

## STUDY PROFILE

### Collection of genetic data at scale for a nationally representative population: the UK Millennium Cohort Study

Emla Fitzsimons, [e.fitzsimons@ucl.ac.uk](mailto:e.fitzsimons@ucl.ac.uk)  
University College London, UK and  
Institute for Fiscal Studies, London, UK

Vanessa Moulton, [vanessa.moulton@ucl.ac.uk](mailto:vanessa.moulton@ucl.ac.uk)  
University College London, UK

David A Hughes, [d.a.hughes@bristol.ac.uk](mailto:d.a.hughes@bristol.ac.uk)  
Sam Neaves, [s.neaves@bristol.ac.uk](mailto:s.neaves@bristol.ac.uk)  
Karen Ho, [Karen.Ho@bristol.ac.uk](mailto:Karen.Ho@bristol.ac.uk)  
Gibran Hemani, [g.hemani@bristol.ac.uk](mailto:g.hemani@bristol.ac.uk)  
Nicholas Timpson, [N.J.Timpson@bristol.ac.uk](mailto:N.J.Timpson@bristol.ac.uk)  
University of Bristol, UK

Lisa Calderwood, [l.caldenwood@ucl.ac.uk](mailto:l.caldenwood@ucl.ac.uk)  
Emily Gilbert, [emily.gilbert@ucl.ac.uk](mailto:emily.gilbert@ucl.ac.uk)  
University College London, UK

Susan Ring, [S.M.Ring@bristol.ac.uk](mailto:S.M.Ring@bristol.ac.uk)  
University of Bristol, UK

A DNA bank has been created from the Millennium Cohort Study (MCS) saliva samples. A total of 23,336 samples are available, from 9,259 cohort members (4,630 males and 4,629 females), 8,898 mothers and 5,179 fathers. There are 4,533 mother, child, father ‘triads’. This paper describes the collection of the saliva samples from cohort members and their biological parents in the MCS. It analyses response rates and predictors of response, and details the DNA extraction, genotyping and imputation procedures performed on the data.

**Key words** genetic data • Millennium Cohort Study • longitudinal • ethnic diversity • nationally representative population

#### Key messages

- It describes the collection of saliva samples in the Millennium Cohort Study, from cohort members and their resident biological parents.
- It analyses response rates and predictors of response.

- It details the DNA extraction, genotyping and imputation procedures performed on the data.

To cite this article: Fitzsimons, E., Moulton, V., Hughes, D., Neaves, S., Ho, K., Hemani, G., Timpson, N., Calderwood, L., Gilbert, E. and Ring, S. (2021) Collection of genetic data at scale for a nationally representative population: the UK Millennium Cohort Study, *Longitudinal and Life Course Studies*, vol XX, no XX, 1–19, DOI: 10.1332/175795921X16223668101602

---

## **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 and Platt, 2014; Joshi and 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, well-being 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 (that is, those who are part of the cohort born at the turn of the millennium) and their resident biological parents, that is, parents who were living with them at the time of interview. 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 of genome-wide association study (GWAS) 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 (Connelly and Platt, 2014; Joshi and Fitzsimons, 2016), also enable studies of the longitudinal or fine-scale phenotypic associations of already-discovered genetic associations from larger cohorts. Its diverse population, including continental and complex ancestries, increases representation of certain groups and increases opportunities to understand how genotype influences phenotype. The MCS is unique in being the only population-based, nationally representative study in the UK containing genetic triads.

The paper proceeds as follows: background; an overview of the fieldwork collection; a discussion of the DNA extraction process; response rates; genotyping, quality control and imputation; illustration of the resource's self-described ethnic diversity and continental genetic ancestry; and accessing the data.

