Discrimination between human populations using a small number of differentially methylated CpG sites: A preliminary study using lymphoblastoid cell lines and peripheral blood samples of Caucasian and Chinese origin.

CURRENT STATUS: UNDER REVIEW

Patrycja Daca-Roszak
Institute of Human Genetics Polish Academy of Science
patrycja.daca-roszak@igcz.poznan.pl
Corresponding Author
ORCiD: https://orcid.org/0000-0003-1094-1333

Roman Jaksik
Silesian University of Technology

Julia Paczkowska
Institute of Human Genetics Polish Academy of Science

Michal Witt
Institute of Human Genetics Polish Academy of Science

Ewa Zietkiewicz
Institute of Human Genetics Polish Academy of Science

DOI: 10.21203/rs.3.rs-20643/v1

SUBJECT AREAS
Epigenetics & Genomics

KEYWORDS
DNA methylation, human population identification, pyrosequencing, population differentiating CpGs
Abstract

Background

Epigenetics is one of the factors shaping natural variability observed among human populations. A small proportion of heritable inter-population differences are observed in the context of both the genome-wide methylation level and the methylation status of individual CpG sites. It has been demonstrated that a limited number of carefully selected differentially methylated sites may allow discrimination between main human populations. However, most of the few published results have been performed exclusively on B-lymphocyte cell lines.

Results

The goal of our study was to identify a set of CpG sites sufficient to discriminate between Caucasian and Chinese populations based on the difference in the DNA methylation profile not only in cell lines but also in primary cell samples. The preliminary selection of CpG sites differentially methylated in these two populations (pop-diff-met CpGs) was based on the analysis of two groups of commercially available ethnically-specific B-lymphocyte cell lines, performed using Illumina Infinium Human Methylation 450 BeadChip Array. A subset of ten pop-diff-met CpGs characterized by the best differentiating criteria (|Mdiff|>1, q<0.05; lack of the confounding genomic features), and ten additional CpGs in their immediate vicinity, were further tested using pyrosequencing technology in both B-lymphocyte cell lines and in the primary samples of the peripheral blood representing two analyzed populations. To assess the population-discriminating potential of the selected set of CpGs (further referred to as “composite pop(CEU-CHB)-diff-met CpG marker”), three classification methods were applied. The predictive ability of the composite 8-site pop(CEU-CHB)-diff-met CpG marker was assessed using 10-fold cross-validation method on two independent sets of samples.

Conclusions

Our results showed that less than 10 pop-diff-met CpG sites may distinguish Caucasian and Chinese populations; importantly, this small composite pop-diff-met CpG marker performs well in both lymphoblastoid cell lines and in non-homogenous blood samples.

Background
Genetic variation of human populations is extensively explored in a variety of fields including epidemiological and medical studies (e.g. population-specific susceptibility to diseases, pharmacogenomics), but also in evolutionary studies and forensics (e.g. populations origin, relationships, identification) (1–5). The relation between the genome variation and population ancestry has been admittedly proven (6–9). A variety of genomic markers (SNPs, CNVs, microsatellites, and mtDNA, Y-chromosome haplotypes) providing accurate ancestry information have been identified, validated and successfully implanted in population-stratification tests [e.g. (10–12)]. The differences between human populations are shaped not only by the genomic DNA variation but also by transcriptomic and DNA methylation variation (13–20). Therefore, besides the most frequently used genomic DNA markers, some “non-classical markers”, representing inter-population differences in the expression and in the DNA methylation level, can potentially be used to discriminate between populations. In fact, a number of population-specific mRNA markers have been identified and tested in both B-cell lines and in a primary biological material, e.g. blood see (21). The use of DNA methylation variation to differentiate populations remains more controversial. It is well known that the majority of differences in the level of DNA methylation are caused by multiple environmental factors (e.g. nutrition, exposure to pollutants, social conditions, etc. (22–24). However, the recent development of high-throughput methods (mainly microarray technology) provided a wealth of data, which have demonstrated that a considerable part of the methylation variance reflects stable and heritable differences. Some of them are inter-individual and some differentiate populations (13, 18–20, 25–27). The inter-population differences are observed in both the genome-wide methylation level and in the methylation status of individual CpG sites (15, 16, 19, 20, 28–30). Compared to the genomic DNA variation, the persistent inter-population differences in the methylation level are rather small; nevertheless, they represent a possible source of markers that could be used for human population stratification. The inter-population differences in the level of methylation have been demonstrated in distinct types of a biological material: B-lymphocyte cell lines [e.g.(19, 20, 31, 32)], skin cells [e.g.(33, 34)], blood samples [e.g. (13, 25)]. Moreover, it has been shown that even a limited number (~ 400 CpGs) of carefully selected differentially methylated CpG
sites may allow discrimination of three main human groups: African, European and Asian (20). The goal of our study was to identify a small set of differentially methylated CpG sites (pop-diff-met CpGs) sufficient to discriminate between Caucasian and Chinese populations, which could be used as an easily manageable, composite pop(CEU-CHB)-diff-met CpG marker for a forensic differentiation between samples based on their population origin.

A set of 14 CpG sites characterized by significant population differences in their methylation (|Mdiff| > 1 at q < 0.05, and the lack of confounding SNPs under Illumina probes) was identified, based on the analysis of 36 commercially available B-lymphocyte cell lines of Caucasian and Chinese origin, performed using Illumina Infinium Human Methylation 450 BeadChip Array. A subset of 10 CpGs characterized by the best criteria, and ten additional CpGs in their immediate vicinity, was further tested in both B-lymphocyte cell lines and in primary samples of peripheral blood. Statistical evaluation of the discriminating potential of the best-performing pop-diff-met CpGs, employing 10-fold cross-validation method, was then performed in two independent sets of samples.

Results

I – Selection of candidate pop-diff-met CpGs

Illumina Infinium Human Methylation 450 BeadChip Array (HM450K array), previously applied to characterize methylation level in B-lymphocyte cell lines representing CEU (n = 18) and CHB (n = 18), revealed a set of 96 CpGs, differentiating the two populations at the significance level p < 0.05, and representing the highest inter-population differences in the average methylation levels (|Mav-diff| > 1; q < 0.05) see (35). From these differentially methylated CpGs, a small set of 14, characterized by the absence of confounding features (lack of SNPs in the studied CpG, lack of frequent SNPs under Illumina probe; no multi-site mapping of the probe), was selected as candidate pop-diff-met CpGs (Table 1).
Eleven of 14 best-differentiating CpGs were located outside CpG islands (in shore or shelf regions, gene body, transcription site start or 5'UTR regions). Three CpG sites, cg04036182, cg07207043 and cg00031303, were located in the genomic island of SHF, RREB1 and SDHAP1 genes, respectively. The highest inter-population differences in the methylation level (~ 40% difference) were observed in cg18136963 and cg26367031 ($M_{av\_diff} \geq 2.7$).

