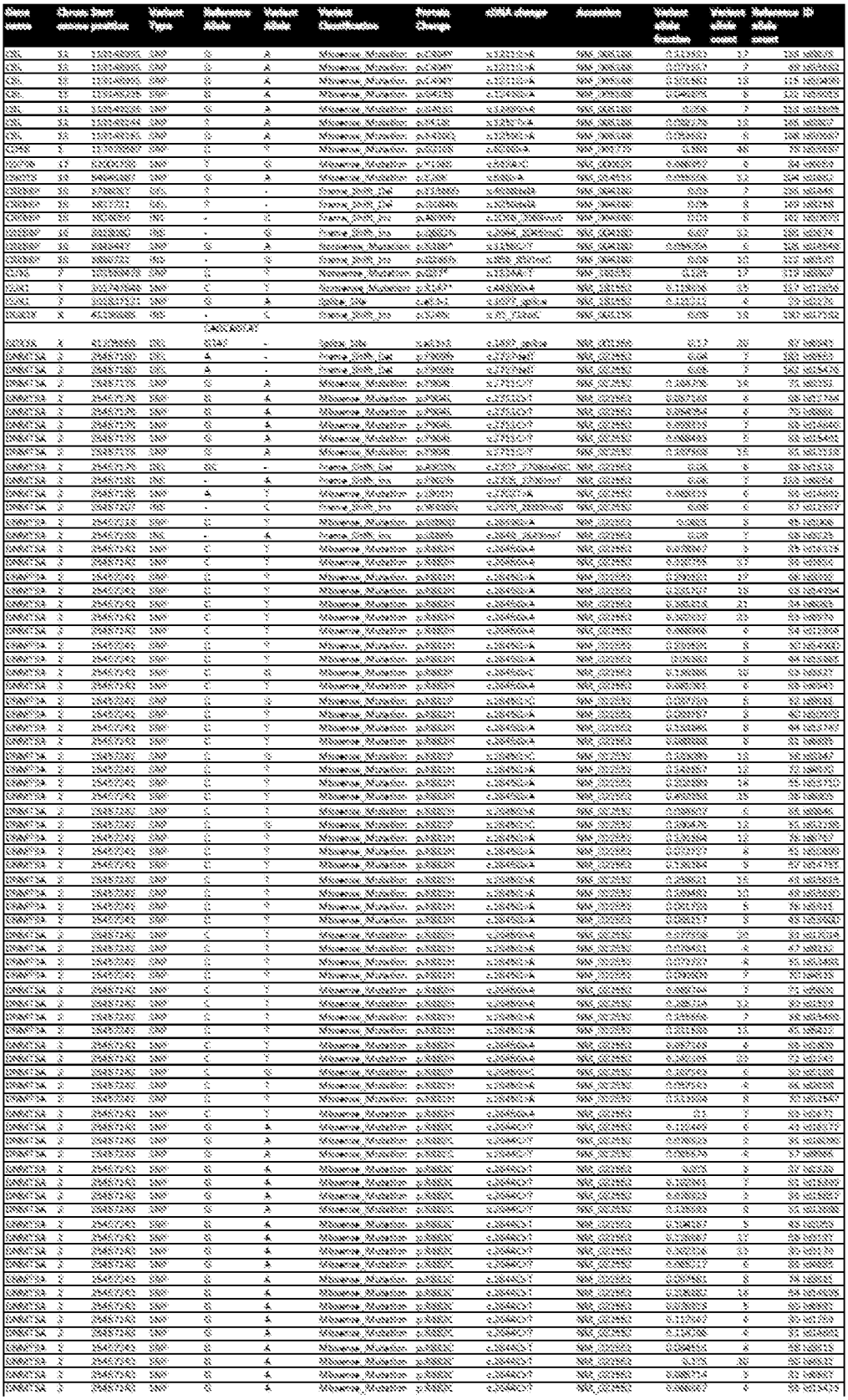
Gene name	Reported mutations used for variant calling	Accession	Number of variants found
ARID1A	Frameshift/nonsense, A339V, M872T	NM_006015	0
ASXL1	Frameshift/nonsense in exon 11-12	NM_015338	62
BCL10	Frameshift/nonsense/splice-site	NM_003921	0
BCL11B	Frameshift/nonsense/splice-site, A360T, D432Y, H445Y, R447H, G452K, H479Y, G595S, L617Q, G647R	NM_138576	1
BCL6	R40H, Y111H, P341H, S850R, R585W, Q678R	NM_001130845	0
BCOR	Frameshift/nonsense/splice-site	NM_001123385	4
BCORL1	Frameshift/nonsense/splice-site	NM_021946	2
BIRC3	Frameshift/nonsense/splice-site exon 2	NM_182962	1
BRAF	G464E, G464V, G466E, G466V, G469R, G469E, G469A, G469V, V471F, V472S, N581S, I582M, I592M, E586K, D594N, D594G, D594V, D594E, F595L, F595S, G596R, L597V, L597S, L597Q, L597R, A598V, V600M, V600L, V600R, V600E, V600K, V600A, V600G, V600D, K601E, K601N, K601*, W604R, W604G, S605G, S605F, S605N, G606E, G606A, G606V, H608R, H608L, G615R, S616P, S615F, L612S, L618W	NM_004333	2
BRCC3	Frameshift/nonsense/splice-site	NM_024332	6
BTG1	M1I, H2Y, P3R, Y5H, S11I, S33A, F25C, R27H, K39Q, L31F, Q36H, I37M, D38E, F40C, E46D, E48Q, A49P, P58L, E59D, I54V, I115V, E117D, N116S	NM_001731	0
BTG2	A45T, A45E	NM_006763	0
CARD11	E93D, G123S, G126D, T128M, F130I, R179W, K182N, M183L, K215M, D230N, L232LI, M240MGLNKM, K244T, S250P, S250P, L251P, L253P, V256*, D338E, T355*, D357V, Y361H, M365K, D387V, E387V, D401V, R418G, R423W, E432K, E626K	NM_032415	1
CBL	RING finger missense p.351-421	NM_005188	12
CBLB	RING finger missense p.373-412	NM_170662	0
CCND3	Frameshift/nonsense/splice-site, T211A, P212L, V215G	NM_001760	0
CD58	Frameshift/nonsense/splice-site, G210C, G210S	NM_001779	1
CD70	L50R, G66R, F186S	NM_001252	0
CD79A	del191-226, del179-226, del191-206	NM_001783	0
CD79B	V9A, D89G, Y92F, D193G, D194G, Y196C, Y197C, Y197D, Y197H, T207fs, Y200*, V212A, del195_197, del196_229, del193_229, del205_229	NM_000626	1
CDKN2A	Frameshift/nonsense/splice-site	NM_000077	0
CDKN2B	Frameshift/nonsense/splice-site	NM_004936	0
CEBPA	Frameshift/nonsense/splice-site	NM_004364	0
CHD2	H620L, F1146L, L1270F	NM_001271	0
CNOT3	Frameshift/nonsense, E20K, R57W, R57Q, E70K	NM_014516	1
CREBBP	Frameshift/nonsense/splice-site, D1435E, R1446L, R1446H, R1446C, Y1450C, P1476R, Y1482H, H1487Y, W1502C, Y1503D, Y1503H, Y1503F, S1680del	NM_004380	6
CRLF2	F232C	NM_022148	0
CSF1R	L301F, L301S, Y969C, Y969N, Y969F, Y969H, Y969D	NM_005211	0
CSF3R	T615A, T618I, truncating p.741-791	NM_000760	0
CTCF	Frameshift/nonsense, R377C, R377H, P378A, P378L	NM_006565	0
CUX1	Frameshift/nonsense	NM_181552	3
DDX3X	R276K, R376S, R376C, D506Y, R528C, R528H, R534S, R534H, P568L	NM_001356	2
DIS3	R780K, R780T, R780H, R780S, D488H, P480Q, D488N, S477R, R467Q, M662R, R689Q, R514K	NM_014953	0
DNMT3A	Frameshift/nonsense/splice-site, P307S, P307R, R326H, R326L, R326C, R326S, R366P, R366H, R366G, A368T, F414L, F414S, F414C, C497Y, Q527H, Q527P, Y533C, G543A, G543S, G543C, L547H, L547P, L547F, M548I, M548K, G550R, W581R, W581G, W581C, G646V, G646E, L653W, L653F, V657A, V657M, R659H, Y660C, R676W, R676Q, G685R, G685E, G685A, D686Y, D686G, G699R, G699S, G699D, P700S, P700R, P700Q, D702N, D702Y, V704M, V704G, I705F, I705T, I705S, C710S, S714C, N717S, N717I, P718L, R720H, R720G, Y724C, R729Q, R729W, R729G, F731L, F752del, F732S, F732L, F734L, F734C, Y735C, Y735N, Y735S, R736H, R736C, R736P, L737H, L737V, L737F, L737R, A741V, R749C, R749L, F751L, F752del, F752C, F752L, F752I, F752V, L754R, L754H, F755S, F755I, F755L, M761I, M761V, G762C, S770W, S770P, R771Q, F772I, F772V, L773R, E774K, E774D, D781G, R792H, G796D, G796V, N797Y, N797H, P799R, P799H, R803S, P804S, P804L, S828N, K829R, Q842E, P849L, D857N, W860R, F868S, G869S, G869V, M880V, S881R, S881I, R882H, R882P, R882C, R882G, Q886R, G890D, L901R, L901H, P904L, F909C, A910P	NM_022552	403
EBF1	Frameshift/nonsense, M1R, S7G, F54S, G143V, G171D, G196F, N237K, S238T, S238Y, R381S	NM_024007	0
EED	Frameshift/nonsense/splice-site, L240Q, I363M	NM_003797	0
EP300	Frameshift/nonsense/splice_site, VF1148_1149del, D1399N, D1399Y, P1452L, Y1467N, Y1467H, Y1467C, R1627W, A1629V	NM_001429	1
ETV6	Frameshift/nonsense/splice-site	NM_001987	1
EZH2	Frameshift/nonsense/splice-site, Q62R, N102S, F145S, F145C, F145Y, F145L, G159R, E164D, R202Q, N322S, E244K, R283Q, H292R, P488S, E457Q, R581H, T568I, K629E, Y641N, Y641H, Y641S, Y641D, Y641F, D659Y, D659G, V674M, A677G, A677V, R679C, R679H, R685C, R685H, A687V, N688I, N688K, H689Y, S690P, I708V, I708T, I708M, E720K, E740K	NM_001203247	2
EZR	Q245M	NM_003379	0
FAM46C	Frameshift/nonsense/splice-site	NM_017709	1
FAS	Frameshift/nonsense/splice-site	NM_000043	0
FBXO11	Frameshift/nonsense/splice-site	NM_001190274	0
FBXW7	Frameshift/nonsense/splice-site, E74A, G161V, F280L, R465H, R505C, G597E, R1165Q	NM_033632	1
FLT3	V579A, V592A, V592I, F594L, M737I, P(590-591)GD	NM_004119	1
FOXP1	Frameshift/nonsense/splice-site	NM_032682	1
FYN	L174R, R176C, Y531H	NM_002037	0
GATA1	Frameshift/nonsense/splice-site	NM_002049	0
GATA2	Frameshift/nonsense/splice-site, R293Q, N317H, A318T, A318V, A318G, G320D, L321F, L321P, L321V, Q328P, R330G, R361L, L359V, A372T, R384G, R384K	NM_001145661	0
GATA3	Frameshift/nonsense/splice-site ZNF domain, R276W, R276Q, N286T, L348V,	NM_001002295	0
GNA13	I34T, G57S, S62F, M68K, D234N, Y145F, L152F, E167D, Q169H, R264H, E273V, V223G, Y362G, L371F	NM_006572	0
GNAS	R201(844)S, R201(844)C, R201(844)H, R201(844)L, Q227(970)K, Q227(970)R, Q227(970)L, Q227(970)H, R374(1017)C	NM_016592	8
GNB1	K57N, K57M, K57E, K57T, I80T, I80N	NM_002074	22
HIST1H1B	S89N, S89R, G101D, G73A, K84N, A123D	NM_005322	0
HIST1H1C	P118L, P129A, K156R, K187R, K/G61/83N/A	NM_005319	1

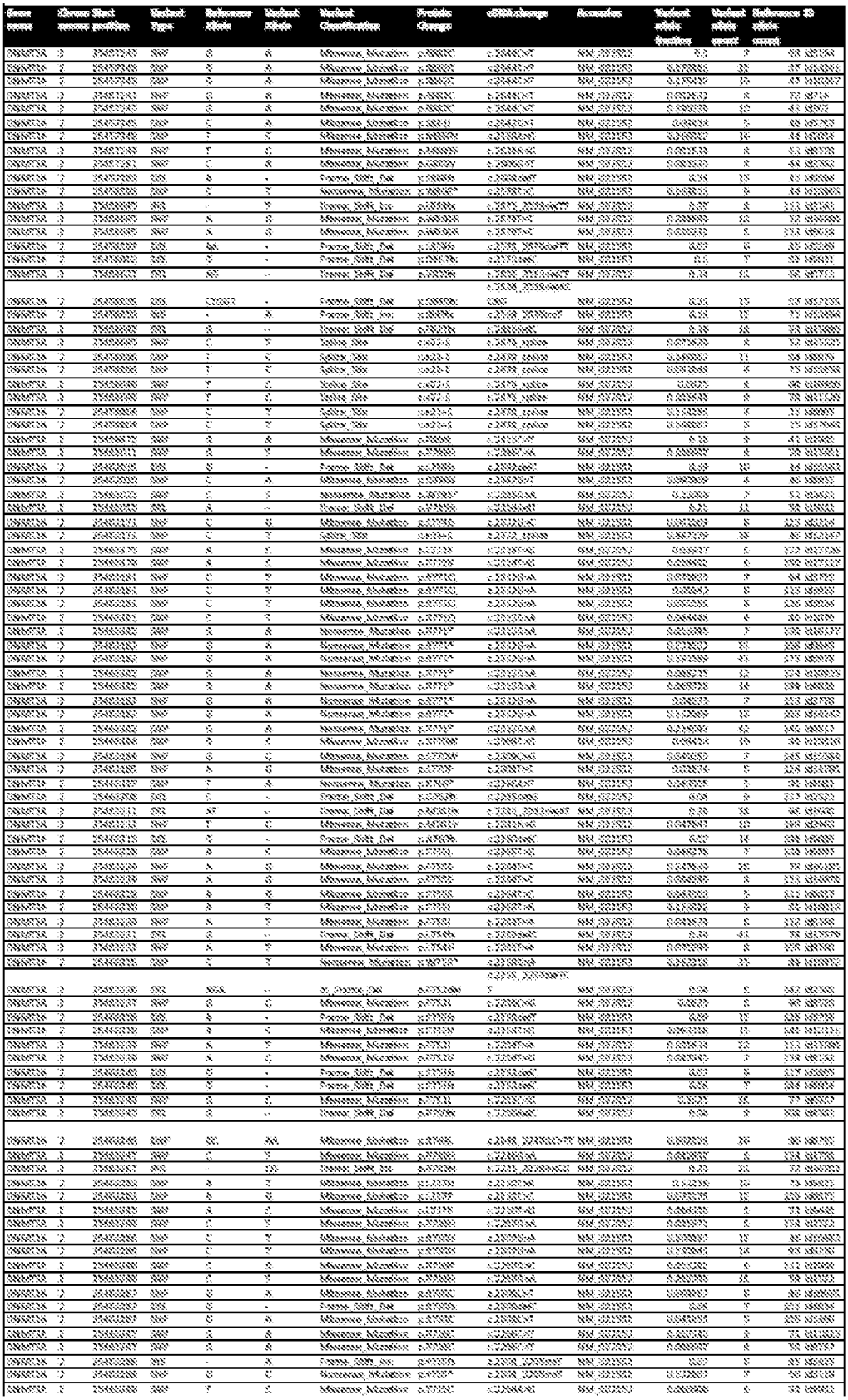
HIST1H1D	Frameshift/nonsense	NM_005320	0
HIST1H1E	R35ST, A187V, R196S, K202E, K205E	NM_005321	0
HIST1H3B	A48S, S87N, S87T	NM_003537	0
HLA-A	Frameshift/nonsense, G124D, A164T	NM_002116	0
ID3	Frameshift/nonsense/splice-site, S89R, V55G, P56S, L64F, S65R, I74V, L80R, M95R,	NM_002167	0
IDH1	R132C, R132G, R132H, R132L, R132P, R132V, V178I	NM_005896	0
IDH2	R140W, R140Q, R140L, R140G, R172W, R172G, R172K, R172T, R172M, R172N, R172S	NM_002168	3
IKBKB	K171E	NM_001556	0
IKZF1	Frameshift/nonsense	NM_006060	1
IKZF2	Frameshift/nonsense	NM_016260	0
IKZF3	Frameshift/nonsense	NM_012481	0
IL7R	exon 6 cysteine insertion	NM_002185	0
INTS12	M445I	NM_020395	0
IRF4	K59, S118T, I32V, L40W, Q60K, Q60H	NM_002460	0
IRF8	Frameshift/nonsense from c.877-426, S34T, S55A, T80A, K108E	NM_002163	0
JAK1	T478A, T478S, V623A, A634D, L653F, R724H, R724Q, T783M, L783F	NM_002227	0
JAK2	N533D, N533Y, N533G, H538R, K539E, P539L, I540T, I540V, V617F, R683S, R683G, del/ins537-539L, del/ins539-539L, del/ins540-540MK, del/ins540-544MK, del/ins541-548K, del542-543, del543-544, ins12E46-547	NM_004972	31
JAK3	M511T, M511I, A572V, A572T, A573V, R657Q, V715I, V715A	NM_000215	3
JARID2	Frameshift/nonsense/splice-site	NM_004973	1
KDM6A	Frameshift/nonsense/splice-site, del61R	NM_021140	3
KIT	ins503, V559A, V559D, V559G, V559I, V560D, V560A, V560G, V560E, del560, E561K, del579, P627L, P627T, R634W, K642E, K642Q, V654A, V654E, H697Y, H697D, E761D, K809R, D816H, D816Y, D816F, D816I, D816V, D816H, del551-559	NM_000222	1
KLHL6	F49L, L55P, T64S, T64I, L65V, L65P, Q81, L90V, R92T, R92W	NM_130446	3
KRAS	G12D, G12A, G12E, G12V, G13D, G13C, G13V, G13F, G13R, G13A, G13V, G13E, T58I, G60D, G60A, G60V, Q61K, Q61E, Q61P, Q61R, Q61L, Q61H, K117E, K117N, A146T, A146P, A146V	NM_033360	3
LEF1	Frameshift/nonsense	NM_016269	0
LRRK2	E155K, I1425	NM_198578	0
LTB	Frameshift/nonsense	NM_002342	0
LUC7L2	Frameshift/nonsense/splice-site	NM_016019	3
MALT1	K39Nf, V750S	NM_006785	0
MAP2K1	F53L, Q56P, K57T, K57N, K57E, I103N, C121S, N122D, F124Q	NM_002755	0
MAPSK14	H66DQ	NM_002754	0
MED12	E33K, L36P, G44S, A59P, R521H, L1224F	NM_005120	0
MEF2B	Frameshift/nonsense/splice-site, I5F, L67R, Y69C, Y69H, E77K, R81K, N83Y, D83N, D83H	NM_001145785	0
MLL	Frameshift/nonsense	NM_005933	0
MLL2	Frameshift/nonsense	NM_003482	4
MPL	S505G, S505N, S505C, L510F, del513, W515A, W515R, W515K, W515G, W515L, A519T, A519V, Y591D, W515-519RT	NM_005373	7
MXRA5	Frameshift/nonsense/splice-site	NM_015419	0
MYD88	V217F, S219C, M240T, S251N, P258, L273P	NM_001172567	2
NOTCH1	Frameshift/nonsense	NM_017617	2
NOTCH2	Frameshift/nonsense	NM_024408	2
NPM1	Frameshift p.W288fs (insertion at c.859_860ins[CCTG])	NM_002520	0
NRAS	G12S, G12R, G12C, G12N, G12P, G12Y, G13D, G12A, G12V, G12E, G13S, G13R, G13C, G13N, G13P, G13Y, G13D, G13A, G13V, G13E, G60E, G60R, Q61R, Q61L, Q61K, Q61P, Q61H, Q61Q	NM_002524	6
P2RY8	M52R, Y82C, R86C, C184Y, M188I, A263T, N264S, F265I, Y266M	NM_178129	0
PAPD5	Frameshift/nonsense	NM_001040284	0
PAX5	Frameshift/nonsense/splice-site, S80N, P80R,	NM_016734	0
PDS5B	Frameshift/nonsense/splice-site, R1293Q	NM_015032	0
PDS5Z	Frameshift/nonsense	NM_020239	1
PHF6	Frameshift/nonsense/splice-site, A40G, M125I, S246Y, F265L, R274Q, C297Y, H302Y, H329L	NM_001015877	1
PIK3CA	S45Y, G545S, G1007D, L1022F, M1043I, D1045N, H1047R	NM_006218	1
POT1	Frameshift/nonsense/splice-site before OB domain (c.374), M1L, Y36N, K90E, Q94R, Y123C, H266L, G272V, C591W	NM_015450	0
POU2AF1	P27A	NM_006235	0
POU2F2	T223A, T223S, T239S, R282H, T287I, G382F	NM_002698	0
PRDM1	Frameshift/nonsense/splice-site	NM_001198	3
PRPF40B	Frameshift/nonsense/splice-site, P15H, M5I, P405L, P562S	NM_001031698	1
PRPF8	M1907I	NM_006445	0
PTEN	Frameshift/nonsense/splice-site, D24G, R47S, F56V, L57W, H61R, R86N, Y68H, C71Y, F93C, Y88C, D92G, D92V, D92E, H93Y, H93D, H93Q, N94I, P95L, I101T, C105F, C105S, C107Y, L112V, H123Y, C124R, C124S, K125E, A126G, K128N, R130G, R130Q, R130L, G132D, I135V, I135K, C136R, C136F, K144Q, A151T, D153Y, G153R, Y155H, Y155C, R159N, R159S, R161H, R161Q, G165R, G165E, S170N, S170I, R173C, Y174D, Y177C, H196Y, R234W, G251C, D252Y, F271S, D326G	NM_000314	0
PTPN1	exon 3 frameshift/nonsense, Q5E	NM_002827	0
PTPN11	G60V, G60R, G60A, D61V, D61N, D61G, Y63C, E69K, E69G, E69D, E69Q, F71L, F71S, A72T, A72V, A72D, T73I, E76K, E76G, E76M, E76A, E76G, E139G, E139D, N308D, N308T, N308S, P491L, S502P, S502A, S502L, G503V, G503S, G503A, G503E, Q506P, T507A, T507K	NM_002834	3
RAD21	Frameshift/nonsense/splice-site, R85Q, R206G, Q476R	NM_006265	6
RBBP4	E380K	NM_005610	0
RHOA	C16R, G17V, G17E, T19I, D120Y	NM_001664	0
RIT1	Frameshift/nonsense, G7S, E81G, E81G, F82L, F82C, F82I, F82V, M90I, R122L	NM_006912	1
RPL10	P98S, Q133P	NM_006013	0
RPL5	Frameshift/nonsense/splice-site, Q65R	NM_000969	0
RPS15	Q132S, K133N, K133G, S138A, K145N	NM_001018	1
RPS2	R209G	NM_002952	0
RUNX1	Frameshift/nonsense/splice-site, S73F, H78Q, H78L, R80C, R80P, R80H, L85Q, P86L, P86H, D114G, D128Y, L156P, R135G, R135K, R156S, R156Q, R162S, A165V, R174Q, R177L, R177Q, A214T, D171G, D171V, D171N, R205W, R206Q	NM_001001890	0

