



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

(21) International Application Number:

PCT/EP2013/071065

(22) International Filing Date:

9 October 2013 (09.10.2013)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

12382391.6 11 October 2012 (11.10.2012) EP

(71) Applicants: **UNIVERSITAT DE BARCELONA** [ES/ES]; Centre de Patents de la UB, Baldri Reixac 4 - Torre D, E-08028 Barcelona (ES). **HOSPITAL CLÍNIC DE BARCELONA** [ES/ES]; Villaroel 170, E-08026 Barcelona (ES). **UNIVERSIDAD DE OVIEDO** [ES/ES]; San Francisco 2, E-33003 Oviedo (ES).

(72) Inventors: **MARTÍN SUBERO, José Ignacio**; Universitat De Barcelona, CEK- Hematologia Oncológica, Rosselló 149, 2n - Desp. 2.6, E-08036 Barcelona (ES). **CAMPO GÜERRI, Elías**; Hospital Clínic De Barcelona-Universitat De Barcelona, Anatomia Patológica, Villaroel 170, 5e, esc. 3, E-08036 Barcelona (ES). **LÓPEZ OTÍN, Carlos**; Universidad De Oviedo Otri, Edificio Severo Ochoa, pl. baja, Campus de "El Cristo", E-33006 Oviedo (ES).

(74) Agent: **ILLESCAS TABOADA, Manuel**; Manuel Illescas y Asociados, S.L. (MIA), Principe de Vergara 197, Oficina 1ªA, E-28002 Madrid (ES).

(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, LT, 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, 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))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: METHOD TO PREDICT THE CLINICAL EVOLUTION OF A PATIENT SUFFERING OF CHRONIC LYMPHO-CYTIC LEUKEMIA (CLL)

(57) Abstract: The present invention is a method of predicting the evolution of a patient suffering of CLL comprising the steps of determining in a biological sample obtained from said patient the DNA methylation level of site cg11472422 of the lymphocyte population of said sample, comparing the result of said methylation level with a standard wherein the methylation level is indicative of a CLL outcome subgroup, and predicting the evolution of said patient evaluating the result of the comparison. In a preferred embodiment, the method comprises additionally, the determination of the methylation level of at least one site selecting from any of the following: cg03462096, cg17014214, cg00869668, cg09637172 or any combination thereof.



**METHOD TO PREDICT THE CLINICAL EVOLUTION OF A PATIENT SUFFERING  
OF CHRONIC LYMPHOCYTIC LEUKEMIA (CLL)**

5    **Field of the invention**

The present invention is comprised within the field of diagnostics in biomedicine, and particularly in relation to the establishment and detection of new prognostic groups in CLL based on epigenetic biomarkers.

10

**Background of the invention**

Chronic lymphocytic leukemia (CLL) is the most common leukemia in human adults. The disease corresponds to a heterogeneous B cell neoplasm, with two well-described clinical and molecular subtypes related to the mutational status of the immunoglobulin heavy chain variable region (*IGHV*). CLL patients bearing high level of *IGHV* somatic hypermutation (M-CLL, less than 98% identity with germline *IGHV*) have a favourable clinical outcome, whereas CLLs with no or low level of *IGHV* mutational load (U-CLL, more than 98% identity) follow a worse evolution.

20

The detection of the *IGHV* mutational status is typically performed by PCR, sequencing and sequence homology analysis, which is labour intensive and expensive. This analysis is ideally performed with mRNA of the CLL sample, which not always is available in good quality; alternatively, the test is performed on DNA from the CLL sample, decreasing the reliability and reproducibility of this type of analysis. Further, the art uses a 98% threshold to classify CLLs into M-CLL or U-CLL, which is problematic because it may lead to misclassification of CLLs with a germline-*IGHV* identity slightly over or under 98%.

25

DNA methylation is a major mechanism involved both in normal cell differentiation and neoplastic transformation. Alterations of this epigenetic mark have been described as potential diagnostic or prognostic parameters in patients of various types of cancer. With respect to CLL, several publications have reported the identification of DNA methylation biomarkers and signatures associated with prognosis.

35

Kanduri et al. in 2010 ("*Differential genome-wide array-based methylation profiles in prognostic subsets of chronic lymphocytic leukemia*". *Blood* 2010; 115: 296-305)

analyzed the DNA methylation profiles of U-CLL and M-CLL, performed with 27k Infinium Illumina methylation arrays. These authors identified 15 genes differentially methylated between these prognostic groups of CLL. It is important to note that the specific CpGs regions disclosed in the present invention to classify CLL samples based on an epigenetic signature, are not included in the 27k infinium Illumina methylation array used by Kanduri et al. Therefore, the CpG sites of the invention could not be identified by Kanduri et al.

International Application WO2006/008650A2 describes the ZAP70 gene methylation status for use in the determination of prognostic groups in CLL. The expression of ZAP70 in CLL is a prognostic factor that divides patients into good prognosis (ZAP70 silenced, mostly present in M-CLL) and worse prognosis (ZAP70 expressed, mostly present in U-CLL). Focusing on the teachings of this application, a recent study by Claus and coworkers (*"Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia"*. *J Clin Oncol* 2012; 30: 2483-91) identifies the DNA methylation status of a unique CpG site in the promoter region of ZAP70 as predictor of the prognosis in CLL patients. Claus and colleagues replace the protein expression by a more reliable method based on DNA methylation to detect CLLs with different prognosis.

Shinawi et al. (*"KIBRA gene methylation is associated with unfavourable biological prognostic parameters in chronic lymphocytic leukemia"*. *Epigenetics* 2012; 7: 211-5) reports that DNA methylation in the KIBRA gene is related to poor prognosis in CLL, mostly due to its presence in the U-CLL group. This document does not disclose any of the CpG sites proposed by the present invention to classify CLL samples based on an epigenetic signature.

Besides, Irving and colleagues (*"Methylation markers identify high risk patients in IGHV mutated chronic lymphocytic leukemia"*. *Epigenetics* 2011; 6: 300-6) identified that the methylation status of CD38, HOXA4 and BTG4 can collectively predict prognosis in CLL, and that this methylation status was independent from the IGHV mutational status in a multivariate statistical analysis.

The International Application WO 2012/031329 A1 describes a method for identifying an epigenetic profile in the genome of a cell in a subject indicative of a cancerous condition or predisposition such as leukemia. The method comprises screening for a

change relative to a control in the epigenetic profile within or proximal to a locus selected from a list of genes associated with the following cell functions: cell fate commitment, transcription factors, DNA binding, subcellular location, and transcription regulator activity. Contrary to this, the present application's method is not based on the detection of epigenetic differences between leukemia samples and controls. Instead, the inventor's innovation is based on the fact that CLL samples can be classified into prognostic groups based on their epigenetic identity with particular subsets of normal B cells. The group of CLL with poor prognosis has come to be similar to normal naive B cells whereas the group of CLL with favourable prognosis, similar to memory B cells.

Other authors, instead of analysing the methylation status of specific genes, have focused on the analysis of the genome as a whole. WO 2007/067695 A2 describes the use of global DNA methylation as a method to decide on the diagnosis, prognosis and treatment of patients suffering from leukemia, including leukemia of lymphoid origin or myeloid origin. The disclosed methods include determining whether leukemia cells of the patient exhibit DNA hypermethylation compared to total cytosine in genomic DNA when compared to normal cells of one or more individuals having similar demographic parameters such as age or sex. The publication also includes additional determinations related to treatment, diagnosis or prognosis of patients before, during or after treatment, such as determining the white blood cell counts, presence of markers Zap-70 or CD38, presence of chromosomal alterations, modification of the histone proteins, measure of the expression levels of several microRNAs, analysis of the level of expression of different tumor suppressor genes, as well as measuring beta-2-microglobulin. Overall, these methods compile the analysis of the global DNA hypermethylation to several other markers not related to the methylation status.

Most of these articles and patents share the common feature of differentiating CLLs into those two above mentioned groups of different prognosis based on the *IGHV* mutational status, namely U-CLL and M-CLL.

The problem of the art is then to find an alternative epigenetic mark for the prognosis of CLL with a higher sensibility than the currently available tests. The solution proposed by the present invention is the identification of new DNA methylation signatures related to the cell of origin of different CLLs, specifically based on the methylation level of specific CpG sites. These signatures are able to classify CLLs not only into two groups with poor and favourable prognosis, but also the identification of a novel third group of the disease with intermediate prognosis.

### **Description of the invention**

The strategies of the art considering the assessment of the methylation of the genome as a whole or the identification of aberrant methylation levels of particular genes are not used in the approach made by the present invention. The main difference proposed in the present application is that the inventors do not aim to detect DNA methylation differences between tumors and healthy control samples. Instead, a novel strategy to classify CLL samples based on an epigenetic signature related to different cells of origin is proposed. The epigenetic profile of CLLs with poor prognosis points towards a naive B cell as the original cell-type of this neoplasm. In contrast, CLLs with favourable prognosis seem to be originated from more mature B cells, such as memory B cells. Thus, the inventors apply biomarkers which methylation level is related to normal naive or memory B cells in order to detect naive-like CLLs (poor prognosis) or memory cell-like CLLs (favourable prognosis), respectively. This strategy has surprisingly achieved a remarkable increase in sensibility. Indeed, the inventors report the presence of a third group of CLL with intermediate prognosis, which may be related with a third type of cell of origin different from naive and memory B cells.

According to this, the present invention is a method to predict the clinical evolution of a patient suffering of CLL comprising the steps of determining in a biological sample obtained from said patient the DNA methylation level of site **cg11472422 (SEQ ID No. 16)** of the lymphocyte population of said sample, comparing the result of said methylation level with a standard wherein the methylation level is indicative of a CLL outcome subgroup, and predicting the evolution of said patient evaluating the result of the comparison. In a preferred embodiment, the method comprises the further determination of the methylation level of at least an additional site selected from: **cg03462096 (SEQ ID No. 17)**, **cg17014214 (SEQ ID No. 18)**, **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)** or any combination thereof, of the lymphocyte population of said sample in the first step and comparing the result of the methylation level of the five sites on with the procedure. In a more preferred embodiment, the method of the invention comprises the determination of the methylation level of the combination of sites **cg11472422 (SEQ ID No. 16)**, **cg03462096 (SEQ ID No. 17)**, **cg17014214 (SEQ ID No. 18)**, **cg00869668 (SEQ ID No. 19)** and **cg09637172 (SEQ ID No. 20)**.

Particularly, the above mentioned method is preferred to determine the DNA methylation level in the cytosines of CpG sites. The specific cytosines whose methylation level has been analysed are showed in bold and underlined in Table 4.

- 5 In the scope of the present application, the term “patient” is defined as a mammal subject, preferably human, that has been independently diagnosed as suffering from leukemia or lymphoma by any conventional diagnosing method.

- 10 In the scope of the present application, CLL refers to Chronic Lymphocytic Leukemia involving any lymphocyte, including but not limited to various developmental stages of B cells and T cells.

- 15 For the purposes of present patent specification, the term “clinical evolution” refers to the development of a disease and its determination is based on the assessment of clinical symptoms and biological parameters that influence clinical interventions. As a consequence, the term “clinical evolution” refers to the assessment in a subject suffering from that disease, of the clinical status, progression, or regression of the aforesaid disease, and/or prognosis of disease’s course in the future.

- 20 As used herein, the term “favourable clinical outcome” means that the patients do not show, in addition to a hemogram with an abnormal population of monoclonal B cells, clinical symptoms as fatigue, autoimmune hemolytic anemia, infections, splenomegaly, hepatomegaly, lymphadenopathy, extranodal infiltrates, or any other clinical symptom typically associated with the disease. This favourable prognostic group frequently  
25 remains untreated for prolonged periods of time, of a median of 8 years after diagnosis, and show a median survival of about 20 years. On the contrary, the poor prognostic group of CLL presents clinical symptoms at an earlier time, with a median time to treatment of 2 years after diagnosis and a median survival time of 8 years.

- 30 In the scope of the present application, the term “DNA methylation” refers to methylation of cytosine residues such as 5-methylcytosine, which may include cytosine residues within CpG sites or outside CpG sites. It is particularly preferred to analyse the methylation level of the cytosine residue within CpG sites which is showed in bold and underlined in Table 4. As used herein, the term “methylation level” means the  
35 methylation average of cytosine in the context of a CpG dinucleotide of DNA. The methylation level is a quantitative variable ranging from 0 for completely unmethylated cytosines and 1 to fully methylated cytosines.

In the scope of the present application, the term "biological sample" is defined as a sample naturally occurring or extracted from a patient in which cellular or genetic material is contained. In a very preferred embodiment of the present invention, these terms refer to peripheral blood lymphocytes, tissue containing B cells or extracted B cells. Within the context of the present invention, the specimen or sample may be used in a crude form, a preserved form (i.e. includes additional additives commonly added to preserve the integrity of the cellular material under environmental stress, such as freezing), a partially purified form, a purified form (e.g., isolated cellular material), or any other common preparatory form.

In a particular embodiment of the present invention, said biological sample is a blood sample or a tissue sample, preferably a lymph node sample or a peripheral blood lymphocyte sample, said peripheral blood lymphocyte sample being most preferred a B-cell purified sample.

In the scope of the present application, the term "peripheral blood lymphocyte sample" is a sample comprising isolated mononuclear blood cells typically isolated by cell density centrifugation, which may contain T cells, B cells, NK cells and monocytes. As compared to peripheral blood samples from healthy donors, CLL patients show a clearly increased proportion of B cells, in some cases reaching almost 100% of the peripheral blood lymphocyte sample.

In the scope of the present application, the term "B-cell purified sample" is a sample with a B-cell content exceeding 70%, either because the peripheral blood lymphocyte sample of the patient contains large proportion of B cells or because the B cells of the CLL sample have been purified by any B cells isolation method, typically by magnetic or fluorescent cell sorting.

Throughout this specification, unless the context requires otherwise, the word "comprise" or variations such as "comprises" or "comprising", will be understood to imply the inclusion of a stated element or integer or method step or group of elements or integers or method steps but not the exclusion of any element or integer or method step or group of elements or integers or method steps. The open term "comprise" and variants thereof will include, for the purposes of present patent specification, therefore, the closed term "consist".

In another preferred embodiment of the present invention, the lymphocyte population from which the methylation level of a site is determined, is a B-cell purified population.

5 In a further preferred embodiment, the standard against which the result of the methylation level of a site is compared, is a B-cell DNA methylation standard related to naive and memory B cells.

The inventors have extensively characterized the DNA methylome of CLLs with mutated or unmutated *IGHV* as well as several mature B cell subpopulations using a  
10 combined approach. The DNA methylome of few samples (n=5) has been first investigated by whole-genome bisulfite sequencing, and then a large series of samples (n=139 CLLs and 26 normal B cell subpopulations) has been analysed by high-density DNA methylation microarrays.