**II – Technical validation**

A subset of ten pop-diff-met CpGs candidates meeting even more stringent statistical criteria ($|M_{av\_diff}| \geq 1.2$ at $q < 0.05$), and ten additional CpGs located in their close proximity, was analyzed using pyrosequencing technique (Table 2).

Table 2. Comparison of DNA methylation levels assessed using Illumina HM450K array and pyrosequencing assays (PyroAssays).

| nb | Candidate pop-diff-met CpGs | Genomic position (GRCh:37) | Locus | Gene region | Type of region | $|M_{av\_diff}|$ | q-value |
|----|----------------------------|-----------------------------|-------|-------------|----------------|----------------|--------|
| 1  | cg18136963                 | chr6:139013                 | FLJ49 | not provided | N_Shore        | 2.950          | 0.0355 |
| 2  | cg26367031                 | chr3:178984                 | KCNM3 | 5'UTR; 1st exon | not provided | 2.775          | 0.0215 |
| 3  | cg03140118                 | chr1:379393                 | ZC3H12A | TSS1500 | N_Shore   | 2.411          | 0.001  |
| 4  | cg23669876                 | chr1:36489276               | AGO3  | Body (LTR)  | not provided | 2.355          | 0.0039 |
| 5  | cg00862290                 | chr3:178984                 | KCNM3 | TSS200      | S_Shore      | 2.247          | 0.008  |
| 6  | cg08979191                 | chr5:132113                 | SEPT8 | TSS200      | S_Shore      | 1.875          | 0.0185 |
| 7  | cg24037715                 | chr14:35203968              | -     | not provided | sea         | 1.691          | 0.0003 |
| 8  | cg07207043                 | chr6:705149                 | RREB1 | not provided | CpG Island  | 1.534          | 0.0345 |
| 9  | cg04036182                 | chr15:45458                 | SHF   | not provided | CpG Island  | 1.451          | 0.0201 |
| 10 | cg00031303                 | chr3:195681400             | SDHAP1| not provided | CpG Island  | 1.359          | 0.005  |
| 11 | cg07904028                 | chr4:632850                 | PPP2R2C | body   | not provided | 1.257          | 0.0145 |
| 12 | cg09972454                 | chr16:15083088             | PDXDC1| body      | N_Shore     | 1.232          | 0.0029 |
| 13 | cg24861686                 | chr8:114180                 | BLK   | body       | N_Shelf     | 1.193          | 0.000  |
| 14 | cg03585734                 | chr1:15398865              | FHAD1 | body       | not provided | 1.123          | 0.0144 |

CpGs selected for pyrosequencing validation are bolded. Shores and shelves are defined in Illumina as regions 0-2 kb and 2-4 kb, respectively, from a CpG island. N – upstream, S – downstream; TSS - transcription site start; LTR - long terminal region.
For cg00862290, which corresponds to the third CpG locus in PyroAssay 3, no reliable pyrosequencing data was obtained. Assays 4 (cg07904028) and 7 (cg26367031) did not pass technical evaluation step.

| CpG name in HM450K array | PyroAssay name | beta_mean_CEU (n=18) | beta_mean_CHB (n=18) | CEU.beta_mean - CHB.beta_mean | q-value | CEU.n (n=10) |
|--------------------------|---------------|---------------------|---------------------|-------------------------------|---------|---------------|
| cg24861686               | 1_CpG1*       | 0.841               | 0.697               | 0.143                         | 0.0000  | 0.813         |
| cg03140118               | 2_CpG1*       | 0.176               | 0.503               | -0.327                        | 0.0010  | 0.131         |
|                         | 3_CpG1        | -                   | -                   | -                             | -       | 0.410         |
|                         | 3_CpG2        | -                   | -                   | -                             | -       | 0.289         |
| cg00862290               | 3_CpG3*       | 0.466               | 0.161               | 0.305                         | 0.0080  | -             |
| cg07904028               | 4_CpG1*       | 0.515               | 0.714               | -0.199                        | 0.0145  | -             |
| cg08979191               | 5_CpG1*       | 0.779               | 0.520               | 0.258                         | 0.0185  | 0.782         |
|                         | 5_CpG2        | -                   | -                   | -                             | -       | 0.609         |
|                         | 5_CpG3        | -                   | -                   | -                             | -       | 0.599         |
|                         | 6_CpG1        | -                   | -                   | -                             | -       | 0.067         |
| cg04036182               | 6_CpG2*       | 0.271               | 0.486               | -0.215                        | 0.0201  | 0.112         |
| cg26367031               | 7_CpG1*       | 0.539               | 0.170               | 0.369                         | 0.0215  | -             |
| cg18136963               | 8_CpG1        | -                   | -                   | -                             | -       | 0.520         |
|                         | 8_CpG2*       | 0.514               | 0.162               | 0.352                         | 0.0355  | 0.498         |
|                         | 8_CpG3        | -                   | -                   | -                             | -       | 0.423         |
| cg07207043               | 9_CpG1*       | 0.625               | 0.820               | -0.195                        | 0.0345  | 0.529         |
|                         | 9_CpG2        | -                   | -                   | -                             | -       | 0.422         |
|                         | 9_CpG3        | -                   | -                   | -                             | -       | 0.480         |
|                         | 10_CpG1       | -                   | -                   | -                             | -       | 0.258         |
| cg23669876               | 10_CpG2*      | 0.368               | 0.728               | -0.360                        | 0.0290  | -             |

Due to technical reason (see Technical section in Supplement for details), some CpGs were excluded, and a subset of 17 CpGs was analyzed in further experiments.

Pyrosequencing results were collected as proportional values, separately for each analyzed CpG site (Table 2). The average value of differences in methylation level between the studied populations was in the range of 0.119 (PyroAssay 6_CpG1) to 0.387 (PyroAssay 2_CpG1). Statistically significant population differences (p < 0.05) were obtained for most of the CpG sites. The results from
pyrosequencing were concordant with the results from HM450K array. The only exception was PyroAssay 5, where no statistically significant population differences in the level of methylation were noted for two out of the three examined CpGs (5_CpG2 and 5_CpG3); nevertheless, this PyroAssay was not excluded from further analyzes.

Figure 2 shows the distribution of methylation levels in individual B-lymphocyte cell lines used in the technical validation phase. Eight PyroAssays (1, 2, 3, 5, 6, 8, 9 and 10) passed the technical validation and were used in the further step of biological validation.

III – Biological validation of population differences in methylation level
Independent B-lymphocyte cell lines
To test the biological validity of population-differentiating methylation status of 17 CpG sites, eight PyroAssays were performed in the independent set of B-lymphocyte cell lines. Statistically significant (p < 0.05) population differences in the mean methylation level were observed for 6 out of 8 tested PyroAssays (covering 12 CpG sites, see Table 3).

