Author's Response To Reviewer Comments

We thank the Reviewers for positive feedback and valuable comments. To address the comments we have made substantial changes to the manuscript, predominantly in the Methods and Results sections. We have also uploaded new publicly available data on GEO. Our point by point response is provided below in sections marked ‘Author Response’.

Response to reviewers:

Reviewer #1: This is an interesting epigenetic study on periodontitis using multiple tissues in a relatively large sample size and applying proper statistical models. The presentation is well organized in a well written paper. I only have some technical comments.

1. Table 1 is only mentioned in the Metabolomics data section. It should be referred to much earlier in the Study population section.

Author Response:

We thank the Reviewer for this comment. As suggested, we now refer to Table 1 earlier in the manuscript in:

- Methods on page 8 lines 114-115 in the manuscript where we state:

“The number of samples across different datasets explored in this study is described in Table 1.”

- Results on page 18 line 312 where we write:

“Epigenome-wide analysis of gingival bleeding was carried out in the blood methylome across 452874 probes in 528 subjects (Table 1).”

2. It seems that the SNP CpGs were not removed from analysis. What happened to the Y-linked CpGs?

Author Response:

Our analyses utilized all data on autosomal and X-chromosome probes. We used the 65 SNP probes to check for sample swaps (there we no sample swaps) and we included these probes in the downstream analysis. We excluded probes on the Y-chromosome, and have updated the manuscript to clarify this in the Methods (page 11, lines 166 and 169) and Results sections (page 18 line 312).

3. Before applying statistics, methylation data were normalized to standard normal distribution. How was this done? On beta value or M value?

Author Response:

Methylation beta values were normalized to the standard normal distribution. This is described in Methods on page 11 line 171-172 where we state:

“The methylation beta levels at each probe were normalized to standard normal distribution (N (0,1))…”

4. In the EWAS model on whole blood, cell composition was included. Was this blood cell count or cell proportion estimates from minfi package? In any case, it should be clearly stated.

Author Response:

Blood cell counts were estimated using the DNA Methylation Age Calculator online (Horvath. 2013. Genome Biology 14:R115) We now clarify this in Methods on page 16, lines 251-252 where we state:

“Blood cell counts were estimated using the DNA Methylation Age Calculator online [21]”

5. In the mixed effects model, family structure (twin pairing) and zygosity were defined as random variables. I would like to know how zygosity is specified as a random factor.

Author Response:

The zygosity variable was introduced as a random effect in the mixed effects model. In the results presented in this manuscript zygosity was specified as a simplified factor variable differentiating MZ complete twin-pairs from all other individuals in the dataset (full DZ twin-pairs and unrelated individuals). This approach makes over-simplified assumptions, including that effects across two twins in a DZ pairs are similar, and that effects across different pairs of MZ twin are similar. We have carried out comparisons with models where zygosity is coded as a factor variable with multiple levels capturing more closely complex effects across different types of twin pairs (i.e. same factor level for any full pair of MZ twins, and different factor levels for two twins in a DZ pair; applied across all twin pairs). Our comparisons show that EWAS results (P-value estimates and regression coefficients) from the simplified and full models are consistent, and also that the simplified model results are consistent with those from a simple linear regression in the largest subset of unrelated subjects in the sample. Therefore, to reduce computational burden we fit EWAS using a simplified zygosity variable.

6. RNAseq was performed on whole blood. Why blood cell composition was not considered in data analysis? In the EWAS model on whole blood, cell count data were adjusted. This data can easily be included in statistical analysis of RNAseq data.

Author Response:

This is because we did not have estimated blood cell counts for all the subjects with available gene expression levels in blood. Specifically, only a small number of samples overlapped and had both gene expression and DNA methylation estimates (N=120 (Gum bleeding), N=112 (Tooth mobility)). Furthermore, even in subjects who had both gene expression and methylation profiles, blood methylation-based cell composition estimates were not compatible for expression analysis, because the methylation and expression samples were collected at different time points in many subjects. Therefore, we conducted gene expression analysis without adjustment for blood cell counts.

7. Some of the candidate genes selected were from GWAS. Replication of these genes would mean cis regulated mQTL. However, the biggest advantage of using twins in EWAS is to control for genetics effects. Could this weaken the replication analysis of GWAS results?