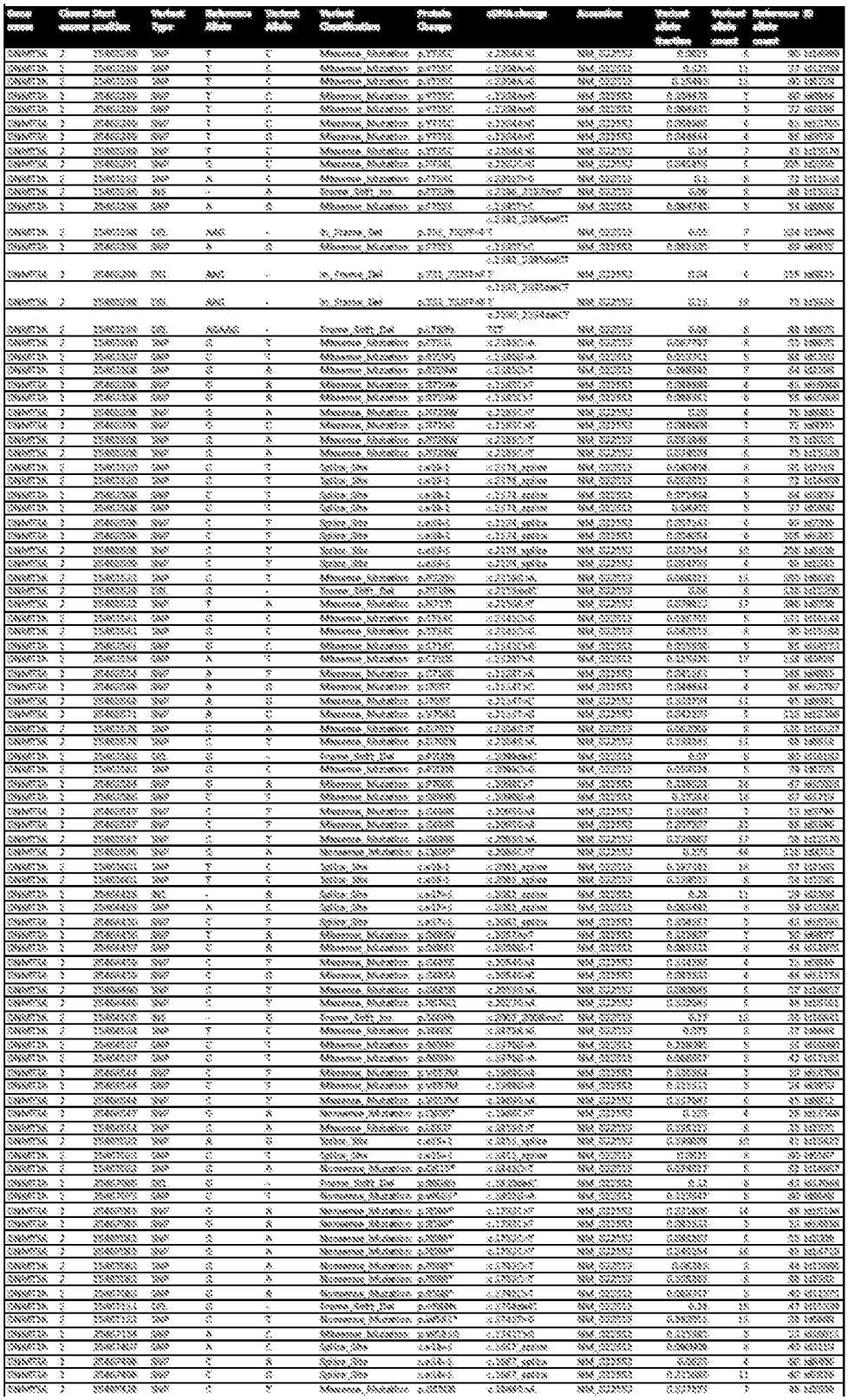
SETBP1	D868N, D868T, S869N, G870S, I871T, D880N, D880G	NM_015559	0
SETD2	Frameshift/nonsense, V1190M	NM_014159	4
SETDB1	Frameshift/nonsense, K715E	NM_001145415	1
SF1	Frameshift/nonsense/splice-site, T454M, T474A, Y476C, A508G	NM_004630	2
SF3A1	Frameshift/nonsense/splice-site, A57S, M117I, K166T	NM_005877	1
SF3B1	G347V, R387W, R527Q, E592K, E622D, Y623C, R625L, R625C, H662Q, H662D, K666N, K666T, K666E, K666R, K700E, V701F, A708T, G740R, G742D, A744P, D781G, E783K	NM_012433	27
SFRS2	Y44H, P95H, P95L, P95T, P95R, P95A, P107H, P95fs	NM_003016	11
SGK1	Frameshift/nonsense/splice-site	NM_001143676	0
SMC1A	K190T, R586W, M689V, R807H, R1090H, R1090C	NM_006306	1
SMC3	Frameshift/nonsense, R155I, Q367E, D362V, K571R, R661P, G662C	NM_005445	1
SOCS1	Frameshift/nonsense/splice-site, R48W	NM_003745	0
SPRY4	S113N, G117R	NM_001127496	0
STAG1	Frameshift/nonsense/splice-site, H1085Y	NM_005862	1
STAG2	Frameshift/nonsense/splice-site	NM_006603	1
STAT3	M206K, G618R, Y640F, N642H, N647I, D661N, D661H, D661Y, D661V	NM_139276	4
STAT5A	N642H	NM_003152	0
STAT5B	N642H, Y665F	NM_012448	0
STAT6	N417Y, N417S, D419H, D419G, D419Y, N421K, N430T, N480S	NM_001178081	0
SUZ12	Frameshift/nonsense	NM_015355	1
SWAP70	Frameshift/nonsense/splice-site	NM_015055	0
TBL1XR1	Frameshift/nonsense/splice-site	NM_024665	1
TCF3	N551K, V557E, V557G, D561E, D561V, D561N, M572K	NM_003200	0
TET1	Frameshift/nonsense/splice-site, V128F, H1297Y, R1656Q, V2120M,	NM_030625	1
TET2	Frameshift/nonsense/splice-site, S282F, N312S, L346P, S460F, D666G, P941S, C1135Y	NM_001127208	72
TMEM30A	Frameshift/nonsense	NM_018247	0
TNF	L47F, N52Y, P58S	NM_000594	0
TNFAIP3	Frameshift/nonsense, D117V, M476I, P574L	NM_006290	3
TNFRSF14	Frameshift/nonsense/splice-site, Y26N, Y26D, C42P, C42W, A102P	NM_003820	2
TP53	Frameshift/nonsense/splice-site, S46F, G105C, G105R, G105D, G108S, G108C, R110L, R110C, T118A, T118R, T118I, L130V, L130F, K132Q, K132E, K132W, K132R, K132M, K132N, C135W, C135S, C135F, C135G, Q136K, Q136E, Q136P, Q136R, Q136L, Q136H, A138P, A138V, A138A, A138T, T140I, C141R, C141G, C141A, C141Y, C141S, C141F, C141W, V143M, V143A, V143E, L145Q, L145R, P151T, P151A, P151S, P151H, P152S, P152R, P152L, T155P, R158H, R158L, A159V, A159P, A159S, A159D, A161T, A161D, Y163N, Y163H, Y163D, Y163S, Y163C, K164E, K164M, K164N, K164P, H168Y, H168P, H168R, H168L, H168G, M169I, M169T, M169V, T170M, E171K, E171Q, E171G, E171A, E171V, E171D	NM_001126112	33
TRAF3	Frameshift/nonsense	NM_145725	0
TYW1	R555Q, E601K	NM_018264	0
U2AF1	D14G, S34F, S34Y, R35L, R156H, R156Q, Q157R, Q157P	NM_006758	5
U2AF2	R18W, Q143L, M144I, L187V, Q190L	NM_007279	0
UBR5	Frameshift/nonsense/splice-site exon 58	NM_015902	0
WT1	Frameshift/nonsense/splice-site	NM_024426	0
XBP1	L167I, P326R	NM_001079539	0
XPO1	E571A, E571K	NM_003400	0
ZNF471	G483, P486, R425	NM_020813	0
ZRSR2	Frameshift/nonsense, R126P, E133G, C181F, H191Y, I202N, F239V, F239Y, N261Y, C280R, C302R, C326R, H330R, N382K	NM_005089	3
Total			805

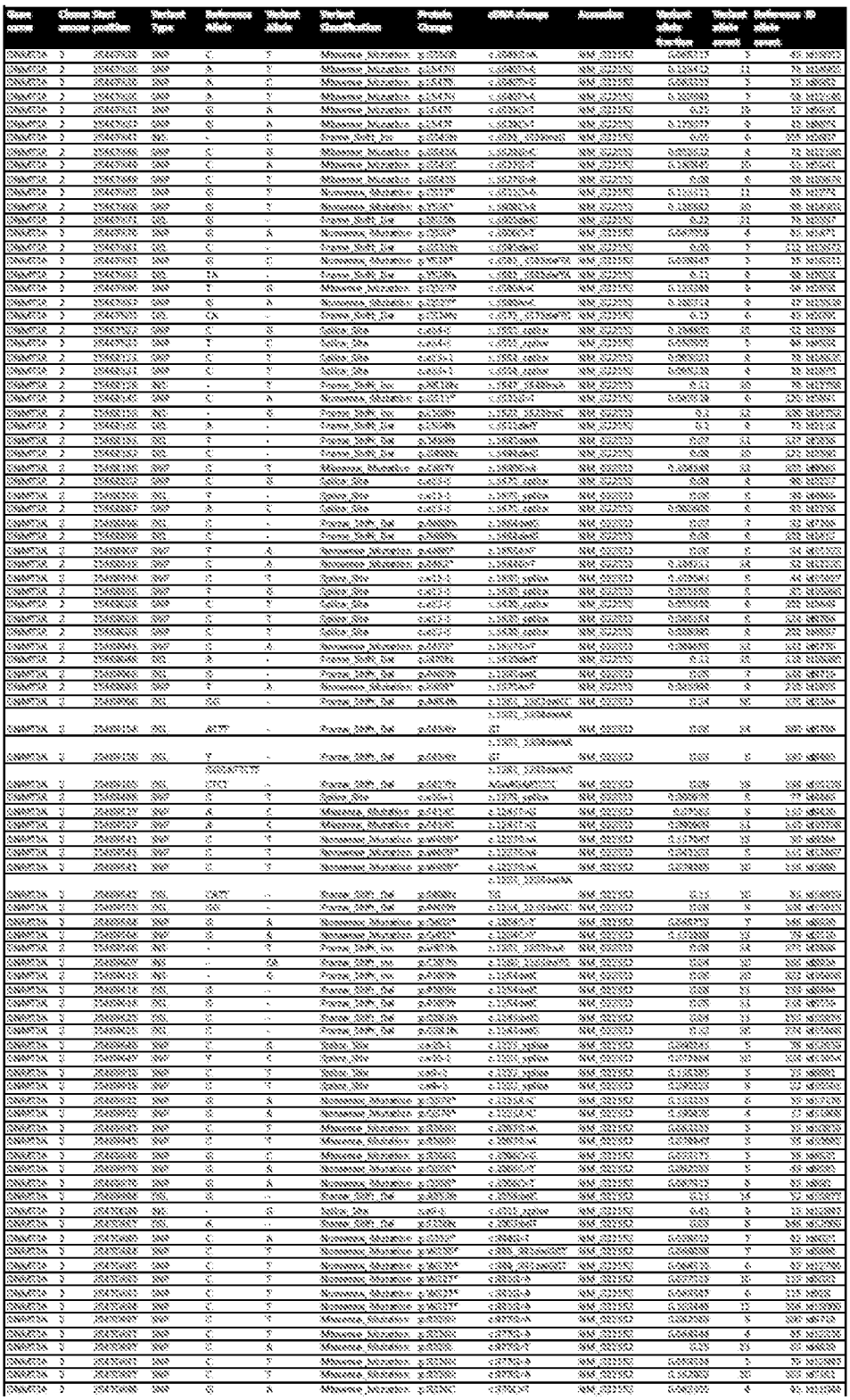
Supplementary Table S3
Called somatic variants in 160 hematopoeitic genes