Table 3. Validation of eight PyroAssays performed in the independent set of B-lymphocyte cell lines.

CpG sites characterized by statistically significant inter-population differences in their methylation level are bolded. padj_beta: p-value after Benjamin Hochberg correction; pop-diff potential: differentiation potential of individual sites: 0-non-differentiating; 1-differentiating.
| PyroAssay number | CEU (n) | CHB (n) | CEU.mean | CHB.mean | CEU.var | CHB.var | CEU.mean - CHB.mean | padj_beta |
|-----------------|--------|--------|----------|----------|---------|---------|----------------------|-----------|
| 1_CpG1          | 34     | 34     | 0.800    | 0.759    | 0.008   | 0.006   | 0.040                | 0.03      |
| 2_CpG1          | 34     | 34     | 0.243    | 0.252    | 0.052   | 0.040   | -0.008               | 0.72      |
| 3_CpG1          | 34     | 34     | 0.246    | 0.222    | 0.069   | 0.051   | 0.024                | 0.82      |
| 3_CpG2          | 34     | 34     | 0.203    | 0.168    | 0.044   | 0.031   | 0.035                | 0.69      |
| 5_CpG1          | 34     | 34     | 0.718    | 0.594    | 0.057   | 0.041   | 0.124                | 0.04      |
| 5_CpG2          | 34     | 34     | 0.561    | 0.420    | 0.046   | 0.046   | 0.141                | 0.04      |
| 5_CpG3          | 34     | 34     | 0.522    | 0.448    | 0.064   | 0.049   | 0.074                | 0.31      |
| 6_CpG1          | 34     | 34     | 0.132    | 0.242    | 0.017   | 0.029   | -0.110               | 0.00      |
| 6_CpG2          | 34     | 34     | 0.236    | 0.343    | 0.036   | 0.031   | -0.107               | 0.01      |
| 8_CpG1          | 35     | 35     | 0.481    | 0.180    | 0.111   | 0.039   | 0.301                | 0.00      |
| 8_CpG2          | 35     | 35     | 0.492    | 0.166    | 0.125   | 0.050   | 0.325                | 0.00      |
| 8_CpG3          | 35     | 35     | 0.459    | 0.193    | 0.108   | 0.050   | 0.267                | 0.00      |
| 9_CpG1          | 34     | 34     | 0.713    | 0.806    | 0.042   | 0.035   | -0.093               | 0.07      |
| 9_CpG2          | 34     | 34     | 0.632    | 0.772    | 0.035   | 0.021   | -0.140               | 0.00      |
| 9_CpG3          | 34     | 34     | 0.657    | 0.784    | 0.049   | 0.030   | -0.127               | 0.01      |
| 10_CpG1         | 30     | 31     | 0.146    | 0.561    | 0.035   | 0.055   | -0.415               | 0.00      |
| 10_CpG2         | 30     | 31     | 0.171    | 0.640    | 0.043   | 0.062   | -0.469               | 0.00      |

In the majority of PyroAssays, the level of methylation was similar across the neighboring CpG sites (Table 3). Only two CpGs (5_CpG3 and 9_CpG1) had distinct methylation level compared to the rest of positions targeted by the respective PyroAssay, with no statistically significant differences between the two populations (Table 3). The highest inter-population differences in methylation level were noted for CpGs covered by PyroAssays 8 and 10 (Table 3, CEUmean-CHBmean column). PyroAssays 2 and 3 didn't reveal any statistically significant population differences in CpG methylation.

**Peripheral blood samples**

To test, whether population differences in the methylation levels of CpGs observed in CEU and CHB cell lines, reflected real differences between the two populations (and were not due to the cell lines’ peculiarities), the second step of biological validation was performed, using a primary biological material, i.e. peripheral blood samples from individuals representing two analyzed populations (n = 40...
from both CEU and CHB).

Overall, PyroAssays revealed similar inter-population differences in the level of CpG methylation in both B-lymphocyte cell lines and in blood samples. Furthermore, similar to the results obtained in B-lymphocyte cell lines, a high consistency in the methylation level among individual CpG sites examined within a given PyroAssay was also observed in blood samples (Fig. 3). The greatest inter-population differences in the level of CpG methylation was observed in PyroAssays 8 and 5. Only few inconsistencies were observed between B-lymphocyte cell lines and blood samples. Population differences in the methylation of 5_CpG3 and 9_CpG1 sites, which did not reach statistical significance in B-cell lines, were statistically significant in blood samples, whereas the inter-population differences in 1_CpG1 were not significant in blood samples. On the other hand, CpG sites targeted by PyroAssay 10, which classified as strongly population-differentiating sites in the B-cell lines, in blood samples were characterized by the lowest average differences in their methylation values.

For the majority of PyroAssays, methylation readouts in individual blood samples were tightly clustered, as opposed to those observed in B-lymphocyte cell lines. The only exception was PyroAssay 8, where the spread of the readouts from blood samples was much larger, and had a clear a tri-modal methylation distribution (see Discussion).

IV- Discriminating potential of the selected pop-diff-met CpGs

Identification of a composite pop(CEU-CHB)-diff-met marker

Pearson correlation analysis was performed using data from B-lymphocyte cell lines analysis (n = 10 CEU; n = 10 CHB) obtained during the technical validation step. Analysis showed a high correlation coefficient (0.8-1) within each of the corresponding PyroAssays, and simultaneously a low correlation (< 0.5) between individual PyroAssays (see Fig. 4 below).

To select the non-redundant set of validated pop-diff-met CpGs, correlated sites identified in the Pearson correlation analysis in each of the PyroAssays were removed. Based on the p-value after Benjamin Hochberg correction (the lowest padj_beta values were selected, see Table 3), a set of eight CpG sites (1_CpG1, 2_CpG1, 3_CpG2, 5_CpG1, 6_CpG2, 8_CpG1, 9_CpG3, 10_CpG1) was selected. This set of eight non-redundant, validated pop-diff-met CpGs formed a composite pop(CEU-CHB)-diff-met
marker, with the potential to discriminate between CEU and CHB populations based on the differences in the level of methylation.

Testing of the composite pop(CEU-CHB)-diff-met marker
To assess the population-discriminating potential of the 8-site composite pop(CEU-CHB)-diff-met marker, three different classification methods were used: support vector machines (SVM) with linear kernel, linear discriminant analysis (LDA) and random forest (RF). The predictive ability of each method was assessed using 10-fold cross-validation, which was repeated 1000 times due to the moderate number of available cases.

The results obtained using each of the classification algorithms (SVM, LDA and RF) were compared in terms of AUC parameter (area under ROC curve) (see Fig. 5).

