Technical and biological sources of unreliability of Infinium type II probes of the Illumina MethylationEPIC BeadChip microarray

Tatiana Nazarenko[1,2,3], Charlotte D. Vavourakis[1,2], Allison Jones[3], Iona Evans[3], Antony W. Watson[4], Kirsten Brandt[4], Clive Carter[5], Alexey Zaikin[3,6], Chiara Herzog[1,2*] and Martin Widschwendter[1,2,3,7*]

[1] Research Institute for Biomedical Aging Research, Universität Innsbruck, Innsbruck, 6020, Austria.
[2] European Translational Oncology Prevention and Screening (EUTOPS) Institute, Milser Str. 10, Hall in Tirol, 6060, Austria.
[3] Department of Women's Cancer, UCL EGA Institute for Women's Health, University College London, Medical School Building, 74 Huntley Street, London, WC1E 6AU, UK.
[4] Human Nutrition Research Centre, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.
[5] Transplant and Cellular Immunology Laboratories, Leeds Teaching Hospital NHS Trust, St James's University Hospital, Leeds, LS9 7TF, UK.
[6] Department of Mathematics, University College London, 25 Gordon Street, London WC1H 0AY, UK
[7] Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

* Correspondence: Email: Chiara.Herzog@uibk.ac.at or Martin.Widschwendter@uibk.ac.at

ABSTRACT

Over recent years countless epigenetic studies have been undertaken to investigate DNA methylation (DNAme) patterns in health and disease, frequently using Illumina Methylation arrays. Here we focus on the reliability of Infinium type II probes (~85% of all probes) on this platform and demonstrate that Illumina MethylationEPIC BeadChip v1.0 data from paired whole blood samples taken six weeks apart display an unexpectedly high level of variability on probes with low normalized mean intensities (MI). We propose a method for estimating probes 'unreliability' scores by simulating the β-value distribution influenced by technical noise and averaging the spread over all samples. We show that these unreliability scores negatively correlated with MI, which is associated with the number of C-bases in the probe sequence. To validate our results, we demonstrate that also for two publicly available HumanMethylation450K BeadChip datasets unreliability increases with increasing variability measured between replicates. Furthermore, we find that removing low MI type II probes during initial pre-processing can improve precision of DNAme derived cell type composition and age predictors, and that most of these less reliable probes are still present on the newer version (v2.0) of the MethylationEPIC BeadChip. Finally, to advance existing pre-processing pipelines, we implemented an R package for estimating MI and probe ‘unreliability’ scores.

INTRODUCTION

DNA methylation (DNAme) is a chemical modification of DNA that entails the addition of a methyl group to a cytosine (C) residue resulting in 5-methylcytosine, and most commonly occurs in the context of CpG dinucleotides in humans (1). The study of epigenetics and DNA methylation has become one of the most topical areas of genomic research in recent years, both from a functional point of view and a clinical perspective, owing to its potential application in cancer risk prediction and early detection strategies (2, 3). The two most widely used techniques to study epigenome-wide DNAme are whole-genome bisulphite sequencing (WGBS) and Illumina methylation arrays. Both technologies require bisulfite pre-treatment of DNA to enable distinction of methylated from unmethylated cytosine residues in the context of CpG dinucleotides. Whereas WGBS provides information regarding the DNAme status of a series of linked CpGs, the Illumina methylation arrays allow a more affordable and high-throughput assessment of the methylation status of a subset single CpGs dinucleotides throughout the genome.

The Illumina BeadArray technology has undergone substantial re-development over the years and the total number of CpGs that can be simultaneously analyzed has increased substantially from ~25,000 in 2008 (HumanMethylation27KBeadChip), to ~485,000 in 2011 (HumanMethylation450K BeadChip), to over ~850,000 CpG sites in 2016 (MethylationEPIC BeadChip v1.0), and finally to over ~935,000 CpG sites in 2022 when MethylationEPIC BeadChip v2.0 was released.

Illumina MethylationEPIC BeadChip microarrays, thereafter, denoted as EPIC arrays, include two different types of bead chemistry, Infinium Type I and II probes (4, 5). Type I probes have two separate probe sequences per CpG dinucleotide (one each for methylated and unmethylated CpGs), whereas Type II probes have just one probe sequence per CpG dinucleotide. Consequently, Type II probes take up much less physical space on the arrays than Type I probes and are the most abundant type on the EPIC array, constituting ~85% of all probes. Discrimination of methylated versus unmethylated alleles is made possible by single nucleotide primer extension which results in the incorporation of Cy3 or Cy5 labelled nucleotides into the target sequence and emitting green or red fluorescence, respectively. The level of methylation at specific CpG sites is expressed as *Beta* (β-value), which represents a constant from '0' (unmethylated) to '1' (fully methylated) and for type II probes can be written as:

$$\beta = \frac{GREEN}{GREEN+RED+\alpha} \; ; \qquad (1)$$

with α a small positive constant (typically 100) added to the equation to avoid dividing by zero when both GREEN and RED signals are equal to 0. If β = 0, then the interrogated CpG is unmethylated (there is no green signal), if β = 1, then the interrogated CpG is methylated (there is no red signal).

When assessing Illumina Methylation array data, basic pre-processing steps would typically include identifying probes and/or samples with a low signal to noise intensity which should be excluded (6), correcting for background intensity and dye bias (7), performing within-array normalization to reduce differences in beta-distributions obtained from Infinium I and II probes (8), and imputing missing data (9). For this, several established methods have been benchmarked and implemented into pre-processing pipelines available as R packages, such as *minfi* (10), *ChAMP* (11) and the latest *ENmix* (12). Additionally, previous studies have identified the necessity to exclude low-specificity probes that

can bind to multiple sequences within the genome, as well as probes that contain genetic variants in their underlying sequence (5, 13). Lastly, a recent study by Sugden *et al.* (14) identified a large set of 'unreliable' probes that poorly reproduced methylation values when samples from the same DNA source were measured either on the HumanMethylation450K or MethylationEPIC BeadChip. However, to date comprehensive understanding of the factors which influence the reliability of Illumina array probes is lacking (where *reliability* refers to the ability to reproduce data), which has substantial implications for the accurate interpretation of array data, especially since a typical experimental design for Illumina Methylation arrays does not include technical or biological replication.

In this study, we present a series of comprehensive analyses that explore yet unidentified factors affecting the validity of Illumina Methylation array data, purposely restricted our assessment to Infinium Type II probes, as they represent by far the most abundant probe type on the latest Illumina MethylationEPIC BeadChips. Paired longitudinal data from 142 paired blood samples (from 71 volunteers) collected 6 weeks apart was generated with the MethylationEPIC BeadChip v1.0, which enabled us to distinguish inter-individual DNAme variability with intra-individual DNAme data over time. Our results reveal new insights into factors affecting the variability of DNAme derived from the EPIC array and we thus propose a novel, data-driven method for the assessment of probe reliability. We expect that these findings will further improve existing pre-processing pipelines and the subsequent interpretation of next-generation Illumina Methylation array results.

