Institute of Education

UCL

Collection of DNA samples and genetic data at scale in the UK Millennium Cohort Study

CLS working paper number 2020/7

By Emla Fitzsimons[a,b], Vanessa Moulton[a], David A Hughes[c,d], Sam Neaves[c,d], Karen Ho[d], Gibran Hemani[c,d], Nicholas Timpson[c,d], Lisa Calderwood[a], Emily Gilbert[a], Susan Ring[c,d]

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

[a]Centre for Longitudinal Studies, University College London

[b]Institute for Fiscal Studies, London

[c]MRC Integrative Epidemiology Unit at University of Bristol, Bristol, BS8 2BN

[d]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN

Corresponding author

Emla Fitzsimons (UCL Centre for Longitudinal Studies)

e.fitzsimons@ucl.ac.uk

This working paper was first published in August 2020 by the UCL Centre for Longitudinal Studies.

UCL Institute of Education

University College London

20 Bedford Way

London WC1H 0AL

www.cls.ucl.ac.uk

Disclaimer

This working paper has not been subject to peer review.

CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Institute of Education, University College London, or the Economic and Social Research Council.

How to cite this paper

Fitzsimons, E., Moulton, V., Hughes, D.A., Neaves, S., Ho, K., Hemani, G., Timpson, N., Calderwood, L., Gilbert, E., Ring, S. (2020) *Collection of DNA samples and genetic data at scale in the UK Millennium Cohort Study.* CLS Working Paper 2020/7. London: UCL Centre for Longitudinal Studies.

Abstract This paper describes the collection of saliva samples from cohort members and their biological parents in the Millennium Cohort Study. It analyses response rates, predictors of response, and details the DNA extraction, genotyping and imputation procedures performed on the data.

1. Introduction

The Millennium Cohort Study (MCS) is a large and nationally representative birth cohort study following the lives of children born across the UK around the turn of the millennium (Connelly & Platt, 2014; Joshi & Fitzsimons, 2016). Data collected through the study include a range of detailed social, behavioural and economic measures; cognitive, educational, emotional and physical development; aspirations, identity, wellbeing and personality. There have been seven waves to date, at ages 9 months, 3, 5, 7, 11, 14 and 17 years. A key feature of the sixth (age 14) wave of the study was the collection of genetic information, via saliva, from cohort members (i.e. those who are part of the cohort born at the turn of the Millennium) and their resident biological parents. This paper provides information on the data collection process, response rates, DNA extraction, genotyping, quality control and imputation and genetic ancestry of this cohort.

The addition of genetic information to this rich, longitudinal data resource will enable discovery genome wide association (GWAS) study analyses based on study focus traits, trajectories and familial phenotypes. We expect this resource to trigger a range of studies into how genetic and environmental factors shape human development across the life course. This includes both the use of novel genetic predictors of early life factors and analyses using genetics as a lever for causality (Mendelian randomisation, MR). Genetic data will, in conjunction with the high-quality phenotype data, also enable studies of the longitudinal or fine-scale phenotypic associations of already-discovered genetic associations from larger cohorts. The MCS is unique in being the only population-based, nationally representative study in the UK containing genetic trios.

The paper proceeds as follows: a background in section 2, an overview of the fieldwork collection in section 3, section 4 discusses the DNA extraction process; section 5 describes response rates, section 6 details genotyping, quality control and imputation, section 7 illustrates the resource’s self-described ethnic diversity and continental genetic ancestry, and section 8 describes accessing the data.

2. Background

The sixth, or age 14, wave of the MCS took place across the UK from January 2015 through March 2016. Interviews were conducted with 11,726 families, and 11,872 cohort members (Fitzsimons, 2017), representing a response rate of 76.3%. This was a particularly extensive and innovative wave, containing several elements including interviews and self-completion questionnaires with resident parents, self-completion with cohort members, physical measurements, cognitive assessments, saliva samples, accelerometer collection and time use diaries. The study obtained ethical approval from London-Central REC (13/LO/1786).