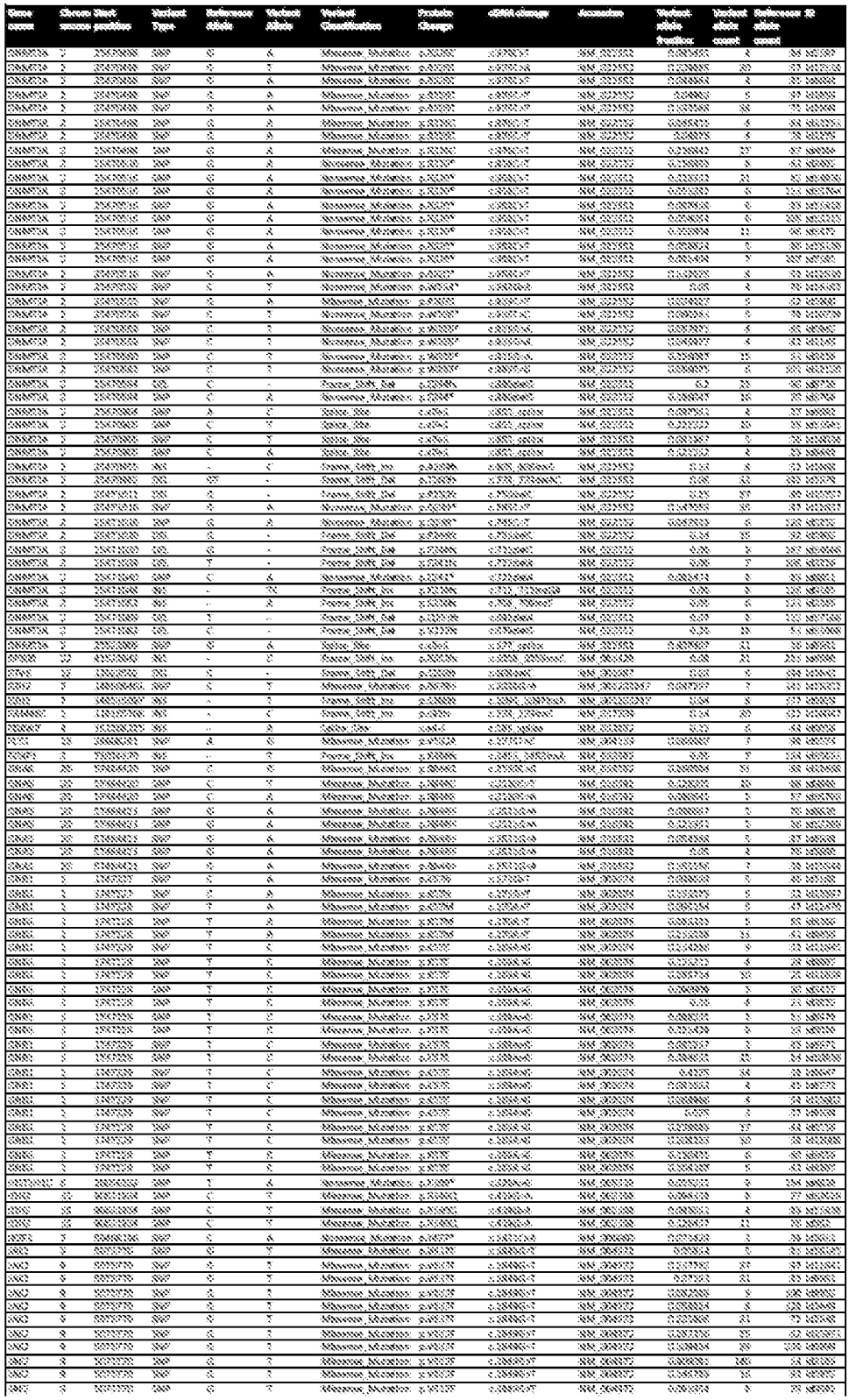


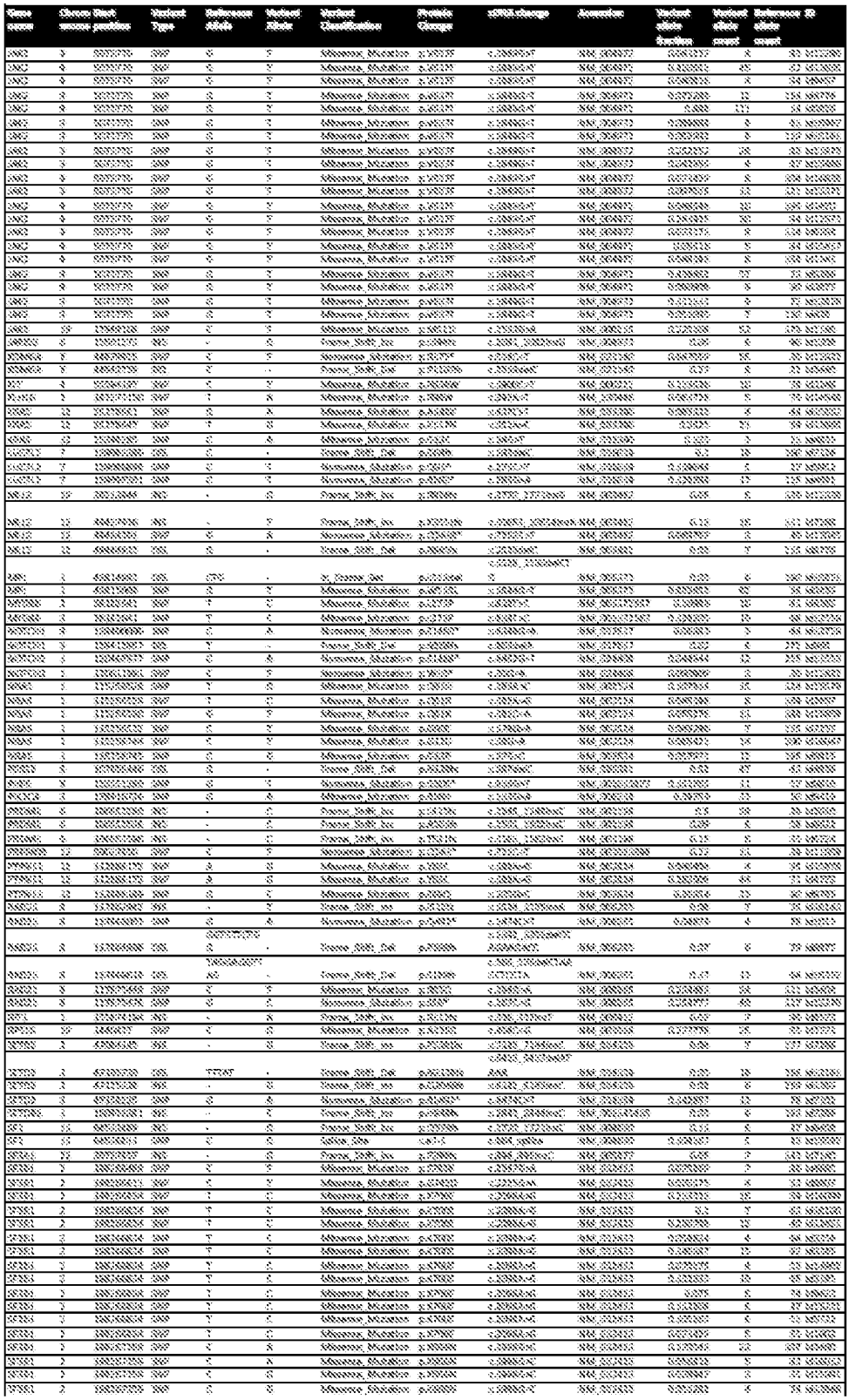


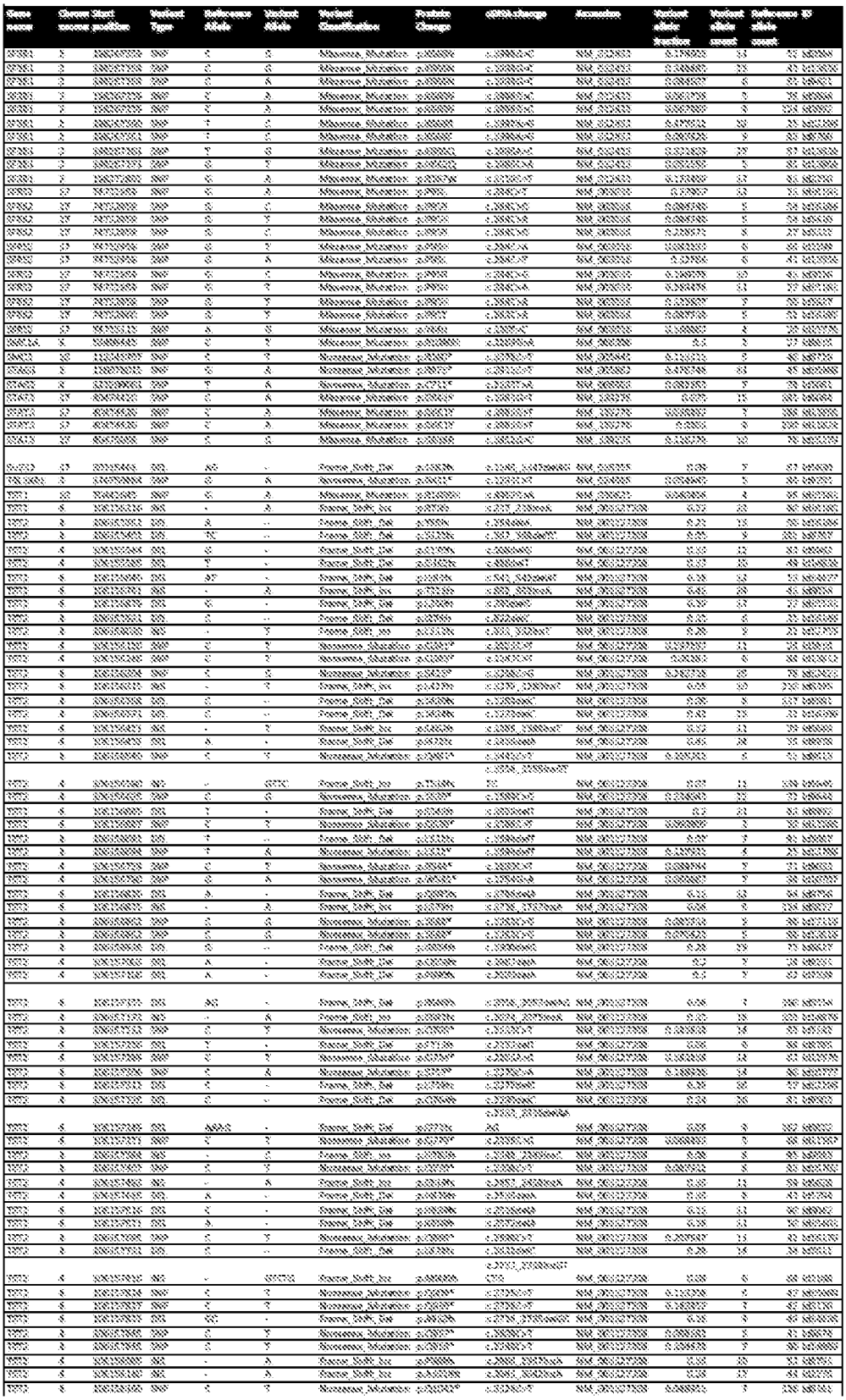


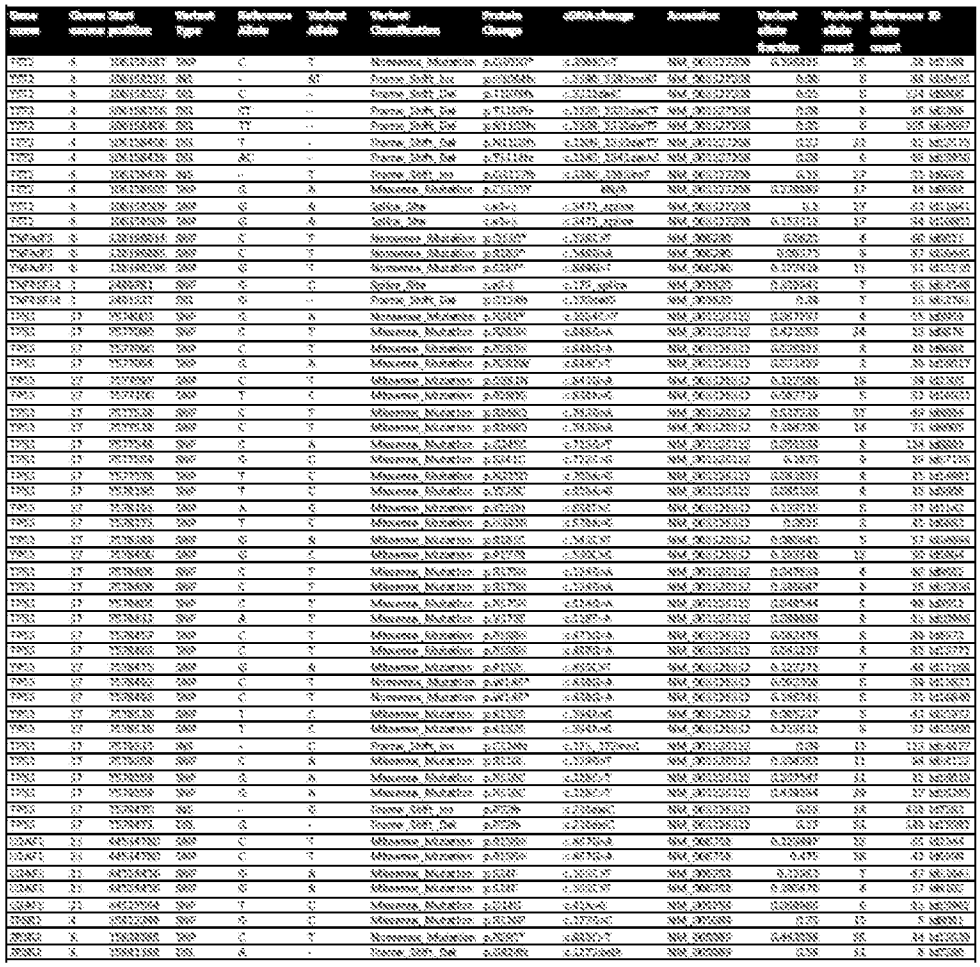












Supplementary Table S4
List of non-hematopoietic genes and variants queried

Gene name	Reported mutations used for variant calling	Accession	Number of variants found
ACVR1B	[illegible]	[illegible]	[illegible]
AKT1	[illegible]	[illegible]	[illegible]
APC	[illegible]	[illegible]	[illegible]
APOL2	[illegible]	[illegible]	[illegible]
ARHGAP35	[illegible]	[illegible]	[illegible]
ARID2	[illegible]	[illegible]	[illegible]
ATP5B	[illegible]	[illegible]	[illegible]
B2M	[illegible]	[illegible]	[illegible]
BAP1	[illegible]	[illegible]	[illegible]
CASP8	[illegible]	[illegible]	[illegible]
CDH1	[illegible]	[illegible]	[illegible]
CDKN1B	[illegible]	[illegible]	[illegible]
CDKN2A	[illegible]	[illegible]	[illegible]
CHD1	[illegible]	[illegible]	[illegible]
CHD4	[illegible]	[illegible]	[illegible]
CTNNB1	[illegible]	[illegible]	[illegible]
EGFR	[illegible]	[illegible]	[illegible]
ERBB2	[illegible]	[illegible]	[illegible]
ERBB3	[illegible]	[illegible]	[illegible]
EZH1	[illegible]	[illegible]	[illegible]
FGFR3	[illegible]	[illegible]	[illegible]
GPS2	[illegible]	[illegible]	[illegible]
HRAS	[illegible]	[illegible]	[illegible]
KDM5C	[illegible]	[illegible]	[illegible]
KEAP1	[illegible]	[illegible]	[illegible]
MAP2K4	[illegible]	[illegible]	[illegible]
MAP3K1	[illegible]	[illegible]	[illegible]
MTOR	[illegible]	[illegible]	[illegible]
NFE2L2	[illegible]	[illegible]	[illegible]
PBRM1	[illegible]	[illegible]	[illegible]
PIK3CA	[illegible]	[illegible]	[illegible]
PIK3R1	[illegible]	[illegible]	[illegible]
RASA1	[illegible]	[illegible]	[illegible]
RB1	[illegible]	[illegible]	[illegible]
SEPT12	[illegible]	[illegible]	[illegible]
SMAD4	[illegible]	[illegible]	[illegible]
SMARCA4	[illegible]	[illegible]	[illegible]
SPOP	[illegible]	[illegible]	[illegible]
STK11	[illegible]	[illegible]	[illegible]
VHL	[illegible]	[illegible]	[illegible]
Total			[illegible]

Supplementary Table S5
Called variants in non-hematopoeitic genes

Gene name	Chromosome	Start position	Variant Type	Reference Allele	Variant Allele	Variant Classification	Protein Change	cDNA Change	Accession	Variant allele fraction	Variant allele count	Reference allele count	ID
APC	5	112175212	DEL	AAAAG	-	Frame_Shift_Del	p.I1307fs	c.3923_38	NM_001127511	0.28	28	74	id3447
CASP8	2	202149654	DEL	C	-	Frame_Shift_Del	p.N306fs	c.918delC	NM_033355	0.51	94	89	id8185
CASP8	2	202149654	DEL	C	-	Frame_Shift_Del	p.N306fs	c.918delC	NM_033355	0.47	95	109	id7570
CASP8	2	202149901	SNP	C	T	Nonsense_Mutation	p.Q389*	c.1165C>T	NM_033355	0.387755	19	30	id16144
CASP8	2	202149901	SNP	C	T	Nonsense_Mutation	p.Q389*	c.1165C>T	NM_033355	0.385714	27	43	id4304
GPS2	17	7217689	SNP	G	G	Nonsense_Mutation	p.Q80*	c.238C>T	NM_004489	0.098361	12	110	id18023
MAP3K1	5	56152535	SNP	G	G	Nonsense_Mutation	p.W137*	c.5918>A	NM_005921	0.422414	49	67	id6400
NFE2L2	2	178098808	SNP	G	A	Missense_Mutation	p.A124V	c.371C>T	NM_006164	0.089787	3	40	id14588
PIK3R1	5	67509168	SNP	C	C	Nonsense_Mutation	p.R356*	c.1156C>T	NM_181523	0.107143	3	25	id11720
VHL	3	10191614	SNP	C	T	Nonsense_Mutation	p.Q28*	c.607C>T	NM_000551	0.333333	13	26	id13948

Supplementary Table S6
Logistic regression for factors associated with clonality

Logistic regression was performed using the variables age (as a continuous variable), ancestry, sex, T2D, and age/sex interaction. Other interaction terms were modeled, but none were significant. Proportion of variance explained is derived by analysis of variance (ANOVA) for the generalized linear model, and is equal to deviance for the variable divided by residual deviance for the null model.

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.07	1.08(1.07-1.09)	<0.001	0.06
European (referent)				0.003
African-American	0.11	1.12(0.9-1.39)	0.3	
East-Asian	0.02	1.02(0.79-1.31)	0.86	
Hispanic	-0.39	0.68(0.55-0.83)	<0.001	
South Asian	-0.2	0.82(0.63-1.05)	0.125	
No T2D (referent)				0.002
Has T2D	0.26	1.3(1.12-1.51)	<0.001	
Male (referent)				0.001
Female	0.83		0.066	
Age:Female	-0.02	0.98(0.97-1)	0.023	0.001

Supplementary Table S7
Logistic regression for factors associated with clonality by ancestry group

Logistic regression was performed using the variables age (as a continuous variable), sex, T2D, and age/sex interaction for each ancestry group. Proportion of variance explained is derived by analysis of variance (ANOVA) for the generalized linear model, and is equal to deviance for the variable divided by residual deviance for the null model.

