Clinical Epigenetics

# DNA methylation outlier burden, health, and ageing in Generation Scotland and the Lothian Birth Cohorts of 1921 and 1936

Anne Seeboth[1†], Daniel L. McCartney[1†], Yunzhang Wang[2], Robert F. Hillary[1], Anna J. Stevenson[1], Rosie M. Walker[1,3], Archie Campbell[1], Kathryn L. Evans[1,3], Andrew M. McIntosh[1,3,4], Sara Hägg[2], Ian J. Deary[3,5] and Riccardo E. Marioni[1,3*]

## Abstract

**Background:** DNA methylation outlier burden has been suggested as a potential marker of biological age. An outlier is typically defined as DNA methylation levels at any one CpG site that are three times beyond the inter-quartile range from the 25th or 75th percentiles compared to the rest of the population. DNA methylation outlier burden (the number of such outlier sites per individual) increases exponentially with age. However, these findings have been observed in small samples.

**Results:** Here, we showed an association between age and $\log_{10}$-transformed DNA methylation outlier burden in a large cross-sectional cohort, the Generation Scotland Family Health Study ($N = 7010$, $\beta = 0.0091$, $p < 2 \times 10^{-16}$), and in two longitudinal cohort studies, the Lothian Birth Cohorts of 1921 ($N = 430$, $\beta = 0.033$, $p = 7.9 \times 10^{-4}$) and 1936 ($N = 898$, $\beta = 0.0079$, $p = 0.074$). Significant confounders of both cross-sectional and longitudinal associations between outlier burden and age included white blood cell proportions, body mass index (BMI), smoking, and batch effects. In Generation Scotland, the increase in epigenetic outlier burden with age was not purely an artefact of an increase in DNA methylation level variability with age (epigenetic drift). $\log_{10}$-transformed DNA methylation outlier burden in Generation Scotland was not related to self-reported, or family history of, age-related diseases, and it was not heritable (SNP-based heritability of 4.4%, $p = 0.18$). Finally, DNA methylation outlier burden was not significantly related to survival in either of the Lothian Birth Cohorts individually or in the meta-analysis after correction for multiple testing ($HR_{meta} = 1.12$; 95% $CI_{meta} = [1.02; 1.21]$; $p_{meta} = 0.021$).

**Conclusions:** These findings suggest that, while it does not associate with ageing-related health outcomes, DNA methylation outlier burden does track chronological ageing and may also relate to survival. DNA methylation outlier burden may thus be useful as a marker of biological ageing.

**Keywords:** Ageing, Epigenetics, Stochastic epigenetic mutations, Epigenetic outliers, Survival, Generation Scotland, Lothian Birth Cohorts

* Correspondence: riccardo.marioni@ed.ac.uk
†Anne Seeboth and Daniel L. McCartney contributed equally to this work.
[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK
[3]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK
Full list of author information is available at the end of the article

# Background

Chronological age is a major risk factor for many diseases. Consequently, as life expectancy increases, so too does the prevalence of age-related diseases. However, there are considerable inter-individual differences in health outcomes [1]. In order to improve risk stratification beyond crude models based on chronological age, much recent work has thus attempted to identify and test the utility of markers of 'biological age'.

One focus of research into ageing biomarkers has been on age-related changes in DNA methylation at cytosine-phosphate-guanine (CpG) dinucleotide sites. DNA methylation is a dynamic epigenetic modification that is involved in the regulation of gene expression and is influenced by both genetics [2] and the environment [3]. Analyses have identified a number of loci where methylation changes consistently across individuals as they age, for example the cg16867657 locus in the ELOVL2 gene [4]. In addition to such sites, methylation levels at other loci have been shown to diverge as individuals grow older [5]. This is consistent with the observation that, on average, inter-individual variability in DNA methylation tends to increase as people get older, a phenomenon referred to as epigenetic drift [5, 6].

Recently, researchers have started to investigate the accumulation of DNA methylation outliers (sometimes referred to as 'stochastic epigenetic mutations'). DNA methylation outliers occur when the DNA methylation level at a specific site in an individual's genome differs greatly from that of the majority of the population at this locus. It is known that abnormal methylation patterns can lead to aberrant gene expression, and DNA methylation outliers have been shown to be associated with the development of certain cancers [7, 8] or neurodevelopmental disorders and congenital anomalies [9].

In addition, many studies have investigated the dysregulation of biological processes with age. This includes innate immune system function [10] but also epigenetic processes such as DNA methylation [11]. DNA methylation outlier burden (i.e. the number of outlier sites per individual), as a novel measure of such dysregulation, may thus also have predictive value as an index of biological ageing.

Indeed, Gentilini and colleagues [12] found an exponential association between age and DNA methylation outlier burden in 170 individuals aged between 3 and 106 years ($r_{\text{log (outlier burden)} - \text{age}} = 0.63$). The authors defined outliers as methylation levels of greater than 3 times the inter-quartile range (IQR) from the 25th and 75th percentiles compared to the rest of the population. More recently, two further studies replicated this finding using data from 658 individuals of a wide age range (mean = 54.3 years, SD = 12.7 years) and longitudinal data from 375 older individuals (48 to 98 years),

respectively [13, 14]. Both studies reported individual outlier burden to be highly variable (range 119–18,308 out of 769,042 CpGs [14]; range 58–26,291 out of 370, 234 CpGs [13]) and non-normally distributed. Following log-transformation, these studies found outlier burden to be higher in older individuals ($z = 8.54$; $p = 1.2 \times 10^{-17}$ [14]; $\beta = 8.3 \times 10^{-3}$, $p = 1.2 \times 10^{-13}$ [13]). In addition, white blood cell proportions were found to be significantly associated with outlier burden in both studies.

Here, we provide a comprehensive characterisation of DNA methylation outlier burden in a large cross-sectional sample of 7010 individuals and in two longitudinal samples of 430 and 898 individuals, respectively. We investigate the association of DNA methylation outlier burden with age, explore the potential confounding effect of epigenetic drift on these findings, relate outlier burden to more than a dozen health- and ageing-related traits and to survival, and determine the genetic contribution to individual differences in outlier burden.

# Results

## Descriptive statistics

Descriptive statistics for age, sex, and outlier burden in the three cohorts are shown in Table 1 (for more detailed descriptive statistics including covariate information, see Supplementary Tables 1A, 1B and 1C, Additional file 1).