A key feature of the sixth wave was the collection of saliva samples from cohort members and their biological parents as part of the home visit carried out by interviewers. Integrating the collection of saliva samples from children and their biological parents into home visits carried out by trained interviewers brought with it several advantages. Associated cost and logistical considerations were particularly important in implementing it at scale across the UK, and in a study involving young people (Sun & Reichenberger, 2014). Whilst lay interviewers are increasingly involved in the collection of biomedical measures using non-invasive methods (McFall, Conolly, & Burton, 2014), our study represented a significant departure from previous longitudinal studies that are tagged on to a clinic visit (e.g. ALSPAC, UK Biobank, NSHD), or taken by nurse interviewers as part of a home visit (e.g. the Health Survey for England, the UK National Diet and Nutrition Survey, the 1958 National Child

Development Study, the 1970 British Cohort Study, and Understanding Society: the UK Household Longitudinal Study (UKHLS)). Follow-up nurse or clinic visits tend to suffer from high dropout rates and are also relatively expensive (Clemens, Given, & Purdon, 2012). Where clinics or nurse visits are not otherwise required, other techniques have been used. For instance, the 1958 National Child Development Study in the UK, and the Wisconsin Longitudinal Study in the US, have included self-administered saliva sample collection posted back by respondents (Calderwood, Rose, Ring, & McArdle, 2014). Most comparable to our approach is the US Fragile Families and Wellbeing Study, which collected saliva samples in the homes at age 9, from cohort members and mothers, achieving compliance rates of 86% and 80% respectively.

Saliva is widely regarded as the preferred minimally-invasive approach to collecting samples to enable genotyping. Compared to the alternative methods for DNA collection, such as blood samples, its advantages include an ability to be collected by interviewers, rather than the clinically trained such as by a phlebotomist or nurse. When collected in an appropriate preservative samples can be stored at ambient temperatures for months, which both alleviates time pressures in terms of delivery to the laboratory and concerns about degradation which can be an issue following freeze/thaw cycles of blood. DNA purification is straightforward and the resulting material, compatible with major genotyping technologies, provide reliable results.

One drawback is that saliva samples collected and processed using methods and as described in this paper, using Oragene kits, are largely limited to (epi-)genomic assays, unlike biosamples such as blood which can provide other measures, such as metabolomics, clinical chemistry measures and proteomics. Another potential disadvantage of saliva collection include lower mean endogenous or host DNA yields because of inclusion of exogenous source material such as oral microbial DNA (Abraham et al., 2012; Bruinsma, Joo, Wong, Giles, & Southey, 2018; Gudiseva et al., 2016a). However previous studies have found that saliva samples provide sufficient DNA for genotyping (Gudiseva et al., 2016b) Indeed (Bruinsma et al., 2018) show that it is possible to obtain a higher quantity of DNA from saliva than whole blood samples of the same volume, consistent with findings reported by (Hansen, Simonsen, Nielsen, & Hundrup, 2007). Saliva has also been found to provide higher quality DNA than other non-invasive methods, such as buccal swabs (Rogers, Cole, Lan, Crossa, & Demerath, 2007).

A DNA bank has been created from the MCS saliva samples. A total of 23,336 samples are available, from 9,259 cohort members, 8,898 mothers and 5,179 fathers. There are 4,533 mother, child, father “trios”. The MCS is the only population-based, nationally representative study in the UK containing genetic trios.

3. Collection of saliva samples in the Millennium Cohort Study

3.1. Collection protocols

Prior to the age 14 fieldwork, interviewers attended a three-day training session covering all aspects of the upcoming survey including consent protocols and saliva collection. Interviewers were provided with detailed instructions on the collection of the samples in the home; packaging and return of samples to the laboratory; strict protocols were provided in order to reduce contamination from foreign DNA by bacteria, fungi and food remnants. This training was followed by accreditation, which was a requirement for interviewers to collect saliva samples in the field.

The Oragene® 500 DNA Self-Collection Kit made by DNA Genotek was used to collect saliva samples (http://www.dnagenotek.com/ROW/products/OG500.html). The Self-Collection Kit is a repository for the collection, preservation, transportation, and purification of DNA from saliva. The kits are routinely used to provide high-molecular-weight DNA and the manufacturer-stated median DNA yield from the kit is 110 ug from a 2ml saliva sample.

All cohort members were eligible to provide a saliva sample, along with their biological mother and father if resident in the household and available.

The consent process was as follows. Written consent was required from parents for their own samples, and for their child to provide a sample, and the 14-year old cohort members themselves had to provide verbal consent. Interviewers collected written consent using carbon-copy consent forms, with parents retaining a copy of the consent they had provided. Once consent forms were received in the office from interviewers, each form was checked to ensure valid consent had been provided. This entailed ensuring the parent had ticked or initialled the appropriate boxes as directed, and had signed the form. In cases where ticks/initials or a signature were missing, consent was deemed invalid and the corresponding saliva sample was destroyed. Consent could be withdrawn at any time, in writing, without providing a reason.

