A comparison of epithelial cell content of oral samples estimated using cytology and DNA methylation

Author list (suggested)

Yen Ting Wong[1], Michael A Tayeb[2], Laurence B Lovat[3], Andrew E Teschendorff[4,5], Rafal Iwasiow[2], Jeffrey M Craig[1,6,7]

Affiliations

[1] IMPACT Strategic Research Centre, School of Medicine, Barwon Health, Geelong, Vic, Australia

[2] DNA Genotek Inc., Ottawa, ON, Canada

[3] Division of Surgery & Interventional Science, UCL, London WC1E 6BT, UK

[4] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, 320 Yue Yang Road, Shanghai 200031, PR China.

[5] UCL Cancer Institute, 72 Huntley Street, University College London, London WC1E 6BT, UK.

[6] Murdoch Children's Research Institute, Department of Paediatrics, The University of Melbourne, Royal Children's Hospital, Melbourne, Vic., Australia.

[7] Corresponding author

jeffrey.craig@deakin.edu.au

Abstract

Saliva and buccal samples are popular for epigenome wide association studies (EWAS) due to their ease of collection compared and their ability to sample a different cell lineage compared to blood. As these samples contain a mix of white blood cells and buccal epithelial cells that can vary within a population, this cellular heterogeneity may confound EWAS. This has been addressed by including cellular heterogeneity obtained through cytology at the time of collection or by using cellular deconvolution algorithms built on epigenetic data from specific cell types. However, to our knowledge, the two methods have not yet been compared. Here we show that the two methods are highly correlated in saliva and buccal samples ($R = 0.84$, $P < 0.0001$) by comparing data generated from cytological staining and Infinium MethylationEPIC arrays and the EpiDISH deconvolution algorithm from buccal and saliva samples collected from twenty adults. In addition, by using an expanded dataset from both sample types, we confirmed our previous finding that age has a significant negative correlation with epithelial cell proportion in both sample types. However, children and adults showed a large within-population variation in cellular heterogeneity. Our results validate the use of the EpiDISH algorithm in estimating the effect of cellular heterogeneity in EWAS and showed DNA methylation generally underestimates the epithelial cell content obtained from cytology.

Keywords Buccal, Cytology, DNA methylation, cell-type heterogeneity, epithelial cell, EWAS, saliva

1. Introduction

Cellular heterogeneity is a major potential confounder of epigenome-wide association studies (EWAS) due to the cell type-specific state of DNA methylation. This is particularly the case in oral samples, which are a mixture of epithelial cells from the ectoderm germ cell lineage and immune cells from the mesoderm lineage (1-3). We and others have found that cellular heterogeneity in oral samples is influenced by the method of sample collection, with buccal swabs containing a much higher proportion of epithelial cells than saliva (2, 3). We have also shown that epithelial cell proportion is also strongly influenced by age and oral health status (3). Deconvolution of cellular heterogeneity can be achieved by measuring the proportion of each cell type using cytology of collected cells or through algorithms that use DNA methylation data from specific cell types to generate estimates (1, 2). Such measures can then be used in EWAS models to correct for cellular heterogeneity. There are studies comapring cytology estimates of tumor purity to DNA methylation based estimates and mRNA expression based {Chakravarthy A #2018, Aran D # 2015} However, to our knowledge, no study has compared these two methods for oral samples. We aimed to compare epithelial cell content of buccal samples via ORAcollect•DNA kits and saliva obtained via passive drool collected in Oragene•DNA kits, measured using cytology and estimated with the reference based EPIDISH algorithm (2). We hypothesised that estimations of epithelial cell content would be highly correlated between the two methods. In a sub-study, using customized ORAcollect•DNA collection instructions, we compared two similar methods of collection differing in collection site and duration.

2 Materials and methods

2.1 Participants

Twenty adult volunteers from Deakin University provided informed consent to collect one saliva sample and two buccal samples. Ethics approval was granted by the Human Research Ethics Committee of the Royal Children's Hospital, Melbourne (#33174) and Deakin University (2018-368). All methods were performed according to relevant protocols and regulations. Participants also completed an oral health questionnaire, which included questions about mouth injuries, oral infections, medications and smoking status (Supplementary Methods).