African-American

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.07	1.08(1.05-1.1)	<0.001	0.08
Female	0.37	1.45(0.21-10.2)	0.7	0.002
Has T2D	0.24	1.27(0.9-1.79)	0.18	0.002
Age:Sex	-0.01	0.99(0.96-1.02)	0.52	0.0003

East Asian

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.1	1.11(1.06-1.16)	<0.001	0.03
Female	1.3	3.65(0.08-172)	0.51	0.003
Has T2D	-0.12	0.89(0.57-1.38)	0.61	0.0003
Age:Sex	-0.03	0.97(0.92-1.03)	0.385	0.0006

European

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.07	1.07(1.06-1.09)	<0.001	0.05
Female	1.73	5.62(1.26-25)	0.023	0.001
Has T2D	0.31	1.36(1.04-1.79)	0.026	0.003
Age:Sex	-0.03	0.97(0.95-0.99)	0.013	0.0033

Hispanic

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.08	1.08(1.06-1.11)	<0.001	0.08
Female	0.69	2(0.27-15.6)	0.5	0.00001
Has T2D	0.22	1.24(0.9-1.72)	0.18	0.001
Age:Sex	-0.01	0.99(0.96-1.02)	0.485	0.0003

South Asian

	Beta coefficient	OR(95 CI)	p-value	Variance explained
Age	0.08	1.08(1.05-1.11)	<0.001	0.07
Female	-0.82	0.44(0.01-12.3)	0.64	0.0001
Has T2D	0.49	1.63(1.04-2.56)	0.033	0.007
Age:Sex	0.01	1.01(0.96-1.07)	0.677	0.0002

Supplementary Table S8
Logistic regression for factors associated with RDW≥14.5%

Individuals were from Jackson Heart Study or UA control cohort.

Covariate	OR (95% CI)	p-value
Age 60-69	1.3(0.9-1.8)	0.17
Age 70-79	1.8(1.2-2.7)	0.002
Age 80-89	2.7(1.3-5.2)	0.005
Age >90	1.7(0.9-3.1)	0.09
Female	1.1(0.8-1.5)	0.76
Has T2D	0.9(0.7-1.2)	0.45
DNMT3a mutation	1.9(1-3.6)	0.048
WBC	1.1(1.0-1.1)	0.092
Hemoglobin	0.7(0.7-0.8)	<0.001
Platelet count	1.0(1.0-1.0)	0.65

Supplementary Table S9
Association of cytopenias with clonality

Individuals were classified as having cytopenia as defined in Methods. Statistical comparisons were performed using Fisher's exact test. WBC - white blood cell count, Hgb – hemoglobin, Plt – platelet count. Individuals were from Jackson Heart Study, Longevity Genes Project, Botnia, Helsinki-sib, and Malmo-sib. Of the 5 individuals with multiple cytopenias and clonal mutations, 2 had 2 detectable mutations, suggesting that these might be undiagnosed cases of MDS.

	Clone	No Clone	
Low WBC	6	59	65
Normal WBC	119	2628	2747
	125	2687	

OR 2.2(0.8-5.2), p=0.066

	Clone	No Clone	
Low Hgb	30	616	646
Normal Hgb	109	2350	2459
	139	2966	

OR 1.0(0.7-1.6), p=0.83

	Clone	No Clone	
Low Plt	4	107	111
Normal Plt	132	2789	2921
	136	2896	

OR 0.8(0.2-2.1), p=0.82

	Clone	No Clone	
Any cytopenia	35	745	780
No cytopenia	104	2224	2328
	139	2969	

OR 1.0(0.6-1.5), p=1

	Clone	No Clone	
2+ cytopenias	5	37	42
1 or 0 cytopenias	134	2932	3066
	139	2969	

OR 3.0(0.9-7.7), p=0.037

Lower limit of normal for blood counts were defined as follows:
White blood cell count:
African-Americans $3.0x10^9$/L, white $4.0x10^9$/L

Hemoglobin:
African-American women 11.2g/dL,
African-American men 12.5 g/dL, white women 11.9 g/dL, white men 13.4 g/dL

Platelet count:
$150x10^9$L

Supplementary Table S10
Association of clonality with known versus unknown causes of anemia

During clinical evaluation, most patients with anemia can be found to have an attributable cause. Using subjects from the Jackson Heart Study, Applicants assessed whether the anemia was attributable to iron deficiency, anemia of chronic inflammation, or renal insuffuciency. Individuals with clonality were less likely to have anemia attributable to one of these causes.

Iron deficiency anemia/microcytic anemia
-mean corpuscular volume<80 fL
-ferritin<20 ng/mL
-ferritin 20-100 ng/mL WITH EITHER total iron binding capacity >370 mcg/dL OR serum iron <50 mcg/dL OR iron saturation< 20%

Anemia of chronic disease
-serum iron <65 mcg/dL WITH total iron binding capacity <250 mcg/dL
-ferritin >350 ng/mL for males
-ferritin >300 ng/mL for females

Renal insufficiency
-estimated glomerular filtration rate <30 mL/min/1.73m^2

	Clone	No Clone	
Known cause	2	360	362
Unknown cause	7	151	158
	9	511	

OR 0.1(0-0.6), p=0.004

Supplementary Table S11
Details on subjects that developed hematologic malignancies

Incident Cases

Age at sampling	Diagnosis	Cohort	Adjudicated	Latency (years)	Mutation on WES (VAF)	Mutations on RHP (VAF)	WBC	HGB	PLT	Death	Cause of death
77	CANCER OF SPLEEN [a]	AJ	No	6	NRAS p.V617F (0.23)	NA	7.8	11	247	Yes	CARCINOMA UNSPECIFIED SITE
66	LEUKEMIA (prior NHL) [b]	AJ	No	7	ASXL1 p.D616fs (0.23)	ASXL1 p.D616fs (0.18)	3.5	12.9	189	No	
57	LYMPHOMA [c]	AJ	No	2	DNMT3A p.R882H (0.29)	NA	14.3 (51.3% lymphocytes) [d]	13	248	No	
85	DLBCL, large intestine [e]	MEC	Yes	5	TET2 p.C1135Y (0.23);ASXL1 p.I959fs (0.27)	TET2 p.C1135Y (0.35);TET2 p.G1192V (0.30);ASXL1 p.I919fs (0.28)				No	
82	MDS-RAEB	MEC	Yes	7	ASXL1 p.T514fs (0.24)	ASXL1 p.T514fs (0.30);TET2 p.S1196fs (0.15)				Yes	Myeloid leukemia
58	LYMPHOMA	AJ	No	4	None	NA	6.3	14.4	220	Yes	COMPLICATIONS DUE TO LYMPHOMA
48	LYMPHOMA	AJ	No	1	None	NA	4.8	14.3	225	Yes	RENAL FAILURE
57	LYMPHOMA	AJ	No	7	None	None	7.3	15.7	230	Yes	HYPOXIC RESPIRATORY FAILURE
62	LEUKEMIA	AJ	No	9	None	NA	2.8	15.8	168	No	
51	BLOOD	AJ	No	8	None	NA	5.7	13.7	315	No	
67	LEUKEMIA	AJ	No	9	None	None				No	
64	MULTIPLE MYELOMA	AJ	No	9	None	NA	4.4	11.4	258	No	
59	LYMPHOMA	AJ	No	8	None	NA		13.2	195	No	
61	ACUTE MYELOID LEUKEMIA	AJ	No	10	None	None		11.7	156	Yes	ACUTE MYELOID LEUKEMIA
51	LEUKEMIA	AJ	No	10	None	NA	4.8	13	334	No	
68	Multiple myeloma	MEC	Yes	8	None	NA				No	

Prevalent Cases

Age at sampling	Diagnosis	Cohort	Adjudicated	Time (years prior)	Mutation on WES (VAF)	Mutations on RHP (VAF)	WBC	HGB	PLT	Death	Cause of death
66	LEUKEMIA	AJ	No	7	None	NA	3.5	12.9	189	No	
48	LYMPHOMA	AJ	No	1	None	NA	4.8	14.3	225	Yes	RENAL FAILURE
76	LEUKEMIA	AJ	No	17	None	NA		11.4	252	No	
63	NHL oropharynx	MEC	Yes	26	None	NA					
64	NHL (subsequent LEUKEMIA, see above)	AJ	No	7	ASXL1 p.D616fs (0.23)	ASXL1 p.D616fs (0.18)	3.5	12.9	189	No	

NHL=Non-Hodgkin's lymphoma, DLBCL=diffuse large B-cell lymphoma, MDS-RAEB=myelodysplastic syndrome, refractory anemia with excess blasts

AJ= Jackson Heart Study, MEC= Multi-ethnic cohort, Hispanics in Los Angeles

WES=whole exome sequencing, RHP=Rapid Heme Panel targeted re-sequencing, VAF=variant allele fraction

WBC=white blood cell count (x 10^9 cells/L), HGB=hemoglobin (g/dL), PLT=platelet count (x10^9cells/L)

NA=Not available

a This is likely splenomegaly secondary to a JAK2-mutated myeloproliferative neoplasm

b This likely represents a therapy related AML/MDS as ASXL1 mutations have never been described in lymphoid malignancies

c DNMT3A mutations have been found in peripheral T-cell lymphomas and early T-progenitor ALL

d This is the second highest absolute lymphocyte count in the cohort of 2,426 subjects who had a WBC differential. The subject with the highest absolute lymphocyte count also had a DNMT3A mutation.

e TET2 has been reported to be mutated in 8-13% of DLBCL, and the mutations are reported to be found in hematopoietic stem cells

Supplementary Table S12
Risks associated with developing incident coronary heart disease and ischemic stroke using traditional risk factors and clonality.

Hazard ratios were estimated using competing risks regression with death as the competing risk. P-values are derived from the Fine-Gray test. Individuals with prior coronary heart disease (CHD) were excluded for CHD analysis, and individuals with prior ischemic stroke were excluded for stroke analysis. Models shown in A or B are from the same population and differ by having covariates either removed or added. A) Coronary heart disease, B) ischemic stroke. Individuals were from Jackson Heart Study and FUSION.

A

n=3,286	Model 1		Model 2		Model 3	
Covariate	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
log(Age)	5.3(2.1-12.8)	<0.001	4.7(1.9-12)	<0.001	4.6(1.9-11)	<0.001
Has T2D	3.3(2.1-5.3)	<0.001	3.4(2.1-5.4)	<0.001	3.5(2.2-5.5)	<0.001
Female	0.7(0.5-1.1)	0.11	0.7(0.5-1.1)	0.13	0.7(0.5-1.1)	0.13
HDL<35 mg/dL	1.1(0.6-2.1)	0.77	1.1(0.6-2.1)	0.81	1.1(0.6-2.1)	0.78
HDL>60 mg/dL	0.7(0.4-1.2)	0.18	0.7(0.4-1.3)	0.25	0.7(0.4-1.3)	0.26
TC >240 mg/dL	2.1(1.3-3.2)	<0.001	2(1.3-3.1)	<0.001	2(1.3-3.1)	<0.001
Former or current smoker	1.6(1.1-2.5)	0.024	1.6(1-2.4)	0.035	1.6(1.1-2.5)	0.02
Hypertension stage II-IV	1.5(1-2.5)	0.06	1.4(0.9-2.3)	0.15	1.4(0.9-2.3)	0.15
BMI>25	1.2(0.6-2.5)	0.55	1.4(0.6-2.8)	0.43	1.3(0.6-2.8)	0.42
Clone present			2.3(1.1-4.8)	0.026		
VAF<0.10					1.4(0.5-4)	0.55
VAF>0.10					4.4(1.9-10.5)	<0.001
Pseudo Log-likelihood		-661		-655		-656
Pseudo likelihood ratio test		103 on 9 df		109 on 10 df		113 on 11 df

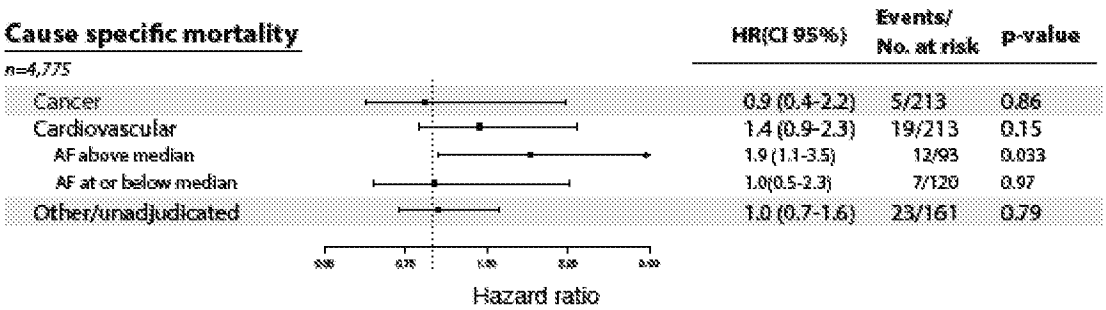
B

n=1,420	Model 1		Model 2		Model 3	
Covariate	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
log(Age)	14.6(4.8-44.7)	<0.001	13.3(4.3-40)	<0.001	13.1(4.3-40)	<0.001
Has T2D	2.9(1.7-5)	<0.001	2.9(1.7-5)	<0.001	3(1.7-5.2)	<0.001
Female	0.8(0.5-1.3)	0.44	0.9(0.5-1.4)	0.53	0.9(0.5-1.4)	0.55
HDL<35 mg/dL	1.2(0.6-2.4)	0.52	1.3(0.7-2.5)	0.45	1.3(0.7-2.5)	0.45
HDL>60 mg/dL	1(0.6-1.9)	0.95	1.1(0.6-2)	0.84	1.1(0.6-2)	0.86
TC >240 mg/dL	1.3(0.8-2.1)	0.29	1.3(0.8-2.1)	0.31	1.3(0.8-2.1)	0.29
Former or current smoker	1.8(1.1-2.9)	0.014	1.8(1.1-2.9)	0.016	1.8(1.1-2.9)	0.014
Hypertension stage II-IV	1.8(1-3.1)	0.037	1.7(0.9-2.9)	0.077	1.7(1-2.9)	0.074
BMI>25	1.5(0.6-3.6)	0.35	1.6(0.7-4)	0.3	1.6(0.7-3.9)	0.3
Clone present			2.2(1.1-4.6)	0.029		
VAF<0.10					1.8(0.7-4.6)	0.2
VAF>0.10					3.1(1.2-8.4)	0.025
Pseudo Log-likelihood		-468		-462		-462
Pseudo likelihood ratio test		79 on 9 df		93 on 10 df		93.5 on 11 df

Supplementary Table S13
Risks associated with developing incident coronary heart disease using traditional risk factors as well as hsCRP, RDW, and clonality.

Hazard ratios were estimated using competing risks regression with death as the competing risk. P-values are derived from the Fine-Gray test. Individuals with prior coronary heart disease (CHD) were excluded for CHD analysis. Models shown in A or B are from the same population and differ by having covariates either removed or added. All individuals were from Jackson Heart Study.

n=1,795	Model 1		Model 2		Model 3	
Covariate	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
log(Age)	3.3(0.6-17)	0.17	2.7(0.5-13)	0.23	2.8(0.6-14)	0.21
Has T2D	2.9(1.4-5.9)	0.004	3(1.5-6)	0.002	3(1.6-5.8)	<0.001
Female	0.7(0.3-1.3)	0.21	0.7(0.3-1.3)	0.21	0.6(0.3-1.2)	0.13
HDL<35 mg/dL	1.6(0.6-4.6)	0.38	1.8(0.6-5.1)	0.29		
HDL≥60 mg/dL	1(0.5-2.1)	0.99	1(0.5-2)	0.9		
TC≥200mg/dL	2.2(1.1-4.1)	0.021	2.3(1.2-4.6)	0.015	2.3(1.2-4.5)	0.013
Former or current smoker	2(1.1-3.7)	0.027	2.1(1.1-3.9)	0.021	2(1.1-3.8)	0.024
SBP≥160 mm Hg	2.4(1.2-4.7)	0.011	2.2(1.1-4.4)	0.023	2.2(1.1-4.3)	0.03
hsCRP > 1.0 mg/L	1.1(0.5-2.5)	0.76	1(0.4-2.3)	0.99		
No clone, RDW≥14.5%			2.3(1.2-4.6)	0.017	2.2(1.1-4.4)	0.02
Clone, RDW<14.5%			0.9(0.1-7.4)	0.95	1.7(0.4-7.7)	0.46
Clone, RDW≥14.5%			9.8(2-48.3)	0.005	9.3(2-43)	0.005
Pseudo Log-likelihood		-273		-268		-278
Pseudo likelihood ratio test		39.1 on 9 df		49.0 on 12 df		48.0 on 9 df



* * *

[00204] Having thus described in detail preferred embodiments of the present invention, it is to be understood that the invention defined by the above paragraphs is not to be limited to

particular details set forth in the above description as many apparent variations thereof are possible without departing from the spirit or scope of the present invention.

WHAT IS CLAIMED IS:

1. A method for identifying and selecting a subject with increased risk of developing a cardiometabolic disease and optionally a hematological cancer, comprising the steps of:

(a) sequencing at least part of a genome comprising one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject,

(b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer.

2. The method according to claim 1, wherein the presence of said mutation(s) also indicates an increase in red blood cell distribution width (RDW).

3. The method according to claim 1, wherein the cardiometabolic disease is atherosclerosis, coronary heart disease (CHD) or ischemic stroke (IS).

4. The method according to claim 1, wherein the hematological cancer is a leukemia, a lymphoma, a myeloma or a blood syndrome.

5. The method according to claim 4, wherein the leukemia is acute myeloid leukemia (AML) or chronic myelogenous leukemia (CML).

6. The method according to claim 4, wherein the blood syndrome is myelodysplastic syndrome (MDS).

7. The method according to claim 1, wherein the one more cells in the blood sample are hematopoietic stem cells (HSCs), committed myeloid progenitor cells having long term self-renewal capacity or mature lymphoid cells having long term self-renewal capacity.

8. The method according to claim 1, wherein the part of the genome is an exome.

9. The method according to claim 1, wherein the sequencing is whole exome sequencing (WES).

10. The method according to claim 1, wherein the subject is a human.

11. The method according to claim 10, wherein the human also exhibits one or more risk factors of being a smoker, having a high level of total cholesterol or having high level of high-density lipoprotein (HDL).