3.2. Transfer protocols

Samples were collected by interviewers from the cohort member and the biological mother and father and were sent to the laboratory by first class post after collection. Samples could be sent from a post office or using post boxes. Samples arrived daily at the Bristol Bioresource Laboratory, with time from sample collection to receipt ranging from a few days to months.

4. DNA Extraction

4.1. Processes

Saliva samples for DNA extraction were received in Bristol between January 2015 and May 2016. Samples were logged on arrival and stored at room temperature. Lists of samples logged were sent weekly to Ipsos MORI and lists of those with consent were returned from Ipsos MORI to the lab weekly. Final lists of samples with confirmed consent and those for disposal were received in October 2016.

4.2. DNA extraction and quantification

DNA was extracted from samples using an automated extraction robot (Tecan Freedom EVO-HSM Workstation using ReliaPrep™ Large Volume HT gDNA Isolation System).

Total DNA was quantified by fluorometic assay, using picogreen (Quant-iT^{TM}PicogreenTMdsDNA reagent (ThermoFisher Scientific)). Assays were performed using a Tecan Freedom Evo liquid handling robot and read on an Infinite F2000 Proreader.

Approximately 90% of samples provided yields of at least 20 µg, sufficient DNA for a range of genetic studies. As expected, for saliva samples, there were large variations in DNA yield. The yield from the child samples was, at 88%, lower than adults (91%), as has been seen in a number of studies e.g. (Gasso et al., 2014).

All samples were extracted well within the recommended storage time of five years for the Oragene kits, with the mean storage time of 304 days, minimum 87 days and maximum 654 days, (http://www.dnagenotek.com/US/pdf/PD-PR-012.pdf).

5. Response

In this section we show response rates for different study participants (cohort members, mothers, fathers), and document the main factors associated with response.

In Table 1 we show how many eligible participants provided a saliva sample, separately by cohort member, mother and father. We then assess how many produced a “useable” sample, i.e. one providing a DNA extraction yield greater than zero. Looking first at the cohort members (i.e. the 14-year olds), of those eligible, 82.65% provided a saliva sample, and DNA was extracted for 78.4% of cohort members. Similar rates are observed for mothers, where just over 83% of those eligible provided a saliva sample, resulting in useable samples for just over 79% of main respondents. Looking at fathers, we see that response rates were around 10 percentage points lower, at 72.1% of eligible, with just over 68% of useable samples received.

Table 1. Overall Response

	Cohort member	% of eligible	Mother	% of eligible	Father	% of eligible	Total
Productive (interviewed) at MCS6	11884		11726		11726		
Eligible for saliva sample collection (biological parent)	11806		11249		7544		30599
Saliva sample collected	9611	81.41%	9221	81.97%	5392	71.47%	24224
Invalid consent	251	2.13%	291	2.60%	191	2.51%	
Saliva sample successful (with DNA extracted)	9259	78.43%	8898	79.10%	5179	68.65%	23336

The vast majority of useable saliva samples are from singletons, with samples from 90 twin pairs, and 6 sets of triplets, as shown in Table 2.

Table 2. Response by singletons, twins and triplets

Successful sample -->	
Singleton	9061
Twins	180
Triplets	18
	9259

Notes: Amongst twins, in 7 cases a saliva sample was provided by one twin only. These are included in the singleton count.

Table 3 shows the distribution of sample volumes, separately for cohort members, mothers and fathers. As the table shows, most samples were of the correct volume. If taken correctly, the final volume should be 4ml: as can be seen from the table, just under 73% of samples had an estimated volume greater than 4ml, with mothers providing the highest quality samples; 7% of samples had an estimated volume of less than 3ml, suggesting that either no sample was given and the tube only contained preservative, or that the preservative solution was not added and only saliva was present. There were discoloured samples ranging from pale yellow to very dark brown and approximately 49% had some food contamination and 17% had marked food contamination. A complete analysis of quality of the data is presented in the next section.

Table 3. Sample volume

	Mother		Father		Cohort member		Total	
Volume:								
Less than 3ml	433	4.87%	520	10.04%	700	7.56%	1653	7.08%
3 - 3.9 ml	1583	17.79%	1123	21.68%	1970	21.28%	4676	20.04%
4 - 4.9 ml	6010	67.54%	3186	61.52%	5910	63.83%	15106	64.73%
5ml or more	872	9.80%	350	6.76%	679	7.33%	1901	8.15%
	8898		5179		9259		23336	

In Table 4, we show combinations of responses within the household. Trios of DNA samples are available for 4,533 households, and mother-child pairs are available for 3,913 households – together representing almost 90% of eligible households.