Author Response:

Our analyses in the current paper are not replication of previously reported GWAS signals. Here, we explored if there was evidence for epigenetic changes in previously identified periodontitis-GWAS candidates. This analysis is in part motivated by our previous results from an epigenetic twin study of type 2 diabetes (T2D). In this previous twin-based study we showed an enrichment for T2D-associated differential methylation signals at T2D-GWAS loci in T2D-discordant MZ twins for T2D, despite controlling for genetic variation using the MZ twin model (Yuan et al. 2014. Nature Communications, 5:5719). We interpret this finding as evidence that GWAS signals target biologically relevant genes for the disease, where multiple layers of biological data are perturbed and show associations with the disease.

8. Concerning heritability estimates, the authors should at least briefly explain how the classical twin model works. Also, which twin model was fitted? ACE or ADE?

Author Response:

We thank the Reviewer for this comment. To address this we have now added an explanation of heritability calculation in the Methods. We write on page 17 line 275-283:

“The phenotypic variance is partitioned into three latent components: additive genetic, shared environmental, and unique environmental factors. In this context, shared environmental factors typically represent familial exposures shared by both twins (for example, in childhood), whereas unique environmental factors are exposures that are specific to each individual subject. Structural equation modeling was used to estimate the best-fitting model for the traits using the likelihood ratio test. The contribution of each latent component to the phenotypic variance was also quantified.”

Reviewer #2: In this manuscript, Dr. Kurushima and colleagues conducted comprehensive two-stages epigenetic-based association study to periodontitis based on hm450 array from UK twins cohort. Both gingival bleeding (N=528) and tooth mobility (N=492) were investigated in the study and several positive association were identified, such as ZNF804A, WHAMM, TMCO6. What's more, the corresponding result were validated by RNA-seq and Metabolomics-assay. The study was performed rigorously and the findings are quite interesting. I only have several tiny concerns for the authors to think about and make further response.

Major Compulsory Revisions

1. In the self-reported questionnaire section, I am wondering why not asking the earliest time for the tooth mobility? Or why not ask the times of the gingival bleeding in the past several years? I don't understand why "difference between the time point of dental phenotype collection and DNA extraction" should be taken as a confounder or factors in the linear regression?

Author Response:

Ideally, DNA collection would be carried out at the same time-point as dental data collection. However, this was not always the case in the TwinsUK cohort existing dataset and the questions that the Reviewer suggests including are not available in the existing data from the TwinsUK cohort. It is beyond the scope of this manuscript to seek funding to carry a new collection of dental phenotype data and epigenetic data from the same time-point.

To take into account the fact that some samples had a longer time gap between DNA collection and dental trait data, and during this time period either methylation or dental traits or both may change, we used the time gap between DNA extraction and dental trait questionnaire data collection as a covariate in the downstream analyses. On the other hand, periodontal traits are unlikely to be very different over 5 years considering the nature of chronic inflammation.

2. How to normalize beta-value to normal distribution (N,0,1)? The details should be provided? Did the author removed the probes containing SNPs? For the RNA-seq analysis, how to deal with the splicing reads should be mentioned?

Author Response:

The Reviewer raised three points in this comment (2.1-2.3, below).

2.1 Quantile normalisation was applied to methylation beta values before downstream analysis – please see response to Reviewer 1 comment 3.

2.2 With regard to the probes containing SNPs, if the Reviewer is referring to the 65 SNPs in the dataset, then please see Response to Reviewer 1 comment 2.

If the Reviewer is referring to probes for which the probe sequence contains SNPs, then we kept these in the analysis because at least one third of Illumina 450K probes contain SNPs in the probe sequence, and this proportion increases if we consider the full spectrum of genetic variant allele frequencies. In the majority of cases, a SNP that is located away from the targeted CpG site will not necessarily have a major impact on hybridization efficiency. We therefore felt that we should not exclude such a large number of data points from the analysis– this practice is in line with the overwhelming majority of published EWAS to date, which do not exclude probes with SNPs in the probe sequence.

2.3 RNA-seq analyses were carried out at the exon level, as we state in the Manuscript. This would capture effects attributed both to differential splicing and to differences in overall gene expression levels.