In Generation Scotland, the mean number of outliers per individual was 1119 (SD = 3361) and highly variable, ranging from 53 to 84,933 (out of 334,352 and 349,027 CpG sites in set 1 and set 2, respectively). The distribution of outlier number per individual was positively skewed, even after $\log_{10}$ transformation. In the combined Generation Scotland dataset, 97% of 361,846 CpG sites had at least one individual with a DNA methylation outlier at present. On average, CpG sites had 22 individuals with DNA methylation outliers at the site (SD = 33, max = 1587, across 7010 individuals). In the Lothian Birth Cohort 1921 (LBC1921) and in the Lothian Birth Cohort 1936 (LBC1936) at wave 1, there were on average 2039 (min = 108; max = 27,676; SD = 3280) and 2197 (min = 119; max = 35,711; SD = 3203) DNA methylation outliers per individual (out of 356,631 CpG sites), respectively. The distribution of outlier number per individual was positively skewed. Following $\log_{10}$ transformation, the distribution looked approximately normal. Sixty-three percent of 356,631 CpG sites were found to have at least one individual with an outlier present in the LBC1921 at wave 1 (83% in the LBC1936). On average, CpG sites had three individuals with outliers at this site (SD = 5, max = 108, across 430 individuals) in the LBC1921 and six individuals with outliers at this site (SD = 10, max = 225, across 898 individuals) in the LBC1936.

**Table 1** Descriptive statistics

**Generation Scotland**

| | | | | |
|---|---|---|---|---|
| *N* | 7010 | – | – | – |
| Females | 4076 (58%) | – | – | – |
| Mean age | 51 (13) | – | – | – |
| Outlier burden | 1119 (3361) | – | – | – |

**LBC1921**

| | Wave 1 | Wave 2 | Wave 3 | Wave 4 |
|---|---|---|---|---|
| *N* | 430 | – | 173 | 82 |
| Females | 260 (60%) | – | 94 (54%) | 44 (54%) |
| Mean age | 79 (0.58) | – | 87 (0.40) | 90 (0.10) |
| Outlier burden | 2039 (3280) | – | 2330 (3901) | 1717 (2029) |

**LBC1936**

| | Wave 1 | Wave 2 | Wave 3 | Wave 4 |
|---|---|---|---|---|
| *N* | 898 | 793 | 607 | 502 |
| Females | 445 (50%) | 376 (47%) | 291 (48%) | 249 (50%) |
| Mean age | 70 (0.83) | 73 (0.70) | 76 (0.67) | 79 (0.62) |
| Outlier burden | 2197 (3203) | 2729 (6020) | 2293 (6861) | 1423 (2845) |

Sample numbers (*N*) alongside their proportion in the entire sample (in %) are reported for categorical variables. Mean (M) and standard deviation (SD) values are reported for continuous variables

### DNA methylation outlier burden and age

In Generation Scotland, there was a small but significant cross-sectional association between age and $\log_{10}$(outlier burden) in all models regardless of adjustments (Supplementary Table 2, Additional file 1) ($\beta = 0.0091$ increase in burden per year, $p < 2 \times 10^{-16}$, fully adjusted model; this corresponded to a 2.1% higher outlier burden per year older in age). The association between $\log_{10}$(outlier burden) and age was non-linear (Fig. 1). There was evidence of heteroscedasticity in the model residuals (Supplementary Figure 1, Additional file 2).

A number of covariates included in the fully adjusted model were significantly associated with DNA methylation outlier burden (Supplementary Table 3A, Additional file 1). B cell, natural killer (NK), CD8$^+$ T cell, and granulocyte proportions were positively associated with outlier burden ($\beta_{\mathrm{raw}} = 0.15$, $p < 2 \times 10^{-16}$; $\beta = 0.075$, $p = 6.6 \times 10^{-10}$; $\beta = 0.084$, $p = 2.2 \times 10^{-16}$; and $\beta = 0.045$, $p = 0.026$; coefficients refer to a unit change in $\log_{10}$(outlier burden) per standard deviation change in the predictor). The proportion of CD4$^+$ T cells was negatively associated with outlier burden ($\beta = -0.078$, $p = 4.5 \times 10^{-9}$). $\log_{10}$(pack years + 1) was positively associated and $\log_{10}$(BMI) was negatively associated with outlier burden ($\beta = 0.025$, $p = 4.3 \times 10^{-3}$; $\beta = -0.013$, $p = 8.0 \times 10^{-3}$). Sex, never vs. ever smoking status, and cancer were not significantly associated with outlier burden.

Longitudinally, and prior to adjustments, DNA methylation outlier burden increased with age in LBC1921 and decreased with age in the LBC1936 (Fig. 1; Fig. 2).

However, after adjustment for covariates, the association between age and $\log_{10}$(outlier burden) was positive in both cohorts (Fig. 1). This association was significant after adjustment for cell counts in the LBC1921 ($\beta = 0.033$, $p = 7.9 \times 10^{-4}$, fully adjusted model; this corresponded to a 7.9% increase in outlier burden per year increase in age) but not in the LBC1936 ($\beta = 7.9 \times 10^{-3}$, $p = 0.074$, fully adjusted model; 1.8% increase per year) (Supplementary Table 2, Additional file 1). Ageing trajectories for those individuals with complete observations, i.e. observations in all waves (63 in the LBC1921 and 337 in the LBC1936), were similar to those observed in the entire sample (Supplementary Figures 2, 3, and 4, Additional file 2).

As in Generation Scotland, estimated cell proportions of B cells, NK cells, and CD8$^+$ T cells in the LBC1921 ($\beta = 0.20$, $p < 2 \times 10^{-16}$; $\beta = 0.13$, $p = 6.7 \times 10^{-6}$; $\beta = 0.082$, $p = 2.8 \times 10^{-3}$) and B cells, NK cells, and CD4$^+$T cells in the LBC1936 were significantly related to outlier burden ($\beta = 0.13$, $p < 2 \times 10^{-16}$; $\beta = 0.056$, $p = 1.0 \times 10^{-4}$; $\beta = -0.059$, $p = 1.1 \times 10^{-3}$). No other covariates were significantly associated with DNA methylation outlier burden in either the LBC1921 or the LBC1936 (Supplementary Table 3B and 3C, Additional file 1).