Table 4. Composition of samples

	N	%
Trio (M, F, CM)	4533	46.78
M, CM, no F	3913	40.38
F, CM, no M	378	3.9
M, F, no CM	186	1.92
M only	266	2.74
F only	82	0.85
CM only	333	3.44
Total	9691	100

Notes: Figures are at the household level; households with multiple cohort members are included once.

Finally, we analyse factors associated with response including: age in months, sex, ethnicity, household highest educational qualification, country. We run four separate models, predicting response for: cohort members, mothers, father, and trios. Estimates are shown in Table 5. Across the board, we see differences in response by education level (those with higher levels more likely to provide a sample) and ethnicity (ethnic minorities less likely to have provided a sample, particularly those from Black African or Black Caribbean backgrounds). There are no differences by education level in the provision of trios, but ethnic differences remain.

Table 5. Predictors of saliva sample

	Mother			Father			Cohort Member			Trio		
Country:(ref: England)												
Wales	-0.03	0.01	*	-0.02	0.02		-0.01	0.01		-0.03	0.02	*
Scotland	0.05	0.01	***	0.03	0.02		0.04	0.01	***	0.04	0.02	*
Northern Ireland	-0.02	0.01		-0.04	0.02		-0.01	0.01		-0.04	0.02	*
Highest parental education: (ref: No qualifications)												
Overseas only	0.04	0.02		-0.01	0.03		0.01	0.03		-0.02	0.05	
NVQ level 1	0.05	0.02	*	0.02	0.03		0.00	0.02		-0.08	0.04	
NVQ level 2	0.07	0.01	***	0.04	0.02		0.03	0.02		-0.01	0.03	
NVQ level 3	0.06	0.02	***	0.06	0.02	**	0.03	0.02		-0.01	0.03	
NVQ level 4	0.07	0.01	***	0.04	0.02		0.04	0.02	**	0.02	0.03	
NVQ level 5	0.06	0.02	***	0.09	0.02	***	0.04	0.02	*	0.04	0.03	
Ethnicity: parent or child (ref: white)												
Mixed	-0.13	0.04	**	-0.22	0.06	***	-0.05	0.02	**	-0.09	0.03	**
Indian	-0.05	0.02	*	-0.10	0.03	***	-0.05	0.03	*	-0.12	0.03	***
Pakistani	-0.11	0.02	***	-0.18	0.02	***	-0.09	0.02	***	-0.23	0.03	***
Bangladeshi	-0.15	0.03	***	-0.19	0.03	***	-0.13	0.03	***	-0.22	0.04	***
Black Caribbean	-0.35	0.04	***	-0.11	0.06		-0.19	0.04	***	-0.30	0.07	***
Black African	-0.24	0.03	***	-0.26	0.04	***	-0.19	0.03	***	-0.30	0.05	***
(inc Chinese, Other)	-0.03	0.03	***	-0.02	0.04		-0.04	0.03		-0.07	0.03	*
Age in months: parent or child	0.00	0.00		0.00	0.00	*	-0.04	0.01	***	-0.06	0.01	***
CM gender: (ref: male)	0.01	0.01		-0.01	0.01		0.00	0.01		-0.01	0.01	
Single: mother	-0.01	0.01										
CM number: (ref: Singleton)												
Twin							-0.09	0.03	**			
Triplet							0.06	0.15				
	11249			7544			11806			7195		

6. Genotyping

6.1. Laboratory procedures

21,432 DNA samples from 21,418 individuals were run on a total of 21,631 arrays (902 chips) (some DNA samples were repeated when quality checks failed). Infinium global screening arrays-24 v1.0 from Illumina, with 24 samples on each chip, were used following manufacturer's instructions. In brief, 200ng of DNA was hybridized to each array following the manufacturer's standard protocol. Samples (10,578) with a concentration below 40ng/ul were concentrated before use by lypholization followed by resuspension in appropriate volume of sterile water. Samples with insufficient DNA (n=1918) were excluded from genotyping.

Arrays were read on an Illumina iScan System (scanner ID N350), using FGPA version 4.0.20 and iScan Control Software version 3.4.8. Hybridization/fluorescence signals were written to idat files. Data for a total of 21,556 arrays derived from 21,368 individuals passed iScan quality control and were passed for genotype calling.