The shape of all presented curves followed the left-hand corner and the top border, indicating the high accuracy of the 8-site composite pop(CEU-CHB)-diff-met marker with a high level of true positive in comparison to false positive results. Similar result was obtained using all three tested classification methods (AUC > 0.9), of which SVM was the most reliable (AUC = 0.996). The SVM validation performed on two independent datasets, B-lymphocyte cell lines (n = 48) and blood samples (n = 40), showed a high accuracy of the classification power in both sets (> 85%) (see Table S4 in Supplement).

Principle Component Analysis was used to assess the potential of the 8-site composite pop(CEU-CHB)-diff-met marker to separate samples from two analyzed populations. While the vast majority of samples clustered according to their population affiliation, two population-specific clusters were located in the close vicinity. The more accurate separation was obtained for blood samples (population-specific clusters were more separated from each other compared to B-cell samples) (Fig. 6A,B).

The variance distribution was attributed to the first (~ 30%) and the second (~ 17%) dimension in both B-lymphocyte cell lines and blood samples. In both PC plots, markers 2_CpG1, 6_CG2, 9_CpG3 and 10_CpG1 correlated with each other and showed higher methylation level in CHB population, whereas markers 1_CpG1, 3_CpG2, 8_CpG1 and 5_CpG1 showed higher methylation level in CEU population. The weight of an individual CpG marker on the principle component was diverse, as
indicated by the vectors length. What is interesting, most CpG markers had similar weight in PC analyzed in B-lymphocyte cell lines (Fig. 6A), while in blood sample, the impact of one marker, 1_CpG1, was distinctly smaller (Fig. 6B).

An additional test was performed to assess the minimal number of pop-diff-met CpGs that would classify Caucasian and Chinese samples with high accuracy. The minimal number of seven unlinked pop-diff-met CpGs (10_CpG1, 6_CpG2, 1_CpG1, 2_CpG1, 9_CpG3, 8_CpG1, 3_CpG2) had a high classification accuracy (AUC ~ 1, and precision > 0.8) (Fig. 7, lower panel) in both B-lymphocyte cell lines and blood samples; discrimination potential obtained in peripheral blood samples (precision = 0.925) was higher in comparison to B-lymphocyte cell lines (precision = 0.854). In order to obtain similar discrimination power in both B-lymphocyte cell lines and peripheral blood samples, we decided to retain the 8-site composite pop(CEU-CHB)-diff-met marker to be used for methylation-based classification of CEU and CHB populations (see Fig below, lower panel).

**Discussion**

The aim of our study was to identify a set of CpG sites characterized by a significant difference in the DNA methylation profile between Caucasian and Chinese populations. Only adult males were analyzed, to avoid gender- and age-related bias in the DNA methylation level. Analysis of 18 CEU and 18 CHB B-lymphocyte cell lines, performed on HM450K array, which measures the methylation of approximately 480,000 CpG sites across the human genome in parallel, revealed only 14 CpG sites with significantly different methylation levels in the studied populations (|Mav_diff| ≥ 1.0 and q < 0.05). According to the literature (37), |Mav_diff| ≥ 1.0 corresponds to a 20% difference in the methylation value. Such a small number of potentially population-differentiating CpG sites, with a relatively low inter-population differences in the methylation level (|M_diff| in the range of 1.1–2.9), stands in line with results from other studies. It has been estimated that only a small fraction of CpGs across the genome stably varies in their methylation among human populations (18, 20, 38, 39).

A subset of ten CpGs preselected in HM450K array experiment (cg24861686, cg03140118, cg00862290, cg07904028, cg08979191, cg04036182, cg26367031, cg18136963, cg07207043, cg23669876), referred to as candidate pop-diff-met CpGs, with the highest inter-population
differences in the mean methylation value ($|M_{av\_diff}| \geq 1.2$ and q-val < 0.05), was selected for further validation. Validation was done using pyrosequencing technique, regarded as a more sensitive method and widely used in DNA methylation studies (40, 41). PyroAssays covered ten candidate pop-diff-met CpGs, and several closely located neighboring CpGs, such that overall 20 CpG sites were tested.

Two-step validation was performed to exclude technical obstacles that could provide faulty results (technical validation in B-lymphocyte cell lines). In the next step, a biological validation in peripheral blood samples was performed to exclude the possibility that the inter-population differences in the methylation level reflected specific conditions related to the maintenance of the CHB and CEU cell lines. Following these two steps. Six out of the eight PyroAssays tested on primary material displayed statistically significant inter-population differences in the methylation level ($p < 0.05$).

These results indicated that the candidate population-differentiating CpG sites selected based on the analysis of B-lymphocyte cell lines, after a proper validation, may be used as population-differentiating markers also in the primary cells (blood samples) [see also (19)].

Our results are especially interesting in the context of a widely discussed suitability of B-lymphocyte cell lines (lymphoblastoid cell lines, LCLs) for population studies on methylation (42, 43). LCLs are a commonly used source of biological material due to their easy availability (Coriell repository resources), tissue homogeneity (exclusively B-lymphocytes) and known population origin. However, some recent studies revealed that laboratory treatment of LCLs, e.g. EBV transformation or specific conditions during culturing (e.g. repeated freeze-thaw cycles), may induce random DNA methylation alterations and thus produce misleading methylation results (42–44). In this context, a comparison of raw methylation readouts collected from B-lymphocyte cell lines and from blood samples in our study revealed interesting observations. A high consistency in the methylation level was observed among CpGs examined within each PyroAssay, both in B-lymphocyte cell lines and in blood samples.

However, the mean values of inter-population differences in blood samples were smaller than in B-lymphocyte cell lines, and did not exceed 30% as opposed to nearly 50% in the cell lines. On the other hand, for the majority of CpGs, the readouts representing methylation in individual cell lines
were scattered, while those representing individual blood samples remained "tightly" clustered around the mean (except for PyroAssay 8, see discussion below). The scattered methylation readouts observed in the cell lines could reflect the lack of homogeneity of technical (cell line maintenance etc.) and/or biological factors (age, and/or lifestyle of cell lines donors). Analysis of the reported age of B-lymphocyte cell lines donors (wherever available) revealed no correlation with the methylation results. Aspects related to the cell line maintenance were beyond our control (cells were purchased from Coriell Repository), but these lines have been used in many studies and to our knowledge no systemic population differences have been reported. The small variance of readouts observed in the primary biological material is more surprising. Knowing that blood is a mixture of different cell types, and that blood donors were not controlled for their lifestyle (e.g. diet, smoking etc.), methylation readouts were expected to be more scattered. On the other hand, the number of blood samples used in the analysis was lower than that of the cell lines, and it is possible that increasing the size of tested group would affect the picture.