2.2 Oral sampling

Oral samples were obtained from participants under supervision of the research team. Participants were advised not to smoke, chew gum, or consume anything apart from water for the 30 minutes prior to providing samples. Ten minutes prior to sample collection, they were asked to rinse their mouth with water. Saliva samples were collected unstimulated via passive drool for three to five minutes to allow sufficient time to collect to the fill line (2mL) of Oragene•DNA collection devices (OG; DNA Genotek Inc, Ottawa, Canada). One hundred microliters of saliva were then smeared onto a microscope slide and immediately fixed with 95% ethanol for 10 minutes and left to dry at room temperature. Oragene DNA-stabilising chemistry contained within the device was then released into the remaining sample. Following collection of saliva, two samples were collected from participants using ORAcollect•DNA (OC, DNA Genotek Inc, Ottawa, Canada), a sponge-tipped oral sample collection kit, sequentially using two collection methods. In the first (OCA) participants gently rubbed the sponge ten times in a back-and-forth motion in the furrow between their lower teeth and inner cheek on one side of their mouth. In the second (OCB), the sponge was rubbed up and down against the inside of the cheek twenty times then rubbed ten seconds in a back-and-forth motion in the furrows between their upper and lower right teeth and inner cheek on the opposite side of their mouth. Each sponge was wiped along the length of a standard size microscope slide

and fixed as outlined for saliva. The sponge was then inserted into the ORAcollect•DNA tube containing DNA stabilising chemistry, capped tightly and mixed by inversion 15 times.

2.3 Slide staining and microscopy

Slides were stained using Diff-Quik as detailed elsewhere (4). All slides were deidentified and analysed by two observers. Cell types were counted via bright field microscopy at 100x magnification in regions with adequate cell density. Counts were used if the discrepancy between observers was less than 10% of the total count of count for each cell type. For counts which discrepancy between observers in more than 10% will be discarded and re-count again. A minimum of 50 epithelial cells and a 100 cells total was counted. Cells were scored as epithelial cells or immune cells, the latter including segmented cells, lymphocytes and monocytes (3).

2.4 DNA extraction

Genomic DNA was extracted from 0.5 mL of each oral sample using ethanol precipitation via prepIT•L2P kits (DNA Genotek Inc, Ottawa, Canada) following the manufacturer's protocol. DNA concentration was measured using PicoGreen (Thermo Fisher Scientific, Canada) in a SpectraMax M2 plate-based fluorimeter (Molecular Devices, CA, USA). DNA quality was measured using a TapeStation (Agilent, Santa Clara, United States).

2.5 DNA Methylation arrays

Following genomic DNA extraction from all the samples, these genomic DNA samples were treated with bisulphite to convert unmethylated cytosine into uracil and transformation of uracil into thymine by amplification. Genome-wide analysis of DNA methylation was assessed using Infinium MethylationEPIC arrays (Illumina, CA, USA) with probes of over 850,000 methylation sites at the GenoFIND Genomic Service Lab (DNA Genotek Inc, Ottawa, Canada). Hybridization and scanning were performed according to manufacturer's instructions.

2.6 Pre-processing of Illumina Infinium array data

MethylationEPIC array analysis was performed using the R statistical programming language (www.R-project.org) and Bioconductor packages (5). Raw intensity data (IDAT) files were imported into R (3.6.3; http://cran.r-project.org/). Data quality was assessed using the *minfi* (v1.34.0) Bioconductor package (5). The MethylationEPIC probes were filtered by removing those with poor signal to noise ratio (mean detection p-value of >0.01), cross-reactivity to multiple genomic locations, containing a single nucleotide polymorphism at the CpG site, or map to sex chromosomes (6). Data was then subjected to subset-quantile within array normalisation (SWAN), (7)) and between-array normalisation (SQN) (8). The HEpiDISH/EpiDISH and Robust partial correlation (RPC) algorithms were applied to estimate proportions of epithelial, fibroblast and immune cells from MethylationEPIC array data (9).