15 Based on the results of the whole genome bisulfite sequencing of CLL samples and controls, the inventors identified that DNA hypomethylation targeting non-promoter regions is the most frequent DNA methylation change both in CLLs as compared to normal B cells, and also in different types of normal B cell subpopulations. To capture the epigenomic heterogeneity of CLL and describe possible clinical associations, the  
20 inventors analysed a large series of 139 CLLs DNA microarrays. Various types of B cells from peripheral blood of age-matched healthy donors were also tested. All the 139 CLLs and control samples contained at least 95% purified B cells.

An unsupervised principal component analysis (PCA) of microarray data indicated that  
25 the DNA methylome of the two molecular subgroups M-CLL and U-CLL is clearly distinct. Comparing the DNA methylome of U-CLL vs. M-CLL, 3,265 differentially methylated CpGs were identified, being 2,091 methylated only in U-CLL and 1,174 methylated only in M-CLL. Hierarchical clustering of these 3,265 CpGs indicated that U-CLL were epigenetically similar to CD5-NBC/NBCs whereas the DNA methylome of  
30 M-CLL resembled both types of memory B cells (MBC). From those CpGs, it was observed that 1,649/3,265 CpGs (50.5%) had similar DNA methylation levels in U-CLL vs. CD5-NBC/NBC and in M-CLL vs. MBC ( $p < 10^{-300}$ ). These data indicate that U-CLLs are epigenetically related to pre-germinal center B cells such as CD5+NBC/NBC whereas M-CLLs are epigenetically similar to germinal center-experienced B cells such  
35 as MBC.

One of the key aspects of the study performed by the inventors is that a DNA methylation signature of 1,649 CpGs is able to define the epigenetically-related cell of origin of CLL subtypes. Interestingly, a consensus clustering analysis with 10,000 permutations of this 1,649 CpG-signature revealed three rather than two groups of CLLs (Figure 1). One group was related to naive B cells and was made mostly of U-CLLs (mean *IGHV* identity of 99.5%). A second group was related to memory B cells and was made mostly of M-CLL with high load of somatic *IGHV* mutation (mean *IGHV* identity of 92.7%) The third consensus cluster showed an intermediate DNA methylation profile between naive and memory B cells and was enriched for M-CLLs with a significantly lower level of somatic *IGHV* mutations (mean *IGHV* identity of 96.1%).

Remarkably, these three groups of CLL identified by their epigenetic signature showed significantly different clinical biological features (Figures 2, 3). Patients in stage A with a MBC-like pattern had a favourable evolution with 20% requiring treatment at 10 years, whereas this proportion was 43% and 100% for patients with intermediate and NBC-like patterns, respectively. A multivariate Cox regression model indicated that this DNA methylation signature ( $p = 3 \times 10^{-4}$ ), CD38 expression ( $p = 0.015$ ) and LDH levels ( $p = 0.047$ ) were the only parameters with independent prognostic value in CLL. The putative cell of origin of the CLL recognized by the DNA methylation imprint is revealed as a relevant factor affecting the clinical outcome of this disease. This finding should offer a new biological approach for the stratification of CLL patients and facilitate the choice of therapeutic strategies.

The DNA methylation pattern of the 1,649 CpGs related with the cell of origin has a high level of redundancy. Therefore, the inventors sought to extract few regions with the ability to classify CLL cases into one of the three novel prognostic groups. Using bioinformatics for strategy, the inventors detected 5 CpG regions that can reliably classify CLL cases into NBC-like CLL, MBC-like CLL and intermediate CLL (Figure 4, Tables 1 and 2).

The method disclosed in the present invention was validated by bisulfite pyrosequencing (BPS) in two independent large series of CLL patients ( $n=211$ , Hospital Clinic, Barcelona, Spain;  $n=97$ , University of Leicester, UK). The inter-laboratory reproducibility of the novel BPS assays and prediction model based on epigenetic biomarkers disclosed in the present invention was also tested.

The method of the invention can be performed by different techniques well known in the art. One embodiment is a method that comprises a polymerase chain reaction, preferably a quantitative polymerase chain reaction. Another embodiment is a method comprising a DNA sequencing reaction, preferably a pyrosequencing reaction. In a preferred embodiment of the invention, said quantitative polymerase chain reaction or sequencing reaction is performed in the presence of a probe specific for the DNA sequence of a leukocyte obtained from a CLL patient. In another preferred embodiment, the DNA of the patient is treated with sodium bisulfite.

Another particular embodiment for performing the method of the invention comprises methylation-specific melting curve analysis. In another preferred embodiment, the method comprises methylation-sensitive restriction enzymes. In a further embodiment, the method comprises mass spectrometry (e.g. based on Sequenom technology) and still in another preferred embodiment the method of the invention comprises immunoprecipitation of methylated DNA.

The most preferred embodiment for the performance of the method of the invention comprises methylation-specific microarrays.

Indeed, a very preferred embodiment of the invention is an array for use in the diagnosis of CLL in a human patient, comprising a probe suitable to detect the DNA methylation of site **cg11472422 (SEQ ID No. 16)** in the lymphocyte genome of a biological sample obtained from said patient. Preferably, said array further comprises probes suitable to detect the DNA methylation of the combination of site **cg11472422 (SEQ ID No. 16)** with at least one additional site selected from: **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)**, **cg03462096 (SEQ ID No. 17)**, **cg17014214 (SEQ ID No. 18)** or any combination thereof. In the Table 3 there are disclosed suitable probes to detect the DNA methylation site of five biomarkers (CpGs sites) as disclosed in the present invention. In a preferred embodiment, the array of the invention comprises the probes suitable to detect the DNA methylation of the sites: **cg11472422 (SEQ ID No. 16)**, **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)**, **cg03462096 (SEQ ID No. 17)** and **cg17014214 (SEQ ID No. 18)**.

In this sense, Bibikova M. and colleagues ("High density DNA methylation array with single CpG site resolution". *Genomic, Academic Press*. 2011. 4:288-295) provide a genome-wide DNA methylation BeadChip which allows high-throughput methylation profiling of the human genome. This high density array can assay over 480k CpG sites

and covers the 99% of RefSeq genes with a multiple probes per gene, 96% of CpG regions from the UCSC database, CpG regions shores and additional content selected from the whole-genome bisulfite sequencing data and input from DNA methylation experts. But it is important to note that in the Bibikova's study neither the CLL is  
5 analysed nor the specific CpGs sites with diagnostic and/or prognostic value are disclosed. Therefore, Bibikova and colleagues do not identify the 5 CpGs sites disclosed in the present invention as biomarkers with clinical value neither in CLL nor in any other disease

## 10 **Brief Description of the Figures**

**Figure 1.** A consensus clustering analysis of CLL cases after 10,000 permutations indicates that the 1,649-CpG signature classifies CLL cases in 3 optimal clusters. The figure shows the proportion of times that a particular case belongs to each cluster after  
15 10,000 permutations.

**Figure 2** shows that the 3 different CLL epigenetic groups have a different clinical progression. In particular, the Kaplan-Meier plot using time to treatment as clinical variable indicates that NBC-like CLLs are more adverse than MBC-like CLLs and that intermediate CLLs clinically behave in between the other two groups.

20 **Figure 3** summarizes the presence of known CLL-related prognostic factors in the 3 new groups of CLL cases based on the epigenetic signature.

**Figure 4** shows the classification accuracy of cases into NBC-like, MBC-like and intermediate CLLs using the 5 CpGs sites of the invention.

25 **Figure 5** shows examples of pyrograms generated by bisulfite pyrosequencing (BPS) using primers to detect DNA methylation levels of the 5 CpG regions used in the present invention. These pyrograms indicate that DNA methylation levels of the CpGs of interest (arrows) can be accurately detected by pyrosequencing in negative and positive controls using the primers shown in Table 3.

30 **Figure 6** shows in scatter plots that the DNA methylation levels of the five biomarkers (CpG sites) of the invention quantified by the novel BPS assays are highly correlated with those detected by 450k microarrays. CpG 1: cg00869668 (SEQ ID No. 19); CpG 2: cg03462096 (SEQ ID No. 17); CpG 3: cg09637172 (SEQ ID No. 20); CpG 4: cg11472422 (SEQ ID No. 16) and CpG 5: cg17014214 (SEQ ID no. 18).

35 **Figure 7.** High inter-laboratory reproducibility of the BPS assays developed for the epigenetic classification of CLL disclosed in the present invention. Scatter plots showing that the DNA methylation values generated by two independent laboratories using 19 CLL cases are highly correlated. CpG 1: cg00869668 (SEQ ID No. 19); CpG

2: cg03462096 (SEQ ID No. 17); CpG 3: cg09637172 (SEQ ID No. 20); CpG 4: cg11472422 (SEQ ID No. 16) and CpG 5: cg17014214 (SEQ ID no. 18).

**Figure 8.** The DNA methylation levels of the five biomarkers of the invention remain stable over time. (A-F) Correlation of DNA methylation values in CLL samples obtained at different time-points (i.e. at diagnosis and a median of 59 months later). (G) Heatmap showing DNA methylation values of the 5 biomarkers (CpG sites) of the invention in 27 CLL patients in which sequential samples were studied (the letters "a" to "c" below the case number point to samples at different time points).

**Figure 9.** Analysis of two series of CLL with the method and classification disclosed in the present invention, into n-CLL, i-CLL or m-CLL. The heatmap in panel A shows the DNA methylation values of the five biomarkers (CpG regions) of the invention in the training series of 211 patients. The lower part of the panel shows the results of the linear discriminant analysis (LDA) prediction model and points to the probability of each sample to belong to each epigenetic subgroup. Panel B shows a similar graphical representation for the validation series of 97 CLL patients. Panel C illustrates that in both training and validation series, n-CLL, i-CLL and m-CLL show a different distribution of *IGHV* somatic mutations.

**Figure 10.** Differences in the distribution of *IGHV* mutation levels in the new CLL subgroups. (A) The distribution of the levels of *IGHV* somatic hypermutation is unimodal in all three CLL subgroups. In particular, the i-CLL subgroup is not made out of mut and unmut-CLLs but rather represents a group with a moderate level of somatic mutation. (B) (left) The percentage of identity to germline *IGHV* in patients with unmutated *IGHV* is significantly lower in i-CLL than in n-CLL (99.2% vs. 99.9%, respectively;  $p < 0.001$ ) (right) In contrast, the degree of identity to germline *IGHV* in patients with mutated *IGHV* is significantly higher in i-CLL than in m-CLL (95.3% vs. 92.6%, respectively;  $p < 0.001$ ). Mutated and unmutated *IGHV* sequences were defined as having  $< 98\%$  and  $\geq 98\%$  identity to germline *IGHV*, respectively.

**Figure 11.** Kaplan-Meier estimates of time to treatment (TTT) and overall survival (OS) in the training and validation CLL series according to the new epigenetic classification. In the training (panels A and B) and validation (panels C and D) series as well as in the two series merged (panels E and F), the three CLL subgroups show a distinct clinical behavior. Using TTT and OS as end-points, the n-CLL subgroup is clinically adverse, the m-CLL subgroup is clinically favorable and the i-CLL subgroup shows a clinical behavior in between n-CLL and m-CLL. The Kaplan-Meier plots for TTT as end-point only include patients at Binet stage A.

## **Examples**

The following examples are included as support of the claimed subject matter, in no case limiting the scope of the present invention.

5

### **Example 1: Description of samples: CLL cases and normal B cell subpopulations**

139 CLL patients were selected for DNA methylation with Illumina Infinium HumanMethylation450 BeadChip high-density DNA microarrays (Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288-95, 2011). Among these patients, which correspond to a population-based cohort, 57 had IGHV-unmutated genes, 76 IGHV-mutated genes (<98% identity). In 6 more cases the IGHV status could not be unambiguously determined. All patients gave informed consent for their participation in the study.

15

After Ficoll-Isopaque density centrifugation CD19+ B cells from CLL samples were isolated by positive magnetic cell separation by using a magnetic-activated cell sorting (AutoMACS system, Milteny Biotec, Auburn, CA). DNA was extracted from these purified cells by using a QIAamp kit (Qiagen) and the DNA quality was assessed by SYBR-green staining on agarose gels and quantified using Nanodrop ND-100 spectrophotometer. The tumor DNA samples for DNA methylation microarrays contained  $\geq 95\%$  neoplastic cells.

Control samples were obtained from buffy coats from healthy adult donors of age-matched controls. After Ficoll-Isopaque density centrifugation CD19+ B cells were isolated by positive magnetic cell separation by using AutoMACS system (Milteny Biotec, Auburn, CA). To isolate different B cell subpopulation, CD19+ cells were labelled with various mAb combinations for 15 min at room temperature in staining buffer (PBS with 0.5% BSA, Sigma). Naive B cells (CD19+/CD27-/IgD+), non-class-switched memory B cells (CD19+/CD27+/IgM+/IgD+) and class-switched memory B cells (CD19+/CD27+/IgA+ or IgG+) were obtained by Fluorescence-Activated Flow Sorting (FACS) on FACSAriaII (BD Biosciences) after labelling with anti-CD27 APC (BD Biosciences, at final concentration 0.3125  $\mu\text{g/ml}$ ), anti-IgD PE-Cy7 (BD Biosciences), anti-IgM PE (BD Biosciences), anti-IgG FITC (BD Biosciences) and anti-IgA FITC (DakoCytomation). CD5+ naive B cells (CD19+/CD5+/CD27-/CD38<sup>low</sup>) were obtained by FACS sorting after labelling with anti-CD5 PE (BD Biosciences), anti-CD27 PerCP-Cy5.5 (BD Biosciences) and anti-CD38 APC (BD Biosciences). The average

35

purity of the control samples was 98%. For DNA methylation analysis, we used the following 26 samples: 14 samples of total B cells (CD19+), 3 CD5+ naïve B cells (CD5+NBC), 3 naïve B cells (NBC), 3 class-switched memory B cells (csMBC) and 3 non-class-switched memory B cells (ncsMBC)

5

### **Example 2: Microarray-based DNA methylation analysis with 450k arrays**

We used the EZ DNA Methylation kit (Cat# 5004 from Zymo Research, CA, USA) for bisulfite conversion of genomic DNAs. For optimized results, we used 500 ng of gDNA and followed the manufacturer's recommendations. Namely, genomic DNA was denatured by addition of NaOH-containing M-Dilution buffer and incubated for 15 min at 37°C. Freshly prepared CT-conversion reagent containing sodium bisulfite was added to the denatured DNA and samples were incubated for 16 h at 50°C in a thermocycler and denatured every 60 min by heating to 95°C for 30 s. After bisulfite conversion, the DNA was bound to a Zymo-Spin™ I-96 Binding Plate, washed with M-Wash Buffer and desulphonated on the binding plate using M-desulphonation reagent. The bisulfite-converted DNA was eluted from the plate wells in 10 ul elution buffer.