The only exception from the generally small variance of the methylation readouts in blood samples was PyroAssay 8, where the distribution of readouts followed a characteristic tri-modal pattern. This pattern, when observed in HM450K array, has been described to reflect the presence of SNP in the examined CpG sites or in sequences targeted by Illumina probes [see(35, 45)]. Although all PyroAssays in this study were designed to avoid SNP-related bias, a tri-modal pattern observed in PyroAssay 8 prompted us to subject it to a careful scrutiny, to exclude the possible impact of the genomic sequences. Both in silico analysis, performed in Genome Browser Database, and Sanger sequencing of several B-lymphocyte cell lines and blood samples, did not reveal any SNPs/indels in either interrogated CpG sites and under the primers used in PyroAssays (data not shown). It is probable that, here also, increasing the number of samples could change this picture. In fact, an indication of a tri-modal distribution in PyroAssay 8 was also detectable in B-lymphocyte cell lines, but the larger number of samples blurred it into a cloudlike pattern (see Fig. 3).

To confirm the discriminating power of the composite pop(CEU-CHB)-diff-met marker, composed of the validated pop-diff-met CpGs, a number of statistical analyzes were performed. All three
algorithms (SVM, RF and LDA) used to test the sensitivity and specificity (ROC and AUC parameters) of population classification worked well in both types of the biological material (B-cell lines and blood samples), revealing high precision (> 90%) of sample population classification.

According to the literature, a subset of population-specific methylation markers (< 500 pop-CpG sites) allows to carry out discrimination of main human populations. The set of eight pop-diff-met CpGs described in our study is, to our knowledge, the smallest methylation-based composite marker able to discriminate two human populations (13, 20, 39). Principal component analysis using the 8-site composite pop(CEU-CHB)-diff-met marker clearly separated Caucasian and Chinese samples with respect to their population affiliation. What is interesting, a better classification was obtained in peripheral blood samples than in LCL material (see Fig. 6).

To better characterize our composite pop(CEU-CHB)-diff-met marker, we analyzed the genomic location of the differentially methylated CpGs. The vast majority of CpGs targeted by PyroAssays in this study were located outside of the, presumably evolutionary-conserved, CpG islands (see Table 1). CpG sites targeted by PyroAssays 5 and 8 were located in the shore regions (~ 2 kb from CpG islands, as defined by Illumina) of Septin8 and FLJ49/FLJ46906 genes, respectively. CpG sites targeted by PyroAssay 10 were situated in the body of AGO3 gene. The only sites located in CpG islands (of SHF and RREB1 genes, respectively) were those targeted by PyroAssays 6 and 9. Our results therefore concord with other studies, which have indicated that inter-population differences in DNA methylation level are enriched outside CpG islands and are concentrated in regions flanking the islands (shores, shelves) or in gene body regions (38, 39).

Genes, in which our pop-diff-met CpGs reside, are involved in various biological processes: apoptosis regulation (SHF), expression regulation (FLJ49/FLJ46906), RNA interference (AGO3); or participate in distinct biological functions: transcription factor (RRB1), nucleotide binding protein (SEPT8). The biological relevance of the level of individual CpG sites methylation is still disputable (46). However, it has been postulated that CpG sites located adjacent to functional genomics areas (CpG islands and/or shores) and representing similar methylation pattern due to potential effect on the chromatin structure, may play an important biological role (47). In search for a putative long-range co-
methylated, we examined five of the studied CpGs (cg08979191, cg04036182, cg18136963, cg07207043, cg23669876). Methylation status of the neighboring CpG sites, located 200 bp up- and downstream from the pop-diff-met CpGs (co-methylation), was examined in samples from both populations (results in Supplementary file, Table S5), using data from our HM450K array study (35), data accessible through GEO Series accession number: GSE73901). Four of five pop-diff-met CpGs, cg08979191, cg04036182, cg18136963, cg07207043, had other Illumina-targeted CpGs in their vicinity (Table S5 in Supplement). CpG sites located as far as 200 bp down- or upstream of two of the “core” CpGs (cg08979191 and cg18136963), displayed statistically significant inter-population differences in the level of methylation ($IM_{av} - \text{diffI}$ in the range 0.8–2.5) (for details see Table S5 in Supplement). Importantly, the "direction" of these differences was the same as in the “core” cg08979191 and cg18136963 (the reduced level of methylation in Chinese in comparison to Caucasian population. All the co-methylated CpG sites were located in the shore regions flanking CpG islands, of SEPT8 and FLJ49/FLJ46906 genes, respectively. A highly correlated methylation level of CpG sites separated by 200 bp suggests that cg08979191 and cg18136963 represent the methylation status of a longer region; this would be similar to the effect of linkage disequilibrium between SNPs in the human genome. However, it has to be kept in mind that Illumina HM450K array probes target a relatively small proportion of CpG sites in the human genome. A much larger number of neighboring CpG sites are present at the closer distance to these and the remaining pop-diff-met CpGs in our study; to examine methylation status of these sites, techniques addressing the whole genome should be employed, e.g. NGS technology.

DNA methylation constitutes an epigenetic switch in gene expression regulation (19, 20, 31, 48). To determine, whether differentially methylated CpGs in our study reflected population differences in gene expression status, we integrated DNA methylation and gene expression data obtained from our previous studies performed on same set of B-lymphocyte cell lines and conducted on two microarrays systems: HM450K array and HumanHT-12v4 Expression BeadChip Kit expression array. A subset of pop-diff-met CpGs, cg08979191 (shore), cg04036182 (CpG island), cg18136963 (shore), cg07207043 (CpG island), cg23669876 (unknown), for which Illumina data on gene expression were available, was
subjected to Pearson correlation analysis. This analysis clearly indicated the lack of any correlation between the methylation and gene expression levels (data not shown). These results are not surprising, since the regulation of gene expression is a complex process involving e.g. transcription factors, histone modification, non-coding RNA regulation (48–51). A straight methylation-expression correlation is rarely observed, or is noted exclusively in individual genes (48). In conclusion, the biological meaning of the differential methylation status observed in the analyzed populations remains to be elucidated.

Conclusions
Our results showed that even a small set of carefully selected differentially methylated CpGs (pop-diff-met CpGs), may be used to distinguish Caucasian and Chinese populations. Importantly, this composite pop(CEU-CHB)-diff-met marker performs well in both lymphoblastoid cell lines and in non-homogenous blood samples. The performance of our composite marker, estimated using different classification methods, was reasonably high for the limited number of examined samples, although this may change (either decrease or increase) when a larger number of sample are analyzed. Also, further studies using samples from other population groups need to be carried out.

The current knowledge regarding relation between epigenetics and environmental factors, as well as a trans-generation inheritance of methylation pattern [e.g.(31, 52, 53)], is still limited. Nevertheless, it seems that discrimination between populations and inference of population origin of a sample, based on DNA methylation markers, is feasible and may add a new, additional dimension to medical and forensic casework, as earlier postulated (1, 53, 54).