3. In the first result section, cg21245277 was identified to be signficiantly associated with gingival blooding. As the author mentioned, it is obviously a probe containing SNPs. Methylation status in the probes containing SNPs usually shown mis-leading methylation signals since the probe matching and extension, therefore, the epigenetic conclusion here should be double-checked with BS-Seq.

Author Response:

We agree with the Reviewer that this result should be interpreted with caution. To address this comment we now state in the Discussion on page 31 line 535 – page 32 line 549:

“A further limitation was the lack of follow-up validation of the most associated signals in our study using a different DNA methylation profiling technology. One of the peak signals in our results, cg21245277 in ZNF804A contains genetic variants in the probe sequence, which may impact probe hybridization efficiency. It would therefore be valuable to validate the observed DNA methylation levels at this CpG-site using a different technology, for example, bisulfite pyrosequencing.”

4. In the table 3. Genomic assemble version should be mentioned, hg19 or hg38. Meanwhile, the full variable information (P-value, beta, SE) for all the confounder (age etc) will be helpful to be shown in the supplementary.

Author Response:

We have now included the genome assembly version (hg19) in the footnote to Table 3.

To address the rest of the Reviewer’s comment we now include three new Supplementary files: Supplementary Tables 1-3 and refer to these in the Results (page 18, line 314; page 20 lines 355-356).

5. the authors thought buccal tissue would be a good choice to replace with gingiva, buccal mucosa and saliva. Is there any evidence to support this hypothesis?

Author Response:

We agree with the Reviewer that choice of appropriate tissue should be discussed. We therefore now state in the Discussion page 31 lines 530-532:

“Epigenetic modifications in the oral cavity are, to date, considered to take place mostly in epithelial tissue [41] and buccal swabs of the oral cavity typically contain the greatest amount of epithelial tissue [42]. “

6. In the Figure 1, How to understand only cg21245277 was significantly associated with gingival bleeding while other CpGs in the promoter region of ZNF804A don't have any significant signals? Another question is that whether the samples in Figure 1B and 1C are same? Can you observe same difference applying same samples?

Author Response:

We thank the Reviewer for raising this point. In Figure 1A bottom panel we can see that the most-associated probe cg21245277 falls in a co-methylation block with nearby probes on either side, specifically the co-methylation block includes the following probes in 5’ to 3’ orientation: cg00554682, cg13107760, cg21245277, cg13536757, and cg16573328. As we can see below, all 5 probes have a consistent direction of association with gingival bleeding, and 4 of the 5 signals are nominally significant. This provides further supporting evidence for the importance of this signal in gingival bleeding.

Probe\_name Linear\_Regression\_Coefficient SE P-value

cg00554682 -0.121938177 0.068031483 0.076443139

cg13107760 -0.163799037 0.059031207 0.005899856

cg21245277 -0.334767594 0.061112138 7.17E-08

cg13536757 -0.177731962 0.071360307 0.013073907

cg16573328 -0.172161861 0.073956972 0.020346199

In terms of the sample numbers used in Figure 1B and 1C: the number of the samples for EWAS was 528, whereas that for TWAS was 342, and the overlapping number between these two samples was 120. These overlapping samples are plotted in the boxplot in Figure 1B and Figure 1C – therefore the samples are indeed the same in those Figures.

7. the authors mentioned 'family structure' was assigned as one of confounders, I am wondering, how to quantitate this variable? And I didn't find the information for family structure in Supplementary Tables 1-2.

Author Response:

This is the family identifier. To address this comment we now add a column “new\_famID” representing family number in Supplementary Tables 4-8. Usage of this variable in the linear mixed effects model is described in the Methods.

8. UK twins dataset have been used to conduct the EWAS for several different traits. What's the second frequently traits or top 10 frequently diseases traits for the current samples, such as obesity, diabetes? Meanwhile, why the author only selected female twins in the current study?

Author Response:

The dataset included in this paper was not ascertained to select for subjects with a particular disease, but is representative of the TwinsUK larger sample of unselected predominantly female (90%) research volunteers from the UK population. To address this comment, we now include information on age and BMI in the Methods on page 8 lines 116-120 where we state:

“Gingival bleeding status was obtained from 528 female individuals (mean age 57.9, age range (18.6-80.9), mean BMI (26.9), BMI range (16.1-49.3)), and tooth mobility data was obtained from 492 female individuals (mean age 58.0, age range (18.6-80.9), mean BMI (26.9), BMI range (16.2-52.7)), for whom blood DNA methylation profiles were also available.”