Genotyping Performed	arrays	biological samples	individuals
placed on an array	21631	21432	21418
array data available	21556	21373	21368
passed genotype QC	21349	21197	21192

6.2. Genotype calling

Genotype calling for all 21,556 arrays was performed using the Genome Studio v2.0.4 graphical user interface, on a single computer running Windows 7, in a single batch. Data for 618,540 variants was written to a final manifest file and plink ped format using the plink export module, and subsequently converted to a binary ped or bed file using plink2 (Purcell et al., 2007). Genotype calls were then quality controlled (QC), for samples and single nucleotide polymorphism (SNPs) using summary estimates generated by QCtools_v2.0.1 (https://www.well.ox.ac.uk/~gav/qctool/), plink, and bespoke R scripts, as follows. First, individuals were jointly identified and subsequently excluded for having a missingness proportion greater than 20% (n = 199), and/or an estimated heterozygosity greater than or less than five standard deviations from the population mean heterozygosity estimate (n = 120). As there was overlap between these groups, this led to 207 individuals in total being excluded. Second, 51 SNPs were excluded for missingness greater than 20%, followed by the exclusion of 1,473 SNPs for genomic mapping duplicity, totalling 1,524 SNP exclusions. Third we estimated a set of unrelated individuals using plink2's greedy, Ajk relatedness estimator using default parameters of the function --rel-cutoff. This identified 11,176 individuals unrelated at roughly the 5th degree (0.025). No exclusions were made with this data, but this list of unrelated individuals is used below. Fourth, we merged the MCS cohort data set with the 1000 Genomes phase three data, which is derived from 26 global populations, providing us with a data set of 278,052 shared and strand matched SNPs. With this temporary, combined data set we estimated principal components using only the data from the 1000 Genomes data and subsequently projected the MCS data onto it. Using the eigenvectors of principle components (PC) one and two we identified the two-dimensional boundary in which the 1000 Genomes GBR (Great Britain) population occupied and then extracted the sample identifiers for all 9,095 unrelated MCS individuals, as identified in step three, that fell within that space. We note that including relatives, 14,657 MCS individuals did fall within this GBR population space, and that no MCS individuals were identified as extreme outliers on the first five principal components as defined using the 1000 Genomes phase 3 data set. Fifth, using these 9,095, putatively unrelated individuals with strong 1000 Genomes GBR population PC1 and PC2 mapping association, we estimated the Hardy-Weinberg (HW) statistic. Steps three and four of this QC protocol were carried out to comply, as best we can, with the randomly mating, non-structured population assumption of the HW principle. The other assumptions of the HW principle that may influence this data - no overlapping generations, no mutation, no selection, equal distribution of alleles among the sexes - are being ignored. Using a p-value of 2.5×10^{-8} 16,371 SNPs were excluded for exhibiting strong deviations from HW equilibrium. A total of 602,181 SNPs and 21,349 arrays, derived from 21,192 individuals passed these quality control measures.

Genotype Data Available	parent	child	
female	8212	4072	
male	4803	4101	8904
relationship totals	13015	8173	**21188**

6.3. Imputation

To prepare for imputation, we followed the instructions on the Michigan Imputation Server (MIS) website (imputationserver.sph.umich.edu) (Das et al., 2016), and using our quality controlled data set we converted the data to VCF format, split by chromosome and sorted by chromosome and position, using plink2.0 and tabix function of VCFtools (Danecek et al., 2011). We subsequently validated the mapping and strand allocation of our variants using prepared scripts (HRC-1000G-check-bim-v4.2.11.zip) from Will Rayner and the McCarthy group (https://www.well.ox.ac.uk/~wrayner/tools/), as instructed by MIS. Prepared data was submitted to the MIS, phased with Eagle.v2.4 (Loh et al., 2016) and imputed to Haplotype Reference Consortium release 1.1 (HRC r1.1; http://www.haplotype-reference-consortium.org)(McCarthy et al., 2016) using Minimac.v4 (Fuchsberger, Abecasis, & Hinds, 2015; Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). Imputation chunk chr14:1-20Gb, corresponding to the short arm of chromosome 14 - a gene desert, failed to impute given the paucity of genotyped data in this region.

7. Individual Ancestry

Genetic ancestry at the continental level was estimated for each MCS individual using the MCS data projected upon the 1000 Genomes data set as described in section 6.2. Principle components one through four and a k (number of clusters) from 2 to 20 were used to identify clusters and estimate the proportion of total variation explained among clusters. At a k of five the proportion of variation explained by clusters plateaus, explaining 92.9% of the total variation (Figure 1A). The five k clusters are largely consistent with the five super-populations identified by the 1000 Genomes project Africa (AFR), America (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS) providing a criteria to assign MCS individuals to a continental ancestry (Figure 1B; https://www.internationalgenome.org). We note that these continental assignments are limited by both the data used (SNPs and 1000 Genomes populations) and the methodology, and are solely intended to provide an overview of the global genetic diversity of this cohort. Given the data 86.75% of individuals are of European ancestry, 10.46% are of South Asian ancestry, 2.37% are of African ancestry, 0.34% is of East Asian ancestry, and 0.07% is of American ancestry (Figure 1B).