MATERIAL AND METHODS

Sample collection and analysis clinical intervention study

93 individuals were recruited to the TACT (Turmeric-Anti-Inflammatory & Cell Damage Trial – clinical trial number NCT02815475) for a 6-week intervention study. There were 3 arms to the study: one group ('Turmeric Capsule' group) received a 400 mg Turmeric capsule providing 0.27g curcuminoids/day, a second group ('Placebo' Group) a sugar placebo (xylitol), and a third group were asked to regularly cook with Turmeric powder ('Turmeric Powder' group) providing 0.24g curcuminoids/day in their food every alternate day, all for a period of 6 weeks. Ethical approval number 16-WAT-23 was granted by Newcastle University's SAgE ethics committee. 71 participants (53 females and 18 males) completed the study and provided full sets of usable 12-h fasting whole blood samples, which were collected at the start and end of the 6-week intervention into PAXGene DNA blood tubes (Becton Dickenson, 761165). Full blood counts were complemented with measurements of lymphocyte subsets (T/B/NK cells) using the fluorescently labelled antibodies CD3, CD4, CD8, CD19, and CD56, targeting T cells, T helper cells, cytotoxic T cells, B lymphocytes and NK cells (CD3 negative), respectively. Briefly, 50 µL blood were added to the antibody mix and incubated for 20' at RT in the dark. To lyse the red blood cells, the mixture was incubated 10' with 3 mL lysis buffer, washed (PBS-1%FBS) and cells were resuspended prior to flow cytometry analysis using a FACSCanto II (Becton Dickinson). DNA from the blood was extracted using the Machery Nagel NucleoMag® Blood 200 µL extraction kit (cat, 744501) and 500 ng total DNA was bisulfite modified using the EZ-96 DNA Methylation-Lightning kit (Zymo Research Corp, cat, D5047). 8 µl of modified DNA was subjected to methylation analysis on the Illumina

Infinium MethylationEPIC BeadChip (Illumina, CA, USA) at UCL Genomics according to the manufacturer's standard protocol.

Normalization MethylationEPIC data and immune cell subtype inference

Along with the raw beta values (obtained by formula (1) using raw intensities), we also use normalized beta-values obtained from 3 different pipelines: *minfi preprocessFunnorm* (10), *ChAMP* (11), and *ENmix* (12).

Immune cell subtypes were inferred using the *EpiDISH* R package (15) ("RCP" method, centDHSbloodDMC.m) with or without the different normalization methods and compared with actual FACS measurements of lymphocytes, monocytes, and granulocytes (with granulocytes = neutrophiles + basophiles + eosinophils). Probes were considered cell type-dependent when FDR < 0.05 for both p-values associated with the slope of two linear regressions, conducted for samples of the first and second visits separately, of beta-values against the FACS-measured neutrophile and lymphocyte cell fractions.

Differences in methylation changes in treatment groups

To investigate differences in methylation changes between two visits across three treatment groups 'Placebo' (24 patients: 16 females, 8 males), 'Turmeric Capsule' (25 patients: 17 females, 8 males), and 'Turmeric Powder' (22 patients: 20 females, 2 males), we utilized two approaches considering differences in the original beta values between visits, and differences in residuals from linear regression models (beta values versus real neutrophile cell fraction, as the largest blood cell fraction) that were fitted on samples from the first visit only and then applied to all samples. We then performed pairwise comparisons of the groups using the Wilcoxon test, as well as a common comparison of all three groups using a Kruskal-Wallis test. However, none of these tests (for any variant of beta values: raw, *minfi* (10), *ChAMP* (11) and *ENmix* (12)), revealed any statistically significant positions (defined as p-values adjusted for a false discovery rate (FDR) <0.05).

Since the proportion of males in the 'Turmeric Powder' group was lower than in the other two groups, we repeated the tests on only female samples from all three group. Despite this additional analysis, we still did not find any statistically significant positions.

SNP analysis

Single nucleotide polymorphisms (SNPs) were identified from probes with underlying genetic sequence variation at target CpG sites listed by Pidsley et al. ((5); Supplementary Table S4) and further divided into subsets of 0- and 1-position SNPs respective to the probe target CpG. Theoretically these two SNP types affect methylation profiles in specific ways. The 'SNP0-effect' associated with a 0-position of a Infinium type II probe (i.e., with C base of target CG pair) will cause (addition of) false Green (if SNP is G nucleotide) or false Red signals (if SNP is T or A nucleotide), whereas the 'SNP1-effect' associated with a 1-position of a type II probe (i.e., with G base of target CG pair) may cause degradation of the total signal intensity, since the sequence of the probe is not complimentary to the target. Notable, SNPs in the other closest positions towards the end of the probe (likely 2-5 bp away) or large inserts and deletions in more distant positions can have the same 'SNP1-effect'.

The 'SNP0-effect' results in a tri-modal distribution of beta-values on probe where each mode is represented by carriers of one of three variants: (CC – C on both chromosomes; C(SNP) or (SNP)C –

C only on one chromosome; or (SNP)(SNP) where are NON C on both chromosome (for details see Fig. 1d,f and SM:Fig. 1). Conversely, the 'SNP1-effect' results in a tri-modal distribution of intensities levels on probe, where each mode is represented by carriers of one of three variants (highest level: GG – G on both chromosomes; middle level: G(SNP) or (SNP)G – C only on one chromosome; lowest level: (SNP)(SNP) where are NON G on both chromosome). Importantly, on the level of beta-values, the 'SNP-1-effect' results in a bi-modal distribution (where one mode corresponding with (SNP)(SNP) variant and second one combines 2 others variants) (for details see Fig. 1d,g).

MI score

Each sample has its own average intensity likely due to the technical processes occurring in each normalization plate or beadchip during processing of the material. To account for this, we calculate the Mean normalized Intensity of each Infinium type probe (MI score) as follows:

For n samples, calculated across all type II or type I probes separately:

$\overline{M_i}$ and $\overline{U_i}$; average raw methylated and unmethylated signal of ith sample (2)

$AIS_i = \overline{M_i} + \overline{U_i}$; Average Intensity of ith Sample (3)

For each ith sample and each jth type II or I probe:

$N_{ij} = \frac{meth\ signal_{ij} + unmeth\ signal_{ij}}{AIS_i}$; Normalized intensity of jth Probe on ith Sample (4)

For each jth type II or I probe:

$MI_j = \frac{1}{n}\sum_{i=1}^{n} NI_{ij}$; MI score (5)

Unreliability score

We propose two options to select noise-generate probes from user input data: (i) probes on the Y-chromosome for female samples (Y-noise), or (ii) probes for which more than 50% of samples have a p-value > 0.01 as calculated with the *detectionP* function (16) from the *minfi* package (10) (p-noise). For each probe type separately, all raw methylated and unmethylated signals of the noise-generate probes are then combined into a two-dimensional Noise Distribution. Both noise-generating methods for the two Infinium probe types are implemented in the R package epicMI (17), but for this study, we focus on the Y-noise method (using only female samples) and the Infinium type II probes, for which the methylated and unmethylated signals correspond to green (G) and red (R) signals, respectively.

For a given set of noise-generating probes, we created a Reliability Map, consisting of a grid of pairs for which 1000 points from the Noise Distribution were randomly selected:

G_k, R_l with k, l = (0, 5000, step = 100) and $(G^1_{err\,(k,l)},\ R^1_{err\,(k,l)})$, …, $(G^{1000}_{err\,(k,l)},\ R^{1000}_{err\,(k,l)})$; (6)

For each pair k and l, the distribution from 1000 β-values and an associated Q-score were calculated:

$\beta^m_{(k,l)} = \frac{G_k + G^m_{err\,(k,l)}}{G_k + G^m_{err\,(k,l)} + R_l + R^m_{err\,(k,l)}}$ with m=1:1000 ; (7)

$Q_{k,l}$ = quantiles difference ($q_{0.975}$ and $q_{0.025}$) of distribution of $\beta^{1\ldots1000}_{(k,l)}$; Q score (8)

Thus, for type II probes the Reliability Map is a two-dimensional grid (G_k, R_l), of Green and Red signals, where each cell is assigned the $Q_{k,l}$ value, which is associated with the unreliability of the beta value

obtained at the corresponding intensities. For each real (i.e., not modelled) pair of $Green_{ij}$ and Red_{ij} signals from the ith sample and each jth probe, we subtracted the average green and red noise intensities and then found the closest point (G*,R*) and associated Q*-value on the Reliability Map (if the intensity values lie outside the grid, then $Q^* = 0$, that is, it is very reliable) and finally calculate the Unreliability score for each jth type II probe across all samples in the data set:

$$\text{Unreliability} = \frac{1}{n}\sum_{i=1}^{n} Q^* \; ; \qquad \text{Unreliability score (9)}$$