9. this manuscript is significantly a statistical analysis manuscript, submitting the corresponding script conducted in the study to github will be helpful for the further repeat and validation.

Author Response:

The methods use standard linear regression models throughout and are straightforward standard approaches that do not necessitate custom scripts. We state in the Methods on page 16 lines 253-255:

“The association between DNA methylation levels and periodontal traits were carried out using lmer within R 3.4.3 in the package “lme4”.”

Minor Essential Revisions and Discretionary Revisions

1.The authors mentioned personalized treatments and prevention in both abstract and background, however, nothing further discussion were provided to illustrate how to apply methylation signals in 'personalized treatments and prevention'.

Author Response:

We thank the Reviewer for the comment. We still need further research (i.e. on tissue specificity, replication, and reverse causality) until personalised treatment and prevention can be realistically explored. At present, we envisage that it may be feasible to use DNA methylation to identify subjects at high risk of periodontal disease and provide advice on exposures of specific environmental factors. To address this comment, we now state in the Discussion on page 33 lines 566-571:

“Longer-term, it may be feasible to use DNA methylation profiles towards personalised treatment and prevention of disease. For example, DNA methylation levels at the most associated signals in our study may be used to identify subjects at high risk of disease, towards modifications in exposures for specific environmental risk factors.”

2.The authors tried to validation the result from whole blood in buccal and adipose tissues, However, as known, the methylation pattern in different tissues usually is different, what's the meaning to validate the result in different tissue for periodontitis.

Author Response:

We use the term validation here to assess if the methylation association with periodontal traits validates across tissues.

3. I want to know why gingival bleeding status and tooth mobility status were obtained from different samples?

Author Response:

This is because the twins answered two questionnaires to assess periodontal traits. Although a large number of twins answered both questionnaires, not all subjects completed both questionnaires. The number of subjects who answered both questionnaires for each ‘omic dataset is clearly described in the last column in Table 1.

4. The author mentioned 'Informed consent was obtained from all participants before sample and data collection'. As we known, UK twin cohort blood extraction is far before the further study design, a detail relationship about current study design and blood sample, as well questionnaire will be helpful.

Author Response:

Information on informed consent (including REC numbers) is included in the Ethics section on page 34 line 599 – page 35 line 603. Detailed description of questions used to assess periodontal traits is included in the Methods “Periodontal phenotypes” section on page 9 line 137 – page 10 line 150.

5. GSE62992 is not exactly what the authors mentioned in the manuscript. GSE62992 only contains 100 normal whole blood samples, please double check it.

Author Response:

We thank the Reviewer for this comment. To address this we have now uploaded a new GEO dataset (GSE121633) and updated the data availability section on page 35 lines 606-616.

In this manuscript, Dr. Kurushima and colleagues conducted comprehensive two-stages epigenetic-based association study to periodontitis based on hm450 array from UK twins cohort. Both gingival bleeding (N=528) and tooth mobility (N=492) were investigated in the study and several positive association were identified, such as ZNF804A, WHAMM, TMCO6. What's more, the corresponding result were validated by RNA-seq and Metabolomics-assay. The study was performed rigorously and the findings are quite interesting. In the first round response, the authors have solved majority of my concerns and I still have some minor concerns before the acceptation:

1. I don’t know how to normalize beta to normal distribution since ‘Methylation Beta’ usually Beta distribution. The authors can check it again in the last version of the manuscript.
2. Suppose the authors didn’t remove the SNPs in the array, I suggest the authors to label whether it is overlapped with common SNPs in any tables of the main-manuscript.
3. As we known, UK-biobank samples have comprehensive phenotype information and phenotype usually is not independent, therefore these variables should be adjusted in the models. The authors adjusted BMI and age, family structure, however, the most relevant phenotypes, such as rheumatoid arthritis, were not mentioned in the manuscript. It might be better to have some discussion in the discussion section.
4. It will be better to show the dot for each sample in Figure 1B and 1C while significant gene symbol could be shown in Figure 2A. In addition, now that Figure 1B and 1C are from same samples, linear regression plot could be shown other than separated methylation and expression plot, right?