2.7 Data analysis

Descriptive statistical analyses were conducted on the age of the participant and proportions of epithelial and immune cells. The assumption of normality of the independent and dependent variables for each cell type was tested using the Shapiro-Wilk test. A Kruskal-Wallis ANOVA analysis was conducted to test for statistically significant differences in cell proportion and DNA yield between collection methods OCA, OCB and OG. In a post-hoc test, the Dunnett's test with Bonferroni correction was applied to identify the relatively small but significant differences among collection methods. Variables collected with insufficient number and information will not be included in the statistical analysis.

Percentage of epithelial cells and estimated cell-type fractions from EpiDISH were graphed using box and whisker plots, which included information on interquartile range (boxes, 25^{th} to 75^{th} percentiles, boxes), median (horizontal lines), data within 5^{th}-95^{th} percentiles (whiskers), outliers (circles), and mean (crosses). The proportion of epithelial cells in oral samples

estimated from cytology and DNA methylation was tested using Pearson correlation coefficient. To investigate the age effect on epithelial cell content estimated using DNA methylation, the buccal and saliva sample data from this study was analysed along with seven of our other studies, three published (3, 10, 11) and four unpublished. These cohorts' details included to investigate the age effect on epithelial cell content was described in supplementary method.

3 Results

3.1 Determination of epithelial and immune cell proportions using cytology

Slides from all twenty adults (mean age 26.9 years, range 21 to 48 years, 60% female) were analysable i.e. had sufficient cells for analysis. Seven individuals reported recent gum bleeding within the seven days preceding their collection day. Examples of microscopic fields of view are shown in **Figure 1**. Epithelial cells were large, with low nuclear to cytoplasmic ratio and immune cells were much smaller with a high nuclear to cytoplasmic ratio. Immune cells included granulocytes with segmented nuclei, lymphocytes with round, dense nuclei surrounded by cytoplasm, monocytes with kidney-shaped nuclei. Between two and twenty fields of view at 100x magnification were required to score the minimum number of cells.

Figure 1 around here

Results for estimations of epithelial cell proportions determined by cytology and DNA methylation analysis are shown in **Supplementary Table 1** and **Figure 2**. The mean proportion of epithelial cells in saliva (58%, SD 17.1%), was significantly lower with than sponge collection methods OCA (86.0%, SD 9.9%) and OCB (87.0%, SD 11.2%), $p < 0.0001$. A

28.5% mean difference with SD 6.5% in compared saliva to cheek swab methods. There was no evidence for a difference in epithelial cell proportions between OCA and OCB ($p = 0.6$). There was also no evidence of an influence of recent gum bleeding (p value =0.5) and sex (p value = 0.9) on epithelial cell proportion across all methods of sampling; results for individual oral collection methods were similar.

Figure 2 around here

3.2 Determination of epithelial cell proportions using DNA methylation analysis

Saliva samples showed a significantly higher mean of total DNA yield per mL (33.7µg, SD 24.2µg) compared to oral sponge collection methods OCA (4.1µg, SD 1.57µg) and OCB (5.9µg, SD 2.71µg), $p < 0.0001$ for both comparisons (**Figure 3**). Although DNA yield was approximately 1.7x higher in OCB compared to OCA, this difference was not significant ($p = 0.083$).

Figure 3 around here

We next used the EpiDISH and robust partial correlation (RPC) algorithm on Infinium MethylationEPIC data to estimate cell type proportions. Although this method calculates proportion of epithelial, immune and fibroblast cell types, we found that the proportion of fibroblasts was negligible (mean = 0.4%) (**Supplementary Table 1**). As this meant that the proportion of immune and epithelial cells had a correlation of -1.0, we limited our analysis to the latter. As with cytology, the mean proportion of buccal epithelial cells determined by DNA

methylation in saliva (25.4%, SD 17.1%), was significantly lower than cheek swab methods OCA (69.5%, SD 18.8%) and OCB (75.5%, SD 17.0%), $p < 0.0001$ (**Figure 2**). A 47.1% mean difference with SD 0.7% in compared saliva to cheek swab methods. There was no evidence for a difference in epithelial cell proportion between OCA and OCB ($p = 0.11$).