## Background

The sixth, or age 14, wave of the MCS took place across the UK from January 2015 to March 2016. Interviews were conducted with 11,726 families, of which there were 11,872 cohort members (Fitzsimons, 2017), representing a response rate of 76.3% (of the eligible sample at age 14). This was a particularly extensive and innovative wave, containing several elements including interviews and self-completion questionnaires with main parent (mostly the mother) and the main parent's resident partner (where applicable), self-completion with cohort members, physical measurements, cognitive assessments, saliva samples, accelerometer collection and time diaries for cohort members to record how they were spending their time. 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 and Reichenberger, 2014). While lay interviewers are increasingly involved in the collection of biomedical measures using non-invasive methods (McFall et al, 2014), our study represented a significant departure from previous longitudinal studies that are tagged on to a clinic visit (for example, the Avon Longitudinal Study of Parents and Children, UK Biobank), or taken by nurse interviewers as part of a home visit (for example, the Health Survey for England, the 1958 National Child Development Study, the 1970 British Cohort Study, and 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 et al, 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. Our approach had been piloted on a small scale a few years previously (Calderwood et al, 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 85% and 78% respectively (Fragile Families, 2019).

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 manner, preservative samples can be stored at ambient temperatures for up to 30 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 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; Gudiseva et al, 2016; Bruinsma et al, 2018). However previous studies have found that saliva samples provide sufficient DNA for genotyping (Gudiseva et al, 2016). 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 et al, 2007). Saliva has also been found to provide higher quality DNA than other non-invasive methods, such as buccal swabs (Rogers et al, 2007).

## **Collection of saliva samples in the Millennium Cohort Study**

### *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. In addition, alongside the in-person training, interviewers were provided with detailed written instructions on the collection of the samples in the home (see Appendix); 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. Interviewers were required to carry out two practice sessions of saliva sample collection, before obtaining accreditation at the end of the training (Ipsos MORI, 2017).

The Oragene® 500 DNA Self-Collection Kit made by DNA Genotek was used to collect saliva samples.<sup>1</sup> The Self-Collection Kit is a repository for the collection, preservation, transportation and purification of DNA from saliva. Benefits of the kit include the fact that they assure no sample degradation, even when stored at room temperature for up to 30 months, and the manufacturer-stated median DNA yield from the kit is 110 µg from a 2 ml saliva sample (Birnboim, 2004).

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.

### *Transfer protocols*

Samples were collected by interviewers from the cohort member and the biological mother and father and were packaged in accordance with the transportation of biological substances regulations, and were sent to the laboratory at the University of Bristol by first-class post. Samples could be sent from a post office or using post

boxes. Samples arrived daily at the Bristol Bioresource Laboratory. Further details are available in section 4.5 of the MCS6 Technical Report on Fieldwork ([Ipsos MORI, 2017](#)).

## DNA extraction

### *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, the agency conducting the fieldwork. This was in order for Ipsos MORI to confirm that the appropriate signed consent had been obtained to collect saliva samples. Lists of those with confirmed 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.

### *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 fluorometric assay, using picogreen (Quant-iT™ Picogreen™ dsDNA 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 (for example, [Gassó et al, 2014](#); [Nishita et al, 2009](#)).