The self-described ethnic group of 7822 children aged 13 to 15 was compared to assigned continental clusters to compare and contrast individual cultural ethnicity with continental genetic ancestry. During the completion of questionnaires individuals were free to describe the ethnic group they belonged to as, amongst others, Bangladeshi, black African, black Caribbean, Indian, mixed, other, Pakistani, white, unknown or free to refuse to answer the question. We would anticipate a non-random overlap of these two categorical traits, but one does not necessitate or dictate the other. We do observe a non-random association between these two variables (hypergeometric test, Monte Carlo simulated p-value = $3.9x10^{-6}$), but there are notable disagreements (Figure 2). For example, 26.9% of black Caribbeans and 16.0% of black Africans are assigned to the EUR genotypic cluster. These observations do not negate an individual's self-described ethnicity, nor do these limited analyses quantify continental genetic ancestry. However, it does exemplify the ethnic and genetic diversity of this data set and highlights the needed careful consideration of both parameters when building analytical models of these data.

8. Accessing the Data

The genetic data will soon be available for access for research, both on its own and alongside the rich phenotype data collected in the MCS. The initial rounds of application will be via METADAC, details here: https://cls.ucl.ac.uk/data-access-training/access-cls-dac/

The application process for subsequent rounds will be detailed on the above web link.

Acknowledgements

We acknowledge core-funding for the sixth sweep of MCS from the Economic and Social Research Council, grant number ES/K005987/1, and co-funding from the following consortium of government departments: Department for Education, Department of Health, Ministry of Justice, Home Office, Department for Transport, Department of Work and Pensions, Welsh Government and Department for Employment and Learning (Northern Ireland).

We are grateful for the voluntary cooperation of the MCS cohort members and their families in providing information to the study.

DNA extraction was carried out in the Bristol Bioresource Laboratories, University of Bristol and genotyped in the Illumina Array Facility, University of Bristol.

9. References

Calderwood L., Rose N., Ring, S. & McArdle W.(2014). Collecting saliva samples for DNA extraction from children and parents: findings from a pilot study using lay interviewers in the UK, *Survey Methods: Insights from the Field.* Retrieved from http://surveyinsights.org/?p=3723

Abraham, J. E., Maranian, M. J., Spiteri, I., Russell, R., Ingle, S., Luccarini, C., . . . Caldas, C. (2012). Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *Bmc Medical Genomics, 5*, 19. doi:10.1186/1755-8794-5-19

Bruinsma, F. J., Joo, J. E., Wong, E. M., Giles, G. G., & Southey, M. C. (2018). The utility of DNA extracted from saliva for genome-wide molecular research platforms. *BMC Res Notes, 11*(1), 8. doi:10.1186/s13104-017-3110-y

Calderwood, L., Rose, N., Ring, S., & McArdle, W. (2014). Collecting saliva samples for DNA extraction from children and parents: findings from a pilot study using lay interviewers in the UK. *Survey Methods: Insights from the Field*. Retrieved from Retrieved from https://surveyinsights.org/?p=3723.

Clemens, S., Given, L., & Purdon, S. (2012). Methods of collecting biological data: Considerations, challenges and implications. *Presentation to the 67th Annual Meeting of the American Association of Public Opinion Research in Orlando, Florida.*

Connelly, R., & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *Int J Epidemiol, 43*(6), 1719-1725. doi:10.1093/ije/dyu001

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet, 48*(10), 1284-1287. doi:10.1038/ng.3656

Fitzsimons, E. e. a. (2017). Millennium Cohort Study, Sixth Survey 2015-2016, User Guide (First Edition).