Methods
STUDY DESIGN
The study consisted of four main phases: selection of candidate pop-diff-met CpGs, two-step validation, and statistical tests (Fig. 1).

DNA samples
DNA samples from unrelated, healthy adult males representing Caucasians and Chinese populations (further referred to as CEU and CHB, respectively), were isolated either from commercially available
B-lymphocyte cell lines (Coriell Cell Repositories) or from samples of peripheral blood (for details see Table S1 and S2 in Supplement).

Both B-lymphocyte cell lines and peripheral blood samples used in this study underwent identical procedures including: DNA isolation (QIAamp DNA Blood Mini Kit, Qiagen), evaluation of its purity (Qubit, DSDNA H5 Assay Kit, Life Technology), and bisulfite treatment (EZ DNA Methylation-Gold Kit, Zymo Research). 500 ng of purified DNA from B-lymphocyte cell lines (n = 90), and peripheral blood (n = 40) was converted with bisulfite solution using EZ DNA Methylation–GoldTM Kit (Zymo Research, Germany), according to the manufacturer’s protocol.

I – Selection of candidate pop-diff-met CpGs: Identification of differentially methylated CpG sites based on Human Methylation 450 BeadChip Array

B-lymphocyte cell lines from CEU (n = 18) and CHB (n = 18) were examined on Illumina Infinium HumanMethylation 450 BeadChip Array (further referred to as HM450K array), according to the manufacturer-specified procedure. All analytical procedures, such as microarray technical quality evaluation, as well as statistical approach implemented in microarray data analysis, have been presented in detail in our previous publication see (35).

II – Technical validation: Pyrosequencing assay design and optimization

Technical validation step was performed in a subset of B-lymphocyte cell lines previously analyzed by HM450K array. Pyrosequencing assays (further referred as PyroAssays) were designed to validate candidate pop-diff-met CpG sites preselected in HM450K array experiment for which effective PyroAssays could be designed (Assay score in PyroMark Assay Design Software ≥ 75, no CpGs under PyroAssay primers); in some cases, PyroAssays covered additional CpGs located in the close proximity (less than 25 bp upstream or downstream) of the selected candidate pop-diff-met CpGs (see Table 2 in Result section).

Wherever possible, PyroAssays were designed to analyze CpGs on the same DNA strand as in the microarray experiment, to eliminate possible differences in the CpG methylation status depending on the DNA strands (the only exceptions were PyroAssays 2, 4, 7) (for details see Table S3 in Supplement).
PCR reaction conditions (PCR program and further sample workflow) are available in Supplementary file (Table S3). Primers for PyroAssays were designed using PyroMark Assay Design Software 2.0.1.15 (Qiagen). Only those PyroAssays, for which specific PCR products were obtained for both bisulfite converted study samples and for methylated/unmethylated controls, were used in further analyses (see Supplementary file, Technical section).

The quality of methylation results collected from pyrosequencing reactions was assessed based on a series of dilution curves obtained for all the PyroAssays (see Supplement Fig.S1).

III – Biological validation: Pyrosequencing assays in independent samples
CpGs that passed technical validation were further tested in two steps. In the first one, PyroAssays were examined in an independent set of B-lymphocyte cell lines from both populations (CEU n = 35; CHB n = 35); in the second step, PyroAssays were tested in peripheral blood samples (CEU n = 20; CHB n = 20) (see Fig. 1).

The same technical conditions (initial sample preparation, PCR reaction, Pyrosequencing process) were applied in both biological validation steps (see Technical section in Supplement).

IV – Statistical analysis
Selection of the best non-redundant pop-diff-met CpGs from among those that passed technical and biological validation steps was conducted using beta regression test from the betareg Bioconductor package (36), with Benjamini Hochberg multiple testing correction.

The selected set of CpGs was then examined for its population-discriminating potential. Sample classification was conducted using three methods: support vector machines (SVM) with linear kernel, random forest (RF) and linear discriminant analysis (LDA). Prior to the classification process, correlated CpGs were removed; it was done by retaining only those with the lowest, adjusted p-values in the beta regression test for the technical validation dataset (20 samples), which was also used to select the best classification method. The predictive ability of the selected set of pop-diff-met CpGs was assessed using each of the classification methods, with 10-fold cross-validation, repeated 1000 times. In all cases classification was conducted using all possible combinations of 1 to 8 CpGs identified as differentially methylated.
The best classification method in terms of AUC (area under ROC curve) was then validated using two independent datasets from 48 B-lymphocyte cell lines and 40 total blood samples; all of the datasets were balanced (equal number of CEU and CHB samples). Classification was conducted in R with caret library and plotROC and ggplot2 used for visualization purposes.

Principal Component Analysis (PCA) was carried out in R using prcomp function from the stats package and visualized with the ggbiplot library.

Abbreviations

CEU
Caucasian population

CHB
Chinese population

HM450K array
Illumina Infinium Human Methylation 450 BeadChip Array

pop-diff-met CpGs
a set of CpG sites differentiating Caucasian and Chinese population based on methylation level differences

PyroAssays
pyrosequencing assays

Declarations

Ethics approval and consent to participate

The study was approved by the Bioethical Committee at the Central Clinical Hospital of the Ministry of Interior in Warsaw (No 67/2010). The samples of peripheral blood were collected from anonymous healthy donors and were obtained with their informed consent. The donors were explicitly informed about the aim of the sample collection by providing them with a short description of the proposed project. The informed consent was taken in verbal form to ensure the anonymity of the donors, who did not wish to disclose their names. The B-lymphocyte lines used in the project were purchased from Corriell depository, and were selected to represent studied populations (Caucasian, CEU; and Chinese, CHB); catalogue numbers are listed in Supplement, Table S1.

Consent for publication

Not applicable.
Availability of data and materials
All the data supporting the conclusions of the study are included in the manuscript and the additional file.

Competing interests
The authors declare that they have no competing interests.

Funding
The study was supported by the grant AriaDNA OR00 0027 12 from the NCRD and MINIATURA (2017 101/XAIZ8|00096) from NCN. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors’ Contributions
PDR participated in the design of the study, pyrosequencing validation, data analysis, and drafted manuscript; RJ performed statistical analysis, J.P participated in pyrosequencing assays designing, data analysis and offered technical training in pyrosequencing technology, MW and EZ participated in the design of the study and critically revised the manuscript.
All authors have read and approved the final manuscript.

Electronic resources cited
GEO database http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73901
ENSEMBL database http://www.ensembl.org/index.html
GeneCards- Human Genome Database http://www.genecards.org/
Genome Browser https://genome.ucsc.edu/
All authors read and approved the final manuscript.

Conflict of Interest
Authors declare that there are no conflicts of interest.