The relationship between outlier burden and three measures of epigenetic age acceleration was examined in Generation Scotland and the Lothian Birth Cohorts: Hannum age acceleration, GrimAge acceleration, and intrinsic epigenetic age acceleration (IEAA) [15–17]. A significant positive association was observed between
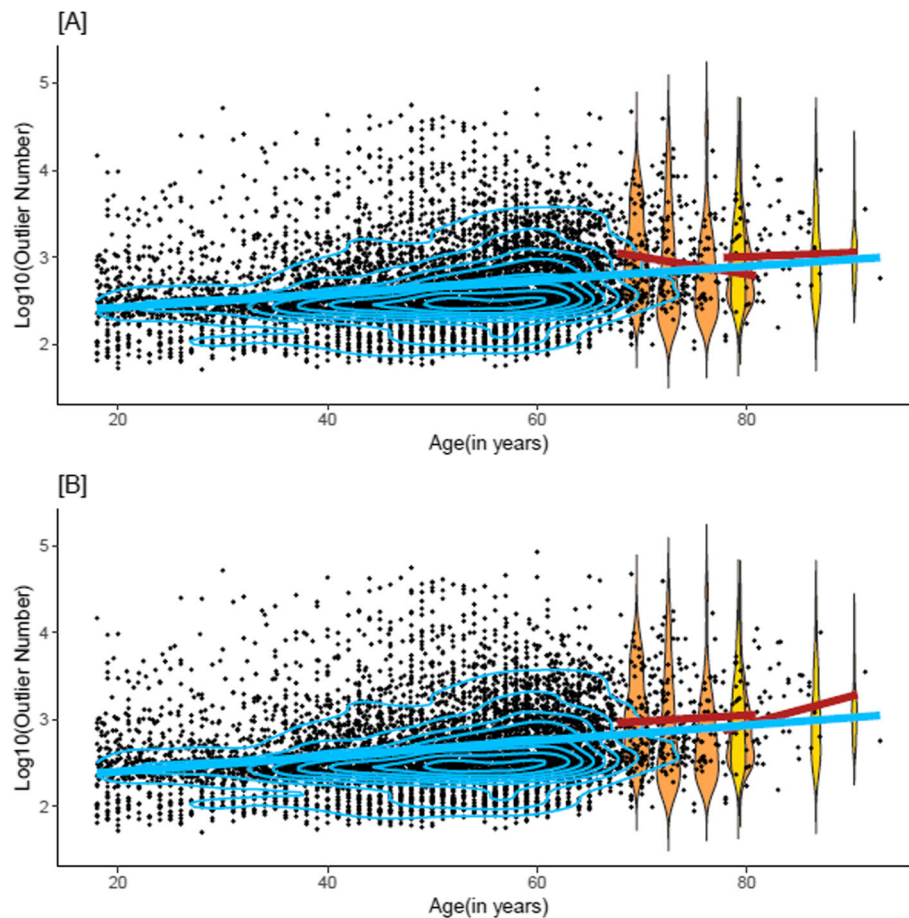
**Fig. 1** DNA methylation outlier burden in Generation Scotland, LBC1921, and LBC1936. Distribution of $log_{10}$(outlier burden) in Generation Scotland (black, blue contour shapes indicating data density) and in the four LBC1936 waves (orange) and three LBC1921 waves (yellow). **a** Linear regression lines in Generation Scotland (blue) and in the LBC1921 and LBC1936 (red) to model the association of $log_{10}$(Outlier Burden) ~ age. **b** Fitted values for the regression of $log_{10}$(outlier burden) ~ age with random factor batch and cell proportions fit to the mean

outlier burden and both Hannum and GrimAge acceleration in LBC1921 after adjusting for sex and estimated cell proportions ($\beta_{GrimAge} = 3.52$, $p = 2.6 \times 10^{-8}$; $\beta_{Hannum} = 5.51$, $p = 2.8 \times 10^{-10}$). Similarly in Generation Scotland, a significant relationship was observed between Hannum and GrimAge acceleration, as well as IEAA ($\beta_{GrimAge} = 1.59$, $p = 2.9 \times 10^{-42}$; $\beta_{Hannum} = 1.32$, $p = 3.1 \times 10^{-52}$; $\beta_{IEAA} = 0.7$, $p = 1.5 \times 10^{-52}$; Supplementary Table 4A, Additional file 1). DNA methylation outlier counts were compared between published constituent clock CpGs (Horvath and Hannum clocks) and non-clock CpGs in Generation Scotland and both LBC cohorts. Hannum clock probes had a lower average outlier count compared to non-clock probes in all cohorts. In Generation Scotland, the average outlier count for Horvath clock probes was significantly greater than that of non-clock probes ($mean_{HorvathClock} = 27.2$, $mean_{NonClock} = 21.7$; $p = 0.002$). There were no significant differences in outlier counts between Horvath clock CpGs and non-clock CpGs in the remaining cohorts (Supplementary Table 4B, Additional file 1).

## DNA methylation outlier burden and epigenetic drift

In Generation Scotland, DNA methylation outlier burden based on an alternative definition of outliers within age groups correlated almost perfectly with outlier burden calculated using the original definition ($r = 0.99$, $p < 2 \times 10^{-16}$). Accounting for epigenetic drift by applying the alternative definition for outliers slightly attenuated the association between DNA methylation outlier burden and age in Generation Scotland (Supplementary Table 5A, Additional file 1) ($\beta = 6.2 \times 10^{-3}$, $p < 10^{-16}$, fully adjusted model; corresponding to a 1.4% higher DNA methylation outlier burden per year older in age). However, the association was still highly significant and the pattern of covariate associations did not change (effect sizes for each covariate presented in Supplementary Table 5B, Additional file 1).

## DNA methylation outlier burden, health, and survival

We did not find evidence for a cross-sectional association between DNA methylation outlier burden and self-
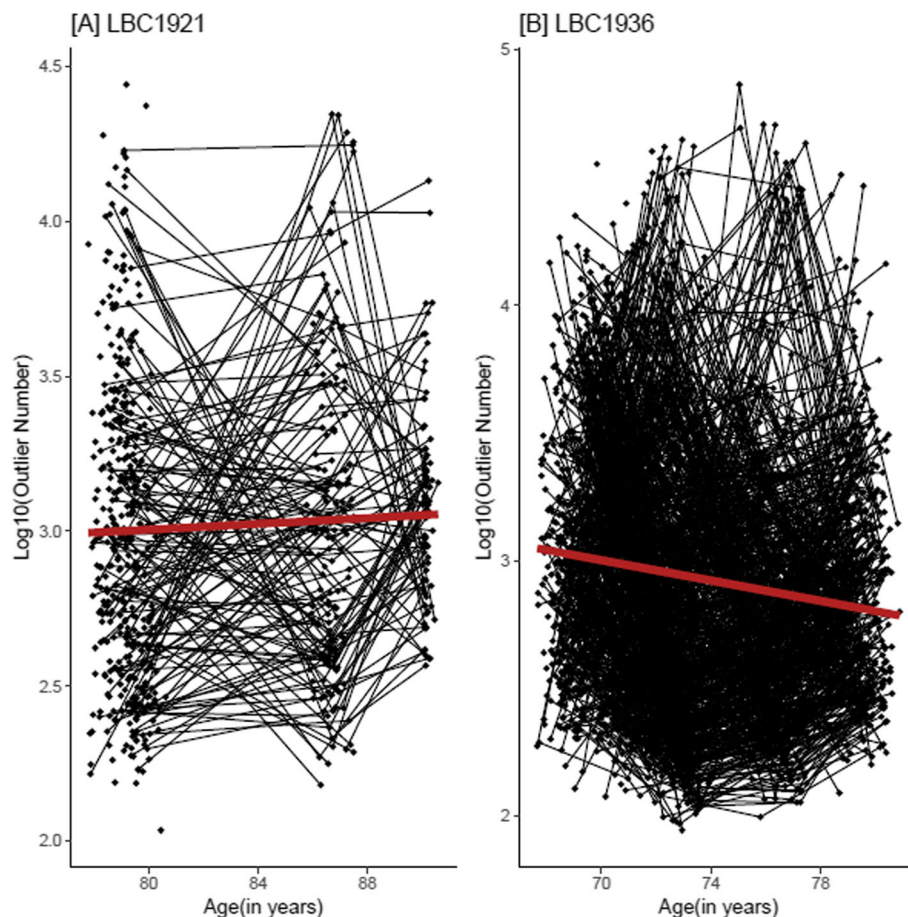
**Fig. 2** Trajectories of DNA methylation outlier burden in the LBC1921 and LBC1936. Longitudinal change in $\log_{10}$(outlier burden) in individuals in the LBC1921 and the LBC1936. The linear regression lines of $\log_{10}$(Outlier Burden) ~ age are shown in red