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics, 31*(5), 782-784. doi:10.1093/bioinformatics/btu704

Gasso, P., Pagerols, M., Flamarique, I., Castro-Fornieles, J., Rodriguez, N., Mas, S., . . . Stop, C. (2014). The effect of age on DNA concentration from whole saliva: implications for the standard isolation method. *Am J Hum Biol, 26*(6), 859-862. doi:10.1002/ajhb.22593

Gudiseva, H. V., Hansen, M., Gutierrez, L., Collins, D. W., He, J., Verkuil, L. D., . . . O'Brien, J. M. (2016a). Saliva DNA quality and genotyping efficiency in a predominantly elderly population. *Bmc Medical Genomics, 9*. doi:ARTN 17 10.1186/s12920-016-0172-y

Gudiseva, H. V., Hansen, M., Gutierrez, L., Collins, D. W., He, J., Verkuil, L. D., . . . O'Brien, J. M. (2016b). Saliva DNA quality and genotyping efficiency in a predominantly elderly population. *Bmc Medical Genomics, 9*, 17. doi:10.1186/s12920-016-0172-y

Hansen, T. V., Simonsen, M. K., Nielsen, F. C., & Hundrup, Y. A. (2007). Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: comparison of the response rate and quality of genomic DNA. *Cancer Epidemiol Biomarkers Prev, 16*(10), 2072-2076. doi:10.1158/1055-9965.EPI-07-0611

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi:10.1038/ng.2354

Joshi, H., & Fitzsimons, E. (2016). The UK Millennium Cohort Study: the making of a multi-purpose resource for social science and policy in the UK. *Longitudinal and Life Course Studies, 7*(4), 409-430. doi:10.14301/llcs.v7i4.416

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Y, A. R., H, K. F., . . . A, L. P. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet, 48*(11), 1443-1448. doi:10.1038/ng.3679

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . Consortium, H. R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics, 48*(10), 1279-1283. doi:10.1038/ng.3643

McFall, S. L., Conolly, A., & Burton, J. (2014). Collecting Biomarkers and Biological Samples Using Trained Interviewers. Lessons from a Pilot Study. *Survey Research Methods, 8*(1), 57-66. Retrieved from <Go to ISI>://WOS:000334031700005

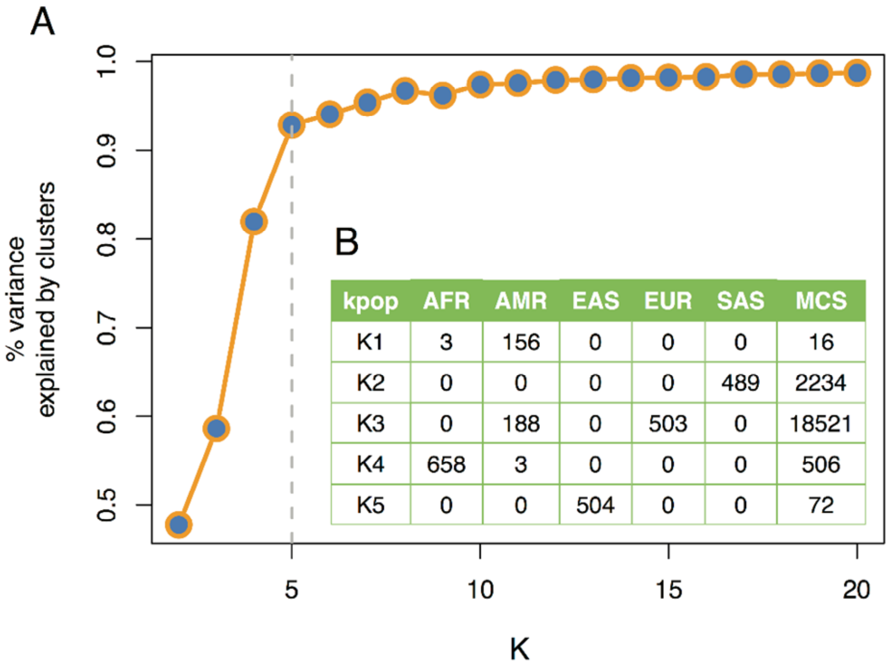
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Rogers, N. L., Cole, S. A., Lan, H. C., Crossa, A., & Demerath, E. W. (2007). New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. *Am J Hum Biol, 19*(3), 319-326. doi:10.1002/ajhb.20586

Sun, F., & Reichenberger, E. J. (2014). Saliva as a source of genomic DNA for genetic studies: review of current methods and applications. *Oral Health Dent Manag, 13*(2), 217-222. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24984625