RESULTS

Unexpected DNAme variability in repeated blood samples

We studied whole blood DNA methylation profiles of 71 volunteers at two time points separated by a six-week interval (n=142). For each CpG locus, we calculated SD β, i.e., standard deviation of the β values within the population at visit 1, and Δβ, i.e., the average (over all individuals) of differences in β values between visits (over time) for the same person. This revealed two distinct groups of CpGs targeted by Infinium Type II probes in terms of DNAme variability (Fig. 1a): first, sites demonstrating a wide range of variability across the population of samples of a single visit (time point 1), and second, sites demonstrating a high degree of variability over time for the same individual. This distinction was evident both for β values calculated from "raw" intensities (Eq. (1)), as for β values normalized with various previously described methods (*minfi preprocessFunnorm* (10), *ChAMP* (11), or *ENmix* (12)). As the interval between obtaining these paired blood samples was only six weeks, we were surprised by the relatively large degree of probe variability. Although participants were allocated to one of three treatment groups ('Placebo', 'Turmeric Capsule' or 'Turmeric Powder'), no significant treatment effects were observed (see Material and Methods section for details). Consequently, we initially checked whether the observed patterns of variability are linked to the genetic background of the targeted sites, in particular with sex chromosomes and single nucleotide polymorphisms (SNPs), both of which are generally recommended to be removed from methylation array data during preprocessing (13).

When grouping probes by their chromosomal location (X, Y, and "other"), sex chromosome-associated probes within our dataset exhibited a high variability across the population, which is partially expected as our participant cohort included both men and women. However, strikingly, methylation values in type II probes targeting sex chromosomes also showed a high variability over time. This effect was less pronounced for β values obtained with *minfi* normalization, which was likely due to the special normalization performed for sex chromosome-associated probes in this pipeline and also for *ENmix* β values (which obtained with using *minfi* pipeline). Red and green signal intensity at sex chromosomes, as expected, was influenced by biological sex and the resulting sex chromosome copy number (Fig. 1b,c): the total signal intensity of an example CpG probe mapping to the X chromosome is higher in females than males, since females have two X chromosomes; conversely, the total signal intensity of an example CpG probe mapping to the Y chromosome is close to '0' since females do not carry a Y chromosome, although some Y-CpGs might be located in pseudoautosomal regions.

Genetic variation and single nucleotide polymorphisms (SNPs) may also influence probe intensity. With respect to the Infinium type II probes, the two most relevant positional types of SNPs occur at position '0' immediately after the 3' end of the probe (SNP 0), where the SNP specifically affects the cytosine residue of the interrogated CpG, or at position '1' at the very end of the 3' end of the probe (SNP 1), where the SNP specifically affects the guanidine residue of the interrogated CpG (Fig. 1d). These two SNPs have distinct impacts on signal: a SNP 0 can result in false red or green signals, depending on the nature of the SNP replacing the 'C' residue, while a SNP 1 impairs hybridization and extension and results in a loss of signal. Other studies (5, 13) previously identified probes on the MethylationEPIC BeadChip v1.0 whose reliability is impacted by SNPs within sequence they target, and we have highlighted these probes in our dataset (Fig. 1e). Interestingly, the subset of probes with the highest Δβ, were not annotated with either SNP 0 or SNP 1. Like the sex-chromosome associated probes, SNP-associated probes contribute to a high degree of variability within the population, but they do not fully explain the high variability in DNAme data over time within individuals. We further find that a SNP 1 results in a bimodal β distribution (corresponding to signal or loss of signal; Fig. 1g), where the "true" variant is represented by the upper layer of intensity. A SNP 0 that gives rise to either false red or green signals manifest as a trimodal pattern of β distribution (mixed, red, green, Fig. 1f), where the nature of the polymorphism – i.e., which base has replaced the cytosine residue – is known and it is possible to determine which particular mode of the β distribution corresponds to the values of the "true" (C/C) variant (see supplementary material SM:Fig. 1b,c for further detail).

Modeling the impact of signal intensity on β value reliability

Since the properties of both SNP-related and sex chromosome-associated probes are closely related to signal intensity (either through DNA quantity, false signals, or a loss of signal), but neither of these genetic-associated probes fully explain the observed variability in repeated blood samples from the same individuals, we further scrutinized the impact of Type II probe intensities on DNAme variability. For each probe on the array, we calculated a mean intensity (MI) value, which represented the corrected mean overall signal strength of the probe, and since it is independent of sample-dependent effects allows for direct comparison of probe quality (see Material and Methods section for details). Overall, Type II probes with the highest level of time-dependent variability have a low MI value (Fig. 2a). Probes on the lower end of the MI scale in our dataset show a low reproducibility in paired blood samples (example cg21373150; Fig. 2b), whereas probes on the higher end of the MI scale a high reproducibility (cg17588455; Fig. 2c). We hypothesize that this high variability at low intensity levels is caused by noise. Therefore, the MI score may potentially allow for the identification of 'noisy' or 'unreliable' probes, i.e., probes which do not provide reproducible results between two timepoints or replicates.

We created a simulation model to estimate the impact of noise on probe "unreliability" (how high the variability of the beta values, given the green and red signals with noise added), leveraging the use of two types of noisy probes generally removed from methylation array data. For the Y-noise method, we repeatedly (1000 times) added a point from the noise distribution obtained from Y chromosome-associated probes in women to a grid with fixed values for green and red signals, calculating 1000 β values each time to finally derive the value "Q", which represented the difference between 0.975 and

0.025 quantiles of this distribution (Fig. 2d). Each pair of green and red signals is therefore assigned a number, Q, that characterizes the "unreliability" of the β value based on it: the higher Q, the higher the β unreliability. Next, we applied this "reliability map" to our data and calculated the average probe "unreliability" by finding the closest point in the grid of the map and its associated Q for each pair of green and red signals of each probe and then averaged these values for each probe. Similarly, for the p-noise method, the reliability map is constructed by using noise distribution from "outlier" (unreliable) probes found by the widely used *detectionP* function (16) of the *minfi* package (10). Both options of selecting noise-probes work somewhat similarly in terms of obtaining unreliability values (SM:Fig. 3 in supplementary materials), but, in this study, we further focus on the Y-noise method applicable to female samples.

Fig. 2e visualizes the dependence of probe unreliability (Q) defined by the Y-noise method on the MI score, and highlights that "outlier" (unreliable) probes identified with the *detectionP* function of the *minfi* package (10) (standard threshold p = 0.01 (red), or p = 1.e-40 (orange; more stringent threshold proposed by (6)) are also probes associated with values of high unreliability. However, the method for detecting outlier probes by p-value only takes into account the proximity of values in the noise distribution, which does not allow for intuitive selection of the cut-off threshold. In contrast, the unreliability (Q) score we present here, is a continuous measure which allows for selection of a threshold based on simple considerations. Probes with an unreliability (Q) less than 0.1 corresponds to probes with an MI>0.35 (Fig. 2e). Importantly, excluding probes based on these criteria virtually removes all probes demonstrating high variability over time (Fig. 2f).

Linking unreliability with probe sequence composition, target sequence copy number and other reliability analyses

Investigating probe composition to identify factors associated with reliability, we found that probes with a low MI score tend to have a lower C content, i.e., G bases in their targets (Fig. 3a). Stronger physical binding between G-C base pairs than A-T base pairs could result in an increase in bound targets and fluorescent signal. Furthermore, probes targeting island, shore, shelf, and open sea regions inherently differ in their CG content (Fig. 3b) and have corresponding average reliabilities. Open sea region probes have a lower MI and higher unreliability scores than probes located in island regions (Fig. 3c,d).

As was previously known, the high intensity of probes depends on the number of copies of their complementary sequences on DNA (13). We analyzed how this is combined with the probe C-base content. In Fig. 3e, we demonstrate that MI (which is analogous to the raw intensities used by Zhou) increases with copy number (for sequences above 30 nucleotides) in each group of probes containing a certain C base in their sequence. Moreover, the intercepts of linear regressions, describing such dependencies, grow synchronously with the increase in the number of C bases in the probes sequence.