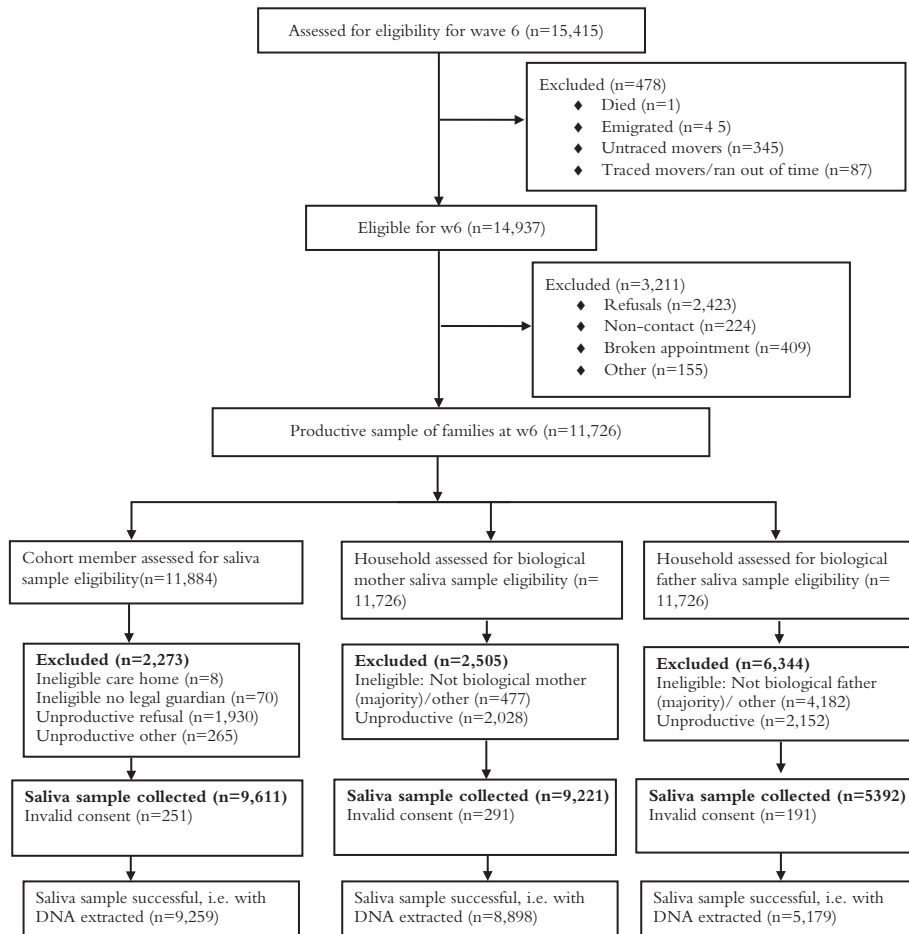
All samples were extracted well within the recommended storage time of five years for the Oragene kits, with the mean storage time of 304 days, minimum 87 days and maximum 654 days.<sup>2</sup>

## Response

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

[Figure 1](#) presents the numbers of eligible participants providing a saliva sample, separately by cohort member, mother and father. It also shows how many produced a 'usable' sample, that is, one providing a DNA extraction yield greater than zero. Looking first at the cohort members (the 14-year-olds), of those eligible, 81.4% provided a saliva sample, and DNA was extracted for 78.4% of eligible cohort members. Similar rates are observed for mothers, where 82% of those eligible provided a saliva sample, resulting in usable samples for 79.1% of eligible main respondents. Looking at fathers, we see that response rates were around 10 percentage points lower than mothers, at 71.5% of eligible (n = 7,544 biological fathers in the household), with 68.7% of usable samples received. A DNA bank has been created from the MCS

**Figure 1: CONSORT 2010 Flow Diagram<sup>3</sup>**



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 ‘triads’.

The vast majority of usable saliva samples are from singletons, with samples from 90 twin pairs, and six sets of triplets, as shown in [Table 1](#).

[Table 2](#) 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 4 ml: as can be seen from the table, just under 73% of samples had an estimated volume greater than 4 ml, with mothers providing the highest quality samples; 7% of samples had an estimated volume of less than 3 ml, 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 (based on visual inspection). However this had no impact on the quality of DNA extracted. A complete analysis of quality of the data is presented in the next section, ‘Genotyping’.

**Table 1:** Response by singletons, twins and triplets

Successful sample	
Singleton	9,061
Twins	180
Triplets	18
	9,259

*Note:* Among twins, in seven cases a saliva sample was provided by one twin only. These are included in the singleton count.

**Table 2:** Sample volume

	Cohort member		Mother		Father		Total	
<b>Volume:</b>								
Less than 3 ml	700	7.56%	433	4.87%	520	10.04%	1,653	7.08%
3–3.9 ml	1,970	21.28%	1,583	17.79%	1,123	21.68%	4,676	20.04%
4–4.9 ml	5,910	63.83%	6,010	67.54%	3,186	61.52%	15,106	64.73%
5 ml or more	679	7.33%	872	9.80%	350	6.76%	1,901	8.15%
	9,259		8,898		5,179		23,336	

**Table 3:** Composition of samples

	N	%
Triad (M, F, CM)	4,533	46.77
M, CM, no F	3,913	40.38
F, CM, no M	378	3.90
M, F, no CM	186	1.92
M only	266	2.74
F only	82	0.85
CM only	333	3.44
Total	9,691	100

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

In [Table 3](#), we show combinations of responses within the household. Triads of DNA samples are available for 4,533 households, and mother–child pairs are available for a further 3,913 households – together representing almost 90% of households who gave any saliva sample.