reported disease (Supplementary Tables 6A and 6B, Additional file 1) or family history of disease (Supplementary Tables 7A and 7B, Additional file 1) in Generation Scotland following Bonferroni correction.

DNA methylation outlier burden was not significantly associated with survival in the LBC1921 or the LBC1936 individually (Fig. 3; Supplementary Table 8, Additional file 1). However, when meta-analysed across the two cohorts, the association between outlier burden and survival was significant at a nominal $p < 0.05$ threshold (hazard ratio$_{meta}$ = 1.12; 95% CI$_{meta}$ = [1.02; 1.21]; $p_{meta}$ = 0.021).

**Heritability of DNA methylation outlier burden**
There was no significant genetic contribution by all common genome-wide SNPs (minor allele frequency > 5%) to phenotypic variability in $\log_{10}$-transformed DNA methylation outlier burden ($h^2$ = 0.044, SE = 0.048, $p$ = 0.18).
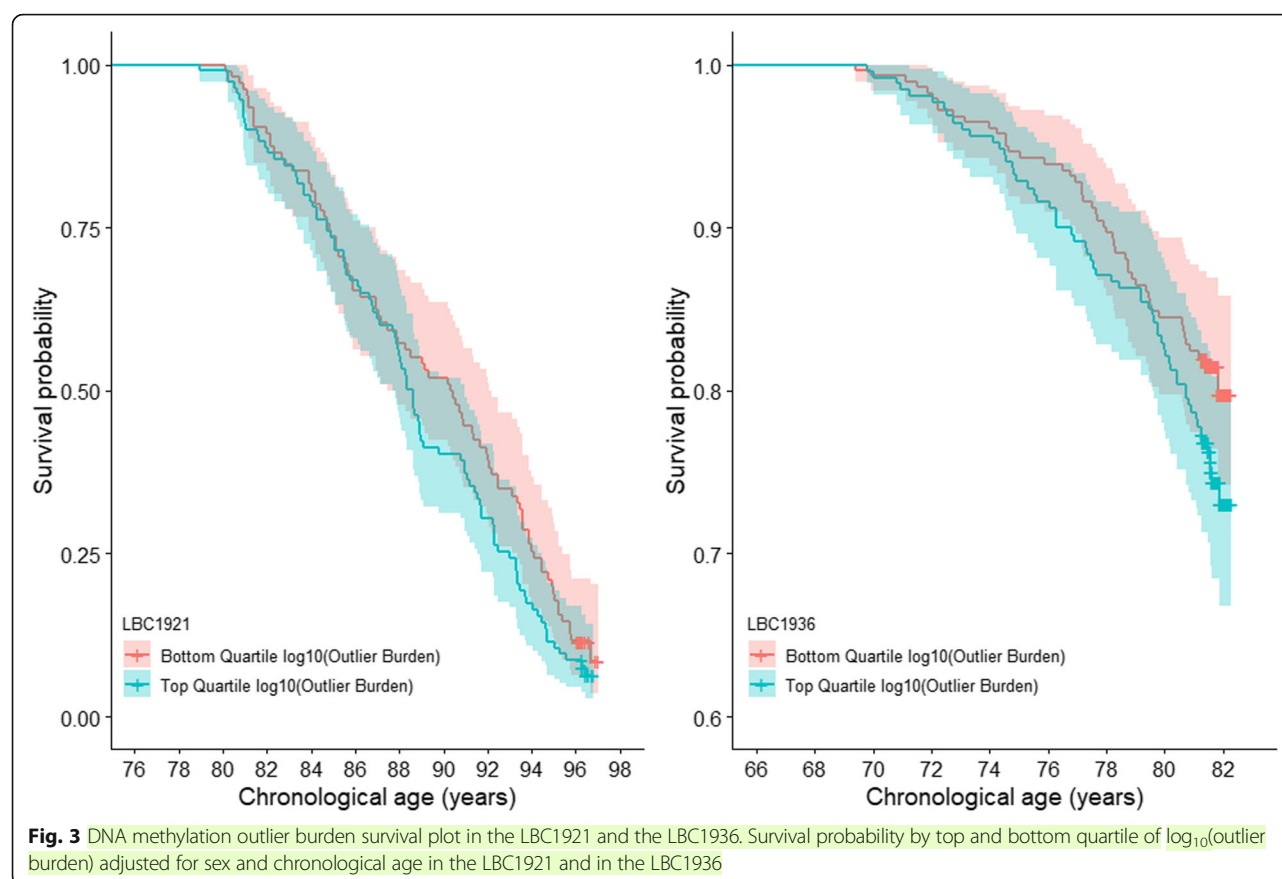
**Directionality of DNA methylation outliers**
Directionality of DNA methylation outliers was assessed in Generation Scotland and baseline Lothian Birth

Cohort samples. Hypermethylated outliers were defined as DNA methylation outliers that were greater than three times the IQR above the upper quartile of a given CpG, whereas hypomethylated outliers were defined as those that were less than three times the IQR below the lower quartile. There were 69,298 hypermethylated outlier probes common to all cohorts, and 61,578 hypomethylated outlier probes. There was a significant difference in the distribution of hyper- and hypomethylated outliers across CpG islands, shores, shelves, and open seas ($\chi^2$ = 46,944; $p < 2.2 \times 10^{-16}$). Hypermethylated outlier CpGs were predominantly located at CpG islands, whereas the majority of hypomethylated outlier sites mapped to open seas (Supplementary Table 9, Additional file 1).