FIGURES

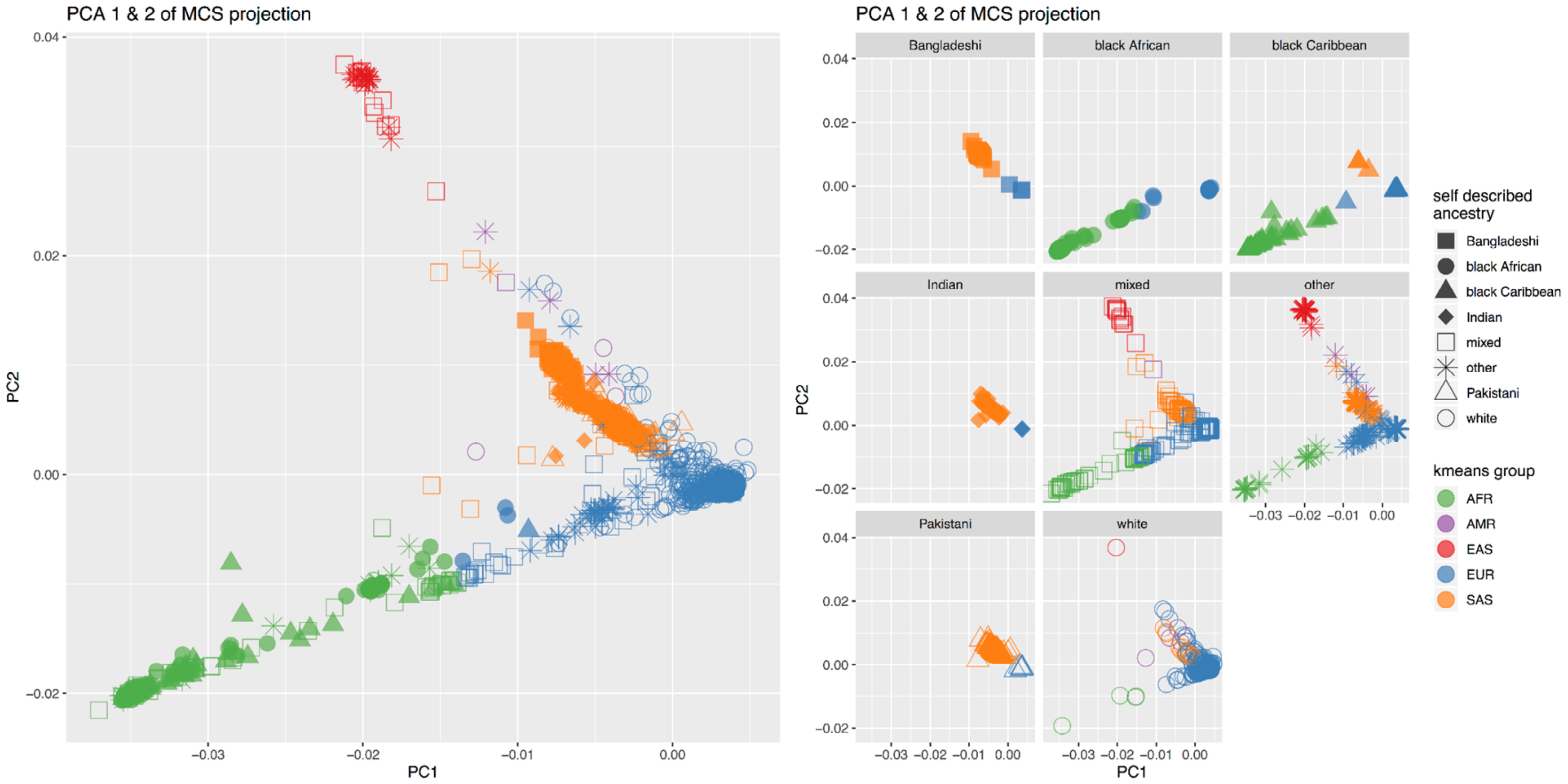
Figure 1. K-means clustering variance explained by clusters



kpop	AFR	AMR	EAS	EUR	SAS	MCS
K1	3	156	0	0	0	16
K2	0	0	0	0	489	2234
K3	0	188	0	503	0	18521
K4	658	3	0	0	0	506
K5	0	0	504	0	0	72

Legend: K-means clustering results. (A) Proportion of total variance explained by groups. X-axis defines the number of k clusters; y-axis defines the proportion of total variance in principle components one through four explained between clusters. (B) A table of the K5 clusters and the grouping of individuals from the 1000 Genomes individual pre-defined super-populations (Africa (AFR), America (AMR), East Asia (EAS), Europe (EUR), South Asia (SAS)) and the Millennium Cohort study samples (MCS).

Figure 2. K-means clustering variance explained by clusters



Legend: MCS young people principle components one and two. Shapes define individual self-described ethnicity; the five colours define k-means continental ancestry. The eight self-described ethnicities are further plotted individually to aid visualization.