Finally, we analyse factors associated with response including: age in months and ethnicity of participants, sex of cohort member, household highest educational qualification, and country. We run four separate linear probability models, predicting response for: cohort members, mothers, fathers and triads. Estimates are shown in [Table 4](#). 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 triads, but ethnic differences remain.

**Table 4:** Correlates of a valid saliva sample, with DNA extracted: by respondent type

	Cohort Member		Mother		Father		Triad	
	b(SE)	95% CIs	b(SE)	95% CIs	b(SE)	95% CIs	b(SE)	95% CIs
<i>Country: (ref: England)</i>								
Wales	-0.01 (.01)	[-0.03, 0.02]	-0.03 (.01)*	[-0.05, 0.00]	-0.02 (.02)	[-0.06, 0.01]	-0.03 (.02)*	[-0.07, 0.00]
Scotland	0.04 (.01)***	[0.02, 0.07]	0.05 (.01)***	[0.02, 0.07]	0.03 (.02)	[-0.00, 0.07]	0.04 (.02)*	[-0.00, 0.07]
Northern Ireland	-0.01 (.01)	[-0.03, 0.02]	-0.02 (.01)	[-0.04, 0.01]	-0.04 (.02)	[-0.07, 0.00]	-0.04 (.02)*	[-0.08, -0.00]
<i>Highest parental education: (ref: No qualifications)</i>								
Overseas only	0.01 (.03)	[-0.05, 0.06]	0.04 (.02)	[-0.00, 0.09]	-0.01 (.03)	[-0.08, 0.05]	-0.02 (.05)	[-0.12, 0.07]
NVQ level 1 (Vocational / GCSE grades D-G)	0.00 (.02)	[-0.04, 0.05]	0.05 (.02)*	[0.01, 0.09]	0.02 (.03)	[-0.03, 0.08]	-0.08 (.04)	[-0.17, -0.00]
NVQ level 2 (GCSE grades A*-C)	0.03 (.02)	[-0.01, 0.06]	0.07 (.01)***	[0.04, 0.10]	0.04 (.02)	[-0.00, 0.08]	-0.01 (.03)	[-0.08, 0.05]
NVQ level 3 (A levels)	0.03 (.02)	[0.01, 0.06]	0.06 (.02)***	[0.03, 0.09]	0.06 (.02)**	[0.02, 0.11]	-0.01 (.03)	[-0.08, 0.05]
NVQ level 4 (degree)	0.04 (.02)**	[0.01, 0.08]	0.07 (.01)***	[0.04, 0.09]	0.04 (.02)	[-0.00, 0.08]	0.02 (.03)	[-0.05, 0.08]
NVQ level 5 (postgraduate)	0.04 (.02)*	[0.01, 0.08]	0.06 (.02)***	[0.03, 0.10]	0.09 (.02)***	[0.04, 0.13]	0.04 (.03)	[-0.02, 0.10]
<i>Ethnicity: parent or child (ref: white)</i>								
Mixed	-0.05 (.02)**	[-0.09, -0.01]	-0.13 (.04)**	[-0.21, -0.05]	-0.22 (.06)***	[-0.34, -0.10]	-0.09 (.03)**	[-0.15, -0.03]
Indian	-0.05 (.03)*	[-0.10, -0.00]	-0.05 (.02)*	[-0.10, -0.00]	-0.10 (.03)***	[-0.16, 0.05]	0.12 (.03)***	[-0.18, -0.06]
Pakistani	-0.09 (.02)***	[-0.13, -0.06]	-0.11 (.02)***	[-0.15, -0.07]	-0.18 (.02)***	[-0.22, -0.13]	-0.23 (.03)***	[-0.29, -0.19]
Bangladeshi	-0.13 (.03)***	[-0.18, -0.07]	-0.15 (.03)***	[-0.20, -0.10]	-0.19 (.03)***	[-0.25, -0.12]	-0.22 (.04)***	[-0.29, -0.16]
Black Caribbean	-0.19 (.04)***	[-0.26, -0.11]	-0.35 (.04)***	[-0.41, -0.28]	-0.11 (.06)	[-0.23, 0.01]	-0.30 (.07)***	[-0.45, -0.16]
Black African	-0.19 (.04)***	[-0.24, -0.13]	-0.24 (.03)***	[-0.29, -0.18]	-0.26 (.04)***	[-0.35, -0.18]	-0.30 (.05)***	[-0.40, -0.21]
(including Chinese, Other)	-0.04 (.03)	[-0.09, 0.01]	-0.03 (.03)***	[-0.08, 0.03]	-0.02 (.04)	[-0.09, 0.05]	-0.07 (.03)*	[-0.14, -0.00]