The Infinium Methylation assay was carried out as referenced in Example 1. In brief, 4 ul of converted product (~150 ng) were used in the whole-genome amplification (WGA) reaction. After amplification, the DNA was fragmented enzymatically, precipitated and re-suspended in hybridization buffer. All subsequent steps were performed following the standard Infinium protocol (User Guide part #15019519 A). Fragmented DNA was dispensed onto the HumanMethylation450 BeadChips, and hybridization performed in hybridization oven for 20 h. After hybridization, the array was processed through a primer extension and an immunochemistry staining protocol to allow detection of a single-base extension reaction. Finally, BeadChips were coated and then imaged on an Illumina iScan.

### **Example 3: Data normalization, quality control and filtering of microarray data**

Data from the 450k Human Methylation Array was analyzed by GenomeStudio (Illumina Inc.) and R ([www.r-project.org](http://www.r-project.org)) using the Lumi package (Du, P. et al. Lumi: a pipeline for processing Illumina microarray. Bioinformatics; 2008; 24: 1547-8,) available through Bioconductor. To exclude technical and biological biases that might produce false results in the further analyses, we developed and optimized an analysis pipeline with several filters.

From 485,577 targets, after eliminating those that represents SNPs (65 targets), we filtered out 1,010 quality CpGs with poor detection  $P$  values ( $P > 0.01$ ) in at least 10% of the samples. To remove inter-array variation we normalized all samples using the quantile normalization method of the lumi R package.

5

As the 450k Human Methylation Array applies two distinct types of chemistry (Infinium I and Infinium II) with slightly different performance, they had to be normalized separately. In the case of Infinium I, which employs two types of beads per each locus (for unmethylated and methylated state) and the intensity of differentially labeled nucleotides is measured, a colour adjustment step is added prior the normalization while for Infinium II, where only one type of beads is used, normalization was performed directly. Next, to remove sex-dependent differential methylation, we excluded 8,894 CpGs on X chromosome with a difference of mean beta values between men and women in controls above 0.1 and statistically significant (Wilcoxon test, FDR<0.05). 68 CpGs on chromosome Y were also filtered out. Finally, we excluded non-informative CpGs, e.g. completely unmethylated or completely methylated in all CLLs and controls. To determine the threshold from which we could consider the locus as methylated or not, we plotted the beta values densities of all samples. As there were slight differences in the distribution of beta values between Infinium I and Infinium II assays, for Infinium I assays we considered CpGs with beta values over 0.8 to be methylated and below 0.2 to be unmethylated whereas for Infinium II the threshold were 0.75 and 0.25, respectively.

Finally, we also included one additional filter and excluded all those non-informative CpGs showing a homogeneous DNA methylation pattern in samples and controls (i.e. standard deviation below 0.1). After this step, 194,753 CpGs were used in further analyses.

#### **Example 4: Detection of epigenetic similarities between CLL subtypes and normal B cells**

30

To identify CpGs differentially methylated between U-CLL and M-CLL that could be explained by an epigenetic imprint of normal B cells, we analysed in the control samples (CD5-NBC/NBC vs. MBC) the methylation values of 3,265 CpGs differentially methylated in U-CLL vs. M-CLL. We used the following step-wise analyses:

35

1. We initially excluded from the analysis all CpGs that did not show any DNA methylation differences between the controls (CD5-NBC/NBC vs. MBC). This

was done to exclude those CpGs in which differential methylation was not related to an imprint of normal B cells but rather to de novo changes specific for U-CLL or M-CLL. Following this strategy, from the initial 3,265 CpGs we selected 2,224 that show significant differences between the controls (FDR<0.05;  $|\Delta\beta|>0.25$ ).

2. Identify similarities between U-CLL or M-CLL and normal B cells. Using 2,224 CpGs from step 1, we performed the following comparisons using a Wilcoxon test:

- a. U-CLL vs. CD5-NBC/NBC,
- b. M-CLL vs. MBC,
- c. U-CLL vs. MBC and
- d. M-CLL vs. CD5-NBC/NBC.

To select CpGs with similar DNA methylation values in each CLL subtype and different normal cell types, we took only those CpGs that did not show any significant differences in each of the 4 comparisons (FDR>0.05;  $|\Delta\beta|<0.25$ ). CpGs that could not be unambiguously assigned to any of the 4 lists were excluded (n = 166). Our data clearly indicate that U-CLL and M-CLL are similar to CD5-NBC/NBC and MBC, respectively (Fisher's exact test  $p < 10^{-300}$ ).

#### **Example 5: Definition of new CLL groups based on consensus clustering**

The number of optimal clusters to classify CLL cases and the consistency of hierarchical cluster analysis to define epigenetic groups in CLL was tested using the package 'clusterCons' in R/Bioconductor. The consensus clustering result was calculated from re-sampled clustering experiments using the following parameters:

- a. average linkage clustering,
- b. Manhattan distance metric,
- c. 10,000 iterations and
- d. a sub-sampling ratio of 0.8.

The results of this analysis are shown in Figure 1.

**Example 6: Prognostic impact of DNA methylation profiles in CLL**

The clinical relevance of the candidate prognostic factors in Binet A CLL patients (n=104 cases with complete dataset) was evaluated using multivariate Cox regression for time to treatment (TTT) using the R Package (version 2.12.1, [www.r-project.org](http://www.r-project.org)). We initially explored a model including the following known prognostic factors as well as the three epigenetic groups identified in this study: lymphocyte count (quantitative), LDH levels (low vs. high), CD38 expression (quantitative), ZAP70 expression (quantitative), mutational IGHV status (mutated vs. unmutated) and 11q deletions (absent vs. present). We stepwise removed the variable with highest P value until we reached a final model in which all the variables were significant. To determine the validity of this fitted Cox regression model, we demonstrated that:

- a. each variable meets the assumption of proportional Hazards (each covariate,  $P > 0.05$ ),
- b. none of the observations is clearly influential individually, and
- c. nonlinearity was ruled out by plotting the martingale residuals against each covariate.

The results of this multivariate analysis demonstrate that the identified DNA methylation signature is the most relevant parameter with independent prognostic impact in CLL ( $P = 3 \times 10^{-4}$ ).

**Example 7: Detection of epigenetic biomarkers to identify the new prognostic groups in CLL**

Considering that the 1649-CpG signature that classifies CLLs into new prognostic groups shows a high level of redundancy, we sought to identify few biomarkers with the power to correctly classify CLL cases. We used the following analytic strategy:

1. We used DNA methylation values of the initial epigenetic signature (1649 CpGs) in a series of 133 cases. According to this signature, cases were classified as Naive B cell-like CLL (NBC-CLL, n=53), Memory B cell-like CLL (MBC-CLL, n=58) or Intermediate CLL (Int-CLL, n=22) using a consensus clustering function (R software).

2. Using the limma package of the R software, we performed 3 different comparisons of DNA methylation data: a) NBC-CLL vs. MBC-CLL+Int-CLL, b) NBC-CLL+Int-CLL vs. MBC-CLL and c) Int-CLL vs. MBC-CLL.

3. Data from each comparison were ranked according to the FDR value.

4. We selected the first 3 CpG from each comparison (a total of 9 CpG regions). We performed a linear discriminant analysis (LDA) with the R software using the Modern Applied Statistics with S (MASS) package available through Bioconductor. For this analysis we used with those 9 CpGs to calculate CpGs that allow us to classify CLLs with the highest confidence. In particular, we performed LDAs of all possible combinations to identify the minimum number of CpGs that correctly classify all the 133 CLL samples.

5. Process outlined in point 4 was repeated until 5 CpG regions were extracted from the initial 1649-CpG signature (Table 1).

**Table 1.** Set of 5 CpGs regions with high classification accuracy.

| Region ID             | SEQ ID No. | Gene Name | Gene-related region | Chromosome | Location (hg19) |
|-----------------------|------------|-----------|---------------------|------------|-----------------|
| cg03462096<br>(CpG 1) | 17         | -         | Intergenic          | 14         | 56157197        |
| cg09637172<br>(CpG 2) | 20         | TNF       | Gene body           | 6          | 31545252        |
| cg11472422<br>(CpG 3) | 16         | B3GNTL1   | Gene body           | 17         | 80926014        |
| cg17014214<br>(CpG 4) | 18         | CTBP2     | Gene body           | 10         | 126713638       |
| cg00869668<br>(CpG 5) | 19         | SCARF1    | Promoter region     | 17         | 1549012         |

Based on these analyses, we used the "predict" function of the LDA analysis for the 5 CpG regions to obtain the probability of each CLL patient to belong to each of the 3 described groups of CLL. Using this approach, we could identify that the methylation levels of each CpGs allowed to classify all the cases with 100% accuracy (Figure 4).

The following table includes the median DNA methylation values for each of the 5 CpG regions in each of the 3 groups of CLLs.

**Table 2.** Average methylation levels of 5 CpGs regions on the invention.

| Region ID             | SEQ ID No. | MBC-CLL<br>(median(range)) | Int-CLL<br>(median(range)) | NBC-CLL<br>(median(range)) |
|-----------------------|------------|----------------------------|----------------------------|----------------------------|
| cg03462096<br>(CpG 1) | 17         | 0.16 (0.11-0.59)           | 0.19 (0.12-0.58)           | 0.85 (0.59-0.92)           |
| cg09637172<br>(CpG 2) | 20         | 0.08 (0.04-0.42)           | 0.20 (0.06-0.87)           | 0.90 (0.06-0.95)           |
| cg11472422<br>(CpG 3) | 16         | 0.16 (0.09-0.89)           | 0.57 (0.14-0.89)           | 0.86 (0.57-0.90)           |
| cg17014214<br>(CpG 4) | 18         | 0.17 (0.11-0.70)           | 0.81 (0.25-0.87)           | 0.85 (0.35-0.88)           |
| cg00869668<br>(CpG 5) | 19         | 0.82 (0.23-0.89)           | 0.33 (0.08-0.82)           | 0.13 (0.06-0.80)           |

5

**Example 8: Validation of the CpG-classifier by bisulfite pyrosequencing (BPS)**

A total of 500 ng of genomic DNA from positive DNA methylation control (CpGenome Universal Methylated DNA, Millipore) and negative DNA methylation controls (either normal naive or memory B cells in which 450k methylation arrays revealed lack of methylation for a particular CpG region) were treated with sodium bisulfite using the EpiTect Plus Bisulfite Conversion Kit (Qiagen) following the manufacturer's instructions. Bisulfite converted DNA was eluted and subjected to PCR amplification of the specific region by using forward (F) and reverse (R) primer sets shown in Table 3 (5 CpG regions). The primers were designed using PyroMark Assay Design software 2.0 (Qiagen). One of the primers was biotin labelled (Table 3). The PCR reaction was performed using the PyroMark Gold Q96 kit. The PCR product was checked by 2.0 % agarose gel electrophoresis to confirm the quality and the size of the product. The specific PCR products were then subjected to quantitative pyrosequencing analysis using a PyroMark Q96 ID (Qiagen) and sequencing (S) primers shown in Table 3 according to manufacturer's instructions. The results were analysed by PyroMark CpG SW 1.0 Software (Qiagen). Negative controls show an average methylation level of 11% whereas positive controls showed an average of 98% methylation (Figure 5). These data confirm the validity of the primers to quantify methylation levels in DNA samples.

25

**Table 3.** List of validated bisulfite pyrosequencing (BPS) primers designed to detect DNA methylation levels in each of the 5 CpG regions of the invention. The DNA methylation level is performed in the specific cytosine within CpG regions (see Table 4).

5

| Primer name    | Sequence     | Modification |
|----------------|--------------|--------------|
| cg17014214-F   | SEQ ID NO:1  |              |
| cg17014214-R   | SEQ ID NO:2  | biotinylated |
| cg17014214-Seq | SEQ ID NO:3  |              |
| cg00869668-F   | SEQ ID NO:4  |              |
| cg00869668-R   | SEQ ID NO:5  | biotinylated |
| cg00869668-Seq | SEQ ID NO:6  |              |
| cg11472422-F   | SEQ ID NO:7  | biotinylated |
| cg11472422-R   | SEQ ID NO:8  |              |
| cg11472422-Seq | SEQ ID NO:9  |              |
| cg09637172-F   | SEQ ID NO:10 |              |
| cg09637172-R   | SEQ ID NO:11 | biotinylated |
| cg09637172-Seq | SEQ ID NO:12 |              |
| cg03462096-F   | SEQ ID NO:13 |              |
| cg03462096-R   | SEQ ID NO:14 | biotinylated |
| cg03462096-Seq | SEQ ID NO:15 |              |

**Table 4.** DNA sequences of 5 epigenetic biomarkers (CpGs region) disclosed in the present invention. The cytosine showed in bold and underlined is the target cytosine whose methylation level is measured in each CpG region.

10

| Region ID             | SEQ ID No. | DNA Sequence                                                                                                                              |
|-----------------------|------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| cg03462096<br>(CpG 1) | 17         | TCAAACCTCTGCATCTAGTTTTCTCCTTGCTAGCTAT<br>GTAAATAAAATTAAAATGTGAC[CG]TGTTAGTCATTTCT<br>TGCATTGCTAAAAAGAAATACCCATGGCAGGGTAATG<br>TATAAGGAA   |
| cg09637172<br>(CpG 2) | 20         | ATCTATCTGGGAGGGGTCTTCCAGCTGGAGAAGGGT<br>GACCGACTCAGCGCTGAGATCAAT[CG]GCCCCGACTAT<br>CTCGACTTTGCCGAGTCTGGGCAGGTCTACTTTGGGA<br>TCATTGCCCTGTG |
| cg11472422<br>(CpG 3) | 16         | CGCAGCCAGCCGTTCCAGTCCTAGACATTTACCCAAG<br>AAAAATGCAAACACACGTCCATC[CG]GCACCATCCAT<br>GTGTGTTACGGCAGCCCCAGCACCCGTCCCCGTGT<br>GTTACAGGCAGTC   |
| cg17014214<br>(CpG 4) | 18         | CGATGGCAGAGAACTTAAGGTGGTCTCTGCAGGGTT<br>GTGAGGTCCAGCGCCTCGGTCCC[CG]CCTACAGCTG<br>CTTCAGGGCTATTTTCAAAAAGGTAGATCTTATTGTTT<br>CATCCTGATTTT   |
| cg00869668<br>(CpG 5) | 19         | TCCAGAGCAGCAGCAGCGGGAGCAGCAGCCCCAGC<br>CCCATGGCAGGCAGCTCGGTGGGAG[CG]CTCGGGTT<br>CGTCTGGCCCCACAGCTCCCAACCCCTCCCTGCTT<br>CCTCTCAGGGCTTTT    |

#### Example 9: Reproducibility of the CpG-classifier of the present invention.

- 5 The accuracy of the new method to predict the clinical evolution of a patient suffering of CLL disclosed in the present invention was tested comparing microarray and BPS methylation values in 15 CLL cases of each epigenetic subgroup (total n=45), and the results were highly comparable (average Pearson coefficient of 0.987 ( $p < 0.001$ ) for each CpG region, **Figure 6**). The LDA-based prediction model using BPS data allowed
- 10 the correct classification of all these 45 cases to their corresponding epigenetic CLL subgroup.