Interestingly, our unreliability measure is not associated with a reliability measure proposed by Sugden *et al.* (14), which was calculated using ICC (Intraclass Correlation Coefficient) on β of repeat measurements of the same DNA samples (Fig. 3f). The reliability measure proposed by Sugden *et al.*

does not correlate with MI (Fig. 3g – lower figure) and SNP-associated probes were deemed the most reliable using this measure (Fig. 3g – upper figure), which seems to be due to the high spread of β values in SNP 0 and SNP 1 (Fig. 3h – upper figures). Of note, probes which have the same Sugden reliability score may have different MI and, therefore, differ level of unreliability (Fig. 3h – lower figures). In contrast to the method proposed by Sugden *et al.,* our method of assessing probe reliability is not based on cross-correlation of samples (which can be different in intensity, and therefore result in β bias), but instead offers insights into the reliability of probes based purely on intensity and noise distribution.

Consistency of unreliability scores and probe MI with respect to biological and technical variation

Our method to estimate unreliability of Infinium type II probes is based on the analytical modeling of the effect of noise on probe intensities and explains the high values of Δβ well. However, since our dataset is comprised of paired biological replicate samples, we further investigate whether changes in methylation values over time could still be explained partially by biological influence, despite the absence of a treatment effect, and not only probe unreliability. Fractions of cell subtypes in the blood can change rapidly, for example the proportion of lymphocytes in peripheral blood can rapidly increase due to acute illness, influencing the β values that integrate the methylation status from all cell subtypes in the samples, and further explaining the variation seen in time for the same individuals. Indeed, the proportion of the two main immune blood cell subtypes, neutrophils and lymphocytes, changed between the two visits (SM:Fig. 4a, b). We therefore selected probes whose β values are strongly influenced by these two immune subtypes (SM:Fig. 4c) and evaluated the variability in these cell type-dependent probes between patients at visit 1 (SD β) and over time (Δ β; SM:Fig. 4d), revealing considerable variability over time even after removing all other variable or unreliable probes (probes on sex chromosomes, SNPs-associated probes, and unreliable probes with MI≤0.35).

Therefore, we further want to demonstrate the consistency of the unreliability scores on true technical replicates using two published datasets GSE174422 (256 female samples: 128 duplicate pairs) and GSE52731 (70 replicates of the same liver sample and 56 replicates of the same placenta sample), both generated with the Illumina Infinium 450k Human Methylation Beadchip v1 (Fig. 4). Using raw beta-values, we calculated a 'β variability' for each probe. As described above, for our dataset this is the average of all beta differences calculated between the same individual over time. For GSE174422, the 'β-variability' was defined as the average of all beta differences calculated between 2 replicates of the same individual, for GSE52731 as the standard deviation of β (for lever and placenta tissue separately). In all datasets, 'β-variability' increased with increasing unreliability (calculated using Y-noise approach since all samples are female (Fig. 4a) and decreased with MI (Fig. 4b). The average high level of 'β-variability' for GSE174422 could be explained by the fact the two duplicates of the same individual were always bisulfite-converted on different plates and assayed on different arrays, in contrast to the current study where both samples of the same individual were treated together and resided on the same beadchip. The processing method is not specified for GSE52731, but since many

replicates were produced for each tissue (70 for liver sample and 56 for placenta sample), we may expect that they were located at least on different beadchips (each of which has 8 positions), which also contributes to an increase in the average 'β-variability' level. Finally, on probes common to Illumina Methylation EPIC and Illumina Methylation 450k arrays, MI values are highly correlated with each other among all considered studies (Fig. 4c), regardless of the average sample intensities (Fig. 4d) and distributions on noise probes (Fig. 4e), suggesting MI can serve as a reproducible score that can be used to filter out unreliable probes.

Removal of unreliable probes can improve deconvolution methods and age clocks

We examined the effect of removing poor quality probes on the accuracy of calculating *EpiDISH* (15) cell fractions by comparing it against data on real cell subtype fractions obtained via flow cytometry. Fig. 5a shows that the successive removal of unreliable probes lead to an improvement in the calculation of the fractions of granulocytes (especially significant (Wilcoxon paired test) improvement for GRA (granulocytes) for *minfi* and raw β values) and for MON (monocytes), a slight trend for improvement can be observed. The sensitivity of EpiDISH algorithms to the removal of unreliable probes may indicate that their removal (or the use of unreliability values obtained in this study as weights in model training) can improve the quality of the algorithm.

Removing unreliable probes determined by our method can also improve the accuracy of popular epigenetic clocks (we used the methylclock package (18) for calculation of BLUP and EN (19), Hannum (20), Horvath (21) and DNAm PhenoAge (22) clocks, and the WIDclocks package (23) for computation of the WID-general clock (24)) trained to predict age (Fig. 5b). We note a significant increase in the accuracy especially for the Hannum and the WID-general clocks after the removal of low-intensity Type II probes with MI≤0.35. BLUP and EN clocks were not affected by the removal of low intensity probes, and we found that in these predictors, the most reliable probes have the highest coefficients, and the proportion of unreliable probes (in this study we define them as probes with an unreliability higher than 0.1) in both sets is not more than 5% (Fig. 5c,d). Since these models were trained on a large set of samples of different quality, with a large spread of their intensities, they were forced to penalize low-quality probes, since on such a set they become particularly noisy.

Probes removed to MethylationEPIC BeadChip v2.0 and probes Type I

A new MethylationEPIC BeadChip v2.0 was released in November 2022. The new manifest reports that a large number of 'underperforming' probes were removed compared to the v1.0 manifest (approximately 140,000, i.e., 23% of all type I probes and 15% of all type II probes). We examined the removed probes but found no evidence of an enrichment for probes with high unreliability or low intensity (see SM:Fig. 5 in supplementary materials). Therefore, we assume that the issues raised in our current study will remain of high importance and also for the new version of EPIC array. We also verified that our method of determining unreliability works for EPIC v1 on average, with the same accuracy as on this array after removing declared probes. As shown in the supplementary materials, SM:Fig.6, after the removal of probes, the outliers of the displayed boxplots become more compact,

which means that the removed probes were outliers in terms of both MI and unreliability. We also note that despite the announced large-scale removal of SNPs, some SNP 0 (~15%) and 1 (~25%) probes remained on v2.0. In addition, we observe that some probes are not marked as containing SNPs (neither by Pidsley (5) and Zhou lists (13), nor by the Illumina Manifest), but clearly demonstrate SNP 0 or 1 behavior (see examples in supplementary materials, SM:Fig. 7 for SNP_0-effect and SM:Fig. 8 for SNP_1-effect).

Our method can be used to calculate the MI and reliability of Type I probes (see description for examples in (17)). But as shown in supplementary materials in SM:Fig. 9, the unreliability values for such probes are sensitive to the noise-method (p-noise or Y-noise) and to the set of the probes (all probes of MethylationEPIC BeadChip v1.0 or on their part not removed from version 2.0). Therefore, we recommend using only MI values for Type I probes and using a threshold for them determined from analysis on Type II probes.

DISCUSSION

Despite considerable investment in improving existing analysis pipelines for popular Illumina methylation arrays, room for improvement remains. To facilitate the generation of meaningful findings that will not only increase our understanding of the epigenome and its relationship with health and disease, but also translate into clinically useful tools, it is vital that we fully understand the variability associated with this type of microarray. Here we show that DNAme data from paired whole blood samples taken from the same individuals six weeks apart display variability across the population studied, but also a large degree of variability over time that cannot be attributed to underlying genetic or biological factors alone. Much of the ‘unexplained’ temporal variability in the current study can be attributed to probe quality, which is primarily dictated by the probe sequence complexity and genome location.