(Continued)



Table 4: (Continued)

	Cohort Member		Mother		Father		Triad	
	b(SE)	95% CIs	b(SE)	95% CIs	b(SE)	95% CIs	b(SE)	95% CIs
<i>Age in years: parent or child (ref: 13)</i>								
Age 14	-0.04 (.01)***	[-0.06, -0.02]	0.00 (.00)	[-0.00, 0.00]	0.00 (.00)*	[-0.00, 0.00]	-0.05 (.01)***	[-0.08, -0.03]
Age 15	-0.11 (.03)**	[-0.18, -0.05]					-0.20 (.05)	[-0.30, -0.10]
CM sex: (ref: male)	0.00 (.01)	[-0.01, 0.01]	0.01 (.01)	[-0.00, 0.03]	-0.01 (.01)	[-0.03, 0.01]	-0.01	[-0.03, 0.01]
Single: mother			-0.01 (.01)	[-0.02, 0.01]				
<i>CM number: (ref: Singleton)</i>								
Twin	-0.09 (.03)**	[-0.16, -0.02]						
Triplet	0.06 (.15)	[-0.24, 0.37]						
N:	11,806		11,249		7,544		7,195	

Notes:

(b) Unstandardised coefficients from linear probability model.

(SE) standard error.

p-value where \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$  and 95% Confidence Interval (CIs).

NVQ National Vocational Qualification and academic equivalents.

## Genotyping

### *Laboratory procedures*

A total of 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), as shown in the first row of [Box 1](#). Infinium global screening arrays-24 v1.0 from Illumina, with 24 samples on each chip, were used following manufacturer's instructions. In brief, 200 ng of DNA was hybridised to each array following the manufacturer's standard protocol. Samples (10,578) with a concentration below 40 ng/ul were concentrated before use by lypholisation followed by resuspension in appropriate volume of sterile water. Samples with insufficient DNA ( $n = 1,918$ ) 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. Hybridisation/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, as shown in the second row of [Box 1](#).

---

### **Box 1: Genotyping statistics**

<b><i>Genotyping performed</i></b>	<b>Arrays</b>	<b>Biological samples</b>	<b>Individuals</b>
Placed on an array	21,631	21,432	21,418
Array data available	21,556	21,373	21,368
Passed genotype quality control	21,349	21,197	21,192

---

### *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,<sup>4</sup> 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-cut-off. This identified 11,176 individuals unrelated at roughly the fifth degree (0.025). No exclusions were made with this data, but this list

of unrelated individuals is used later. 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 principal components (PC) one and two we identified the two-dimensional boundary in which the 1000 Genomes GBR (British in England and Scotland) 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 PC 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  $2.5 \times 10^{-8}$ , 16,371 SNPs were excluded for exhibiting strong deviations from HW equilibrium. No exclusions have been made regarding mismatches between the self-reported and genotype predicted sex. A total of 602,181 SNPs and 21,349 arrays, derived from 21,192 individuals passed these QC measures. A breakdown of how many arrays, samples and individuals that went through genotyping is shown in [Box 1](#). In addition, the number of males, females, parents and children in the cohort with genotype data is provided in [Box 2](#).