12. A method for identifying and selecting a subject with an increased risk of developing a cardiometabolic disease and optionally a hematological cancer and providing a personalized medicine method, said method comprising the steps of

(a) sequencing at least part of a genome comprising one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1* of one or more cells in a blood sample of the subject,

(b) identifying from said sequencing one or more mutations in one or more genes selected from the group consisting of *DNMT3A*, *TET2*, *ASXL1*, *TP53*, *JAK2* and *SF3B1*, wherein presence of said mutation(s) indicates an increased risk of developing a cardiometabolic disease and optionally a hematological cancer, and

(c) initiating a treatment or monitoring regimen to suppress said mutation(s) in the subject, thereby decreasing risk of developing a cardiometabolic disease and optionally a hematological cancer.

13. The method according to claim 12, wherein the presence of said mutation(s) also indicates an increase in red blood cell distribution width (RDW).

14. The method according to claim 12, wherein the cardiometabolic disease is coronary heart disease (CHD) or ischemic stroke (IS).

15. The method according to claim 12, wherein the hematological cancer is a leukemia, a lymphoma, a myeloma or a blood syndrome.

16. The method according to claim 15, wherein the leukemia is acute myeloid leukemia (AML) or chronic myelogenous leukemia (CML).

17. The method according to claim 15, wherein the blood syndrome is myelodysplastic syndrome (MDS).

18. The method according to claim 12, wherein the one more cells in the blood sample are hematopoietic stem cells (HSCs), committed myeloid progenitor cells having long term self-renewal capacity or mature lymphoid cells having long term self-renewal capacity.

19. The method according to claim 12, wherein macrophages with said mutation(s) are treated with an agent to increase reverse cholesterol transport, reduce inflammation, or both.

20. The method according to claim 12, wherein the part of the genome is an exome.

21. The method according to claim 12, wherein the sequencing is whole exome sequencing (WES).

22. The method according to claim 12, wherein the subject is a human.

23. The method according to claim 22, wherein the human also exhibits one or more risk factors of being a smoker, having a high level of total cholesterol or having high level of high-density lipoprotein (HDL).

24. A method according to any one of claims 1-23, wherein the one or more mutations are frameshift mutations, nonsense mutations, missense mutations or splice-site variant mutations.

25. A method according to any one of claims 1-24, wherein the mutation in *DNMT3A* is a mutation in exons 7 to 23.

26. A method according to any one of claims 1-24, wherein the mutation in *DNMT3A* is a mutation selected from the group consisting of P307S, P307R, R326H, R326L, R326C, R326S, R366P, R366H, R366G, A368T, F414L, F414S, F414C, C497Y, Q527H, Q527P, Y533C, G543A, G543S, G543C, L547H, L547P, L547F, M548I, M548K, G550R, W581R, W581G, W581C, G646V, G646E, L653W, L653F, V657A, V657M, R659H, Y660C, R676W, R676Q, G685R, G685E, G685A, D686Y, D686G, G699R, G699S, G699D, P700S, P700R, P700Q, D702N, D702Y, V704M, V704G, I705F, I705T, I705S, C710S, S714C, N717S, N717I, P718L, R720H, R720G, Y724C, R729Q, R729W, R729G, F731L, F732del, F732S, F732L, F734L, F734C, Y735C, Y735N, Y735S, R736H, R736C, R736P, L737H, L737V, L737F, L737R, A741V, R749C, R749L, F751L, F752del, F752C, F752L, F752I, F752V, L754R, L754H, F755S, F755I, F755L, M761I, M761V, G762C, S770W, S770P, R771Q, F772I, F772V, L773R, E774K, E774D, D781G, R792H, G796D, G796V, N797Y, N797H, P799R, P799H, R803S, P804S, P804L, S828N, K829R, Q842E, P849L, D857N, W860R, F868S, G869S, G869V, M880V, S881R, S881I, R882H, R882P, R882C, R882G, Q886R, G890D, L901R, L901H, P904L, F909C and A910P.

27. A method according to any one of claims 1-24, wherein the mutation in *TET2* is a mutation selected from the group consisting of S282F, N312S, L346P, S460F, D666G, P941S, and C1135Y.

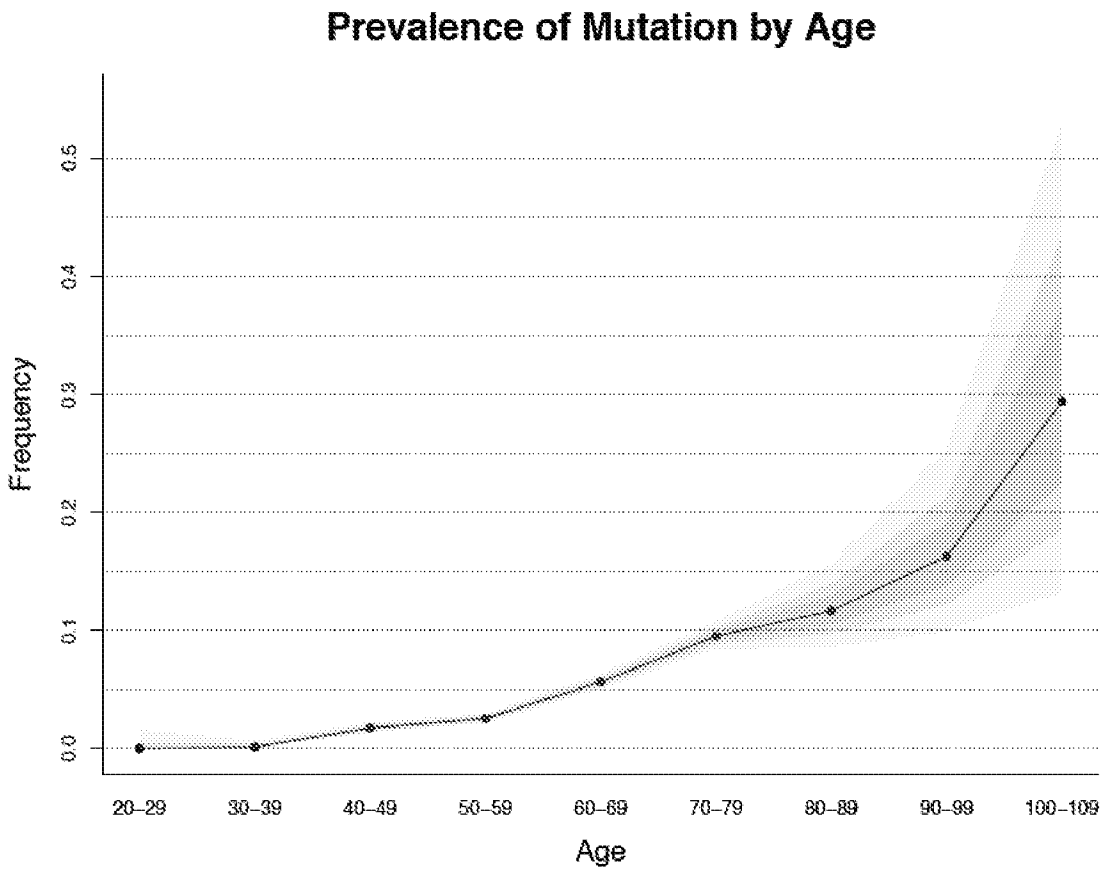
28. A method according to any one of claims 1-24, wherein the mutation in *ASXL1* is a mutation in exon 11-12.

29. A method according to any one of claims 1-24, wherein the mutation in *TP523* is a mutation selected from the group consisting of S46F, G105C, G105R, G105D, G108S, G108C, R110L, R110C, T118A, T118R, T118I, L130V, L130F, K132Q, K132E, K132W, K132R, K132M, K132N, C135W, C135S, C135F, C135G, Q136K, Q136E, Q136P, Q136R, Q136L, Q136H, A138P, A138V, A138A, A138T, T140I, C141R, C141G, C141A, C141Y, C141S, C141F, C141W, V143M, V143A, V143E, L145Q, L145R, P151T, P151A, P151S, P151H, P152S, P152R, P152L, T155P, R158H, R158L, A159V, A159P, A159S, A159D, A161T, A161D, Y163N, Y163H, Y163D, Y163S, Y163C, K164E, K164M, K164N, K164P, H168Y, H168P, H168R, H168L, H168Q, M169I, M169T, M169V, T170M, E171K, E171Q, E171G, E171A, E171V and E171D.

30. A method according to any one of claims 1-24, wherein the mutation in *JAK2* is a mutation selected from the group consisting of N533D, N533Y, N533S, H538R, K539E, K539L, I540T, I540V, V617F, R683S, R683G, del/ins537---539L, del/ins538---539L, del/ins540---543MK, del/ins540---544MK, del/ins541- -543K, del542---543, del543---544 and ins11546---547.

31. A method according to any one of claims 1-24, wherein the mutation in *SF3B1* is a mutation selected from the group consisting of G347V, R387W, R387Q, E592K, E622D, Y623C, R625L, R625C, H662Q, H662D, K666N, K666T, K666E, K666R, K700E, V701F, A708T, G740R, G740E, A744P, D781G and E783K.

Figure 1



No. with mutation	0	1	50	138	282	219	37	14	5
Total	240	885	2894	5441	5002	2300	317	86	17

Figure 2

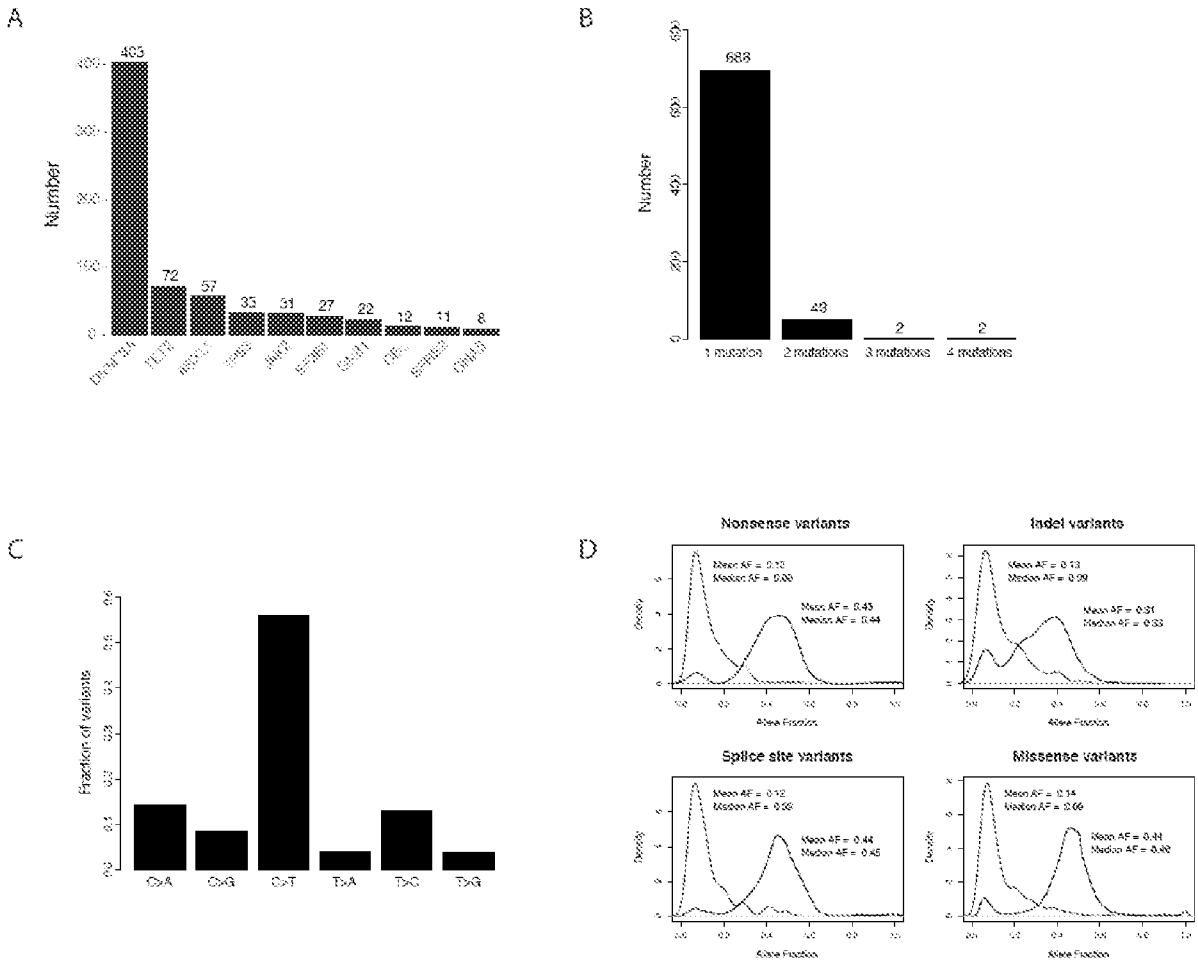

A
Number
403
72
57
35
31
27
22
12
11
8
B
Number
688
48
2
2
1 mutation
2 mutations
3 mutations
4 mutations
C
Fraction of variants
C>A
C>G
C>T
T>A
T>C
T>G
D
Nonsense variants
Indel variants
Splice site variants
Missense variants
Density
Allele Fraction

Figure 3

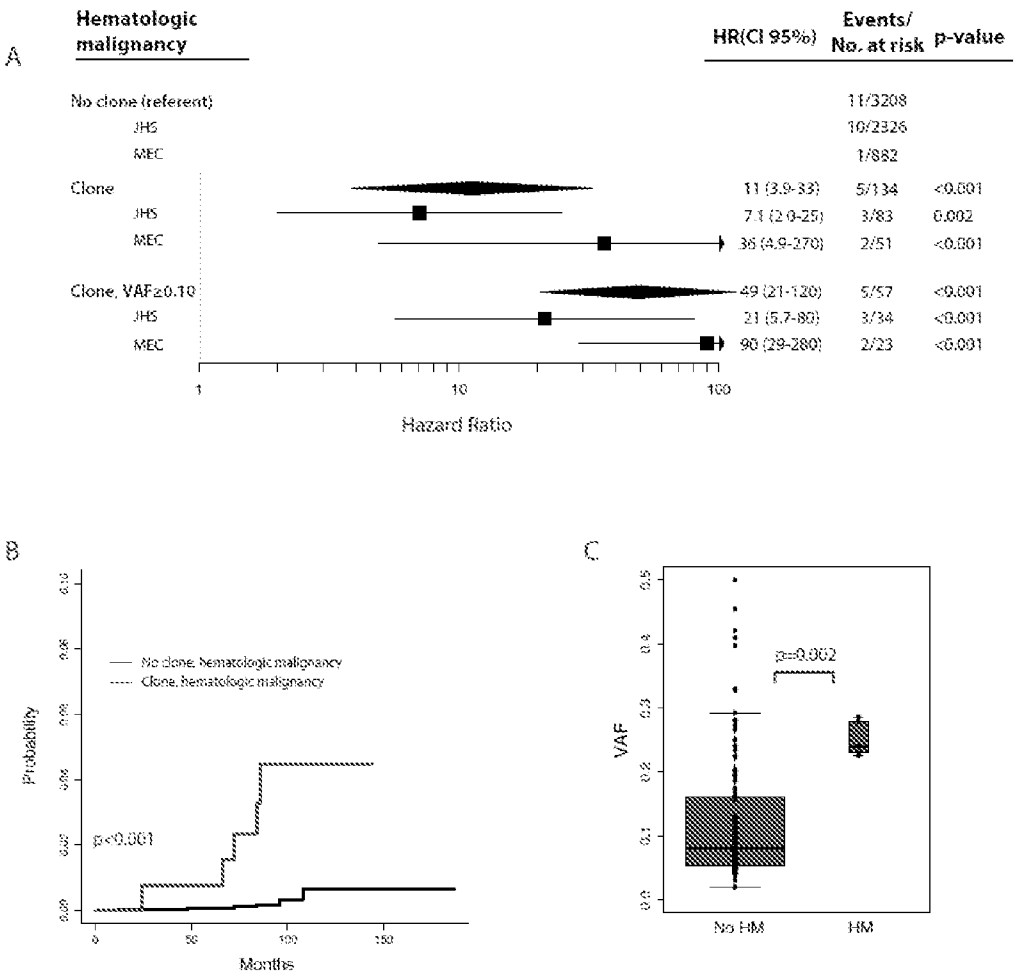

A
Hematologic malignancy
HR(CI 95%)
Events/ No. at risk
p-value
No clone (referent)
JHS
MEC
11/3208
10/2326
1/882
Clone
JHS
MEC
11 (3.9-33)
7.1 (2.0-25)
36 (4.9-270)
5/134
3/83
2/51
<0.001
0.002
<0.001
Clone, VAF≥0.10
JHS
MEC
49 (21-120)
21 (5.7-80)
90 (29-280)
5/57
3/34
2/23
<0.001
<0.001
<0.001
1
10
100
Hazard Ratio
B
No clone, hematologic malignancy
Clone, hematologic malignancy
Probability
p<0.001
Months
C
p=0.002
VAF
No HM
HM