We developed an approach for assessing the unreliability of beta-values, which is based not on the proximity to noise (as the p-value methods do (16)), but on contribution of noise to any of the final signal intensities. Our probe unreliability values not only correlated with variability in time on our dataset, but also with variability measured between replicates, including in independent publicly available datasets. We also demonstrated that removing unreliable, low-intensity Type II probes from the data not only adequately removes most of the unexplained DNAme variability in samples from the same individuals but also improves existing deconvolution methods and age clocks.

We show that noise affects methylation values differently at different intensity levels; noise has a dramatic effect on beta values at low intensity, while at high intensities, the signal cancels out the effect of noise. We disprove the earlier assumption (through the p-value method (16)) that there are only 'bad' probes (those that completely lose signal) and all others are equally good. Even for probes with a low p-value, we demonstrate that the ability of probes to accumulate a signal is not the same, and for example, depends on the C-content in the sequence. On the one hand, this observation can help Illumina to achieve leveling of such an effect within the technological process, on the other hand, it will allow scientists to provide more qualitative comparisons on different regions of DNA (on Islands and

the Open Sea, for example). Excluding low-intensity probes from the analysis (or at least including MI and/or 'unreliability' scores as additional features or weights) can help increase the accuracy of existing (as we have shown) and emerging predictive models on this type of data. At the same time, we note that some existing models (as was shown for age predictive models EN and BLUP (19)) can independently penalize less reliable probes more strongly, but only if they were trained on large datasets of different quality samples (from different experiments), which are not always available to researchers.

We merely suggest excluding low-intensity probes from methylation analyses, but do not exclude that correction or normalization methods may emerge in the future based on the results of this work that can salvage the data generated from these probes. Also, while we could not fully unravel the detailed relationship between probe and sample intensity here, we note that samples from different studies tend to be of different quality, yielding different average intensities, which in turn can also affect the reproducibility of beta-values on probes with different intensities and estimated probe reliability. We also note the importance of the still unexplored question of which factors contribute more strongly to the observed intensities in samples from different experiments, such as sample storage or the exact hybridization and staining procedure followed when processing EPIC beadchips in the wet laboratory.

We implemented our methods to calculate MI unreliability scores into an R package, epicMI (17), which may be used in congruence with existing preprocessing pipelines for Illumina methylation array data. We believe that taking into account raw intensity- and not just beta-values during initial quality control may greatly improve the reproducibility and repeatability of the downstream analyses and modeling performed on this type of data.

AVAILABILITY

GSE52731 and GSE174422 datasets are available on the GEO portal. Raw DNA methylation data from the TACT study cannot be deposited as volunteers did not specifically consent to make data containing genetic information (that could identify individuals) publicly available; preprocessed data with probes at SNP positions (genetic information) can be made available upon reasonable request from researchers in countries compliant with GDPR law only.

The R code for calculation of the MI and unreliability scores is available in a GitHub repository (https://github.com/ChVav/epicMI) (17) and on FigShare (doi: 10.6084/m9.figshare.22270870).

ACCESSION NUMBERS

GSE52731 and GSE174422 datasets are available on GEO portal.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank all volunteers who provided samples for this research.

FUNDING

This work was supported by funding from the European Union's Horizon 2020 European Research Council Program, H2020 BRCA-ERC [Grant Agreement No. 742432]; the charity, The Eve Appeal (https://eveappeal.org.uk/); and the European Union's Horizon 2020 HEAP research program [Grant Agreement No 874662]. AZ and TN acknowledge support by the MRC grant MR/R02524X/1.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Varley,K.E., Gertz,J., Bowling,K.M., Parker,S.L., Reddy,T.E., Pauli-Behn,F., Cross,M.K., Williams,B.A., Stamatoyannopoulos,J.A., Crawford,G.E., *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*, 23, 555–567.

2. Barrett,J.E., Herzog,C., Jones,A., Leavy,O.C., Evans,I., Knapp,S., Reisel,D., Nazarenko,T., Kim,Y.N., Franchi,D., *et al.* (2022) The WID-BC-index identifies women with primary poor prognostic breast cancer based on DNA methylation in cervical samples. *Nat Commun*, 13, 449.

3. Barrett,J.E., Jones,A., Evans,I., Reisel,D., Herzog,C., Chindera,K., Kristiansen,M., Leavy,O.C., Manchanda,R., Bjorge,L., *et al.* (2022) The DNA methylome of cervical cells can predict the presence of ovarian cancer. *Nat Commun*, 13, 448.

4. Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L., *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, 98, 288–295.

5. Pidsley,R., Zotenko,E., Peters,T.J., Lawrence,M.G., Risbridger,G.P., Molloy,P., Djik,S.V., Muhlhausler,B., Stirzaker,C. and Clark,S.J. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*, 17, 208.

6. Heiss,J.A. and Just,A.C. (2019) Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clin Epigenetics*, 11, 15.

7. Triche,T.J., Weisenberger,D.J., Berg,D.V.D., Laird,P.W. and Siegmund,K.D. (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res*, 41, e90–e90.

8. Niu,L., Xu,Z. and Taylor,J.A. (2016) RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. *Bioinformatics*, 32, 2659–2663.

9. Lena,P.D., Sala,C., Prodi,A. and Nardini,C. (2020) Methylation data imputation performances under different representations and missingness patterns. *Bmc Bioinformatics*, 21, 268.

10. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30, 1363–9.

11. Morris,T.J., Butcher,L.M., Feber,A., Teschendorff,A.E., Chakravarthy,A.R., Wojdacz,T.K. and Beck,S. (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, 30, 428–430.

12. Xu,Z., Niu,L. and Taylor,J.A. (2021) The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines. *Clin Epigenetics*, 13, 216.

13. Zhou,W., Laird,P.W. and Shen,H. (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*, 45, e22.

14. Sugden,K., Hannon,E.J., Arseneault,L., Belsky,D.W., Corcoran,D.L., Fisher,H.L., Houts,R.M., Kandaswamy,R., Moffitt,T.E., Poulton,R., *et al.* (2020) Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. *Patterns*, 1, 100014.

15. Zheng,S.C., Webster,A.P., Dong,D., Feber,A., Graham,D.G., Sullivan,R., Jevons,S., Lovat,L.B., Beck,S., Widschwendter,M., *et al.* (2018) A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics*, 10, 925–940.

16. detectionP: detection p-values for all probed genomic positions https://rdrr.io/bioc/minfi/man/detectionP.html.

17. Nazarenko,T. and Vavourakis,C. (2023) epicMI: Calculation of MI and unreliability of MethylationEPIC BeadChip array probes. *Github*, https://github.com/ChVav/epicMI.

18. Pelegí-Sisó,D., Prado,P. de, Ronkainen,J., Bustamante,M. and González,J.R. (2020) methylclock : a Bioconductor package to estimate DNA methylation age. *Bioinformatics*, 37, 1759–1760.

19. Zhang,Q., Vallerga,C.L., Walker,R.M., Lin,T., Henders,A.K., Montgomery,G.W., He,J., Fan,D., Fowdar,J., Kennedy,M., *et al.* (2019) Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med*, 11, 54.

20. Hannum,G., Guinney,J., Zhao,L., Zhang,L., Hughes,G., Sadda,S., Klotzle,B., Bibikova,M., Fan,J.B., Gao,Y., *et al.* (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, 49, 359–367.

21. Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol*, 14, R115.

22. Levine,M.E., Lu,A.T., Quach,A., Chen,B.H., Assimes,T.L., Bandinelli,S., Hou,L., Baccarelli,A.A., Stewart,J.D., Li,Y., *et al.* (2018) An epigenetic biomarker of aging for lifespan and healthspan. *Aging Albany Ny*, 10, 573–591.

23. Herzog,C. and Widschwendter,M. (2022) WIDclocks: Calculation of epigenetic WID clocks. *Github*, 10.5281/zenodo.5521015.