3.3 Comparison between epithelial cell proportions estimated using cytology and DNA methylation

To address our hypothesis that proportions of epithelial cells present in oral samples estimated using DNA methylation analysis represented the cell proportions as measured by cytology, we pooled all samples and compared both methods (**Figure 4**). The two methods were strongly correlated ($R = 0.84$, $P < 0.0001$). However, the intercept of the line of best fit (methylation % = [1.32x cytology %] – 45%) on the x axis was 34%. A 20% mean difference of DNA methylation (SD 17.7%, IQR 63.7%) compared to cytology (SD 12.7%, IQR 50.8%) (**Supplementary Table 1**). Methods correlated similarly in buccals ($R = 0.75$, $P < 0.0001$) and saliva ($R = 0.72$, $P < 0001$) (**Supplementary Figure 1**).

Figure 4 around here

3.4 An age effect on epithelial cell content in saliva

In our previous paper, we observed that epithelial cell content of buccal swabs and saliva was lower in adults compared to children (3). To investigate a possible age effect using epithelial cell content estimated using DNA methylation, we combined buccal swab and saliva data from this study with seven of our other Infinium array studies, including three published (3, 10, 11) and four unpublished (**Figure 5, Supplementary Table 2**). We found a moderate negative

correlation between age and epithelial cell content estimated by DNA methylation (R = -0.72, p <0.0001, 0.59% of epithelial cell content estimated from DNA methylation), with age accounting for 14% of the variation in epithelial cell content. We found a stronger relationship in buccals (R = -0.85, P <0.0001) compared to saliva (R = -0.28, P <0.0001) (**Supplementary Figure 2**).

4 Discussion

4.1 Influences on epithelial content of oral samples

Buccal and saliva samples have a proven utility for epigenomics (12, 13) and other cell-based omics (14, 15). As these samples are mixtures of epithelial and immune cells, deconvolution of these cellular mixtures is of utmost importance. Although cellular deconvolution algorithms based on reference sample types have been applied to epigenomic studies (2), to our knowledge, the validity of such algorithms has not yet been tested using cytology of primary samples. We aimed to address this issue.

Our cytological analysis of adults with a mean age of 26 years showed that the epithelial content of ORAcollect•DNA (OC) samples was 86.5%, similar to the 83.4% we previously obtained using Copan flocked swabs in adults 16 years older (3). In the present study, epithelial cell content of saliva, but not buccal samples, was significantly higher than in our earlier study, which agrees with our previous finding that age has a much greater effect on saliva than on buccal samples. We also found that the epithelial content of ORAcollect•DNA collected samples was around 47% higher than that of saliva. This difference was 11% larger than that of our previous study, which may also reflect an age effect.

Our findings also suggest that the type of buccal collector has minimal influence on the proportion of epithelial cells collected and this may also be one reason why increasing

collection time for OC sponge did not increase the proportion of epithelial cells collected, nor did it significantly increase DNA yield. However, future, larger studies are needed to further test our hypotheses. However, there may be a danger that longer collection times penetrate blood capillaries within the inner cheek, which would increase the proportion of immune cells, which may negate any rise in epithelial cell numbers.

We found no evidence that recent gum bleeding influenced the proportions of epithelial cells with either mode of sample collection. This disproved our hypothesis that gum bleeding would decrease proportions of epithelial cells, possibly because the severity and temporary nature of bleeding may be insufficient to cause a significant impact on immune cell numbers and possibly because of our relatively small sample size.