Moreover, the inventors tested the inter-laboratory reproducibility of the new BPS assays and epigenetic prediction model. A total of 19 CLL samples were blindly processed in two different laboratories, Barcelona (Spain) and Kiel (Germany),

15 including cell separation from different aliquots of the same tumor sample, DNA isolation and BPS. It was observed that the DNA methylation levels derived from the five biomarkers (CpGs region) disclosed in the present invention and the epigenetic classification of the cases as n-CLL, m-CLL or i-CLL using the LDA prediction model were highly reproducible between both institutions (Pearson correlation coefficient of

20 0.976, and weighted Cohen's kappa coefficient of 0.923,  $p < 0.001$ , respectively, **Figure 7**).

Additionally, a prerequisite to evaluate the clinical impact of the epigenetic classification based on the five biomarkers (CpGs regions) of the present invention is that the DNA methylation levels remain stable over time. To analyze this aspect, the inventors selected samples from 27 CLL patients with a median sampling difference between samples of 59 months (range, 5 to 114) and obtained at two or three different time points. Data showed that DNA methylation values of the 5 CpG regions of the invention in the sequential samples were highly concordant (Pearson correlation coefficient of 0.984,  $p < 0.001$ ). This finding confirms that the DNA methylation levels of the 5 biomarkers of the invention and the epigenetic subgroup do not change during the course of the disease (**Figure 8**).

**Example 10: Validation of the epigenetic classification based on five biomarkers (CpGs region) methylation levels of the invention in a large series of CLL patients.**

Once we confirmed the accuracy and reproducibility of the method disclosed in the present invention, the inventors applied the method of the invention to a large series of 211 CLL patients (training series) which comprised 84 women and 127 men with a median age of 61 years (range, 34 to 86). These cases are included in the CLL genome project of the International Cancer Genome Consortium (ICGC). In addition, we tested the inter-laboratory reproducibility of the BPS assays and epigenetic prediction model based on five epigenetic biomarkers disclosed in the present invention in a validation series of 97 CLL patients (validation series) comprised of 31 women and 66 men with a median age of 65 years (range, 33 to 91) from an independent Institution (University of Leicester, UK).

The main clinico-biological features of patients from the training and validation series are shown in **Table 5**. The diagnosis of CLL was established according to the WHO criteria. Clinical and biological data at diagnosis, treatment and follow-up were recorded for further statistical analysis. Patients were treated according to the criteria from the International Workshop on CLL (IWCLL).

The data obtained in the analysis of training series of CLL patients with the method disclosed in the present invention based on the methylation levels of five biomarkers disclosed herein classified 90 cases (43%) as n-CLL, 28 (13%) as i-CLL, and 93 (44%) as m-CLL (**Figure 9 A**). The clinico-biological features of the CLL epigenetic subgroups are listed in **Table 6**. The percentage of cases with unmutated *IGHV* was 97% for n-CLL, 25% for i-CLL and 3% for m-CLL ( $p < 0.001$ ). Moreover, the levels of *IGHV* somatic

mutation in these epigenetic subgroups were clearly different, i.e. the mean percentage of identity to *IGHV* germline was 99.7% for n-CLL, 96.3% for i-CLL and 92.8% for m-CLL ( $p < 0.001$ , **Figure 9C**).

5 **Table 5.** Clinico-biological characteristics and outcome of patients from the training and validation series.

| Parameter                        | Category        | Training series  | Validation series | P Value            |
|----------------------------------|-----------------|------------------|-------------------|--------------------|
| Gender                           | Male/Female     | 127/84 (60%)     | 66/31 (68%)       | ns                 |
| Age (years)*                     | 61 (34-86)      | 65 (33-91)       | 0.003             |                    |
|                                  | A               | 196 (93%)        | 71 (73%)          |                    |
| Binet stage                      | B               | 12 (6%)          | 16 (17%)          | $6.10^{-6}$        |
|                                  | C               | 3 (1%)           | 10 (10%)          |                    |
|                                  | 0               | 138 (66%)        | 47 (49%)          |                    |
| Rai stage                        | I-II            | 68 (32%)         | 40 (41%)          | 0.001              |
|                                  | III-IV          | 5 (2%)           | 10 (10%)          |                    |
| Lymphocytes ( $\times 10^9/L$ )* | 13 (2.6-111.3)  | 16.1 (2.8-610)   | 0.007             |                    |
| Hemoglobin (g/L)*                | 140 (45-175)    | na               |                   |                    |
| Platelets ( $\times 10^9/L$ )*   | 203 (92-470)    | na               |                   |                    |
| LDH                              | >UNL            | 17/200 (9%)      | 21/75 (28%)       | $1 \times 10^{-4}$ |
| Beta <sub>2</sub> microglobulin  | >UNL            | 49/186 (26%)     | na                |                    |
| <i>IGHV</i>                      | Unmutated**     | 97/211 (46%)     | 43/92 (47%)       | ns                 |
| <i>IGHV</i> germline identity    | Percent         | 96.2 ( $\pm 4$ ) | 96.0 ( $\pm 4$ )  | ns                 |
| CD38                             | $\geq 30\%$     | 51/194 (26%)     | 20/77 (26%)       | ns                 |
| ZAP-70                           | $\geq 20\%$     | 56/190 (30%)     | na                |                    |
|                                  | Del(13q)(q14.3) | 86/199 (43%)     | 44/95 (46%)       | ns                 |

| Parameter              | Category        | Training series | Validation  |                    |
|------------------------|-----------------|-----------------|-------------|--------------------|
|                        |                 |                 | series      | P Value            |
| Genetics               | Del(11q)(q22.3) | 25/197 (13%)    | 17/94 (18%) | ns                 |
|                        | Trisomy 12      | 32/200 (16%)    | 18/90 (20%) | ns                 |
|                        | Del(17p)(p13.1) | 7/198 (4%)      | 2/93 (2%)   | ns                 |
| <i>NOTCH1</i>          | Mutated         | 25/211 (12%)    | 10/80 (12%) | ns                 |
| <i>SF3B1</i>           | Mutated         | 17/208 (8%)     | 12/83 (14%) | ns                 |
| <i>MYD88</i>           | Mutated         | 3/204 (2%)      | na          |                    |
| <i>TP53</i>            | Mutated         | 6/144 (4%)      | na          |                    |
| Untreated patients     | 108/211 (51%)   | 41/96 (43%)     | ns          |                    |
| 10-year TTT (95% CI)   | Binet stage A   | 52% (44-60)     | 55% (41-69) | ns                 |
| 10-year DLBCL (95% CI) | All             | 9% (4-14%)      | na          |                    |
| 10-year OS (95% CI)    | All             | 77% (69-85%)    | 60% (49-71) | $8 \times 10^{-5}$ |

\* expressed as median (range), \*\*  $\geq 98\%$  identity with germline *IGHV*

Analyzing the *VH* usage of each subgroup we observed that the most frequent rearrangements were *VH1-69* in n-CLL (27%), *VH4-34* in m-CLL (20%) as well as *VH3-23* (14%) and *VH1-18* (11%) in i-CLL (**Table 6**). *NOTCH1* mutations were mainly observed in the n-CLL subgroup whereas *SF3B1* mutations were more frequently observed in i-CLL. Moreover, a significant difference in Binet and Rai stages at diagnosis, with early stages more frequently seen in m-CLL, was detected.

Data clearly show that n-CLL, i-CLL and m-CLL represent different subgroups of the disease. Additionally, the inventors further analyzed the specific features of the i-CLL cases. This subgroup contained cases showing a unimodal distribution of moderately mutated *IGHV* (**Figure 10**), particular *VH* usage (11% *VH1-18* vs. 2% in n-CLL and m-CLL,  $P=0.05$ ), as well as a higher incidence of *SF3B1* mutations (21% vs. 6% in n-CLL and m-CLL,  $P=0.01$ ) and other non-hematological neoplasms (8/28 (29%) vs. 22/183 (12%) in n-CLL and m-CLL,  $P=0.04$ ). Collectively, these data suggest that the i-CLL subgroup represents a distinct clinico-biological CLL subgroup rather than a mere mixture of n-CLL and m-CLL cases.

**Table 6.** Main clinico-biological characteristics of the training series (n=211) according to the epigenetic classification.

| Parameter                             | Category       | n-CLL<br>(n=90) | i-CLL<br>(n=28) | m-CLL<br>(n=93) | P Value               |
|---------------------------------------|----------------|-----------------|-----------------|-----------------|-----------------------|
| Gender                                | Male/Female    | 59/31 (66%)     | 16/12 (57%)     | 52/41 (56%)     | ns                    |
| Age (years)*                          | 60 (36-86)     | 58 (34-82)      | 63 (36-82)      | ns              |                       |
|                                       | A              | 78 (87%)        | 26 (93%)        | 92 (99%)        |                       |
| Binet stage                           | B              | 10 (11%)        | 1 (4%)          | 1 (1%)          | 0.02                  |
|                                       | C              | 2 (2%)          | 1 (4%)          | 0 (0%)          |                       |
|                                       | 0              | 51 (57%)        | 13 (46%)        | 74 (80%)        |                       |
| Rai stage                             | I-II           | 35 (39%)        | 14 (50%)        | 19 (20%)        | 0.002                 |
|                                       | III-IV         | 4 (4%)          | 1 (4%)          | 0 (0%)          |                       |
| Lymphocytes<br>(x10 <sup>9</sup> /L)* | 13.6 (2.6-111) | 11.7 (3.8-46)   | 13.1 (4-80)     | ns              |                       |
| Hemoglobin (g/L)*                     | 142 (86-166)   | 138 (45-167)    | 141 (113-175)   | ns              |                       |
| Platelets (x10 <sup>9</sup> /L)*      | 200 (92-470)   | 210 (111-456)   | 203 (100-376)   | ns              |                       |
| LDH                                   | >UNL           | 12/85 (14%)     | 3/27 (11%)      | 2/88 (2%)       | 0.02                  |
| Beta <sub>2</sub> microglobulin       | >UNL           | 31/77 (40%)     | 7/26 (27%)      | 11/83 (13%)     | 0.001                 |
| IGHV mutational<br>status             | Unmutated**    | 87/90 (97%)     | 7/28 (25%)      | 3/93 (3%)       | 7.0x10 <sup>-37</sup> |
| IGHV identity                         | Percent        | 99.7 (±1.0)     | 96.3 (±2.5)     | 92.8 (±3.1)     | 4.7x10 <sup>-48</sup> |
|                                       | VH1-18         | 2/89 (2%)       | 3/28 (11%)      | 2/93 (2%)       | 0.06                  |
| VH usage                              | VH1-69         | 24/89 (27%)     | 2/28 (7%)       | 1/93 (1%)       | 1.0x10 <sup>-6</sup>  |
|                                       | VH3-23         | 2/89 (2%)       | 4/28 (14%)      | 9/93 (10%)      | 0.04                  |
|                                       | VH4-34         | 4/89 (4%)       | 0/28 (0%)       | 19/93 (20%)     | 3.7x10 <sup>-4</sup>  |
| CD38                                  | ≥ 30%          | 38/80 (48%)     | 6/26 (23%)      | 7/88 (8%)       | 4.2x10 <sup>-8</sup>  |
| ZAP-70                                | ≥ 20%          | 48/79 (61%)     | 2/25 (8%)       | 6/86 (7%)       | 1.5x10 <sup>-14</sup> |

| Parameter                 | Category        | n-CLL<br>(n=90) | i-CLL<br>(n=28) | m-CLL<br>(n=93) | P Value               |
|---------------------------|-----------------|-----------------|-----------------|-----------------|-----------------------|
| CD49d                     | ≥ 30%           | 16/28 (57%)     | 4/9 (44%)       | 8/50 (16%)      | 0.001                 |
| Genetics                  | Del(13q)(q14.3) | 27/81 (33%)     | 10/27 (63%)     | 42/91 (46%)     | 0.02                  |
|                           | Del(11q)(q22.3) | 20/79 (25%)     | 3/27 (11%)      | 2/91 (2%)       | 3.6x10 <sup>-5</sup>  |
|                           | Trisomy 12      | 21/82 (26%)     | 2/27 (7%)       | 9/91 (10%)      | 0.008                 |
|                           | Del(17p)(p13.1) | 4/80 (5%)       | 1/27 (4%)       | 2/91 (2%)       | ns                    |
| NOTCH1                    | Mutated         | 20/90 (22%)     | 1/28 (4%)       | 4/93 (4%)       | 3.1x10 <sup>-4</sup>  |
| SF3B1                     | Mutated         | 8/88 (9%)       | 6/28 (21%)      | 3/92 (3%)       | 0.008                 |
| MYD88                     | Mutated         | 0/87 (0%)       | 1/28 (4%)       | 2/89 (2%)       | ns                    |
| TP53                      | Mutated         | 2/54 (4%)       | 1/20 (5%)       | 3/70 (4%)       | ns                    |
| Untreated patients        | All             | 22/90 (24%)     | 16/28 (57%)     | 70/93 (75%)     | 4.3x10 <sup>-11</sup> |
| 10-year TTT<br>(95%CI)    | Binet stage A   | 97% (91-100)    | 42% (20-64)     | 22% (12-32)     | 4.6x10 <sup>-18</sup> |
| 10-year DLBCL<br>(95% CI) | All             | 18% (7-29)      | 6% (0-17)       | 0%              | 6.9x10 <sup>-4</sup>  |
| 10-year OS<br>(95%CI)     | All             | 53% (39-67)     | 85% (69-100)    | 95% (89-100)    | 9.3x10 <sup>-9</sup>  |

\* expressed as median (range); \*\* ≥ 98% identity with germline IGHV; ns: not significant; del: deletion; UNL: upper normal value; TTT: time to treatment; DLBCL: diffuse large B cell lymphoma; OS: overall survival; 95% CI: 95% confidence interval.