24. Barrett,J.E., Herzog,C., Kim,Y.N., Bartlett,T.E., Jones,A., Evans,I., Cibula,D., Zikan,M., Bjorge,L., Harbeck,N., *et al.* (2022) Susceptibility to hormone-mediated cancer is reflected by different tick rates of the epithelial and general epigenetic clock. *Genome Biol*, 23, 52.

FIGURES

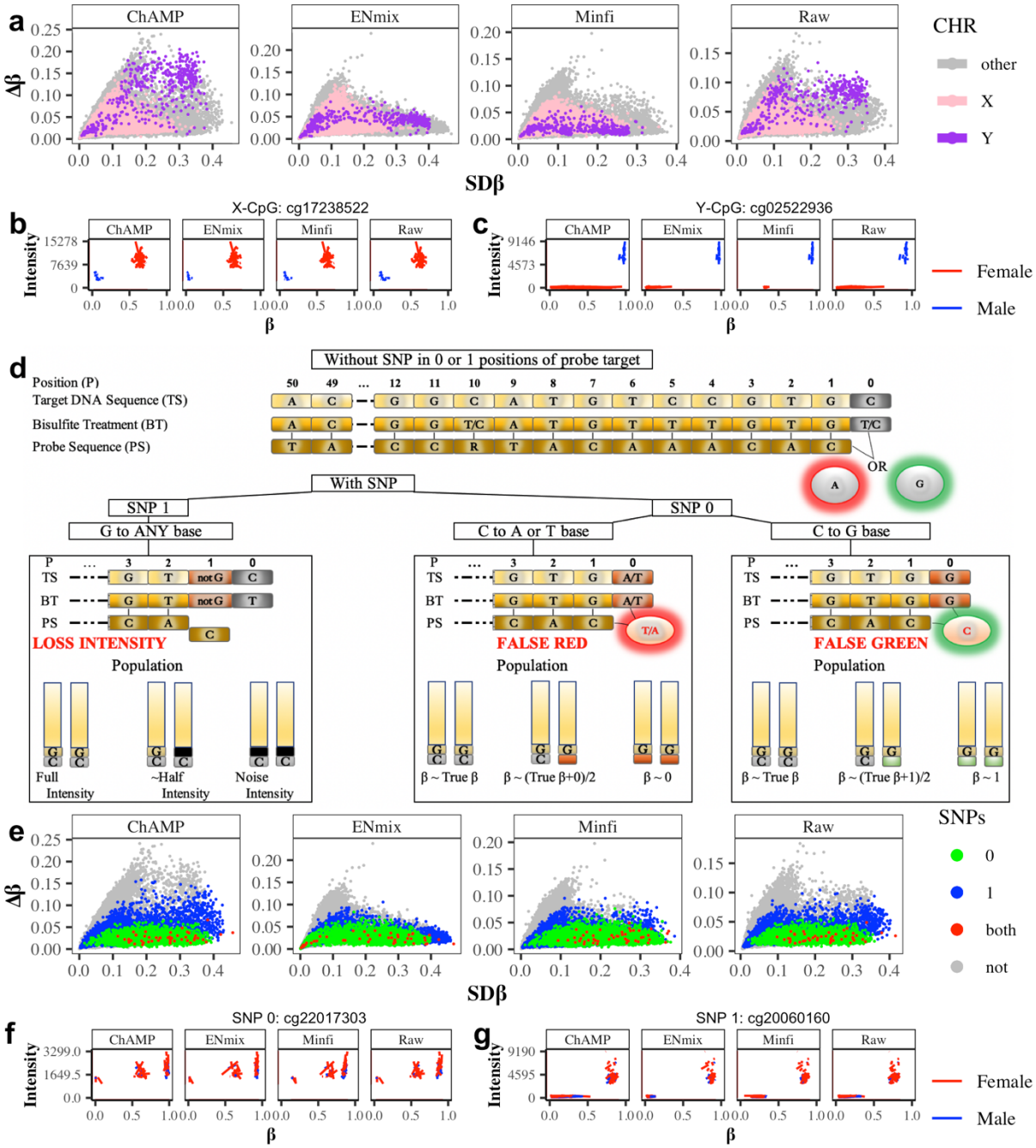

Fig. 1 Variability associated with genetic factors (sex and genetic variants). (a) Sex chromosome-associated probes demonstrate high variability both within population at visit 1 (SD β, x-axis) and over time (Δ β, y-axis); **(b)** example of β value and intensity of an X-chromosome-associated probe in males and females; **(c)** example of β value and intensity of a Y-chromosome-associated probe in males and females. **(d)** Two types of SNPs and their impact on signal: SNPs in position 0 result in false green or red signals, while SNPs in position 1 result in a loss of signal; **(e)** SNPs in position 0 and 1 both demonstrate high variability in the population at visit 1 (SD β), but SNPs in position 1 furthermore demonstrate high variability over time (Δ β). **(f)** Example of probe with 0-position SNP shows tri-modal b-values distribution; **(g)** Example of probe with 1-position SNP shows tri-modal intensities distributions.