4.2 Comparison of epithelial cell proportions using cytology and DNA methylation

Using the EpiDISH algorithm (2) on DNA methylation data generated by Infinium MethylEPIC arrays, we estimated that epithelial cell proportion was lower in saliva compared to OC-collected samples by an average of 47%, a larger magnitude than that shown using cytology. Across all samples, the correlation between the two methods of epithelial cell estimation was very high (R=0.84). Taken together, these findings prove our primary hypothesis and imply that post hoc deconvolution accurately estimates cellular heterogeneity in oral samples. In a study comparing proportions of various blood cell types estimated using flow cytology and a blood-specific DNA methylation-based algorithm, a wide range of correlations, between 0.51 and 0.97 were observed (16). Our line of best fit showed that DNA methylation underestimated the epithelial cell content determined using cytology by 34% at 0% epithelial cells and by 13% at 87% epithelial cells (**Figure 4**). This is larger than the over- or under-estimations of up to 10% observed in the previous study of blood (16). This discrepancy could be for a number of reasons. The reference dataset was derived from Illumina InfiniumHM450 array data from 11 different epithelial cell lines (2) which may not

accurately represent buccal epithelial cells. We also cannot rule out the possibility that buccal and immune cell types may have been differentially applied to slides prior to cytological examination. Nevertheless, the high correlation between epithelial cell proportions based on cytology and DNA methylation should still be sufficient to use the latter to generate estimations across a set of biosamples for adjustment within EWAS.

4.3 Investigating an age effect on the proportion of immune cells in buccal swabs

We found a significant negative correlation of epithelial content in buccal swabs and saliva with age (**Figure 5, Supplementary Figure 2**). This agrees with our previous study that showed an effect in the same direction with buccal swabs and saliva in children and adults (3). In our earlier study, we showed that epithelial cell proportion was significantly lower in children with gingivitis. As gingivitis and other oral inflammatory pathologies such as periodontitis increase in prevalence with age, this may result in an increase in immune cell content of oral samples and a corresponding decrease in epithelial cell content.

4.4 Strengths, limitations and future studies

To our knowledge, this study is the first to analyse the correlation between cell proportions in oral samples estimated using cytology and DNA methylation. Another strength is our longitudinal analysis showing a decline of epithelial content of buccal swabs and saliva with age. However, our sample size (n=20 for both buccal samples and saliva; all studies, n=579) is relatively small, although our sample size for the study of age effects (n=579) was much larger. Future, larger-scale studies that compare estimates of cell proportion using both cytology and DNA methylation are required to validate our findings. Such studies should include a wider age-group and measures of oral health.

5 Acknowledgements

The authors thank Ziad Marroushi and Christina Dillane for technical assistance and Elizabeth Firth for her comments on the manuscript. We thank Evie Muggli, Jane Halliday, Lata Vadlamudi and Tim Silk for permission to include unpublished data.

6 Disclosure of interest

The authors report no conflict of interest.

7 References

1. Teschendorff AE, Zheng SC. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics. 2017;9(5):757-68.

2. Zheng SC, Breeze CE, Beck S, Dong D, Zhu T, Ma L, et al. EpiDISH web server: Epigenetic Dissection of Intra-Sample-Heterogeneity with online GUI. Bioinformatics. 2019.

3. Theda C, Hwang SH, Czajko A, Loke YJ, Leong P, Craig JM. Quantitation of the cellular content of saliva and buccal swab samples. Sci Rep. 2018;8(1):6944.

4. Yang GC, Papellas J, Wu HC, Waisman J. Application of Ultrafast Papanicolaou stain to body fluid cytology. Acta cytologica. 2001;45(2):180-5.

5. 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.

6. Chen Y-a, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9.

7. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome biology. 2012;13(6):R44.

8. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4(3):325-41.

9. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017;18(1):105.

10. Martino D, Loke YJ, Gordon L, Ollikainen M, Cruickshank MN, Saffery R, et al. Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. Genome Biol. 2013;14(5):R42.

11. Mohandas N, Loke YJ, Hopkins S, Mackenzie L, Bennett C, Berkovic SF, et al. Evidence for type-specific DNA methylation patterns in epilepsy: a discordant monozygotic twin approach. Epigenomics. 2019;11(8):951-68.

12. Langie SA, Moisse M, Declerck K, Koppen G, Godderis L, Vanden Berghe W, et al. Salivary DNA Methylation Profiling: Aspects to Consider for Biomarker Identification. Basic Clin Pharmacol Toxicol. 2016.

13. Farah R, Haraty H, Salame Z, Fares Y, Ojcius DM, Said Sadier N. Salivary biomarkers for the diagnosis and monitoring of neurological diseases. Biomed J. 2018;41(2):63-87.