- 5 Potential prognostic factors for two end-points, i.e. time to treatment (TTT) and overall survival (OS), were evaluated in multivariate analyses using Cox proportional hazard regression. The qualitative covariates with more than one category were converted into a set of dummy variables omitting one category (Klein JP, Moeschberger ML. Survival analysis. Techniques for censored and truncated data. New York: Springer Verlag; 10 1997). The assumption of proportional hazard was tested for each variable using a graphical and a mathematical method. If such assumption could not be confirmed, the potential prognostic factors were entered as time-dependent variables by creating interactions of the predictors and a function of survival time. Factors significant at the

0.05 level were kept in the final model. These computations were made using the statistical package IBM-SPSS Statistics version 20.

On the one hand, the prognostic value of the new epigenetic classification was analyzed together with the Binet clinical stage and the *IGHV* mutational status in the training series of 211 CLL patients as well as in a validation series of 97 CLL patients. On the other hand, to evaluate the prognostic value of mutations in *NOTCH1* and *SF3B1*, these factors together with a broad group of well-known prognostic variables were also tested in the Cox proportional hazard regression. Backward stepwise Wald selection of factors which were significant at the 0.05 level in univariate analysis for both end-points was used. The following ten variables were tested: age <70 vs. ≥70 years, sex female vs. male, epigenetic classification, Binet clinical stage, *IGHV* mutational status present vs. absent, *NOTCH1* mutation present vs. absent, *SF3B1* mutation present vs. absent, adverse cytogenetics [del(11q) and/or del(17p) present vs. absent], LDH normal vs. high and CD38 positive cells <30 vs. ≥30%. Such analysis was performed in a group of 308 CLL patients collected from both training and validation series.

Surviving training series CLL patients had a median follow-up of 7.9 years (range, 0.4 to 24), with 103 patients (49%) requiring treatment. Median TTT for n-CLL, i-CLL and m-CLL patients in Binet stage A was 3.4 years, 12.3 years and non-reached, respectively ( $p < 0.001$ , **Figure 11A**). Other variables predicting shorter TTT were unmutated *IGHV* ( $p < 0.001$ ), high ZAP-70 ( $p < 0.001$ ), high CD38 ( $p < 0.001$ ), mutated *NOTCH1* ( $p < 0.001$ ), mutated *SF3B1* ( $p < 0.001$ ), elevated LDH ( $p < 0.002$ ), elevated  $\beta 2M$  ( $p < 0.003$ ), and adverse cytogenetics [del(11q) and/or del(17p)] ( $p < 0.001$ ), advanced Binet stage ( $p < 0.001$ ), and advanced Rai stage ( $p < 0.001$ ). A multivariate Cox analysis demonstrated the independent value of the epigenetic classification to predict TTT (**Table 7**). Treatment given to the patients did not significantly differ among the epigenetic subgroups.

Thirty-seven training series CLL patients died during the follow-up, with a 10-year OS of 53%, 85% and 95% for n-CLL, i-CLL and m-CLL, respectively ( $p < 0.001$ , **Figure 11B**). Other variables predicting shorter OS were age ≥70 years ( $p < 0.001$ ), advanced Binet stage ( $p < 0.001$ ), advanced Rai stage ( $p < 0.001$ ), unmutated *IGHV* ( $p < 0.001$ ), high CD38 ( $p < 0.001$ ), high ZAP-70 ( $p = 0.001$ ), adverse cytogenetic aberrations ( $p < 0.001$ ), mutated *SF3B1* ( $p < 0.001$ ), elevated LDH ( $p = 0.01$ ), and elevated  $\beta 2M$  ( $p < 0.001$ ). In the multivariate analysis, Binet stage and *IGHV* mutational status retained the independent value for prognosis OS (**Table 7**).

Finally, 12 CLL patients developed transformation to DLBCL, with a 10-year risk of cumulative incidence of transformation of 18%, 6% and 0% for n-CLL, i-CLL and m-CLL, respectively ( $p < 0.001$ ).

- 5 **Table 7.** Cox proportional hazard regression analysis performed in 211 patients collected from the training series.

| Variable                            |         | Relative risk | 95% CI       | Wald Chi-square | Degrees of freedom | P Value               |
|-------------------------------------|---------|---------------|--------------|-----------------|--------------------|-----------------------|
| <b>End-point: Time to treatment</b> |         |               |              |                 |                    |                       |
| Epigenetic subgroup                 | overall |               |              | 65.156          | 2                  | $7.1 \times 10^{-15}$ |
|                                     | ESG (1) | 8.536         | 5.032-14.548 | 63.252          | 1                  | $1.8 \times 10^{-5}$  |
|                                     | ESG (2) | 2.572         | 1.274-5.190  | 6.949           | 1                  | 0.008                 |
| Binet stage                         | overall |               |              | 37.293          | 2                  | $8.0 \times 10^{-9}$  |
|                                     | BS (1)  | 0.145         | 0.044-0.478  | 16.065          | 1                  | 0.002                 |
|                                     | BS (2)  | 1.277         | 0.345-4.731  | 0.134           | 1                  | 0.714                 |
| <b>End-point: Overall survival</b>  |         |               |              |                 |                    |                       |
| IGHV mutated                        |         | 0.105         | 0.040-0.278  | 20.665          | 1                  | $5.5 \times 10^{-6}$  |
| Binet stage                         | overall |               |              | 24.176          | 2                  | $5.6 \times 10^{-6}$  |
|                                     | BS (1)  | 0.060         | 0.017-0.217  | 18.395          | 1                  | $1.8 \times 10^{-5}$  |
|                                     | BS (2)  | 0.352         | 0.083-1.494  | 2.003           | 1                  | 0.157                 |

ESG: Epigenetic subgroup. BS: Binet Stage.

Codification of categorical variables ESG and BS:

- 10 ESG(1): n-CLL=1, i-CLL=0, m-CLL=0; ESG(2): n-CLL=0, i-CLL=1, m-CLL=0  
BS(1): A=1, B=0, C=0; BS(2): A=0, B=1, C=0

#### **Example 11: Validation of the epigenetic classification in an independent series of patients (validation series)**

15

The clinico-biological characteristics and the prognostic value of the epigenetic classification disclosed in the present invention were assessed in an independent

series of 97 CLL patients from a different geographical origin (University of Leicester, UK). The simplified BPS analysis and LDA prediction model was applied to this validation series of patients and classified 36 cases (37%) as n-CLL, 17 (17%) as i-CLL, and 44 (45%) as m-CLL (**Figure 9B**). As shown in **Table 8**, the great majority of the clinico-biological features of the epigenetic subgroups observed in the training series were confirmed in this validation set. Among those, an identical distribution of *IGHV* mutational levels, with a mean identity to germline of 99.3% in n-CLLs, 96.7% in i-CLLs and 93% in m-CLLs was observed ( $p < 0.001$ , **Figure 9C**). In addition, a similar impact of the epigenetic classification was observed on TTT (10-year TTT of 92% vs. 54% vs. 14% for n-CLL, i-CLL and m-CLL, respectively;  $p < 0.001$ ; **Figure 11C**) and on OS (10-year OS of 28% vs. 65% vs. 83% for n-CLL, i-CLL and m-CLL, respectively;  $p < 0.001$ ; **Figure 11D**). In the validation series, the following variables predicted shorter TTT in the univariate analysis: advanced Rai stage ( $p < 0.001$ ), advanced Binet stage ( $p < 0.001$ ), unmutated *IGHV* ( $p < 0.001$ ), mutated *NOTCH1* ( $p = 0.001$ ), elevated LDH ( $p = 0.007$ ), and adverse cytogenetic aberrations ( $p = 0.001$ ). In the multivariate analysis, Binet stage and epigenetic subgroup showed independent prognostic value for TTT (**Table 9**). In the validation series, the univariate analysis of factors predicting worse OS were: age  $\geq 70$  ( $p < 0.001$ ), advanced Rai stage ( $p < 0.001$ ), advanced Binet stage ( $p < 0.001$ ), unmutated *IGHV* ( $p < 0.001$ ), high CD38 ( $p = 0.003$ ), adverse cytogenetic aberrations ( $p = 0.007$ ), and mutated *SF3B1* ( $p = 0.023$ ). In the multivariate analysis, Binet stage and *IGHV* retained the independent value for prognosis (**Table 9**).

**Table 8.** Main clinico-biological characteristics of the validation series (n=97) according to the epigenetic subgroup.

| Parameter                          | Category         | n-CLL<br>(n=36) | i-CLL<br>(n=17) | m-CLL<br>(n=44)      | P Value               |
|------------------------------------|------------------|-----------------|-----------------|----------------------|-----------------------|
| Gender                             | Male/Female      | 29/7 (81%)      | 11/6 (65%)      | 26/18 (59%)          | ns                    |
| Age (years)*                       | 64 (47-87)       | 61 (33-80)      | 66(41-91)       | ns                   |                       |
| Binet                              | A                | 20 (56%)        | 12 (71%)        | 39 (89%)             | 0.008                 |
|                                    | B                | 8 (22%)         | 4 (24%)         | 4 (9%)               |                       |
|                                    | C                | 8 (22%)         | 1 (6%)          | 1 (2%)               |                       |
| Rai stage                          | 0                | 12 (33%)        | 5 (29%)         | 30 (68%)             | 0.001                 |
|                                    | I-II             | 16 (44%)        | 11 (65%)        | 13 (30%)             |                       |
|                                    | III-IV           | 8 (22%)         | 1 (6%)          | 1 (2%)               |                       |
| Lymphocytes (x10 <sup>9</sup> /L)* | 50.6 (3.3-610)   | 9.0 (3.8-138)   | 13.3 (2.8-404)  | 3.7x10 <sup>-4</sup> |                       |
| LDH                                | UNL              | 10/27 (37%)     | 5/14 (36%)      | 6/34 (18%)           | ns                    |
| IGHV                               | Unmutated*       | 32/33 (97%)     | 7/16 (44%)      | 4/43 (9%)            | 2.9x10 <sup>-13</sup> |
| IGHV identity                      | Percent          | 99.3 (±1.0)     | 96.7 (±2.6)     | 93.0 (±3.7)          | 1.2x10 <sup>-13</sup> |
|                                    | VH1-18           | 1/34 (3%)       | 1/17 (6%)       | 0/43 (0%)            | ns                    |
|                                    | VH1-69           | 12/34 (25%)     | 0/17 (0%)       | 3/43 (7%)            | 4.8x10 <sup>-4</sup>  |
|                                    | VH3-23           | 3/34 (9%)       | 1/17 (6%)       | 4/43 (9%)            | ns                    |
|                                    | VH4-34           | 1/34 (3%)       | 0/17 (0%)       | 9/43 (21%)           | 0.01                  |
| CD38                               | ≥30%             | 9/28 (32%)      | 2/14 (14%)      | 9/36 (26%)           | ns                    |
| Genetics                           | Del(13q) (q14.3) | 16/36 (44%)     | 8/17 (42%)      | 20/42 (46%)          | ns                    |
|                                    | Del(11q) (q22.3) | 14/36 (39%)     | 2/17 (12%)      | 1/41 (2%)            | 1.4x10 <sup>-4</sup>  |
|                                    | Trisomy 12       | 10/33 (30%)     | 4/16 (25%)      | 4/41 (10%)           | 0.08                  |
|                                    | Del(17p) (p13.1) | 1/35 (3%)       | 1/17 (6%)       | 0/41 (2%)            | ns                    |
| NOTCH1                             | Mutated          | 8/31 (26%)      | 1/11 (9%)       | 1/38 (3%)            | 0.014                 |
| SF3B1                              | Mutated          | 7/31 (23%)      | 3/14 (21%)      | 2/38 (5%)            | 0.09                  |
| Untreated patients                 | All              | 4/36 (11%)      | 4/17 (23%)      | 33/43 (77%)          | 6.8x10 <sup>-9</sup>  |
| 10-year TTT (95% CI)               | Binet stage A    | 92% (78-100)    | 54% (24-84)     | 14% (0-28)           | 1.4x10 <sup>-8</sup>  |
| 10-year OS (95% CI)                | All              | 28% (8-48)      | 65% (40-80)     | 83% (71-95)          | 3x10 <sup>-6</sup>    |

\* expressed as median (range), \*\* ≥ 98% identity with germline *IGHV*

**Table 9.** Cox proportional hazard regression analysis performed in 97 patients collected from the validation series.

| Variable                            |         | Relative risk | 95% CI         | Wald Chi-square | Degrees of freedom | P Value              |
|-------------------------------------|---------|---------------|----------------|-----------------|--------------------|----------------------|
| <b>End-point: Time to treatment</b> |         |               |                |                 |                    |                      |
| Epigenetic subgroup                 | overall |               |                | 25.804          | 2                  | $2.5 \times 10^{-6}$ |
|                                     | ESG (1) | 9.693         | 4.035 - 23.286 | 25.800          | 1                  | $3.7 \times 10^{-7}$ |
|                                     | ESG (2) | 4.494         | 1.774 - 11.388 | 10.035          | 1                  | 0.002                |
| Binet stage                         | overall |               |                | 17.691          | 2                  | $1.4 \times 10^{-4}$ |
|                                     | BS (1)  | 0.121         | 0.045 - 0.325  | 17.526          | 1                  | $2.8 \times 10^{-5}$ |
|                                     | BS (2)  | 0.233         | 0.082 - 0.661  | 7.493           | 1                  | 0.006                |
| <b>End-point: Overall survival</b>  |         |               |                |                 |                    |                      |
| <i>IGHV</i> mutated                 |         | 0.277         | 0.142 - 0.538  | 14.339          | 1                  | $1.5 \times 10^{-4}$ |
| Binet stage                         | overall |               |                | 15.478          | 2                  | $4.3 \times 10^{-4}$ |
|                                     | BS (1)  | 0.198         | 0.088 - 0.446  | 15.274          | 1                  | $9.3 \times 10^{-5}$ |
|                                     | BS (2)  | 0.410         | 0.162 - 1.034  | 3.570           | 1                  | 0.059                |

5

**Example 12: Epigenetic classification as the strongest independent prognostic factor for TTT in CLL**

After validating the clinico-biological significance of the epigenetic classification in the training and validation series of CLL patients, the inventors merged them in order to gain statistical power to define the main features predicting TTT and OS of the patients. Kaplan-Meier plots of TTT and OS for the complete series (training + validation patients) showed that the epigenetic classification has a major impact on CLL prognosis (**Figure 11E and 11F**). Additionally, we performed a Cox multivariate analysis including the following variables: gender, age, Binet stage, epigenetic classification, *IGHV* mutational status, adverse cytogenetic alterations, *NOTCH1* mutations, *SF3B1* mutations as well as CD38 and serum LDH values. The final model, including 308 patients, showed that the epigenetic signature related to the cellular origin of CLL was the most important variable to predict TTT, together with Binet stage, CD38 expression, LDH levels and *SF3B1* mutations (**Table 10**).