**Identification of common DNA methylation outliers**
We investigated the presence of common DNA outliers between cohorts, comparing common outliers (i.e. those present in at least 5% of Generation Scotland individuals; $N \geq 351$ participants) to rare outliers (present in < 5% of

**Fig. 3** DNA methylation outlier burden survival plot in the LBC1921 and the LBC1936. Survival probability by top and bottom quartile of $\log_{10}$(outlier burden) adjusted for sex and chronological age in the LBC1921 and in the LBC1936

Generation Scotland individuals). Of 208 common outlier CpGs in Generation Scotland, data for 187 were available for the LBC cohorts. The proportion of individuals with 'common' outliers in the LBC studies was significantly greater than the proportion of individuals with 'rare' outliers (as defined in Generation Scotland; $p \leq 3.64 \times 10^{-22}$; Supplementary Table 10A, Additional file 1). The distribution of common outliers across CpG islands, shelves, shores, and open seas was significantly different to that of rare outliers ($\chi^2 = 82.58$; $p = 2.42 \times 10^{-16}$; Supplementary Table 10B, Additional file 1).

### Identification of outlier individuals
Eight outlier individuals were identified in Generation Scotland set 2, all of whom had an excess of DNA methylation outliers. Outlier individuals were defined as those having an outlier burden ± 3 interquartile ranges from the upper/lower quartile of outlier counts in the study population. No outlier individuals were identified in Generation Scotland set 1 or at any wave of the Lothian Birth Cohorts. Two outlier CpG sites were common to the eight outlier individuals (cg05941108 and cg02279719). While outlier individuals were older than non-outlier individuals on average, this difference was not significant, suggesting the excess of outliers in these

individuals is not age-related (mean$_{\text{outlier}}$ = 57.0 years, SD$_{\text{outlier}}$ = 15.8; mean$_{\text{non-outlier}}$ = 51.3 years, SD$_{\text{non-outlier}}$ = 13.2; $p = 0.34$).

## Discussion
Here, we characterised DNA methylation outlier burden, which has been suggested as a promising new marker of biological age, in three large cohorts: the cross-sectional Generation Scotland cohort of individuals aged 18 to 92 years and the Lothian Birth Cohorts of 1921 and 1936, which are longitudinal studies of older adults aged, on average, 79 and 69 years at baseline, respectively. We analysed methylation levels at those sites present on the Illumina Infinium HumanMethylation450k array, to permit comparison with findings by Wang and colleagues [13] in the Swedish Adoption/Twin Study of Aging (SATSA), a longitudinal cohort of older individuals in their 70s.

### DNA methylation outlier burden
There was high variability in DNA methylation outlier number between individuals. DNA methylation outlier burden was highly positively skewed and consequently a $\log_{10}$-transformed version of outlier burden used in all analyses. This is consistent with previous findings [13, 14].

When looking at outlier burden per CpG site rather than per individual, Wang and colleagues [13] reported that at 64% of CpG sites, at least one out of 385 individuals in SATSA had an outlier. Here, we found a similar number of CpG sites (63%) that were an outlier in at least one out of the 430 individuals in the LBC1921. However, our findings demonstrate that with increasing sample size, almost all sites will be an outlier at least once (83% of CpG sites in 898 individuals in the LBC1936 and > 96% in 7010 individuals in Generation Scotland). This may suggest that the accumulation of DNA methylation outliers with age is a stochastic process, affecting most CpG sites randomly. However, we observed differences in the distribution of outliers across open seas and CpG islands, consistent with previous reports [13].

## DNA methylation outlier burden and age

DNA methylation outlier burden was found to be positively associated with age in Generation Scotland and in the Lothian Birth Cohorts. The association between age and $\log_{10}$(outlier burden) was significant in Generation Scotland independent of the level of adjustment and in the LBC1921 after adjustments for cell counts. A similar association was seen in the LBC1936 although it was not significant. Effect sizes were comparable to those reported by Wang and colleagues ($\beta = 8.3 \times 10^{-3}$, $p = 1.2 \times 10^{-13}$ [13]).

DNA methylation outlier burden measures were strongly associated with cell proportions and other technical factors, again, confirming previous findings [13, 14]. Moreover, in this study, batch effects were confounded by wave of data collection in the LBC1921 and LBC1936. Outliers were thus defined as values ± 3 interquartile ranges from the upper/lower quartile in each wave rather than at baseline, in order to reduce this confounding effect and obtain more stable definitions.

There was no difference in DNA methylation outlier burden between men and women in any of the three cohorts. This is consistent with findings reported by Curtis and colleagues [14] but not with those by Wang and colleagues [13] who found women to have a slightly higher outlier burden than men. To minimise confounding by sex, Wang and colleagues consequently defined outliers according to methylation level variability in women and men separately.

## DNA methylation outlier burden and epigenetic drift

We hypothesised that rather than merely being a consequence of an increase in outlier burden with age, the age-related increase in DNA methylation variability (epigenetic drift) may in fact contribute to the identification of larger numbers of outliers. An older individual's methylation level might be classified as an outlier when compared to methylation levels in the entire sample, but it may not when compared to that of other older individuals (amongst whom variability is higher) (Supplementary Figure 5, Additional file 2).

Here, we found that controlling for epigenetic drift, by defining outliers based on variability in age deciles, only slightly attenuated the association between DNA methylation outlier burden and age in Generation Scotland.

Of course, defining outliers in even narrower age ranges could have further attenuated the association between outlier burden and age. However, there also was an association between outlier burden and age in the Lothian Birth Cohorts, despite outliers being defined within waves of data collection with very narrow age ranges. Taken together, these findings suggest that the age-related increase in extreme methylation levels (that is, epigenetic outliers) and the age-related increase in methylation level variability (epigenetic drift) appear to be, at least in part, two separate phenomena.

This is also consistent with findings by Wang and colleagues [13] who demonstrated that CpG sites where DNA methylation outliers occur frequently do not tend to be the same CpG sites where methylation level variability increases with age. Of 1185 CpG sites with outliers in more than 50 individuals, only two had previously been identified as being variably methylated.

## DNA methylation outlier burden, health, and survival

In contrast to previous studies, we did not find DNA methylation outlier burden to be associated with self-reported or family history of cancer [7, 13]. Furthermore, it was not associated with a comprehensive list of other self-reported disease outcomes. The prevalence of self-reported disease outcomes in Generation Scotland was variable, ranging from five cases of Alzheimer's disease to 1065 cases of hypertension. It is possible that for several disease outcomes, there was insufficient statistical power to detect associations with DNA methylation outlier burden due to limited observations. While other work has linked extreme DNA methylation patterns to congenital abnormalities [9] and low birth weight [18], the findings of the present study do not support strong links between DNA methylation outlier burden and age-related health outcomes.

In addition, we found mixed evidence regarding the association of DNA methylation outlier burden and lifestyle variables. DNA methylation outlier burden has previously been linked to smoking [19] but not to BMI [12, 19]. In Generation Scotland, higher DNA methylation outlier burden was found to be significantly positively associated with smoking pack-years and significantly negatively associated with BMI. There were no significant associations between these variables in the LBC1921 and LBC1936.