14. Hassaneen M, Maron JL. Salivary Diagnostics in Pediatrics: Applicability, Translatability, and Limitations. Front Public Health. 2017;5:83.

15. Shah S. Salivaomics: The current scenario. J Oral Maxillofac Pathol. 2018;22(3):375-81.

16. Gervin K, Page CM, Aass HC, Jansen MA, Fjeldstad HE, Andreassen BK, et al. Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. Epigenetics. 2016;11(9):690-8.

Figure Legends

Figure 1. Examples of cellular morphology in oral samples. Representative fields of view from Diff-Quik staining of (A) saliva, 100x magnification and (B) OCA buccal sample, 400x magnification. Both samples contain large epithelial cells (Epi) with dense nuclei, and smaller immune cells, exemplified by lymphocytes (Lym), segmented cells (Seg) and monocytes (Mono).

Figure 2. Comparison of the percentage proportion of epithelial cells in oral samples, estimated using cytology and DNA methylation arrays, collected using three different methods (OCA, OCB and OG) estimated. Means are indicated with crosses. The p value of percentage of epithelial cell between OCA and OCB is $p > 0.05$. The p value of between buccal sample collection (OCA and OCB) compared to saliva (OG) is $p < 0.0001$.

Figure 3: Range of DNA yields for each oral sample type. Box and whisker plots from saliva (OG) and the two methods of buccal sample collection (OCA and OCB). Means are indicated with an X.

Figure 4: Comparison of the proportion of epithelial cells in oral samples estimated from cytology and DNA methylation arrays.

Figure 5: Epithelial cell content of oral samples as a function of age in six studies.

Supplementary Figure 1: Comparison of the proportion of epithelial cells in oral samples estimated from cytology and DNA methylation arrays. (A) Data from saliva epithelial cells; (B) Data from buccal epithelial cells.

Supplementary Figure 2: Epithelial cell content of buccal and saliva samples as a function of age. (A) Buccal data is from five studies (n = 344); (B) saliva data is from three studies (n = 234).

Supplementary Table 1: Estimation of cell proportions and for each collection method.

Supplementary Table 2: Epithelial proportions of the buccal and saliva sample data from this study and five of our other studies (n= 753).

8 Supplementary methods and results

8.1 Questionnaire

Before sample collection, each adult participant was asked to complete an oral health questionnaire. The participant was given a unique ID number, other details included birth year, collection date and time, and sex were recorded. The oral health of each participant was recorded via a questionnaire which asked about whether they had bleeding gums when brushing their teeth, mouth ulcers, other mouth lesions, a cold, a sore throat, or other mouth infection during the past week. Participants were also asked whether they used an inhaler, took antibiotics, anti-inflammatories or blood thinners and whether they smoked.

8.2 Study cohorts

As stated in the method, to investigate the age effect on epithelial cell content estimated using DNA methylation, the buccal and saliva sample data from this study was analysed along with three published (3, 10, 11) and four unpublished of our other studies. First, the child participants are recruited from part of the Peri/postnatal Epigenetic Twins Study (PETS) cohort, an Australian twin birth research study based in Melbourne. These participants are involved in a longitudinal study of DNA methylation at birth (n= 29, age 0 year) and age 18 months (n= 24, age 1.5 years) from buccal swabs (10). Second, Theda et al. study, their saliva and buccal samples were collected from ten pairs of twins (n= 20, age range 6.4-7.1 years) from PETS cohort and adult volunteers (n = 23, age range 23-59 years) (3). Third, an epilepsy cohort consisted of monozygotic twin pairs (n= 28) age range of 14-67, who were discordant for epilepsy without a known acquired cause (11).

The four unpublished studies were an Australian longitudinal study of community-based children with ADHD cohort with match age (n=175, 10.4 years). A population subset of the AQUA (Asking QUestions about Alcohol in pregnancy) cohort study (n=187 of neonatal cheek swabs, match age at 0 year). Two data sets collected from buccal samples in the year

2015 (n= 63, age range 28-87 years) and year 2019 (n= 111, age range 35-90 years) provided by a collaborator from University College London.