20

**Table 10.** Cox proportional hazard regression analysis performed in 214 patients collected from training and validation series with data available for all the independent variables.

| Variable                     |         | Relative risk | 95% CI         | Wald Chi-square | Degrees of freedom | P Value               |
|------------------------------|---------|---------------|----------------|-----------------|--------------------|-----------------------|
| End-point: Time to treatment |         |               |                |                 |                    |                       |
| Epigenetic subgroup          | overall |               |                | 47.744          | 2                  | $4.3 \times 10^{-11}$ |
|                              | ESG (1) | 6.345         | 3.735 – 10.779 | 46.711          | 1                  | $8.2 \times 10^{-12}$ |
|                              | ESG (2) | 2.267         | 1.150 - 4.473  | 5.580           | 1                  | 0.018                 |
| Binet stage                  | overall |               |                | 20.748          | 2                  | $3.1 \times 10^{-5}$  |
|                              | BS (1)  | 0.136         | 0.051 - 0.360  | 16.077          | 1                  | $6.1 \times 10^{-5}$  |
|                              | BS (2)  | 0.368         | 0.119 - 1.136  | 3.021           | 1                  | 0.082                 |
| CD38 high                    |         | 1.990         | 1.225 – 3.157  | 8.550           | 1                  | 0.003                 |
| SF3B1 mutated                |         | 1.890         | 1.074 – 3.325  | 4.873           | 1                  | 0.027                 |
| LDH high                     |         | 2.000         | 1.071 – 3.734  | 4.734           | 1                  | 0.030                 |

95% CI: 95% confidence interval.

ESG: Epigenetic subgroup. BS: Binet Stage. Codification of categorical variables ESG and BS:

ESG(1): n-CLL=1, i-CLL=0, m-CLL=0; ESG(2): n-CLL=0, i-CLL=1, m-CLL=0

BS(1): A=1, B=0, C=0; BS(2): A=0, B=1, C=0

Data disclosed in the present invention stated a clinically applicable strategy to track the cellular origin of CLL based on epigenetic biomarkers and identified three CLL subgroups. These CLL subgroups have distinct *IGHV* mutational load, different *VH* usage and varying proportions of somatic mutations in *NOTCH1* and *SF3B1*. This new categorization of patients also has a major clinical impact confirmed in both training and validation series. Out of an extensive DNA methylation signature, the invention disclosed specific biomarkers (CpGs region) with high classification power and developed assays with high inter-laboratory reproducibility. The methylation levels of these 5 biomarkers (CpG regions) (**Table 1**) are modulated during B-cell differentiation, where they seem to act as enhancers. In CLL, the varying methylation levels of these biomarkers are not translated into gene expression changes, showing that may not have a functional impact but represent a stable molecular mark of the cellular origin of each CLL subgroup.

The n-CLL subgroup has an epigenetic imprint of naive B cells and m-CLL of memory B cells. These two subgroups show a major overlap with unmut-CLL and mut-CLL, respectively. A major finding of our invention is the identification of i-CLLs as a third subgroup of CLLs, which comprises about 15% of all cases. Such category does not  
5 seem to be a grey zone between m-CLL and n-CLL, but a distinct subgroup with a differential cellular origin and well-defined clinico-biological characteristics.

Furthermore, we observed that epigenetics is the strongest predictor for TTT, among other well know prognostic factors (**Table 10**), overcoming the role of *IGHV*. The three  
10 epigenetic subgroups predict prognosis in CLL in a more accurate manner than the two subgroups based on *IGHV* mutational status.

In conclusion, the present invention has established a simplified and reproducible method to categorize CLL into three epigenetic subgroups. As these subgroups show  
15 differential biological features and clinical behavior, our data may form the basis for a new classification of CLL that utilizes epigenetic biomarkers to track the cellular origin.

**CLAIMS**

1. Method to predict the clinical evolution of a patient suffering of chronic lymphocytic leukemia, comprising the following steps:
  - 5 - to determine in a biological sample obtained from said patient the DNA methylation level of site **cg11472422 (SEQ ID No. 16)** of the lymphocyte population of said sample,
  - comparing the result of said methylation level with a standard wherein the methylation level is indicative of a chronic lymphocytic leukemia outcome subgroup, and
  - 10 - predicting the evolution of said patient evaluating the result of the previous step.
2. A method according to claim 1, comprising to determine additionally, the  
15 methylation level of at least one of the following CpG sites: **cg03462096 (SEQ ID No. 17)**, **cg17014214 (SEQ ID No. 18)**, **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)** or any combination thereof, of the lymphocyte population of said biological sample in the first step.
- 20 3. A method according to any of the claims 1 or 2, which comprises to determine the methylation level of **cg11472422 (SEQ ID No. 16)**, **cg03462096 (SEQ ID No. 17)**, **cg17014214 (SEQ ID No. 18)**, **cg00869668 (SEQ ID No. 19)** and **cg09637172 (SEQ ID No. 20)**, of the lymphocyte population of said biological sample.  
25
4. A method according to anyone of the preceding claims, in which said biological sample is a blood sample or a tissue sample.
5. A method according to claim 4, in which said tissue sample is a lymph node  
30 sample.
6. A method according to claim 4, in which said blood sample is a peripheral blood lymphocyte sample.
- 35 7. A method according to claim 6, in which said peripheral blood lymphocyte sample is a B-cell purified sample.

8. A method according to anyone of claims 1 to 7, in which said lymphocyte population is a B-cell purified population.

9. A method according to anyone of claims 1 to 8, in which said standard is a B-cell DNA methylation standard related to naive and memory B cells.

10. A method according to anyone of claims 1 to 9, comprising methylation-specific microarrays.

11. Array for diagnosis of chronic lymphocytic leukemia in a patient, comprising at least a probe selected from SEQ ID Nos: 7-9 suitable to detect the DNA methylation level of site **cg11472422 (SEQ ID No. 16)** in the lymphocyte genome of a biological sample from said patient.

12. An array according to claim 11, further comprising at least one of the probes selected from SEQ ID Nos: 4-6, 10-12, 13-15 and 1-3, suitable to detect additionally DNA methylation level of any of the sites **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)**, **cg03462096 (SEQ ID No. 17)** and **cg17014214 (SEQ ID No. 18)**, respectively.

13. An array according to any of the claims 11 or 12 comprising the probes selected from SEQ ID Nos 7-9, 4-6, 10-12, 13-15 and 1-3 suitable to detect the DNA methylation level of sites: **cg11472422 (SEQ ID No. 16)**, **cg00869668 (SEQ ID No. 19)**, **cg09637172 (SEQ ID No. 20)**, **cg03462096 (SEQ ID No. 17)** and **cg17014214 (SEQ ID No. 18)**, respectively.