While there was a positive correlation between DNA methylation outlier burden and GrimAge acceleration, an established predictor of mortality, there was no significant association between higher outlier burden and a higher risk of mortality in the LBC cohorts after correction for multiple testing.

### Heritability of DNA methylation outlier burden
Consistent with findings in twins by Wang and colleagues [13], there was no evidence that outlier burden is heritable.

### Strengths and limitations
Here, we provide a comprehensive characterisation of DNA methylation outlier burden in three large independent cohorts. The sample size of the Generation Scotland cohort is an order of magnitude greater than previous studies [12–14]. In addition, this is the first study to systematically relate DNA methylation outlier burden to a range of diseases and to survival.

There are limitations to this study. First, the distribution of $\log_{10}$(outlier burden) in Generation Scotland was still slightly positively skewed after $\log_{10}$ transformation and its association with age was thus non-linear. Disease status, including cancer information, in the Generation Scotland study was measured by self-report which can be unreliable. In addition, this data was cross-sectional, and for some diseases, prevalence rates were low (e.g. Alzheimer's disease and Parkinson's disease; Supplementary Tables 6 and 7, Additional file 1) which may have limited the present study's ability to detect significant associations. Finally, we did not run analyses separately on outliers based on their direction, that is, 'high' outliers with abnormally high methylation levels and 'low' outliers with lower than average methylation levels [13]. We also did not investigate the functional pathways in which outliers may cluster (some work on this reported in [13, 14, 19]).

### Conclusions
The findings of our study are consistent with previous work, demonstrating an increase in the number of DNA methylation outliers in individuals as they age. This suggests that DNA methylation outliers are unlikely to be just technical artefacts. Furthermore, our findings show that the accumulation of DNA methylation outliers with age is unlikely to be an artefact of epigenetic drift and that it does indeed appear to be driven by stochastic processes. We also demonstrate that the measurement of DNA methylation outliers is heavily associated with cell counts and technical factors. We do not find DNA methylation outlier burden to be robustly associated with age-related diseases, and future studies in larger samples are needed to confirm whether it can predict

survival. Based on the findings reported here, DNA methylation outlier burden may be useful as a marker of biological age and predict survival, but more work will be needed to establish whether DNA methylation outliers can offer insights into the associations between ageing, health, and lifestyle.

## Methods
Here, we used data from the cross-sectional Generation Scotland cohort, as well as from the longitudinal Lothian Birth Cohorts of 1921 and 1936.

### The Generation Scotland cohort
Generation Scotland is a family-based cohort consisting of individuals aged 18 to 98 years, living across Scotland. Participants were initially recruited from individuals registered at GP surgeries and then asked to invite first degree relatives to join the study, resulting in a final sample size of 23,960 individuals. For these participants, genetic as well as clinical, lifestyle, and sociodemographic information are available. Further details on the cohort can be found elsewhere [20, 21].

The Generation Scotland Scottish Family Health Study has been ethically approved and granted research tissue bank status by the NHS East of Scotland Research Ethics Service (REC reference numbers: 05/S1401/89 and 15/ES/0040, respectively). All study participants provided informed written consent.

### The Lothian Birth Cohorts of 1921 and 1936
The Lothian Birth Cohorts are two longitudinal cohort studies of ageing in individuals born in 1921 (LBC1921, $n$ = 550) or in 1936 (LBC1936, $n$ = 1091). At age 11, as part of the Scottish Mental Health Surveys of 1932 and 1947, respectively, most of these individuals completed the Moray House Test of general intelligence. Decades later, those living in Edinburgh and the surrounding Lothian region were re-contacted and invited to participate in the Lothian Birth Cohort studies. Recruitment and baseline testing (wave 1) took place between 1999 and 2001 (mean age ~ 79), and between 2004 and 2007 (mean age ~ 70) for the LBC1921 and LBC1936, respectively. Since then, detailed physical, cognitive, psychosocial, and lifestyle information was collected roughly every 3 years in 4 subsequent waves of testing. In addition, genetic and longitudinal epigenetic profiling is available in both the LBC1921 and the LBC1936. More detail on recruitment and testing can be found elsewhere [22, 23].

Ethical approval for the first wave of Lothian Birth Cohort studies was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/

1998/4/183; LREC/2003/2/29). All participants provided written informed consent.

## DNA methylation data

Genome-wide DNA methylation in Generation Scotland was measured from blood samples collected between 2006 and 2011 (at the time of baseline appointment) using the Illumina Infinium HumanMethylationEPIC BeadChip at > 850,000 CpG sites. The methylation profiling was carried out in two sets, here referred to as set 1 and set 2. Set 1 consisted of 5200 individuals (2586 of whom were genetically unrelated to each other at a relatedness threshold of < 0.025); set 2 consisted of a further 4683 unrelated individuals—both to others in set 2 and all of those in set 1. Methylation profiling in set 1 and set 2 was carried out in 31 batches each. Full details of DNA methylation quality control steps can be found in Additional file 2 (Supplementary Note 1, Additional file 2).

DNA methylation in the Lothian Birth Cohorts was measured repeatedly using the Illumina HumanMethylation450K array at > 450,000 CpG sites. Methylation data are available in three waves in LBC1921 (wave 1, 3, and 4) and in four waves in LBC1936 (wave 1, 2, 3, and 4). Samples in LBC1936 were processed in three separate batches (Supplementary Table 11, Additional file 1) on 13 dates, using 41 plates and 309 microarrays in total. All samples in LBC1921 were processed in one batch on 7 dates, using 11 plates and 76 microarrays. Sample collection and quality control steps have been described in greater detail elsewhere [24, 25]. For a brief description, see Supplementary Note 1, Additional file 2.

Following quality control, methylation $\beta$-values were calculated using the m2beta function in *lumi* [26]. Probes targeting polymorphic sites in the European population, ch-probes, and probes predicted to hybridise to multiple genomic regions (identified for the EPIC array by McCartney et al.) were removed [27]. In addition, probes on the X and Y chromosomes and probes with known meQTLs were excluded. Note that the meQTL probes (both *cis* and *trans*) were identified using the Illumina Infinium HumanMethylation450 array [28] and that no comprehensive list is currently available for the EPIC array.

Following post-QC filtering, the datasets comprised the following sites and samples: 356,631 sites in 436 samples in the LBC1921 and in 906 samples in the LBC1936; 678,519 sites in 2586 unrelated individuals in Generation Scotland set 1; and 724,207 sites in 4450 individuals in Generation Scotland set 2. Here, we limit our analyses to the 334,352 and 349,027 sites measured in Generation Scotland set 1 and set 2, respectively, which are also present on the HumanMethylation450 array. This will make results from Generation Scotland more comparable to those obtained in the Lothian Birth Cohorts.