Figure 4

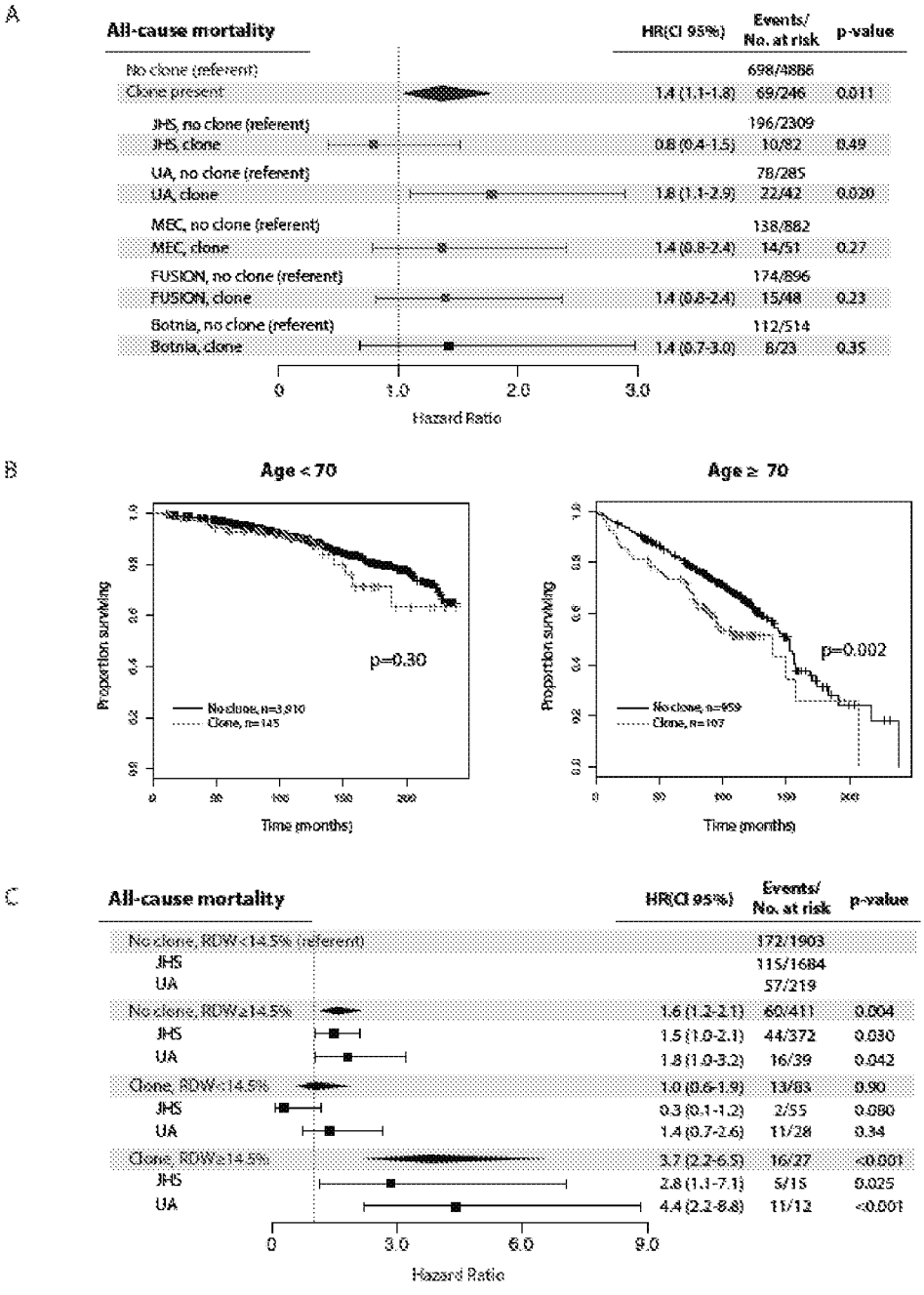

A
All-cause mortality
HR(CI 95%)
Events/ No. at risk
p-value
No clone (referent) 698/4886
Clone present 1.4 (1.1-1.8) 69/246 0.011
JHS, no clone (referent) 196/2309
JHS, clone 0.8 (0.4-1.5) 10/82 0.49
UA, no clone (referent) 78/285
UA, clone 1.8 (1.1-2.9) 22/42 0.020
MEC, no clone (referent) 138/862
MEC, clone 1.4 (0.8-2.4) 14/51 0.27
FUSION, no clone (referent) 174/896
FUSION, clone 1.4 (0.8-2.4) 15/48 0.23
Botnia, no clone (referent) 112/514
Botnia, clone 1.4 (0.7-3.0) 8/23 0.35
0
1.0
2.0
3.0
Hazard Ratio
B
Age < 70
Age ≥ 70
Proportion surviving
p=0.30
p=0.002
Time (months)
C
All-cause mortality
HR(CI 95%)
Events/ No. at risk
p-value
No clone, RDW<14.5% (referent) 172/1903
JHS 115/1684
UA 57/219
No clone, RDW≥14.5% 1.6 (1.3-2.1) 60/411 0.004
JHS 1.5 (1.0-2.1) 44/372 0.030
UA 1.8 (1.0-3.2) 16/39 0.042
Clone, RDW<14.5% 1.0 (0.6-1.9) 13/83 0.90
JHS 0.3 (0.1-1.2) 2/55 0.080
UA 1.4 (0.7-2.6) 11/28 0.34
Clone, RDW≥14.5% 3.7 (2.2-6.5) 16/27 <0.001
JHS 2.8 (1.1-7.1) 5/15 0.025
UA 4.4 (2.3-8.8) 11/12 <0.001
0
3.0
6.0
9.0
Hazard Ratio

Figure 5

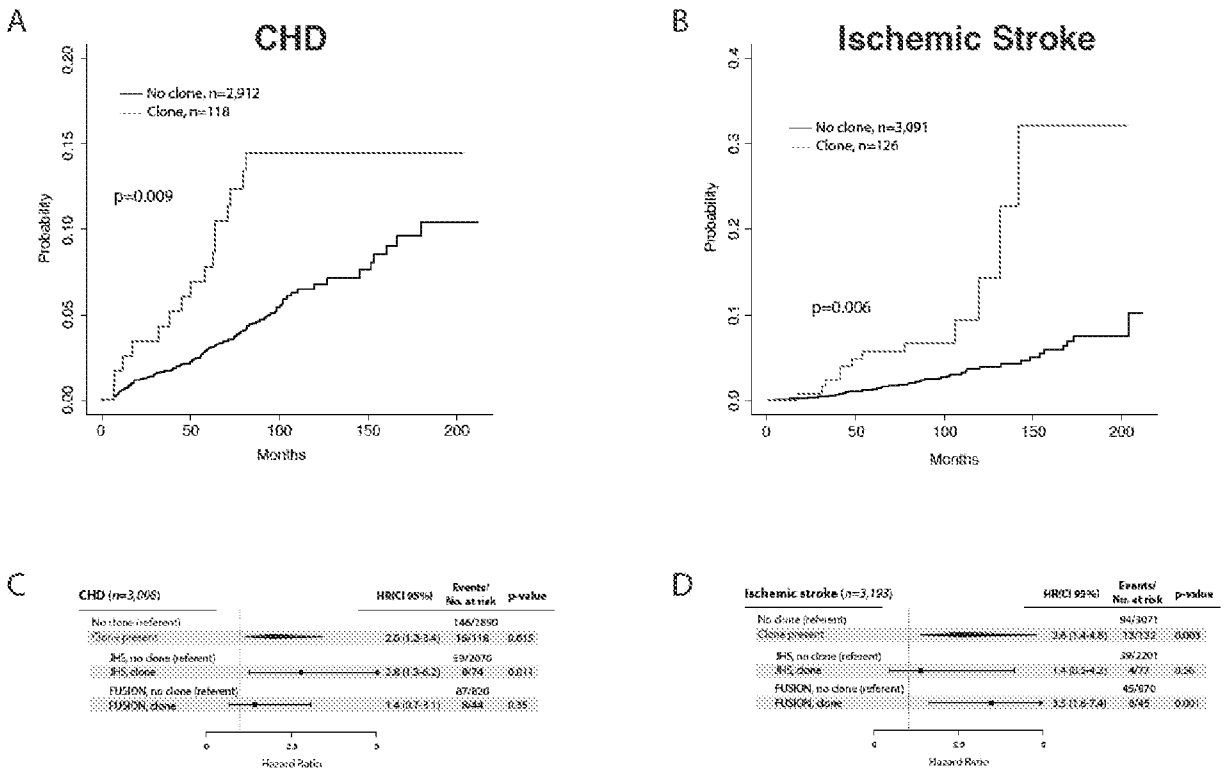

A
CHD
No clone, n=2,912
Clone, n=118
p=0.009
Probability
Months
B
Ischemic Stroke
No clone, n=3,091
Clone, n=126
p=0.006
Probability
Months
C
CHD (n=3,068)
HR(CI 95%)
Events/ No. at risk
p-value
No clone (referent)
146/2890
Clone present
2.0 (1.3-3.4)
16/118
0.015
IHS, no clone (referent)
69/2870
IHS, clone
2.8 (1.3-6.2)
8/74
0.011
FUSION, no clone (referent)
87/820
FUSION, clone
1.4 (0.7-2.1)
8/44
0.35
Hazard Ratio
D
Ischemic stroke (n=3,123)
HR(CI 95%)
Events/ No. at risk
p-value
No clone (referent)
94/3071
Clone present
3.8 (1.4-4.8)
13/132
0.003
IHS, no clone (referent)
39/2201
IHS, clone
1.4 (0.5-4.2)
4/77
0.56
FUSION, no clone (referent)
45/870
FUSION, clone
3.5 (1.8-7.4)
8/45
0.001
Hazard Ratio

Figure 6

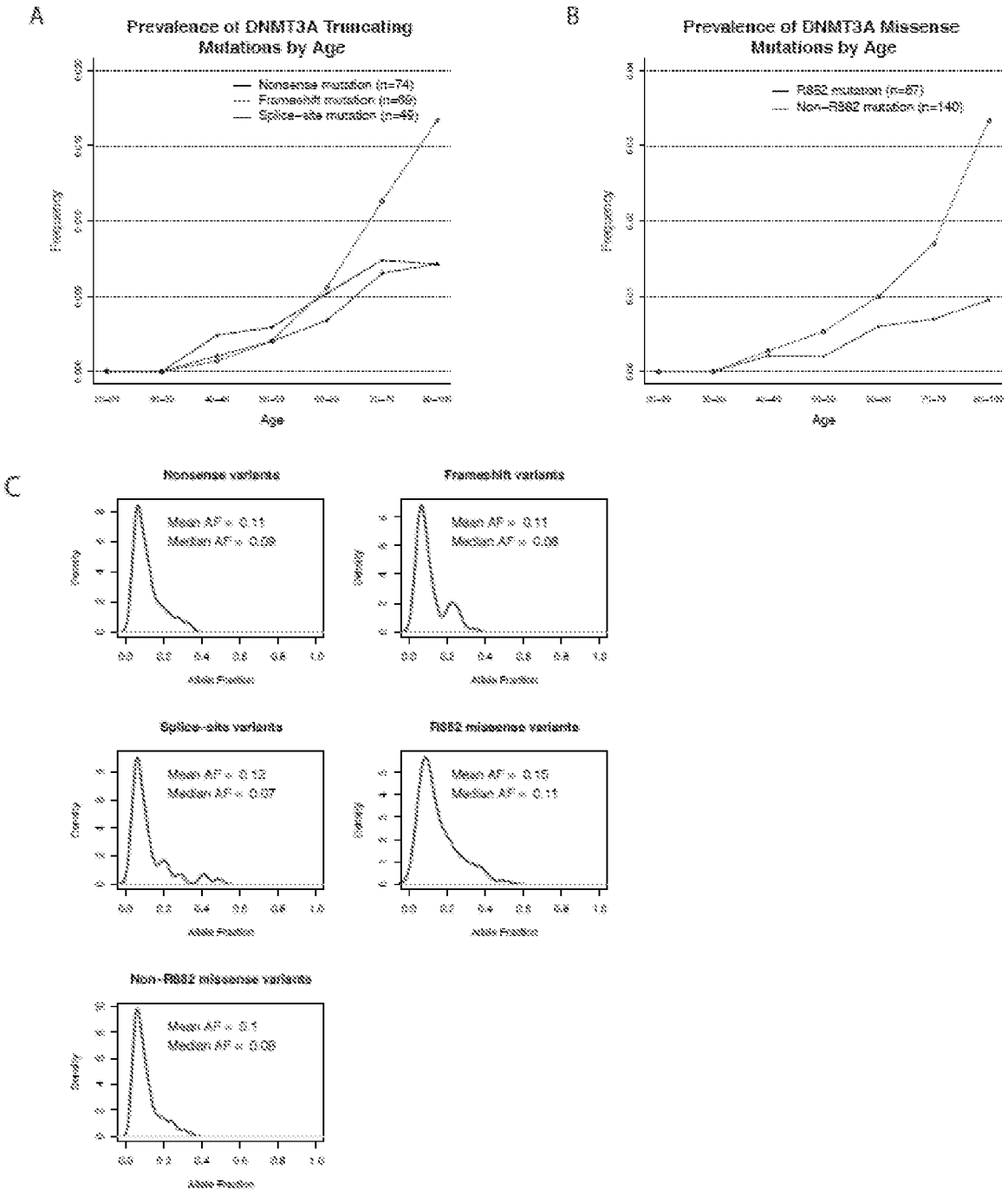
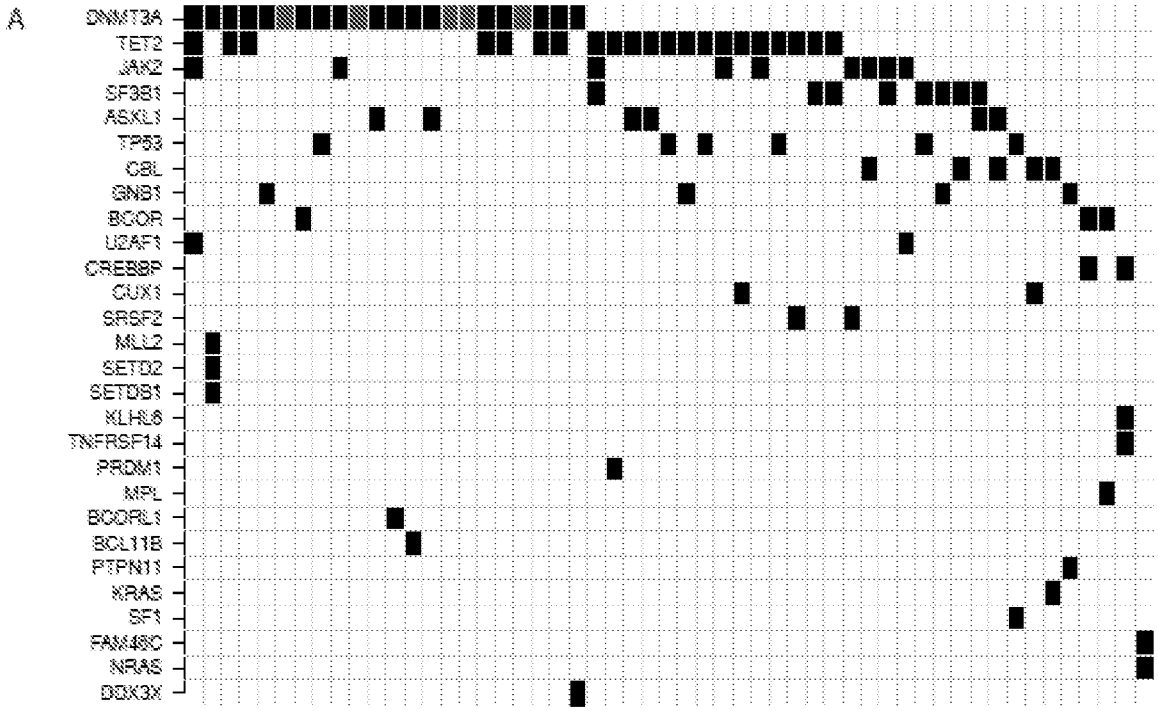

A
Prevalence of DNMT3A Truncating Mutations by Age
Nonsense mutation (n=74)
Frameshift mutation (n=66)
Splice−site mutation (n=49)
Frequency
Age
B
Prevalence of DNMT3A Missense Mutations by Age
R882 mutation (n=87)
Non−R882 mutation (n=140)
Frequency
Age
C
Nonsense variants
Mean AF = 0.11
Median AF = 0.09
Frameshift variants
Mean AF = 0.11
Median AF = 0.08
Splice−site variants
Mean AF = 0.12
Median AF = 0.07
R882 missense variants
Mean AF = 0.15
Median AF = 0.11
Non−R882 missense variants
Mean AF = 0.1
Median AF = 0.08
Density
Allele Fraction

Figure 7


A
DNMT3A
TET2
JAK2
SF3B1
ASXL1
TP53
CBL
GNB1
BCOR
U2AF1
CREBBP
CUX1
SRSF2
MLL2
SETD2
SET3B1
KLHL6
TNFRSF14
PRDM1
MPL
BCORL1
BCL11B
PTPN11
KRAS
SF1
FAM46C
NRAS
DDX3X

B

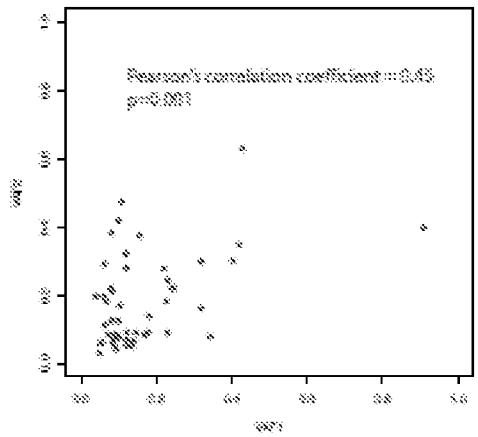

Pearson's correlation coefficient = 0.45
p=0.001

Figure 8

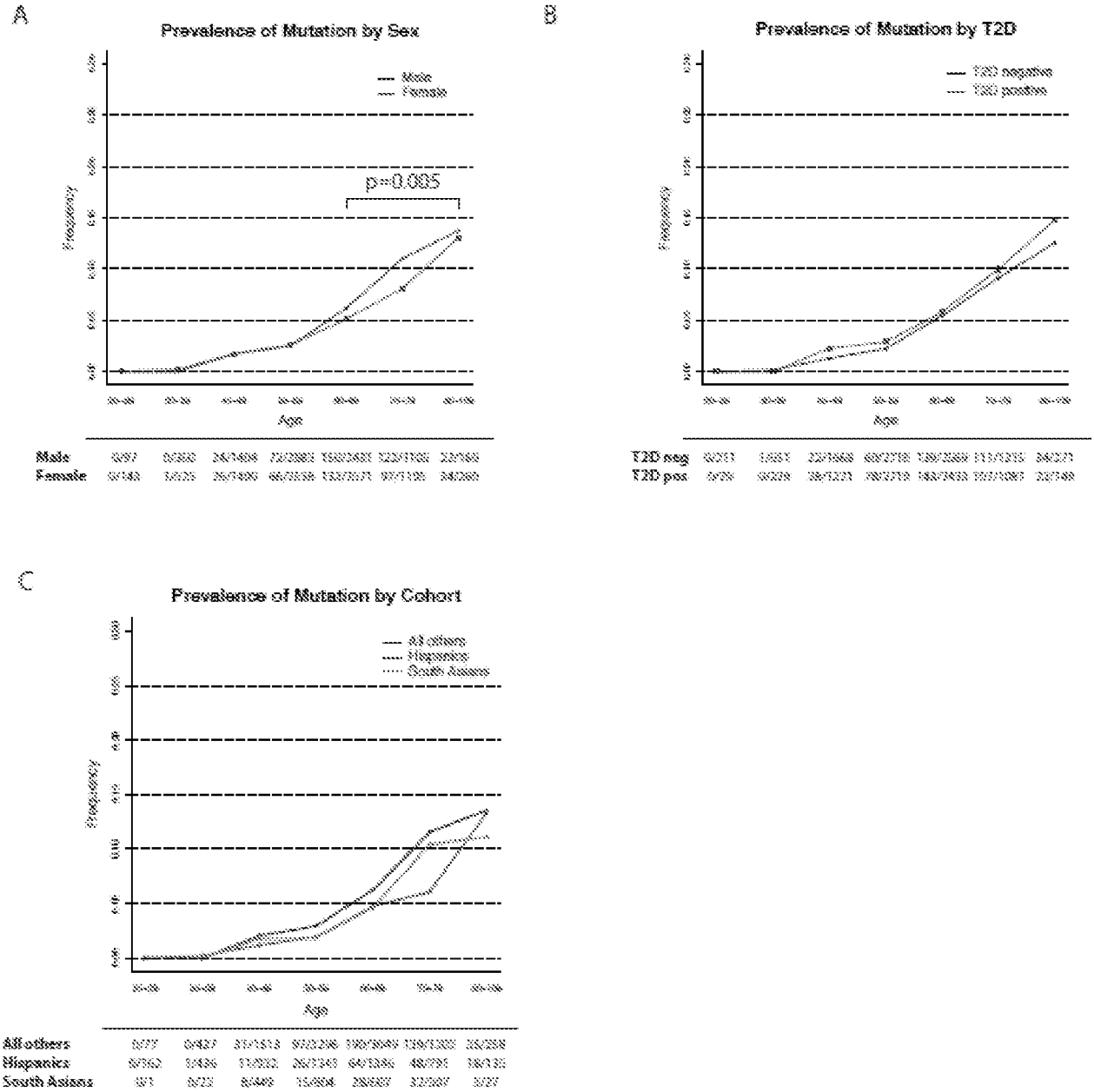

A
Prevalence of Mutation by Sex
Male
Female
p=0.005
Age
Male
Female
B
Prevalence of Mutation by T2D
T2D negative
T2D positive
Age
T2D neg
T2D pos
C
Prevalence of Mutation by Cohort
All others
Hispanics
South Asians
Age
All others
Hispanics
South Asians