Figure 1

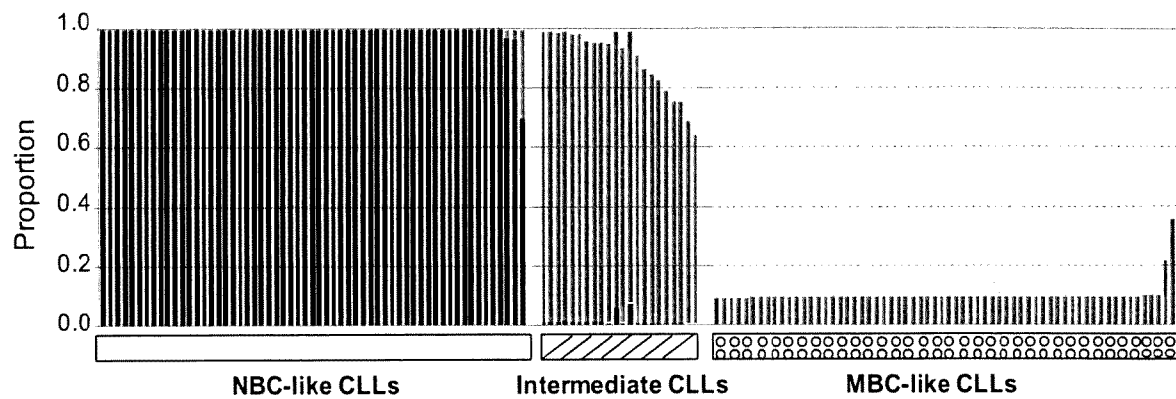


Figure 2

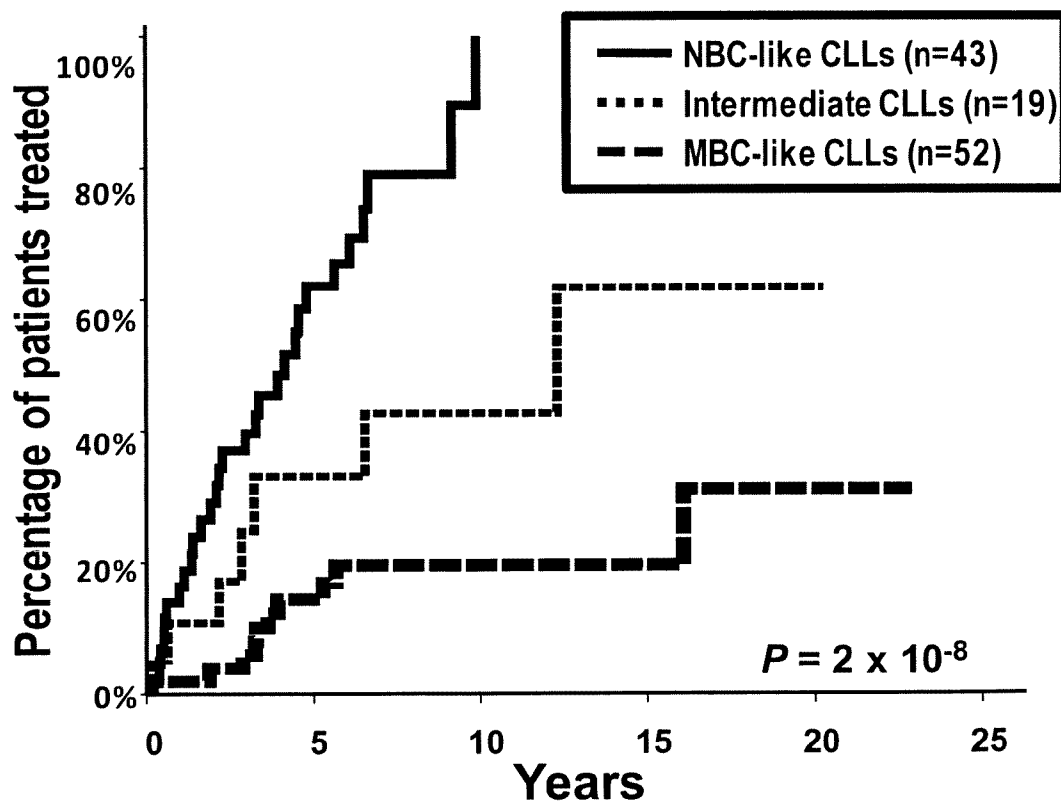


Figure 3

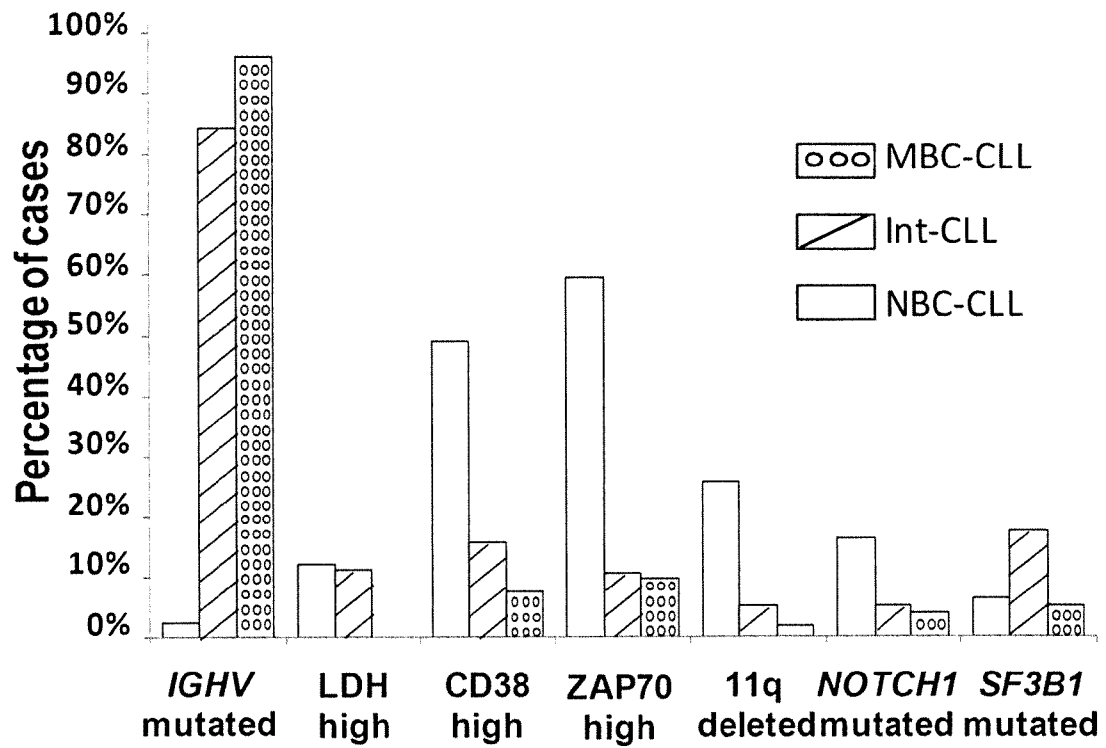
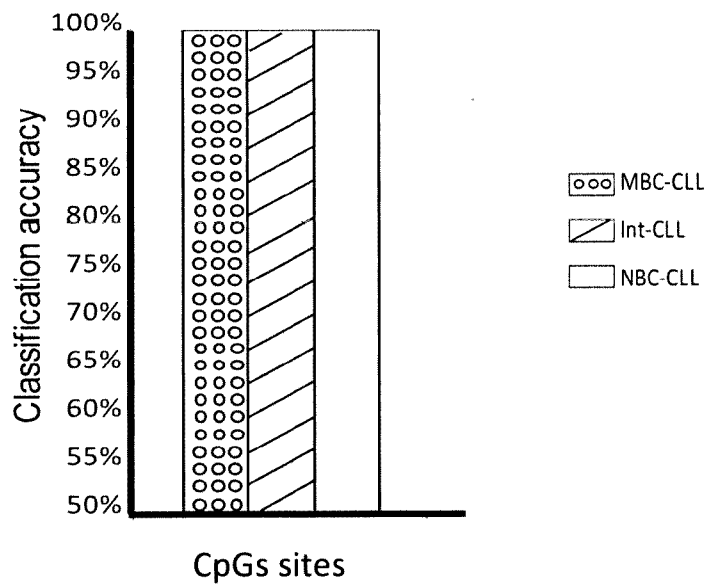
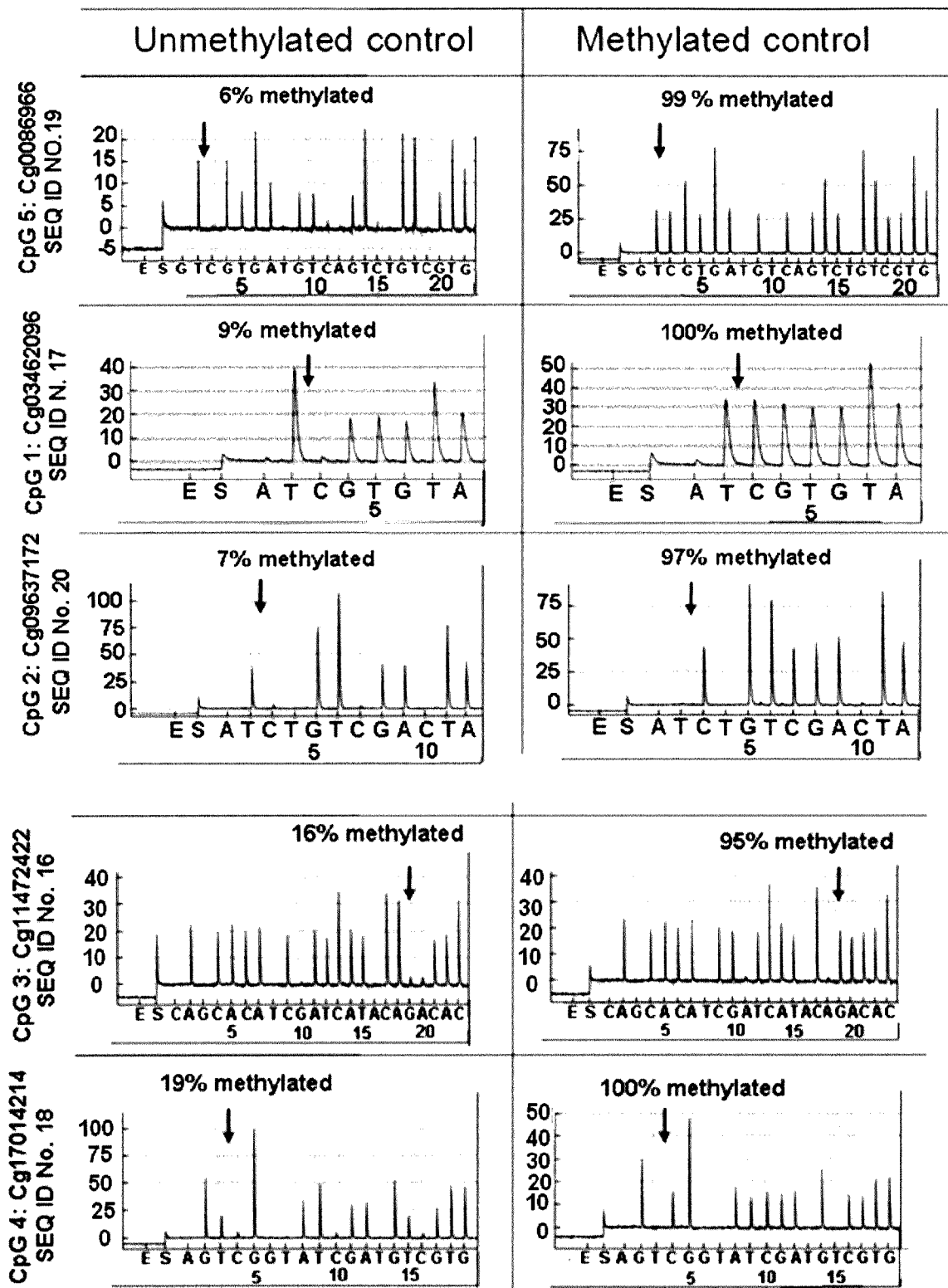


Figure 4



3/9

Figure 5



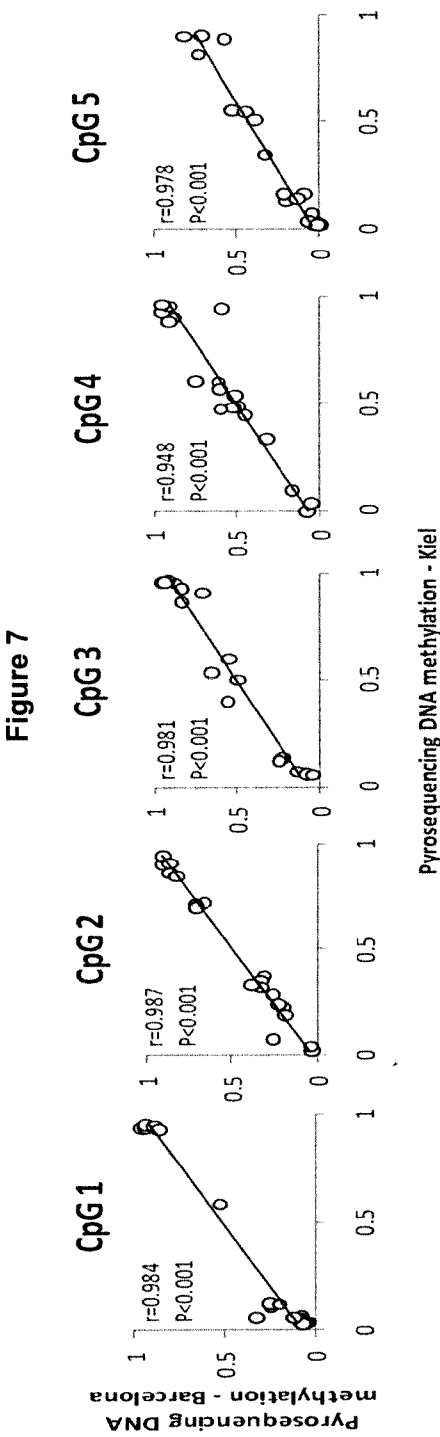
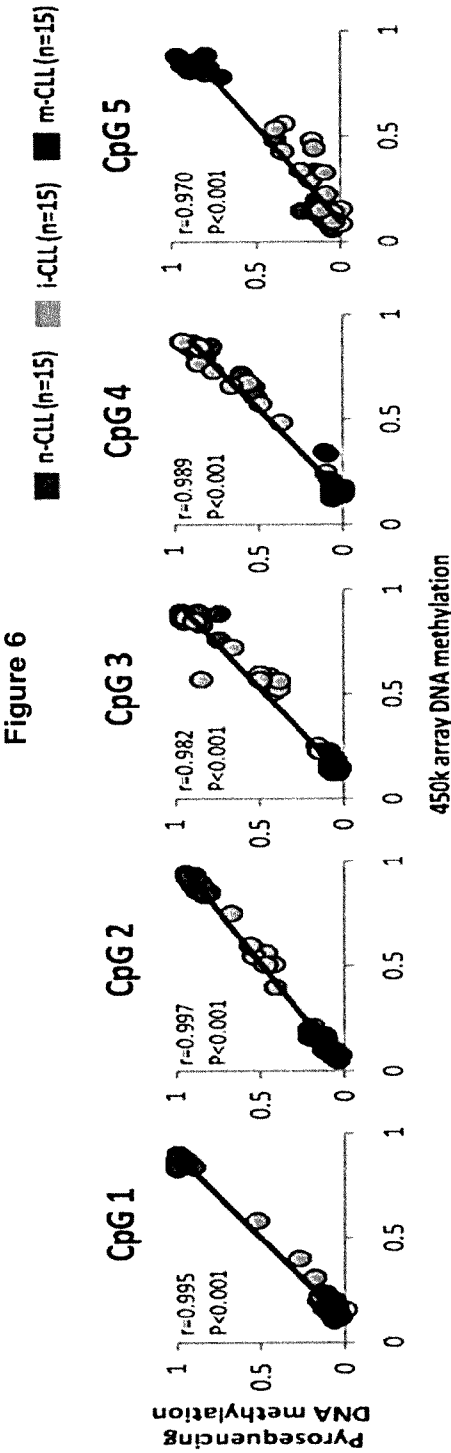