ACKNOWLEDGMENTS
We thank Prof. Giefing, Prof. Słomski, and Prof. Szalata for offering excellent training in pyrosequencing technology, opportunity to use their Pyrosequencers and for valuable scientific discussions.
References

1. Kader F, Ghai M. DNA methylation and application in forensic sciences. Forensic Sci Int. 2015;249:255-65.

2. Hernando-Herraez I, Garcia-Perez R, Sharp AJ, Marques-Bonet T. DNA Methylation: Insights into Human Evolution. PLoS Genet. 2015;11(12):e1005661.

3. Levenson VV, Melnikov AA. DNA methylation as clinically useful biomarkers-light at the end of the tunnel. Pharmaceuticals (Basel). 2012;5(1):94-113.

4. Unnikrishnan A, Freeman WM, Jackson J, Wren JD, Porter H, Richardson A. The role of DNA methylation in epigenetics of aging. Pharmacol Ther. 2019;195:172-85.

5. Spolnicka M, Pospiech E, Peplonska B, Zbiec-Piekarska R, Makowska Z, Pieta A, et al. DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups. Int J Legal Med. 2018;132(1):1-11.

6. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98-U5.

7. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4.

8. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, et al. Discerning the ancestry of European Americans in genetic association studies. Plos Genetics. 2008;4(1):9-17.

9. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253).

10. Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, et al. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set.
Forensic Science International-Genetics. 2014;11:13–25.

11. Daca-Roszak P, Pfeifer A, Zebracka-Gala J, Jarzab B, Witt M, Zietkiewicz E. 
EurEAs_Gplex-A new SNaPshot assay for continental population discrimination and 
gender identification. Forensic Science International-Genetics. 2016;20:89–100.

12. Brandstatter A, Salas A, Niederstatter H, Gassner C, Carracedo A, Parson W. 
Dissection of mitochondrial superhaplogroup H using coding region SNPs. 
Electrophoresis. 2006;27(13):2541–50.

13. Husquin LT, Rotival M, Fagny M, Quach H, Zidane N, McEwen LM, et al. Exploring the 
genetic basis of human population differences in DNA methylation and their causal 
impact on immune gene regulation. Genome Biol. 2018;19(1):222.

14. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, et al. Factors 
underlying variable DNA methylation in a human community cohort. Proc Natl Acad 
Sci U S A. 2012;109(Suppl 2):17253–60.

15. Adkins RM, Krushkal J, Tylavsky FA, Thomas F. Racial differences in gene-specific 
DNA methylation levels are present at birth. Birth Defects Res A Clin Mol Teratol. 
2011;91(8):728–36.

16. Terry MB, Ferris JS, Pilsner R, Flom JD, Tehranifar P, Santella RM, et al. Genomic DNA 
methylation among women in a multiethnic New York City birth cohort. Cancer 
Epidemiol Biomarkers Prev. 2008;17(9):2306–10.

17. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, et al. Significant 
differences in global genomic DNA methylation by gender and race/ethnicity in 
peripheral blood. Epigenetics. 2011;6(5):623–9.

18. Song MA, Brasky TM, Marian C, Weng DY, Taslim C, Dumitrescu RG, et al. Racial 
differences in genome-wide methylation profiling and gene expression in breast 
tissues from healthy women. Epigenetics. 2015;10(12):1177–87.
19. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol. 2012;13(2):R8.

20. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, et al. DNA methylation contributes to natural human variation. Genome Res. 2013;23(9):1363-72.

21. Daca-Roszak P, Swierniak M, Jaksik R, Tyszkiewicz T, Oczko-Wojciechowska M, Zebracka-Gala J, et al. Transcriptomic population markers for human population discrimination. BMC Genet. 2018;19:1471-2156. (Electronic).

22. King K, Murphy S, Hoyo C. Epigenetic regulation of Newborns' imprinted genes related to gestational growth: patterning by parental race/ethnicity and maternal socioeconomic status. J Epidemiol Community Health. 2015;69(7):639-47.

23. Mozhui K, Smith AK, Tylavsky FA. Ancestry Dependent DNA Methylation and Influence of Maternal Nutrition. Plos One. 2015;10(3).

24. Plusquin M, Guida F, Polidoro S, Vermeulen R, Raaschou-Nielsen O, Campanella G, et al. DNA methylation and exposure to ambient air pollution in two prospective cohorts. Environ Int. 2017;108:127-36.

25. Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakuri N, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife. 2017;6:24.

26. Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, et al. DNA methylation dynamics of the human preimplantation embryo. Nature. 2014;511(7511):611-5.

27. Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. Genome-wide methylation data mirror ancestry information. Epigenetics Chromatin. 2017;10:1.
28. Kwabi-Addo B, Wang S, Chung W, Jelinek J, Patierno SR, Wang BD, et al. Identification of differentially methylated genes in normal prostate tissues from African American and Caucasian men. Clin Cancer Res. 2010;16(14):3539-47.

29. Liu J, Morgan M, Hutchison K, Calhoun VD. A study of the influence of sex on genome wide methylation. PLoS One. 2010;5(4):e10028.

30. Xia Y-y, Ding Y-b, Liu X-q, Chen X-m, Cheng S-q, Li L-b, et al. Racial/ethnic disparities in human DNA methylation. Biochimica Et Biophysica Acta-Reviews on Cancer. 2014;1846(1):258–62.

31. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology. 2011;12(1).

32. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-Wide Variation of Cytosine Modifications Between European and African Populations and the Implications for Complex Traits. Genetics. 2013;194(4):987-+.

33. Winnefeld M, Brueckner B, Groenniger E, Staeb F, Wenck H, Lyko F. Stable Ethnic Variations in DNA Methylation Patterns of Human Skin. Journal of Investigative Dermatology. 2012;132(2):466–8.

34. Bormann F, Rodriguez-Paredes M, Hagemann S, Manchanda H, Kristof B, Gutekunst J, et al. Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. Aging Cell. 2016;15(3):563–71.

35. Daca-Roszak P, Pfeifer A, Zebracka-Gala J, Rusinek D, Szybinska A, Jarzab B, et al. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies. Bmc Genomics. 2015;16.

36. Francisco C-N. Beta Regression in R. In: Zeileis A, editor Journal of Statistical
37. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333–46.

38. Byun H-M, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, et al. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. Hum Mol Genet. 2009;18(24):4808–17.

39. Carja O, MacIsaac JL, Mah SM, Henn BM, Kobor MS, Feldman MW, et al. Worldwide patterns of human epigenetic variation. Nature Ecology Evolution. 2017;1(10):1577–83.

40. Potapova A, Albat C, Hasemeier B, Haeussler K, Lamprecht S, Suerbaum S, et al. Systematic cross-validation of 454 sequencing and pyrosequencing for the exact quantification of DNA methylation patterns with single CpG resolution. BMC Biotechnol. 2011;11:6.