Figure 9

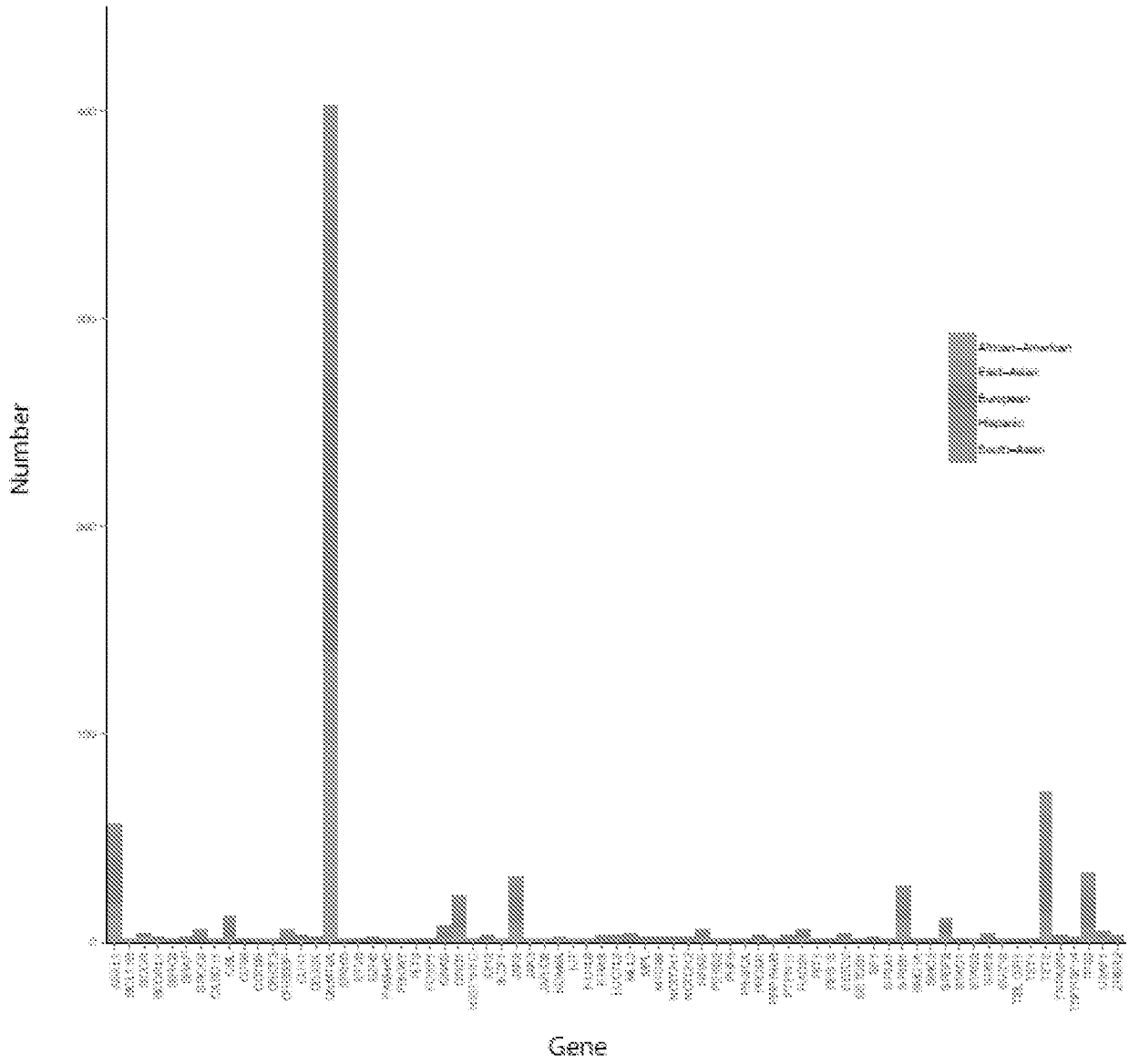


Figure 10

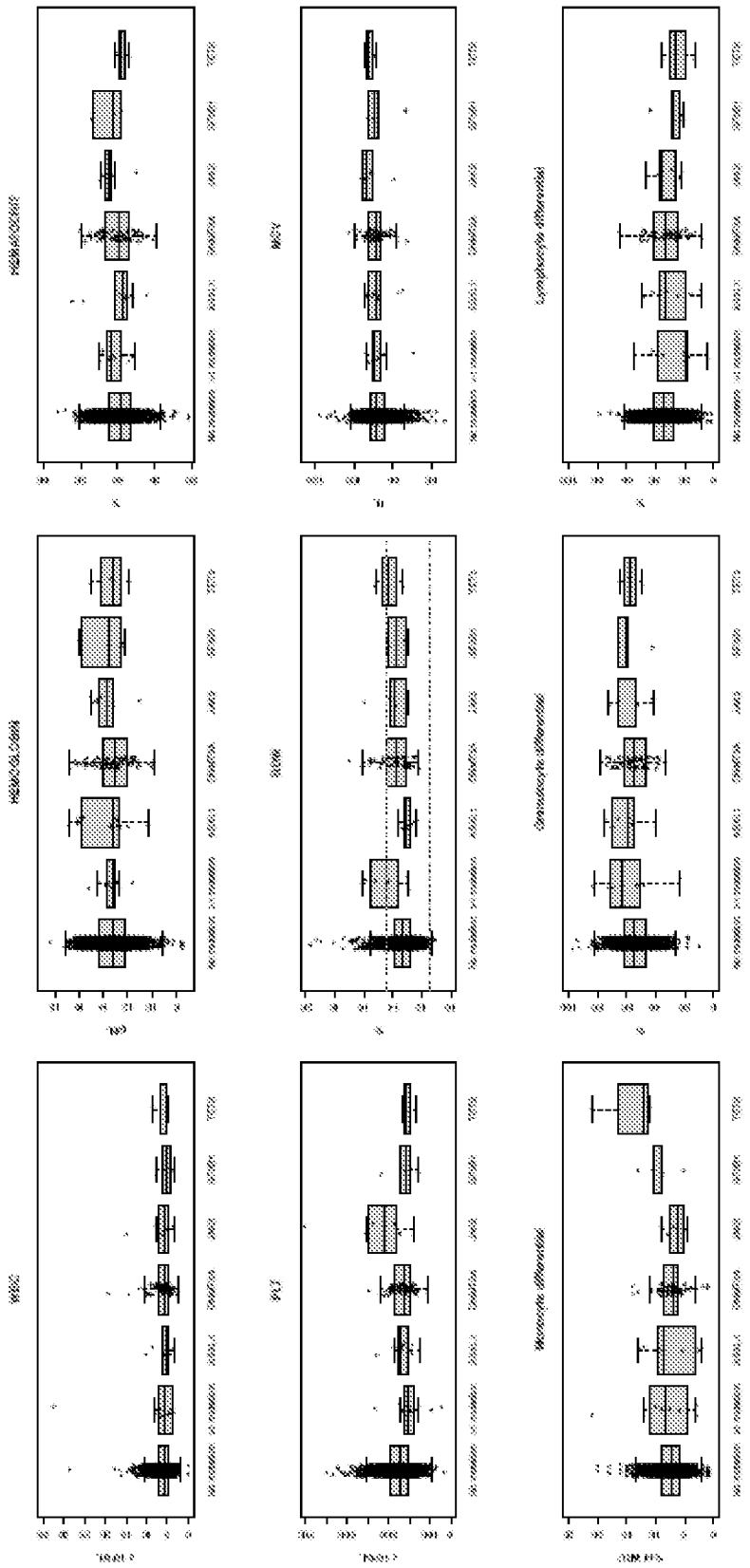


Figure 11

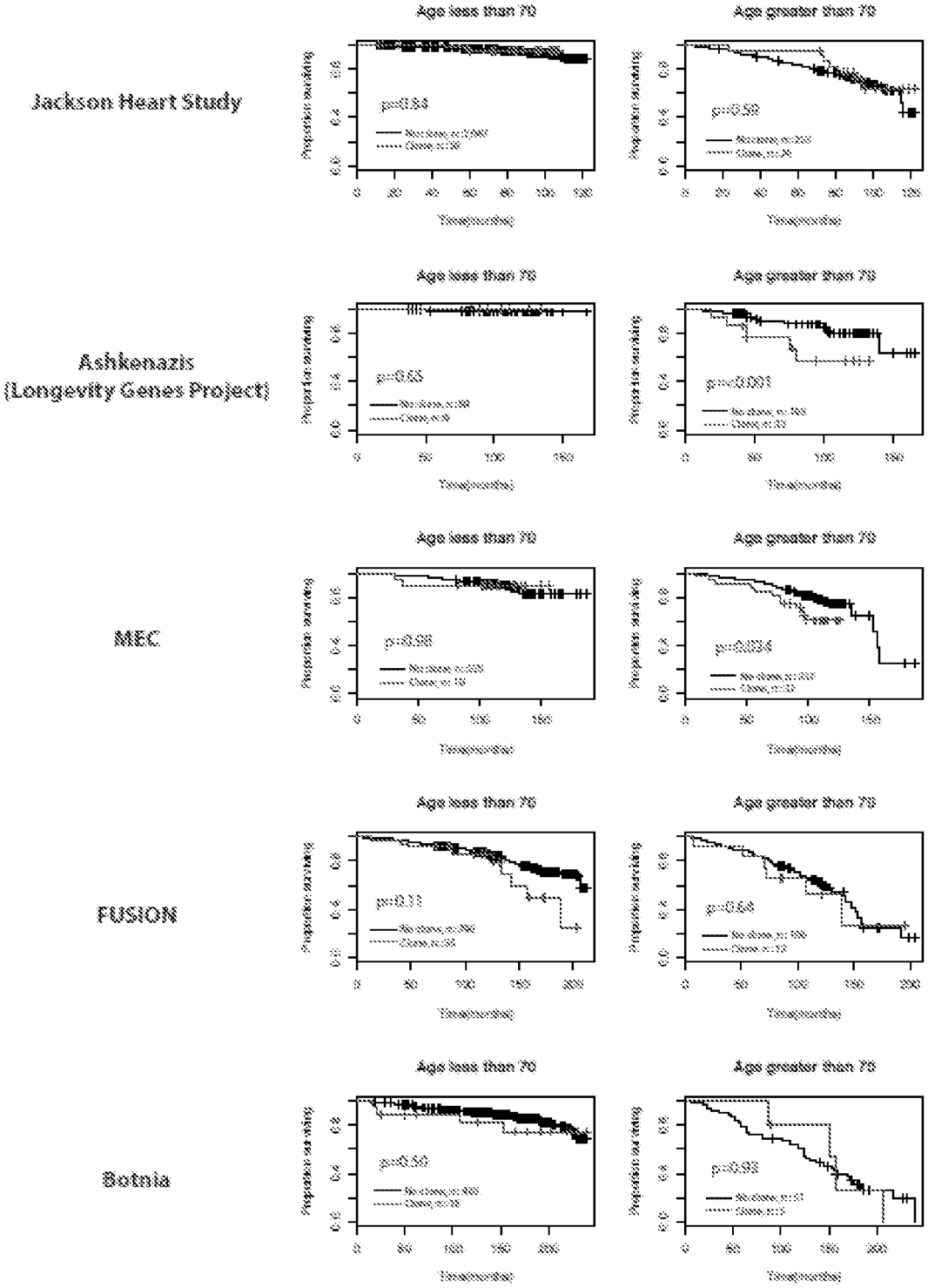

Jackson Heart Study
Age less than 70
Age greater than 70
Ashkenazis
(Longevity Genes Project)
Age less than 70
Age greater than 70
MEC
Age less than 70
Age greater than 70
FUSION
Age less than 70
Age greater than 70
Botnia
Age less than 70
Age greater than 70

Figure 12

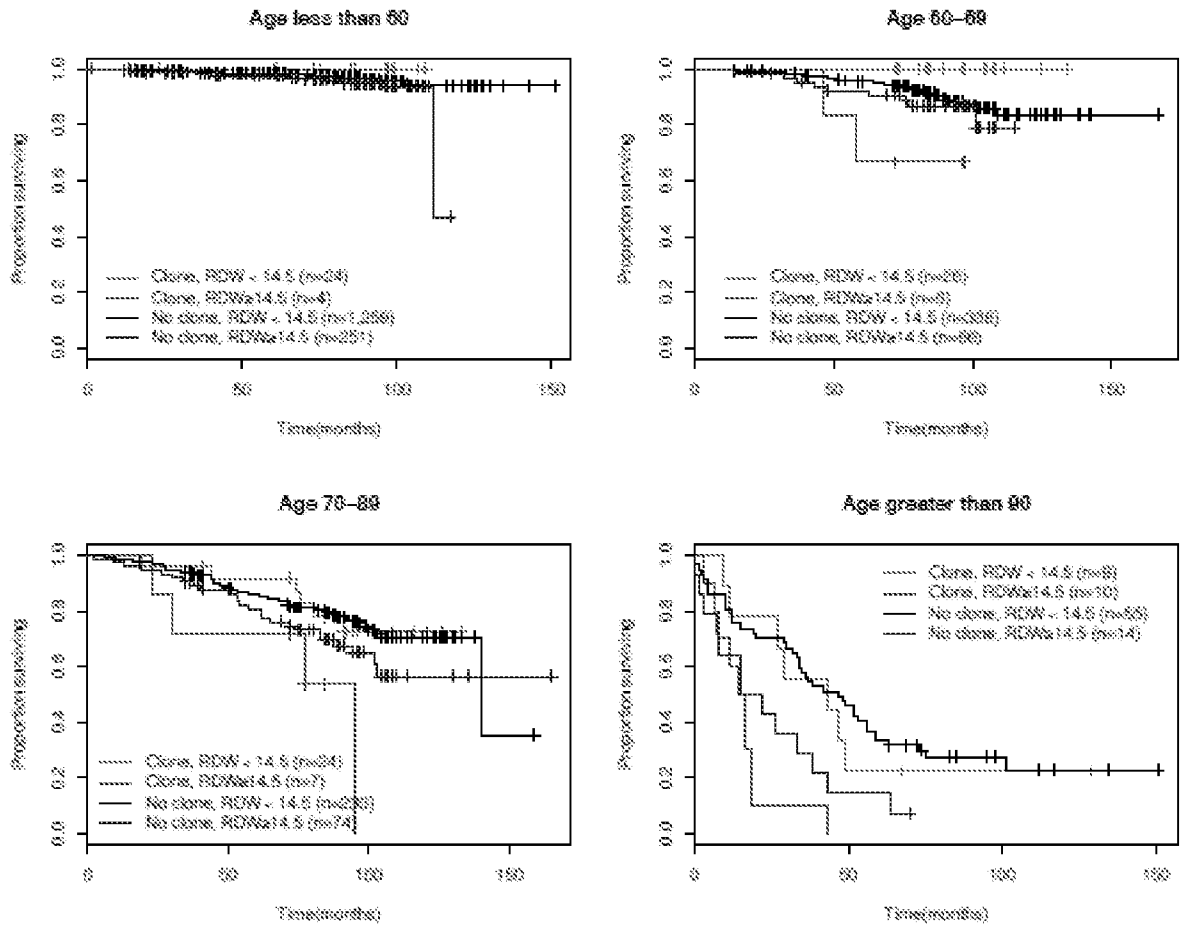

Age less than 60
Age 60–69
Age 70–89
Age greater than 90
Time(months)