Figure 8

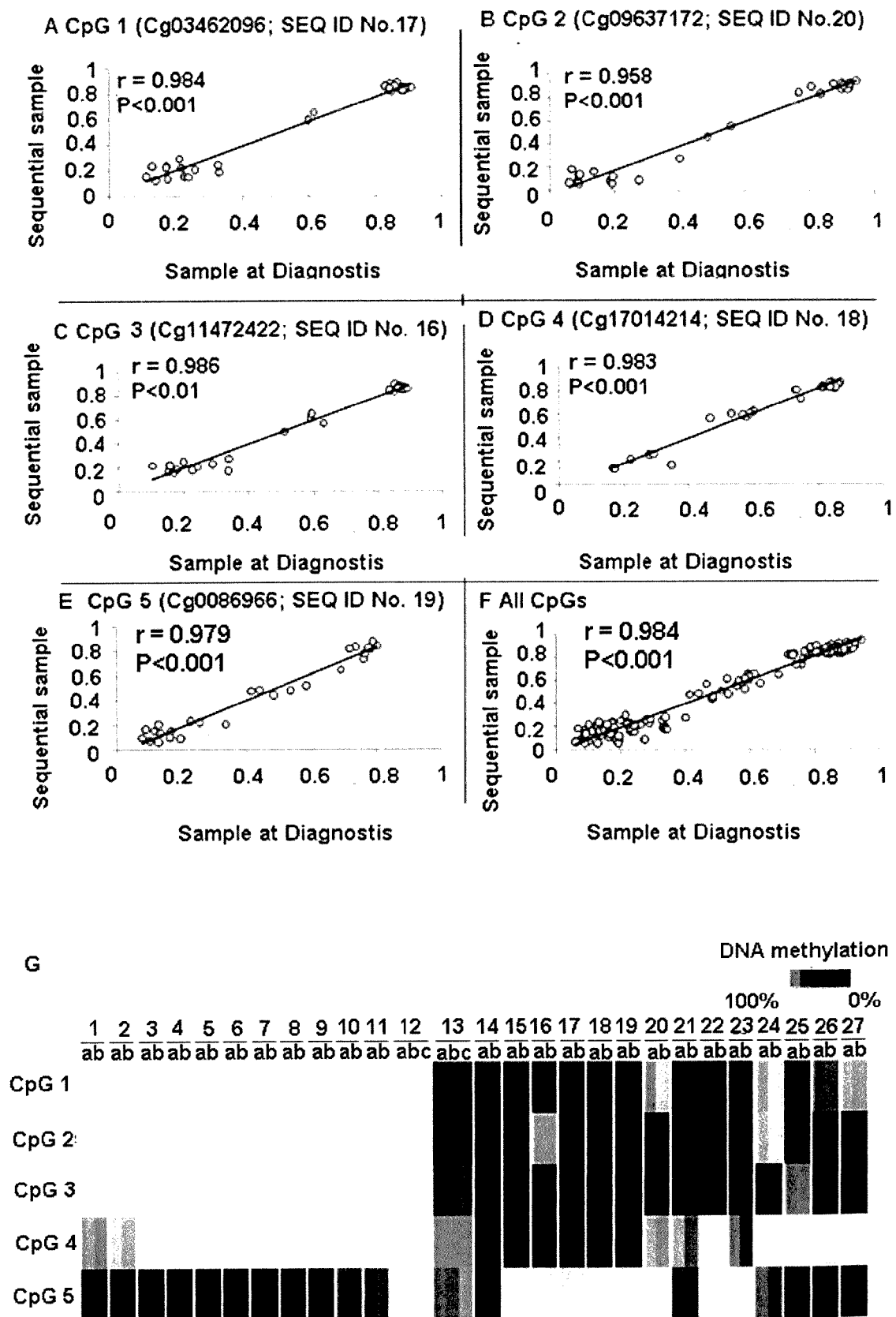


Figure 9

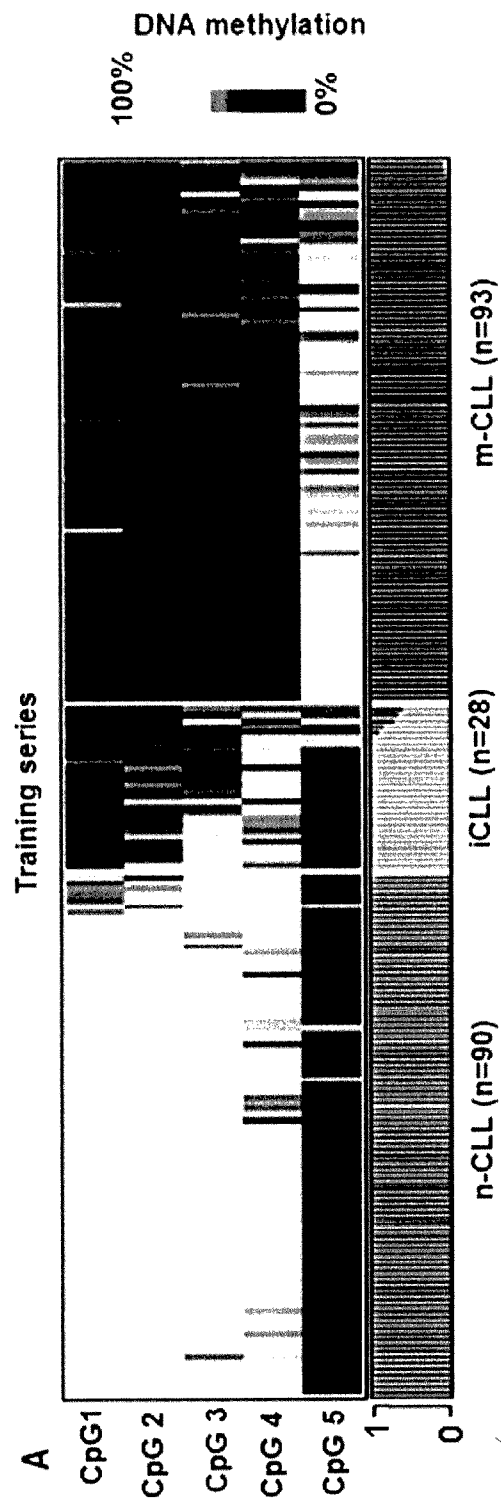
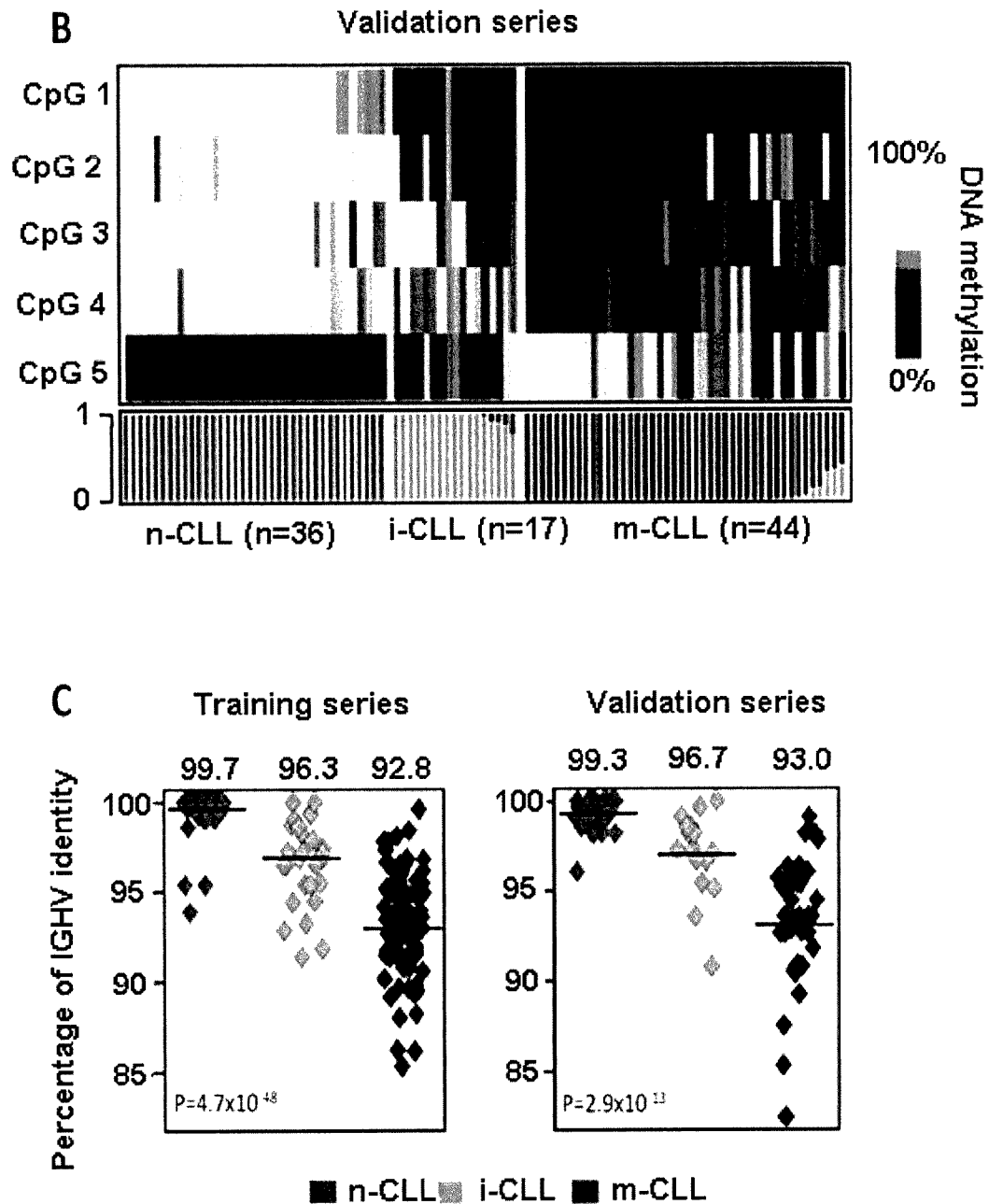


Figure 9 (cont)



8/9

Figure 10

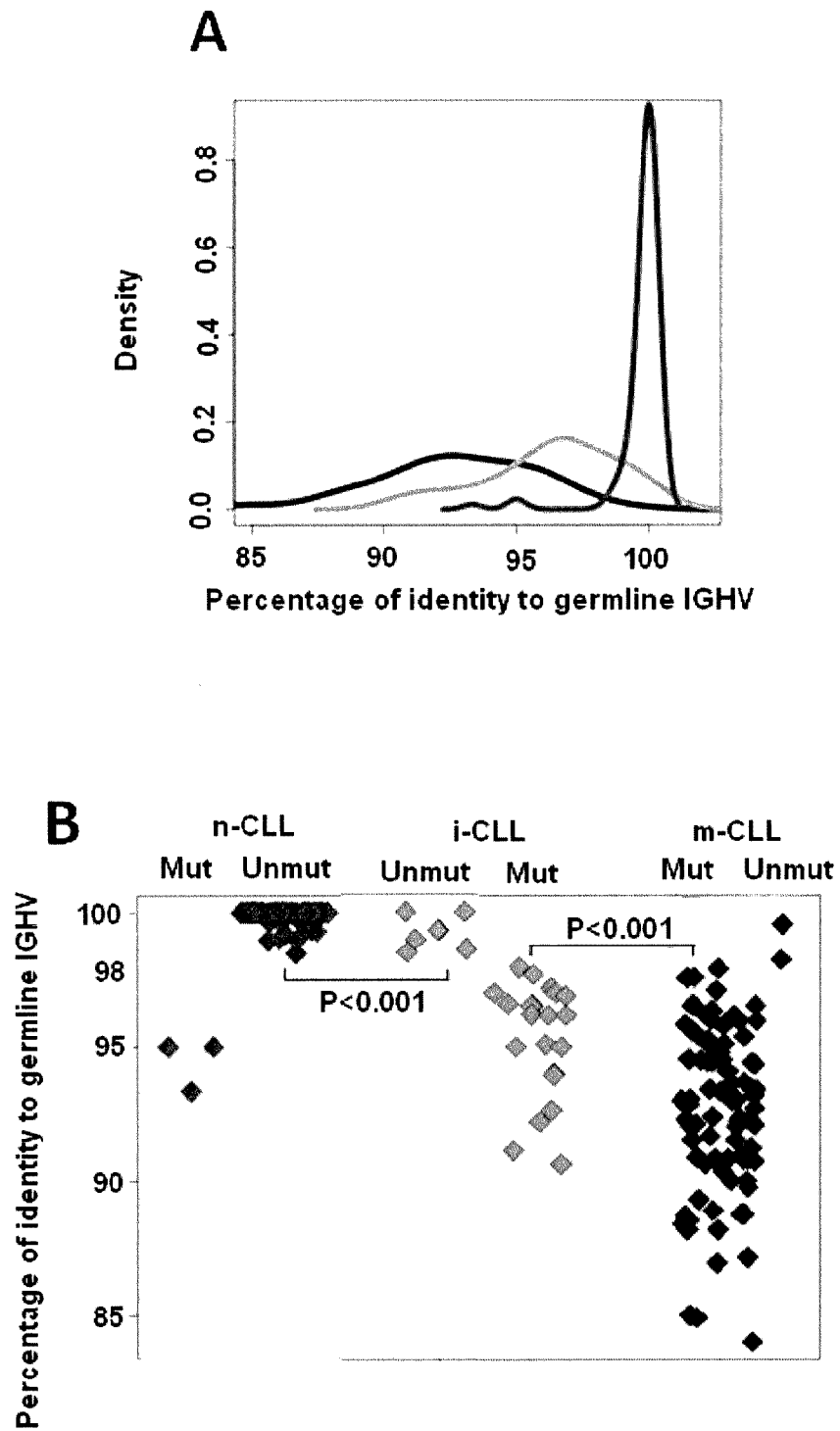
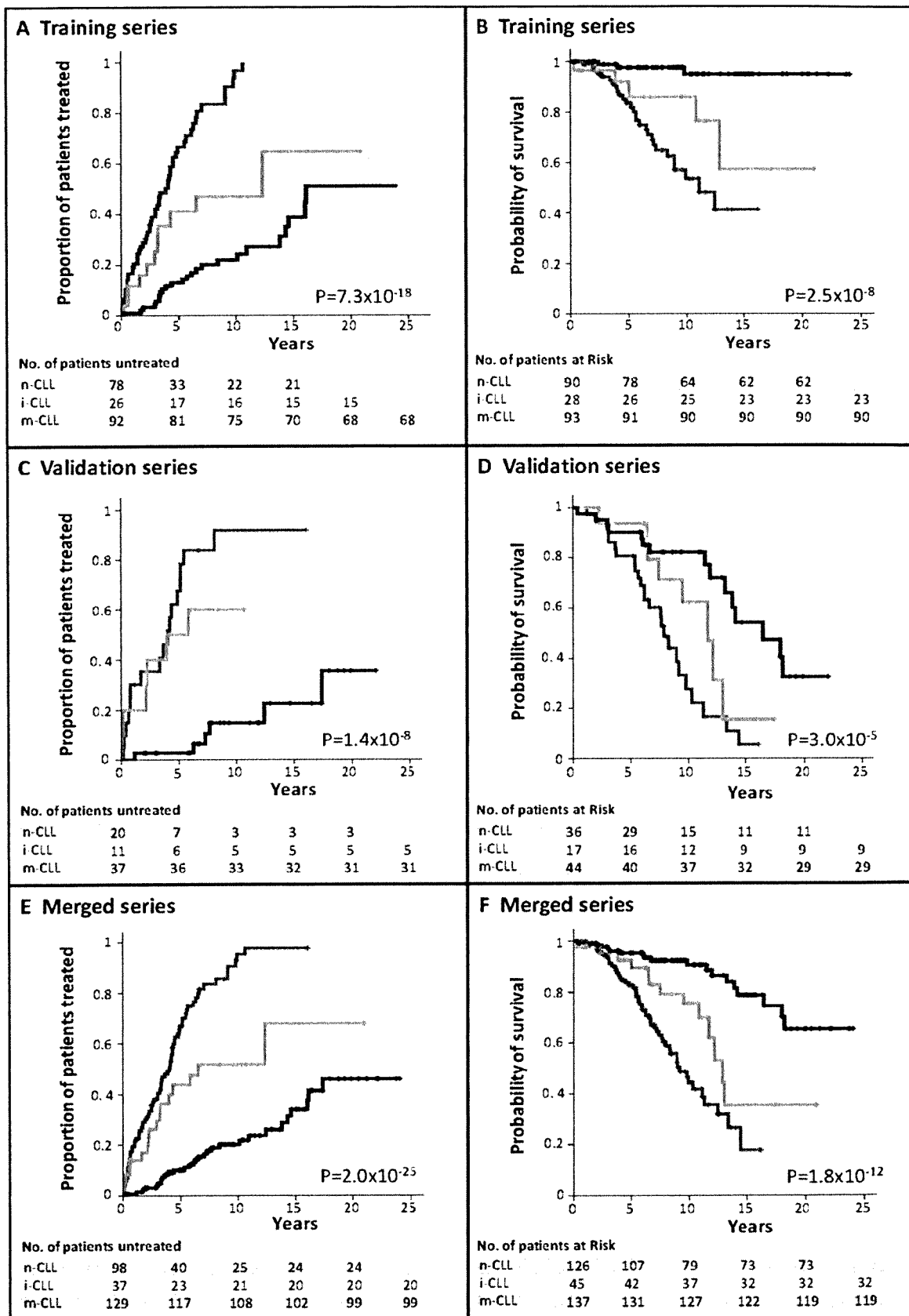


Figure 11



## INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2013/071065

## 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, MEDLINE, EMBASE, 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         | MARINA BIBIKOVA ET AL: "High density DNA methylation array with single CpG site resolution",<br>GENOMICS, ACADEMIC PRESS, SAN DIEGO, US,<br>vol. 98, no. 4, 26 July 2011 (2011-07-26),<br>pages 288-295, XP028304083,<br>ISSN: 0888-7543, DOI:<br>10.1016/J.YGENO.2011.07.007<br>[retrieved on 2011-08-02]<br>the whole document<br>-----<br>-/-- | 11-13                 |



Further documents are listed in the continuation of Box C.



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

"&amp;" document member of the same patent family

Date of the actual completion of the international search

25 November 2013

Date of mailing of the international search report

29/11/2013

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040,  
Fax: (+31-70) 340-3016

Authorized officer

Botz, Jürgen

## INTERNATIONAL SEARCH REPORT

International application No  
PCT/EP2013/071065

| C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT |                                                                                                                                                                                                                                                                                                                                                        |                       |
|------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Category*                                            | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                     | Relevant to claim No. |
| A                                                    | CAHILL NICOLA ET AL: "Genome-Wide Methylation Arrays Reveal Distinct Methylation Profiles in Prognostic Subsets of Chronic Lymphocytic Leukemia", BLOOD, vol. 114, no. 22, November 2009 (2009-11), page 154, XP008159632, & 51ST ANNUAL MEETING OF THE AMERICAN-SOCIETY-OF-HEMATOLOGY; NEW ORLEANS, LA, USA; DECEMBER 05 -08, 2009 the whole document | 1-13                  |
| A                                                    | KANDURI MEENA ET AL: "Differential genome-wide array-based methylation profiles in prognostic subsets of chronic lymphocytic leukemia", BLOOD, vol. 115, no. 2, January 2010 (2010-01), pages 296-305, XP002691004, the whole document                                                                                                                 | 1-13                  |
| A                                                    | RAHMATPANAH FARAHNAZ B ET AL: "Large-scale analysis of DNA methylation in chronic lymphocytic leukemia", EPIGENOMICS, vol. 1, no. 1, October 2009 (2009-10), pages 39-61, XP008159646, the whole document                                                                                                                                              | 1-13                  |
| A                                                    | HALLDORSDDOTTIR ANNA MARGRET ET AL: "Mantle cell lymphoma displays a homogenous methylation profile: A comparative analysis with chronic lymphocytic leukemia", AMERICAN JOURNAL OF HEMATOLOGY, vol. 87, no. 4, April 2012 (2012-04), pages 361-367, XP002691006, the whole document                                                                   | 1-13                  |
| X,P                                                  | KULIS MARTA ET AL: "Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia", NATURE GENETICS, vol. 44, no. 11, 14 October 2012 (2012-10-14), pages 1236-1242, XP002691005, the whole document                                                                                                            | 1-13                  |