41. Brakensiek K, Wingen LU, Langer F, Kreipe H, Lehmann U. Quantitative high-resolution CpG island mapping with Pyrosequencing reveals disease-specific methylation patterns of the CDKN2B gene in myelodysplastic syndrome and myeloid leukemia. Clin Chem. 2007;53(1):17–23.

42. Grafodatskaya D, Choufani S, Ferreira JC, Butcher DT, Lou Y, Zhao C, et al. EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. Genomics. 2010;95(2):73–83.

43. Sun YV, Turner ST, Smith JA, Hammond PI, Lazarus A, Van De Rostyne JL, et al. Comparison of the DNA methylation profiles of human peripheral blood cells and transformed B-lymphocytes. Hum Genet. 2010;127(6):651–8.

44. Caliskan M, Cusanovich DA, Ober C, Gilad Y. The effects of EBV transformation on
gene expression levels and methylation profiles. Hum Mol Genet. 2011;20(8):1643–52.

45. Naeem H, Wong NC, Chatterton Z, Hong MKH, Pedersen JS, Corcoran NM, et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. Bmc Genomics. 2014;15.

46. Wessely F, Emes RD. Identification of DNA methylation biomarkers from Infinium arrays. Front Genet. 2012;3:161.

47. Giuliani C, Sazzini M, Bacalini MG, Pirazzini C, Marasco E, Fontanesi E, et al. Epigenetic Variability across Human Populations: A Focus on DNA Methylation Profiles of the KRTCAP3, MAD1L1 and BRSK2 Genes. Genome Biology Evolution. 2016;8(9):2760–73.

48. Zhong H, Kim S, Zhi D, Cui X. Predicting gene expression using DNA methylation in three human populations. PeerJ. 2019;7:e6757.

49. Verdin E, Ott M. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. Nat Rev Mol Cell Biol. 16. England2015. p. 258–64.

50. Janowski BA, Huffman KE, Schwartz JC, Ram R, Hardy D, Shames DS, et al. Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. Nat Chem Biol. 2005;1(4):216–22.

51. Kaikkonen MU, Lam MT, Glass CK. Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res. 2011;90(3):430–40.

52. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 2014;157(1):95–109.

53. Kader F, Ghai M. DNA methylation-based variation between human populations. Mol Genet Genomics. 2017;292(1):5–35.
54. Vidaki A, Kayser M. Recent progress, methods and perspectives in forensic epigenetics. Forensic Sci Int Genet. 2018;37:180–95.

Figures

| Phase                  | Material            | Methods                      | Results                      |
|------------------------|---------------------|------------------------------|------------------------------|
| **Selection**          | B-CELL LINES        | ILLUMINA HM450K ARRAY        | CANDIDATE POPULATION-         |
| Identification of     | CEU<sub>n</sub>=18  | ~450 000 CpG sites           | -DIFFERENTIATING CpGs        |
| candidate pop-diff-metCpG | CHB<sub>n</sub>=18  |                              | (14 CpGs)                    |
|                        |                     |                              | | Max_diff | >1         |
|                        |                     |                              | p<0.05                       |
| **Technical Validation** | B-CELL LINES       | PYROSEQUENCING               | 8 PYROASSAYS                 |
| PyroAssays design &   | CEU<sub>n</sub>=10  | 10 PyroAssays (10 candidate  | (17 CpGs)                    |
| optimization           | CHB<sub>n</sub>=10  | candidate pop-diff-met CpGs +| p<0.05                       |
|                        |                     | 10 neighboring CpGs)         |                              |
| **Biological Validation** | STUDY 1* B-CELL LINES* | PYROSEQUENCING               | 6 PYROASSAYS                 |
| Validation of pop-diff-metCpG | CEU<sub>n</sub>=35  | 8 PyroAssays (17 CpG)        | (12 CpGs)                    |
| in independent samples | CHB<sub>n</sub>=35  |                              | p<0.05                       |
| **Statistical Analysis** | B-CELL LINES*       | STATISTICAL METHODS          | 8 POP-DIFF-MET CpG           |
| Discriminating        | CEU<sub>n</sub>=24  | SVM                          | p<0.05                       |
| potential of the      | CHB<sub>n</sub>=24  | LDA                          |                              |
| selected pop-diff-metCpG |                     | RF                           |                              |
| **Final Outcome**      | PERIPHERAL BLOOD    |                              |                              |
|                        | CELu<sub>n</sub>=20 | CANDIDATE POP(CEU-CHB)-DIFF-MET-CpG MARKER |                              |
|                        | CHBl<sub>n</sub>=20 |                              |                              |

Figure 1

Study design. * cell lines other than those used in Illumina study.
Results of the technical validation of eight PyroAssays. 20 B-lymphocyte cell lines (10 from each population) were tested. The originally selected candidate pop-diff-met CpGs targeted in each PyroAssay are marked with *. Green – CEU population; blue – CHB population. Dots represent methylation levels in individual samples. Box plots denote mean value (lines inside the boxes) and standard deviation. Statistically significant (p<0.05) population differences in the methylation level are marked in red.

Biological validation of the methylation level at 12 CpG sites, performed in B-lymphocyte cell lines (upper panel) and blood samples (lower panel). Dots represent methylation level in the individual samples. Box plots denote mean value (lines inside the boxes) and standard deviation. Statistically significant (p<0.05) population differences in the methylation level are marked in red.
Correlation matrix showing the results of Pearson correlation analysis. Analysis was performed using data from PyroAssays performed in 20 B-lymphocyte cell lines (n=10 from CEU, n=10 from CHB population). Pearson correlation coefficient values and directions are marked with different colors; positive correlation (from white to red on the color scale); negative correlation (from white to blue) (see color-bar next to the matrix).
Figure 5

Accuracy of the classification using three different classification methods A ROC curve and AUC parameter were calculated for: support vector machines (SVM; blue line), linear discriminate analysis LDA (red line), and random forest (RF; green line). Results were obtained based on B-lymphocyte cell lines (n=20 from CEU and CHB). The ROC curve was created by plotting the true positive fraction against the false positive fraction at various threshold settings.
Figure 6

PC analysis separating samples from two populations. Analysis was performed on B-lymphocyte cell lines (n = 48 from CEU and CHB; A) and blood samples (n = 40 from CEU and CHB; B). Vectors length denotes the power of influence of each feature on principal component, while vectors location show how variables correlate with one another. Dots represent individuals from CEU (green points) and CHB (blue points) population.
Relation between the number of CpG used in a training group and the quality of classification. Upper panel: The value of AUC parameter obtained in the training group (B-cell lines: CEU, n=10; CHB, n=10) depending on the number of CpG studied. Lower panel: Precision of sample classification depending on the number of features (CpGs) analyzed in two tested groups: B-lymphocyte cell lines (B1, red color, n=48) and blood samples (B2, blue color, n=40).

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

Human population identification_supplementary.docx