Figure 13

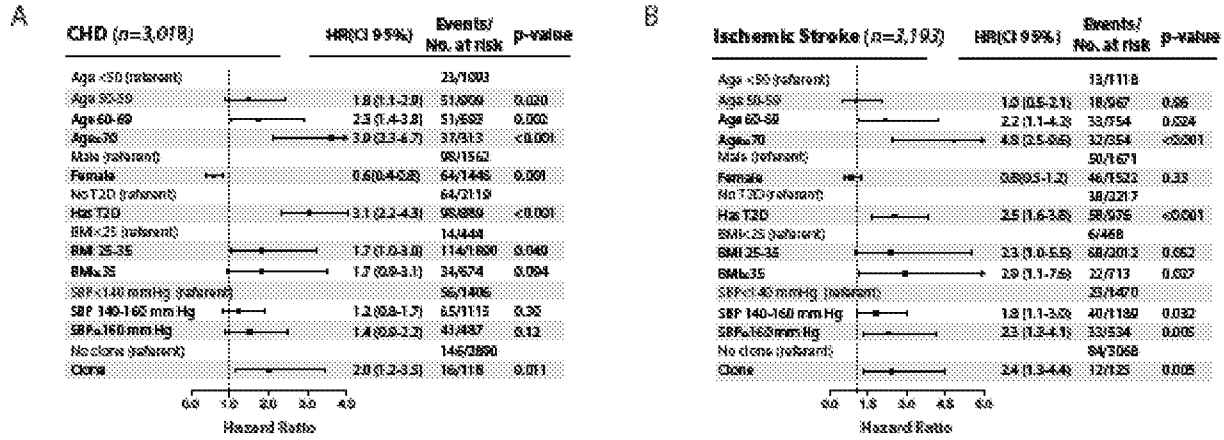

A
CHD (n=3,018)
HR(CI 95%)
Events/ No. at risk
p-value
Age <50 (referent)
Age 50-59
Age 60-69
Age≥70
Male (referent)
Female
No T2D (referent)
Has T2D
BMI<25 (referent)
BMI 25-35
BMI≥35
SBP<140 mmHg (referent)
SBP 140-160 mm Hg
SBP≥160 mm Hg
No clone (referent)
Clone
Hazard Ratio
B
Ischemic Stroke (n=3,193)
HR(CI 95%)
Events/ No. at risk
p-value
Age <50 (referent)
Age 50-59
Age 60-69
Age≥70
Male (referent)
Female
No T2D (referent)
Has T2D
BMI<25 (referent)
BMI 25-35
BMI≥35
SBP<140 mmHg (referent)
SBP 140-160 mm Hg
SBP≥160 mm Hg
No clone (referent)
Clone
Hazard Ratio

Figure 14

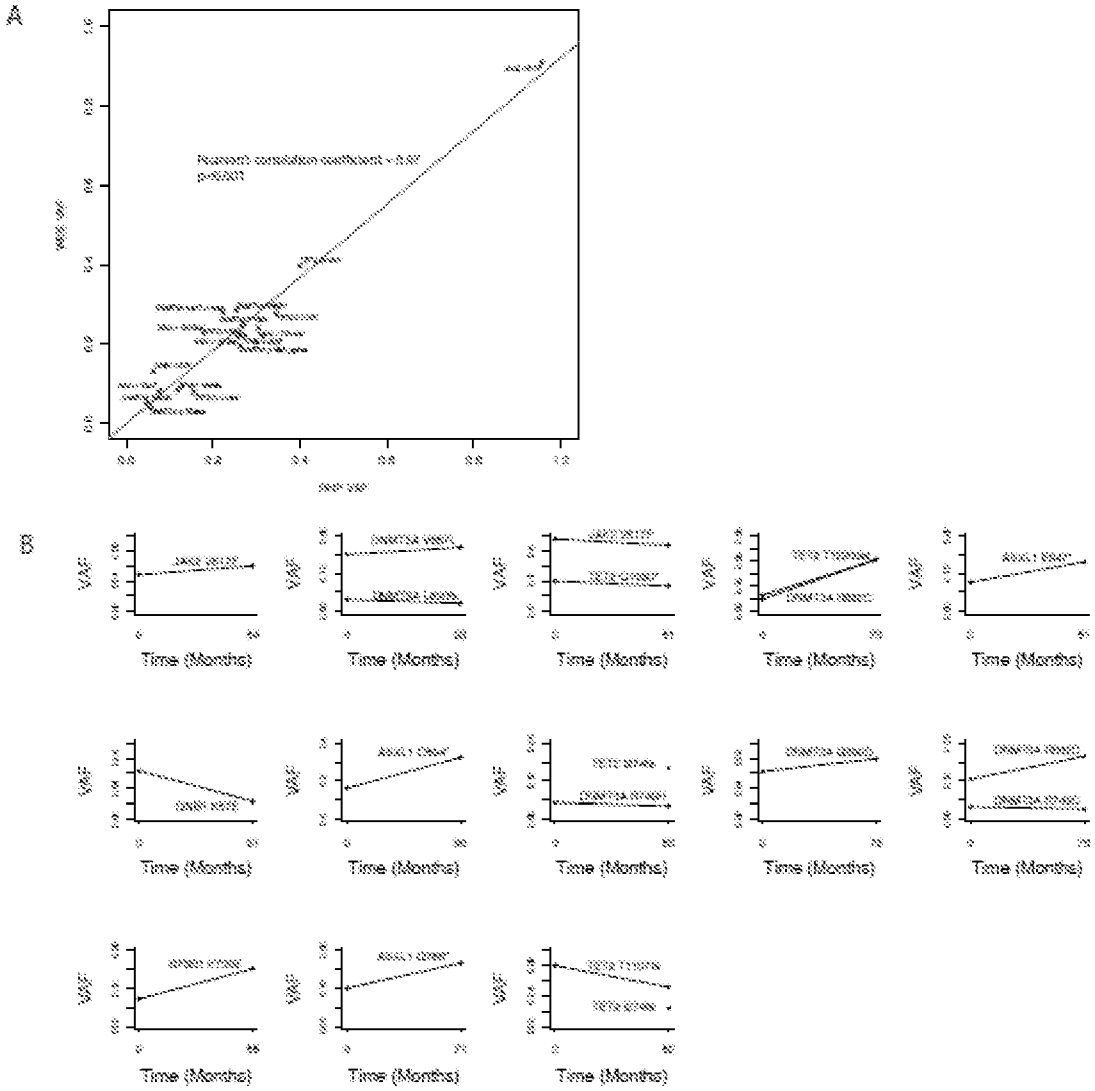


Figure 15

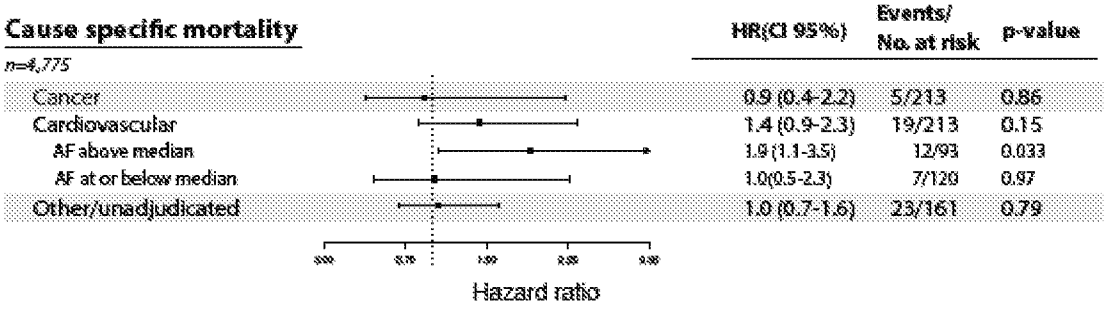

Cause specific mortality
n=4,775
HR(CI 95%)
Events/ No. at risk
p-value
Cancer
0.9 (0.4-2.2)
5/213
0.86
Cardiovascular
1.4 (0.9-2.3)
19/213
0.15
AF above median
1.9 (1.1-3.5)
12/93
0.033
AF at or below median
1.0(0.5-2.3)
7/120
0.97
Other/unadjudicated
1.0 (0.7-1.6)
23/161
0.79
Hazard ratio

Figure 16

A
Clonal hematopoiesis and MI
No. with mutation/ No. at risk
OR (95%)
p-value
ATVB, no MI, <40 (reference)
2/672
ATVB, MI, <40
83 (20-74)
18/724
0.0005
ATVB, no MI, 40-45 (reference)
6/918
ATVB, MI, 40-45
34 (13-110)
23/1042
0.006
PROMIS, no MI, 40-45 (reference)
2/366
PROMIS, MI, 40-45
27 (0.66-241)
21/1420
0.20
PROMIS, no MI, 46-50 (reference)
11/1308
PROMIS, MI, 46-50
29 (1.4-603)
37/1175
0.001
1
2
4
8
16
32
Odds Ratio
B
Gene
Controls (n=2006)
Cases (n=4105)
DNMT3A
15
31
TET2
0
10
ASXL1
0
6
JAK2
0
15
TET2 plus another mutation
1
2

Figure 17

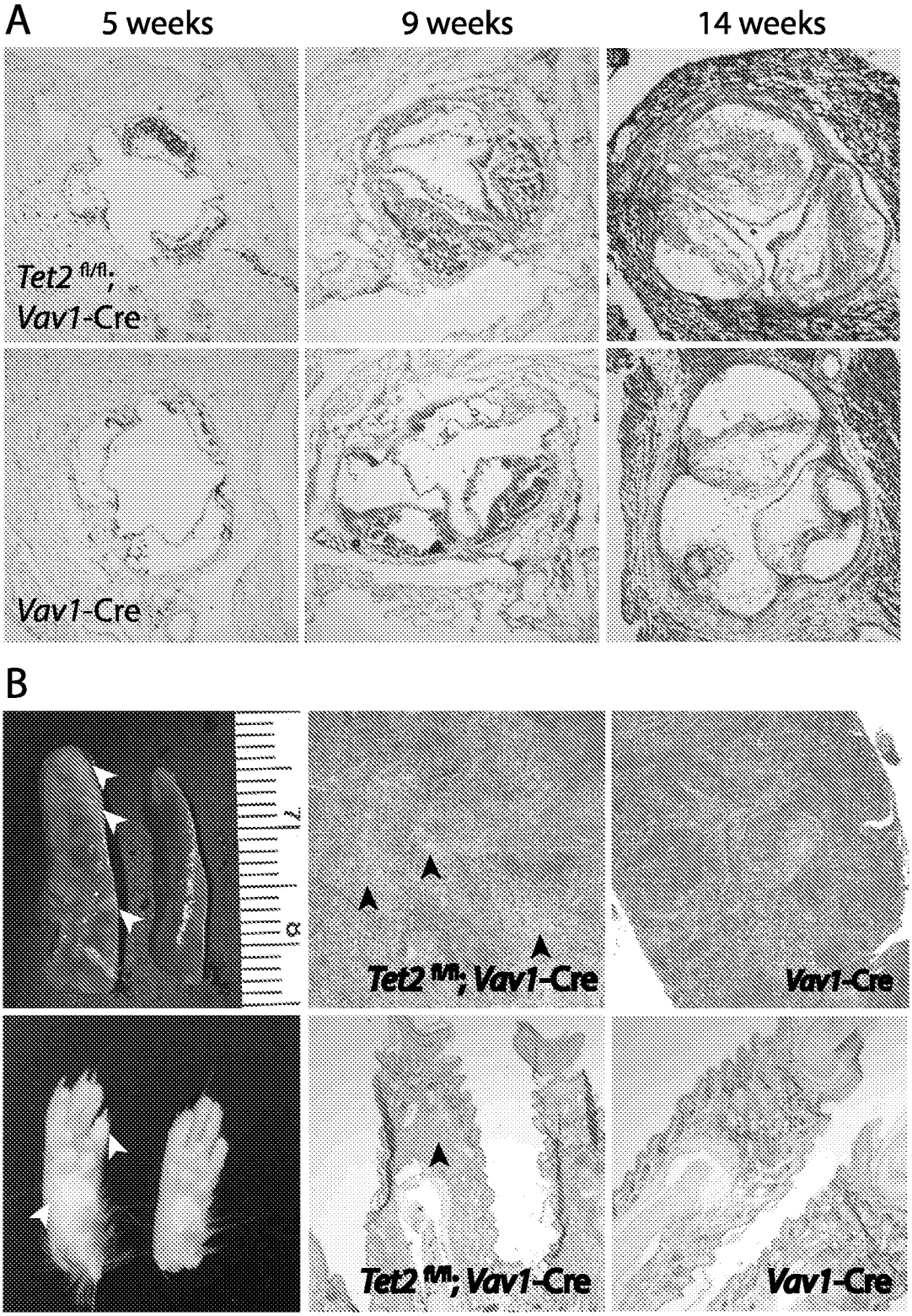

A
5 weeks
9 weeks
14 weeks
Tet2 fl/fl; Vav1-Cre
Vav1-Cre
B
Tet2 fl/fl; Vav1-Cre
Vav1-Cre
Tet2 fl/fl; Vav1-Cre
Vav1-Cre

Figure 18

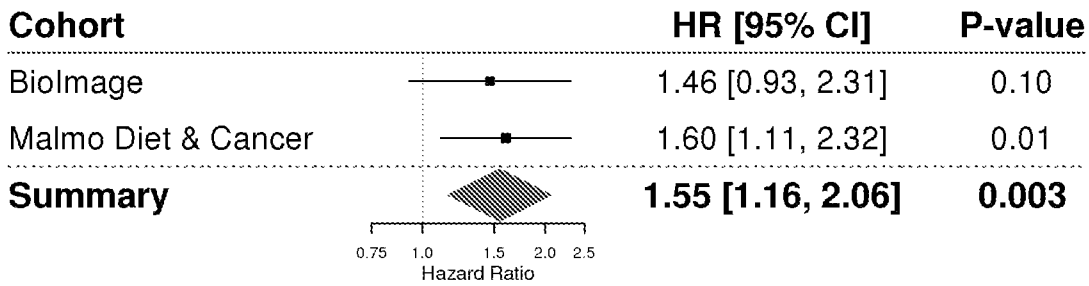
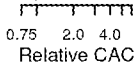

Cohort
HR [95% CI]
P-value
BioImage
1.46 [0.93, 2.31]
0.10
Malmo Diet & Cancer
1.60 [1.11, 2.32]
0.01
Summary
1.55 [1.16, 2.06]
0.003
0.75
1.0
1.5
2.0
2.5
Hazard Ratio

Figure 19

Carrier Status		Relative CAC [95% CI]	P-value	CAC Median [IQR]
Non-Carrier		1		154 [14, 535]
Carrier		2.60 [1.23, 5.47]	0.01	310 [91, 864]



INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/062787

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, EMBASE, FSTA, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	AXEL MUENDLEIN ET AL: "Occurrence of the JAK2 V617F mutation in patients with peripheral arterial disease", AMERICAN JOURNAL OF HEMATOLOGY, vol. 90, no. 1, 26 October 2014 (2014-10-26), pages E17-E21, XP055246813, US ISSN: 0361-8609, DOI: 10.1002/ajh.23874	1-25
Y	abstract; p. e17, para. "introduction"; p. e18, right-hand col., 4th para.	26
Y	US 2010/197518 A1 (XU HUICHUN [US] ET AL) 5 August 2010 (2010-08-05) para. 6, 18, 77, 122; p. 16, table 3	26

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 February 2016	04/05/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ripaud, Leslie

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/062787

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	SIDDHARTHA JAISWAL ET AL: "Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes", NEW ENGLAND JOURNAL OF MEDICINE, vol. 371, no. 26, 26 November 2014 (2014-11-26), pages 2488-2498, XP055246853, US ISSN: 0028-4793, DOI: 10.1056/NEJMoa1408617 the whole document -& SIDDHARTHA JAISWAL ET AL: "Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes - Online supplementary appendix", NEW ENGLAND JOURNAL OF MEDICINE, vol. 371, no. 26, 26 November 2014 (2014-11-26), pages 2488-2498, XP055246869, US ISSN: 0028-4793, DOI: 10.1056/NEJMoa1408617	1-26
A	AMÉLIE BONNEFOND ET AL: "Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications", NATURE GENETICS., vol. 45, no. 9, 14 July 2013 (2013-07-14), pages 1040-1043, XP055246874, NEW YORK, US ISSN: 1061-4036, DOI: 10.1038/ng.2700 cited in the application the whole document	1-26
A	MINGCHAO XIE ET AL: "Age-related mutations associated with clonal hematopoietic expansion and malignancies", NATURE MEDICINE., vol. 20, no. 12, 19 October 2014 (2014-10-19), pages 1472-1478, XP055246879, US ISSN: 1078-8956, DOI: 10.1038/nm.3733 the whole document	1-26
A	JP 2005 151854 A (JAPAN SCIENCE & TECH AGENCY) 16 June 2005 (2005-06-16) the whole document	1-26

1

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/062787

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

25, 26(completely); 1-24(partially)

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 25, 26(completely); 1-24(partially)

   concerns a method of identifying a subject with increased risk of developing a cardiometabolic disease and optionally a hematological cancer comprising sequencing at least part of a gene from blood cells and identifying the presence of at least one mutation in said gene, wherein said gene is DNMT3A.

   ---

2. claims: 27(completely); 1-24(partially)

   idem invention 1, wherein said gene is TET2.

   ---

3. claims: 28(completely); 1-24(partially)

   idem invention 1, wherein said gene is ASXL1.

   ---

4. claims: 29(completely); 1-24(partially)

   idem invention 1, wherein said gene is TP53.

   ---

5. claims: 1-24, 30(all partially)

   idem invention 1, wherein said gene is JAK2 and the mutation is N533D, N533Y or N533S.

   ---

6. claims: 1-24, 30(all partially)

   idem invention 1, wherein said gene is JAK2 and the mutation is H538R

   ---

7. claims: 1-24, 30(all partially)

   idem invention 1, wherein said gene is JAK2 and the mutation is K539E or K539L

   ---

8. claims: 1-24, 30(all partially)

   idem invention 1, wherein said gene is JAK2 and the mutation is I540T or I540V

   ---

9. claims: 1-24, 30(all partially)

   idem invention 1, wherein said gene is JAK2 and the mutation

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

is V617F

---

10. claims: 1-24, 30(all partially)

idem invention 1, wherein said gene is JAK2 and the mutation is R683S or R683G

---

11. claims: 1-24, 30(all partially)

idem invention 1, wherein said gene is JAK2 and the mutation is del/ins537---539L or del/ins538---539L

---

12. claims: 1-24, 30(all partially)

idem invention 1, wherein said gene is JAK2 and the mutation is del/ins540---543MK, del/ins540---544MK, del/ins541---543K, del542---543 or del543---544

---

13. claims: 1-24, 30(all partially)

idem invention 1, wherein said gene is JAK2 and the mutation is ins11546---547

---

14. claims: 31(completely); 1-24(partially)

idem invention 1, wherein said gene is SF3B1.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/062787

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010197518 A1	05-08-2010	CA 2685382 A1	13-11-2008
		CN 101688245 A	31-03-2010
		EP 2152907 A1	17-02-2010
		EP 2311981 A1	20-04-2011
		HK 1156368 A1	31-07-2015
		US 2010197518 A1	05-08-2010
		WO 2008137465 A1	13-11-2008
JP 2005151854 A	16-06-2005	NONE	