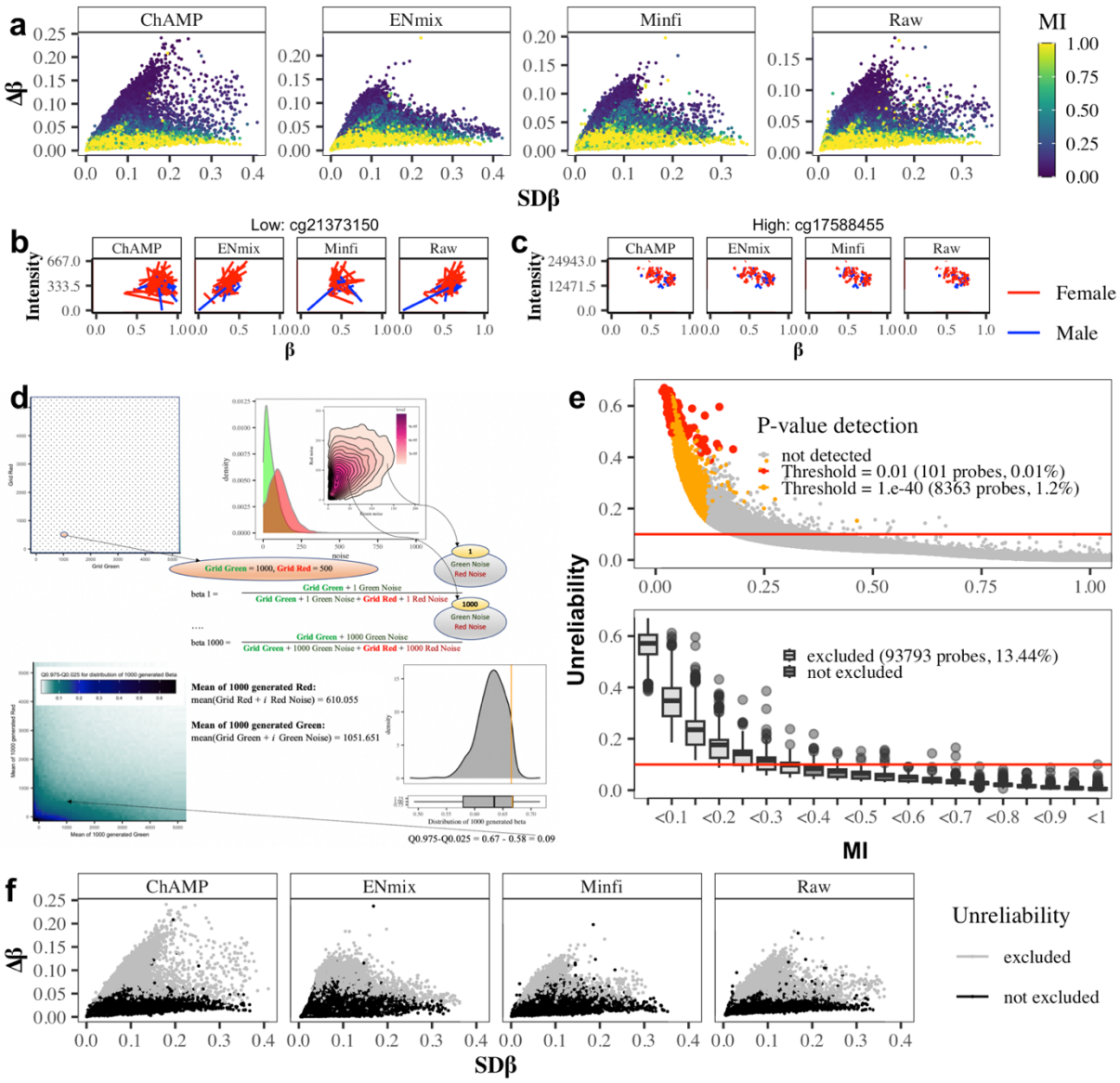

Fig. 2 Impact of probe intensity on β value reliability. (a) Association of probe variability within participants at visit 1 (SD β) and over time (Δ β) with mean intensity (MI). Probes with low MI have high variability over time; Examples of a probe with low MI **(b)** or high MI **(c).** Lines are connected by points corresponding to the individual in two different visits. β values from the same individual are closer for a probe with a high MI than on a probe with a low MI; **(d)** Reliability map of β values. **(e)** dependence of the probe unreliability (Q) on mean intensity. Probes detected using the p-value method are highlighted in orange or red, respectively (standard threshold p = 0.01; corrected p = 1.e-40). As shown here, our unreliability method allows to detect probes more sufficiently with low quality. **(f)** Using an unreliability threshold of 0.1, including only probes in which, on average, β values vary by no more than 0.1 interval. This threshold corresponds to MI probes for which 0.35; **(g)** Probes with high time variability are highly unreliable probes.

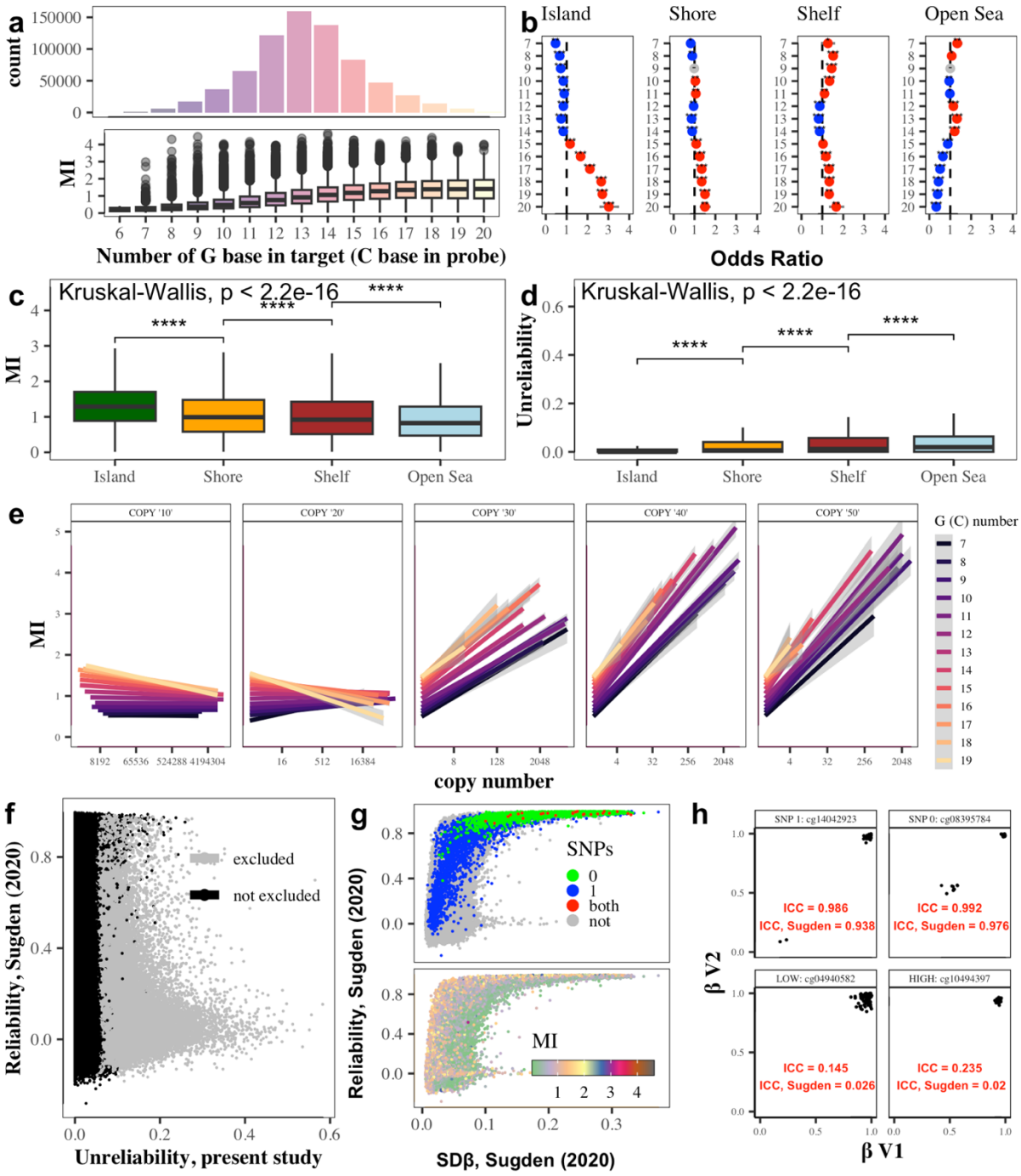

Fig. 3 Probe features impacting DNAme analysis reliability. (a) Dependence of MI on C content in the probe sequence; **(b)** Odds ratio for C enrichment in different DNA regions: Islands, due to their inherently high CG content, have a strong enrichment of probes with a large C content, which leads to the fact that on average MI is higher **(c)** and Unreliability **(d)** is lower for Islands probes than for probes in other regions; **(e)** Dependence of the probe intensities on both C content and copy number (13); **(f)** reliability measure proposed by Sugden *et al.* (14) and the unreliability measure of the present study were not correlated; **(g)** According to Sugden's measure, SNPs are the most reliable probes, and Sugden's reliability measure does not correlate with MI; **(h)** examples of probes showing that the ICC (the measure on which Sugden's reliability is based) is highly related to the variance of β values which

is particularly high in the SNP-associated probes and does not capture the unreliability effects relative to intensity which were investigated in this study.