## Covariates

In addition to technical factors (batch and set in Generation Scotland; date, plate, and array in the Lothian Birth Cohorts), a number of phenotypes were included as covariates in the analyses. These included white blood cell proportions of granulocytes, natural killer (NK) cells, B cells, CD4+T cells, and CD8+T cells which had been estimated from DNA methylation using the Houseman method [29] as implemented in the *estimateCellCounts* function in *minfi* [30]. Furthermore, a binary ever-never smoking variable, a $\log_{10}$-transformed continuous smoking pack years variable, a binary self-report cancer diagnosis (yes/no) variable, and a $\log_{10}$-transformed continuous measure of body mass index (BMI) were included as covariates. Variables were $\log_{10}$-transformed in order to minimise positive skew. Full details can be found in Supplementary Notes 1 and 2, Additional file 2.

## Health and survival data

In Generation Scotland, DNA methylation outlier burden was related to more than a dozen binary measures of self-reported disease and self-reported family history of disease. In the Lothian Birth Cohorts, DNA methylation outlier burden was related to survival. Mortality data was provided by the General Register Office for Scotland (National Records Scotland) through data linkage to the National Health Service Central Register. The data used in the analyses reported here are correct as of January 2018 (LBC1921) and April 2018 (LBC1936). At time of last censor (approximately 18 years from baseline for the LBC1921 and approximately 14 years from baseline for the LBC1936), 37 and 680 participants of the LBC1921 and LBC1936 were alive, respectively.

## DNA methylation outlier burden

In Generation Scotland, DNA methylation outliers were defined as per Gentilini and colleagues [12], that is, as sites with methylation levels greater than three times the interquartile range (IQR) from the upper or lower quartile. Outliers were calculated within each set separately. Data from Generation Scotland set 1 and set 2 were then combined for all following analyses. In the longitudinal Lothian Birth Cohort studies, DNA methylation outliers were calculated in each wave of data collection based on the IQR calculated in that wave (rather than the IQR in wave 1, as done by Wang and colleagues [13]). The reason for this is that there are strong technical effects on DNA methylation measures and that in the Lothian Birth Cohorts, technical factors are confounded by wave of sample collection. For instance, processing set is confounded by wave of sample collection in the LBC1936 (Supplementary Table 11, Additional file 1). In

addition, factors such as plate and date are confounded by wave even within processing sets. By defining outliers within each wave, we hoped to minimise these batch effects, resulting in more stable outlier definitions. Finally, outlier burden per individual was calculated as the total number of outlier sites for each individual. In addition, outliers per CpG site were calculated as the total number of individuals with an outlier per site. Following this, samples with extreme values for any of the estimated cell proportions (> 5 standard deviations from the mean) were removed (5 and 13 samples in Generation Scotland set 1 and set 2, and 7 and 32 samples in LBC1921 and LBC1936, respectively). In addition, a further eight individuals from Generation Scotland set 1 were excluded from the analyses because they had responded with 'yes' to every self-reported disease phenotype included on the questionnaire.

A combined dataset of 7010 individuals remained for analysis in Generation Scotland (2573 in set 1 and 4437 in set 2). The final Lothian Birth Cohort samples consisted of 430 individuals (685 measurements) in the LBC1921 and 898 individuals (2801 measurements) in the LBC1936.

## Analyses

All analyses were performed in R [31]. The *lme4* [32], *lmerTest* [33], and *survival* [34] packages were used to construct linear mixed models and Cox regression models. The *ggplot2* [35] and *survminer* [36] packages were used to create graphs. For a flowchart displaying the analysis steps, see Fig. 4.

### DNA methylation outlier burden and age

Cross-sectional associations between a $log_{10}$-transformed version of DNA methylation outlier burden and age in Generation Scotland were obtained by fitting three linear models. The basic model included sex as a covariate. The second model included additional adjustments for estimated white blood cell proportions (granulocytes, natural killer (NK) cells, B cells, $CD4^+T$ cells, and $CD8^+T$ cells) and technical variables (batch and set). The third (fully adjusted) model also controlled for smoking (ever/never), $log_{10}$(pack years + 1), cancer diagnosis (yes/no), and



**Fig. 4** Flowchart of analyses in the Generation Scotland and Lothian Birth Cohorts. Methylation profiling in Generation Scotland was carried out in two separate sets. DNA methylation outliers were calculated within each set. Analyses were then carried out on a combined dataset. In the Lothian Birth Cohorts of 1921 and 1936, DNA methylation outliers were calculated within each wave of data collection. Analyses were carried out in the LBC1921 and the LBC1936 separately. Results from the survival analyses in both cohorts were meta-analysed

$\log_{10}$(BMI). The first model was a linear regression, whereas models 2 and 3 were linear mixed effects models with a random intercept term fitted for batch.

Longitudinal analyses of outlier burden in the Lothian Birth Cohorts were run by fitting the following three linear mixed models with random intercepts for ID and batch variables. The most basic model included sex as a fixed effect. The adjusted model included sex, estimated white blood cell proportions (granulocytes, natural killer (NK) cells, B cells, CD4$^+$T cells, and CD8$^+$T cells), and technical factors (date, plate, and array) with final adjustments made for smoking (ever/never), $\log_{10}$(pack years + 1), cancer diagnosis (yes/no), and $\log_{10}$(BMI).

Measures of epigenetic age acceleration for each cohort were obtained from the DNA methylation age calculator (https://dnamage.genetics.ucla.edu/). The relationship between outlier burden and intrinsic epigenetic age acceleration, Hannum age acceleration, and GrimAge acceleration was assessed using linear models [15–17].

### DNA methylation outlier burden and epigenetic drift
Variability in methylation levels is known to increase over time (epigenetic drift) [5, 6]. In order to investigate whether this affects the association between outlier burden and age, we re-calculated outliers in Generation Scotland in an alternative fashion. We defined outliers as extreme methylation levels with respect to variability between individuals of the same age group rather than variability overall (Supplementary Figure 5, Additional file 2). We defined outliers as methylation levels greater than three times the IQR above or below the upper and lower quartile based on the variability within age deciles.

### DNA methylation outlier burden, health, and survival
Logistic regression models with basic adjustments for age and sex, followed by full adjustment for cell count proportions, smoking status and log(pack years), cancer, and log(BMI) were fit to investigate the association between DNA methylation outlier burden and binary measures of self-reported disease in the large Generation Scotland dataset. As prevalence rates for some diseases were low, analyses were repeated using self-reports of family history (in mother or father) of disease. We corrected for multiple testing using Bonferroni correction ($p < 0.05/13 = 3.9 \times 10^{-3}$ for self-reported disease and $p < 0.05/16 = 3.1 \times 10^{-3}$ for self-reported family history of disease).

Multivariate Cox proportional hazards models with two levels of adjustments (as per the logistic models described above) were fit to study the association between survival and outlier burden at baseline in the Lothian Birth Cohorts of 1921 and 1936.