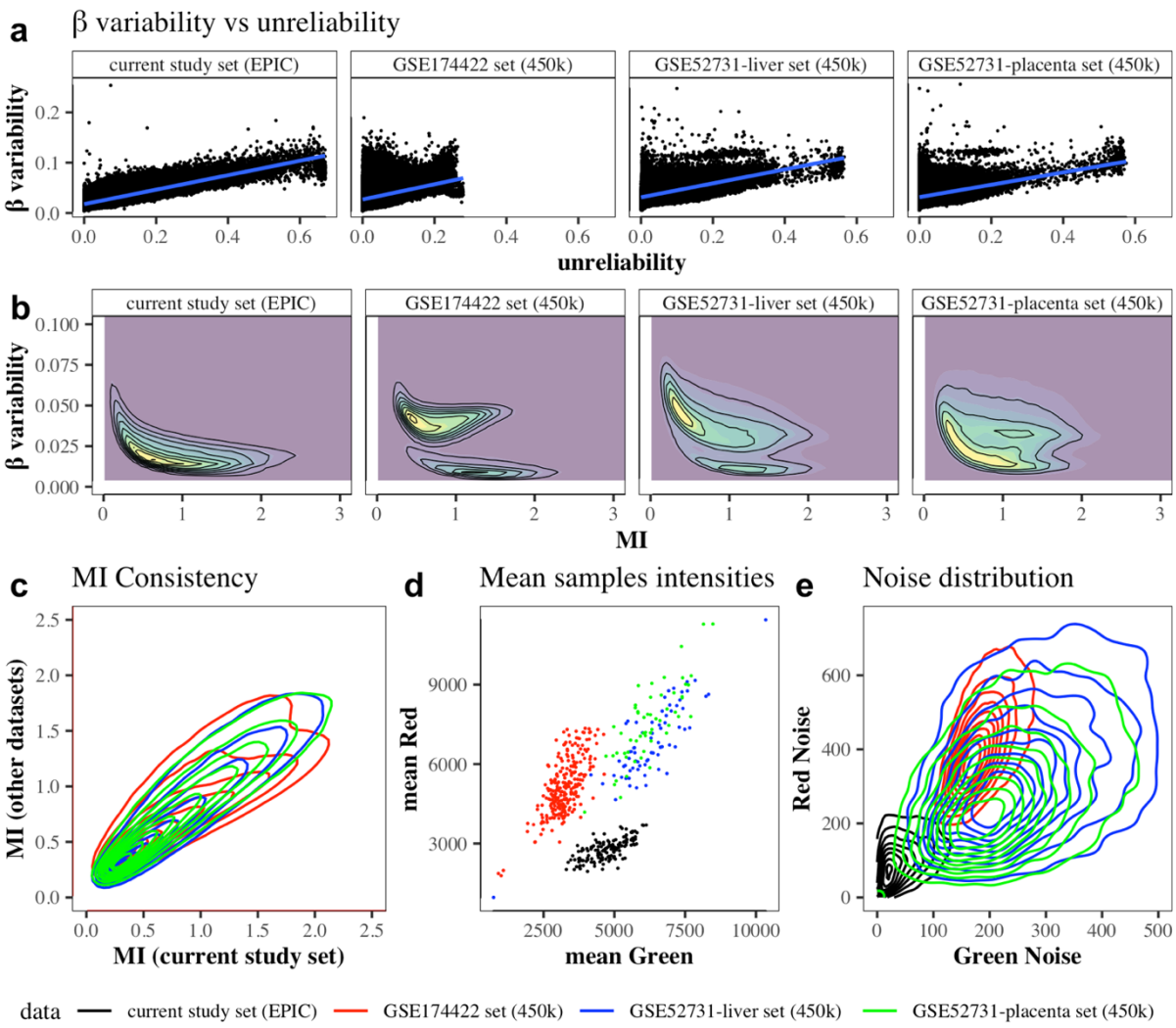

Fig. 4 Consistency of unreliability analysis on replicative samples (a) Unreliability values correlate with the main trend of high beta (raw values) variability – 'β variability': for the current study - average of all beta differences calculated between the same individual over the time; for GSE174422 – average of all beta differences calculated between 2 replicates of the same individual; for GSE52731 – standard deviation of β (for lever and placenta tissue separately). **(b)** Low MI values explain high level of 'β variability' The average high level of 'β variability' for GSE174422 is since the two duplicates of the same individual were always bisulfite-converted on different plates and assayed on different arrays. In contrast to current studies where both samples of the same individual were treated together and resided on the same beadchip. The processing method is not specified for GSE52731, but since many replicates were produced for each tissue (70 for liver and 56 for placenta), we can predict that they were located at least on different beadchips (each of which has 8 positions), which also contributes to an increase in the average 'β variability' level. **(c)** On probes common to EPIC and 450k arrays, MI values are highly correlated with each other in all considered studies, regardless of the average sample intensities **(d)** and distributions on noise probes **(e)**.

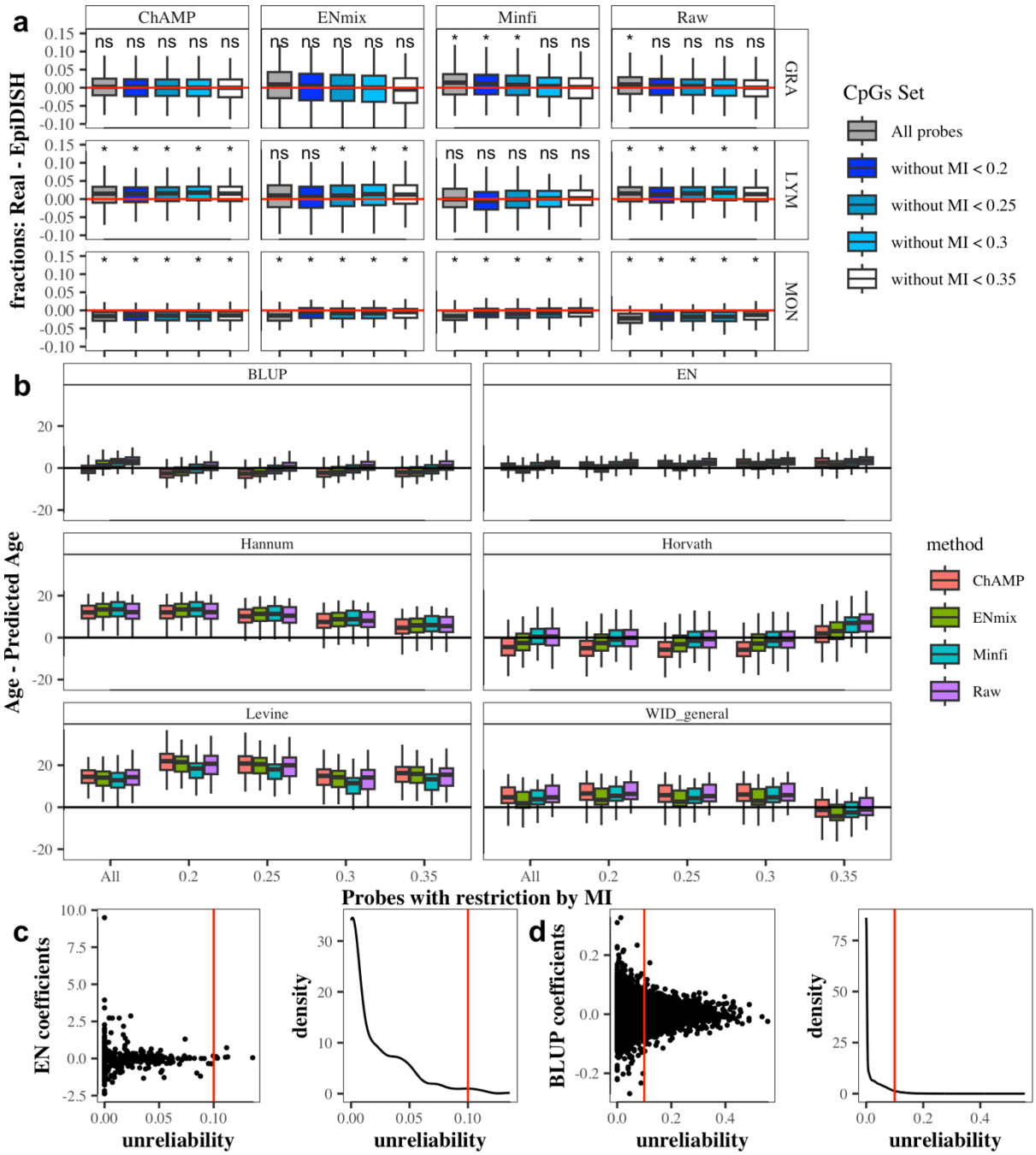

Fig. 5 Predicted cell composition and Age Demonstration of how the sequential removal of poor-quality probes based on various thresholds by MI leads to a more accurate calculation of the *EpiDISH* (15) fractions **(a)** and epigenetic age **(b)** estimated according to BLUP and EN (19), Hannum (20), Horvath (21), Levine (DNAm PhenoAge) (22) and WID-general clocks (24). **(c)** The most reliable probes of BLUP and EN models have the highest coefficients and the proportion of unreliable probes in both sets is not more than 5%.