A meta-analysis of hazard ratios obtained in the LBC1921 and the LBC1936 was run using a fixed effect model as implemented in the rma.uni function in the *metafor* package [37]. We corrected for multiple testing using Bonferroni correction ($p < 0.05/3 = 0.017$).

### Identification of outlier individuals
Outlier individuals were defined as those with an outlier burden ± 3 interquartile ranges from the upper/lower quartile of the outlier burden in each cohort.

### Heritability of DNA methylation outlier burden
Genotype data at > 700,000 SNPs in Generation Scotland was generated using the Illumina OMNIExpress chip. Briefly, SNPs with more than 2% missingness and a Hardy-Weinberg equilibrium test $p < 10^{-6}$ were excluded during quality control. This has been described in greater detail elsewhere [38]. The genetic contribution by all genome-wide SNPs to phenotypic variance in $\log_{10}$(outlier burden) was estimated using genome-wide complex trait analysis (GCTA), GCTA-GREML [39].

## Supplementary information
**Supplementary information** accompanies this paper at https://doi.org/10.1186/s13148-020-00838-0.

---

**Additional file 1: Supplementary Tables 1–11.**

**Additional file 2: Supplementary Figures 1–5.** and **Supplementary Notes 1 and 2.**

---

Seeboth *et al. Clinical Epigenetics*     (2020) 12:49

Page 12 of 13

### Availability of data and materials

According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to access@generationscotland.org. Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh. Lothian Birth Cohort data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

### Ethics approval and consent to participate

Generation Scotland has been ethically approved and granted research tissue bank status by the NHS East of Scotland Research Ethics Service(REC reference numbers: 05/S1401/89 and 15/ES/0040, respectively). Ethical approval for the first wave of Lothian Birth Cohort studies was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). All participants provided written informed consent.

### Consent for publication

Not applicable.

### Competing interests

AMM has received research support from Eli Lilly, Janssen and the Sackler Trust. AMM has also received speaker fees from Janssen and Illumina. The other authors declare that they have no competing interests.

### Author details

[1]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [3]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK. [4]Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh EH10 5HF, UK. [5]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK.

### References

1. Lowsky DJ, Olshansky SJ, Bhattacharya J, Goldman DP. Heterogeneity in healthy aging. Journals Gerontol Ser A Biomed Sci Med Sci. 2013;69(6):640–9.
2. Bell JT, Spector TD. DNA methylation studies using twins: what are they telling us? Genome Biol. 2012;13(10):172.
3. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2012;13(2):97.
4. Johansson Å, Enroth S, Gyllensten U. Continuous aging of the human DNA methylome throughout the human lifespan. PLoS One. 2013;8(6):e67378.
5. Slieker RC, van Iterson M, Luijk R, Beekman M, Zhernakova DV, Moed MH, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol. 2016.
6. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, et al. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci. 2005;102(30):10604–9.
7. Teschendorff AE, Jones A. Widschwendter M. BMC Bioinformatics: Stochastic epigenetic outliers can define field defects in cancer; 2016.
8. Teschendorff AE, Gao Y, Jones A. Ruebner M. Wachter DL, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun: Beckmann MW; 2016.
9. Barbosa M, Joshi RS, Garg P, Martin-Trujillo A, Patel N, Jadhav B, et al. Identification of rare de novo epigenetic variations in congenital disorders. Nat Commun. 2018;9(1):2064.
10. Shaw AC, Goldstein DR, Montgomery RR. Age-dependent dysregulation of innate immunity. Nat Rev Immunol. 2013;13(12):875.
11. Thompson RF, Fazzari MJ, Greally JM. Experimental approaches to the study of epigenomic dysregulation in ageing. Exp Gerontol. 2010/01/10. 2010 Apr;45(4):255–68.
12. Gentilini D, Garagnani P, Pisoni S, Bacalini MG, Calzari L, Mari D, et al. Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. Aging (Albany NY). 2015;7(8):568–78.
13. Wang Y, Karlsson R, Jylhävä J, Hedman ÅK, Almqvist C, Karlsson IK, et al. Comprehensive longitudinal study of epigenetic mutations in aging. Clin Epigenetics. 2019.
14. Curtis SW, Cobb DO, Kilaru V, Terrell ML, Marder ME, Barr DB, et al. Exposure to polybrominated biphenyl and stochastic epigenetic mutations: application of a novel epigenetic approach to environmental exposure in the Michigan polybrominated biphenyl registry. Epigenetics. 2019:1–16.
15. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G. Sadda SV, et al. Mol Cell: Genome-wide methylation profiles reveal quantitative views of human aging rates; 2013.
16. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A. Raj K, et al. Aging (Albany NY): DNA methylation GrimAge strongly predicts lifespan and healthspan; 2019.
17. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol [Internet]. 2013;14(10):R115. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4015143&tool=pmcentrez&rendertype=abstract.
18. Ghosh J, Mainigi M, Coutifaris C, Sapienza C. Outlier DNA methylation levels as an indicator of environmental exposure and risk of undesirable birth outcome. Hum Mol Genet. 2015;25(1):123–9.
19. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging (Albany NY). 2019;11(7):2045.
20. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet. 2006;7.
21. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol [Internet]. 2013;42(3):689–700. Available from. http://www.ncbi.nlm.nih.gov/pubmed/22786799.
22. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: The lothian birth cohorts of 1921 and 1936. Int J Epidemiol. 2012;41(6):1576–84.
23. Taylor AM, Pattie A, Deary IJ. Cohort profile update: the Lothian birth cohorts of 1921 and 1936. International Journal of Epidemiology. 2018.
24. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome Res. 2014;24(11):1725–33.
25. Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. Genome Med. 2018;.
26. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547–8.
27. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genomics Data. 2016;9:22–4.
28. McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, et al. Identification of 55,000 replicated DNA methylation QTL. Sci Rep. 2018;8(1):17605.
29. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics [Internet]. 2012;13(1):86. Available from: http://www.biomedcentral.com/1471-2105/13/86.

30.  Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–9.
31.  Team CR. R a Lang Environ Stat Comput; 2015.
32.  Bates D, Maechler M, Bolker B, Walker S. lme4: linear mixed-effects models using Eigen and S4. R Packag version. 2014;1(7):1–23.
33.  Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw. 2017;82(13).
34.  Therneau TM. A package for survival analysis in S, v. 2.38. See https//CRAN R-project org/package= Surviv. 2015;.
35.  Wickham H. The tidyverse. R Packag ver 11. 2017;1.
36.  Kassambara A, Kosinski M, Biecek P, Fabian S. Package 'survminer.' 2017.
37.  Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1–48.
38.  Amador C, Huffman J, Trochet H, Campbell A, Porteous D, Wilson JF, et al. Recent genomic heritage in Scotland. BMC Genomics. 2015;16:437.
39.  Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.

## Publisher's Note