---

### Box 2: Genotype data availability

<i>Genotype data available</i>	Parent	Child	Step-parent	Sex totals
Female	8,212	4,072	0	12,284
Male	4,803	4,101	4	8,908
Relationship totals	13,015	8,173	4	<b>21,192</b>

---

### *Imputation*

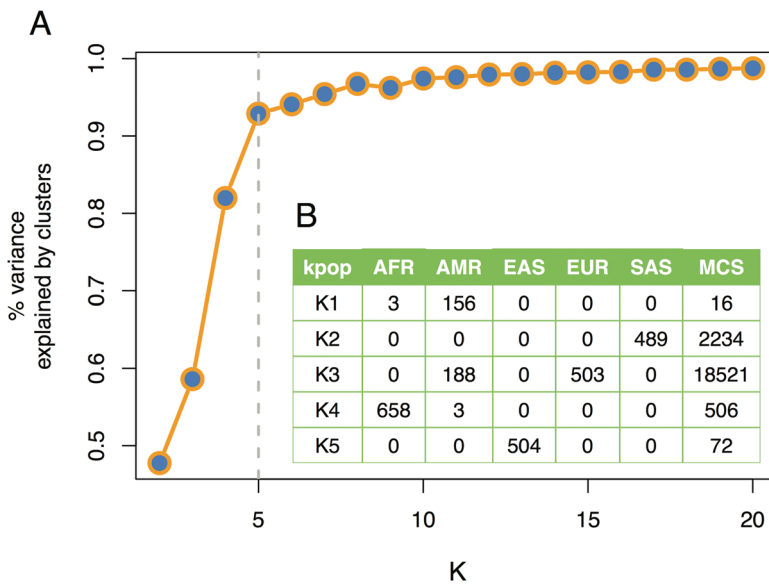
To prepare for imputation, we followed the instructions on the Michigan Imputation Server (MIS) website<sup>5</sup> ([Das et al, 2016](#)), and using our QC data set we converted the data to VCF format, split by chromosome and sorted by chromosome and position, using plink2.0 and tabix function of VCFtools ([Danecek et al, 2011](#)). We subsequently validated the mapping and strand allocation of our variants using prepared scripts<sup>6</sup> from Will Rayner and the McCarthy group,<sup>7</sup> 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<sup>8</sup> (HRC r1.1; [Haplotype Reference](#)

Consortium, 2016) using Minimac.v4 (Howie et al, 2012; Fuchsberger et al, 2015). 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.

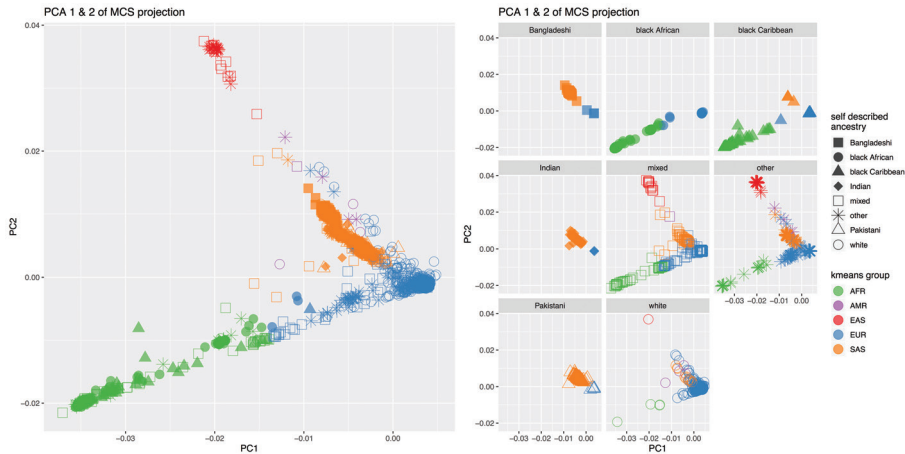
## Individual ancestry

Genetic ancestry at the continental level was estimated for each MCS individual using the MCS data projected on the 1000 Genomes data set as described in the earlier subsection called ‘Genotype Calling’. These analyses use principal components that reduce the genotype data into orthogonal eigenvectors (principal components or PCs) describing how the data covaries. The first PC explains the largest proportion of variation present in the data, and each subsequent one explains a smaller and smaller proportion of the total variation. PC 1–4 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 2A). 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 2B).<sup>9</sup> We note that these continental assignments are limited by both the data used (SNPs and 1000 Genomes populations), the assumption of a single ancestry

**Figure 2:** K-means clustering variance explained by clusters



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 principal components 1–4 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 3:** K-means clustering variance explained by clusters

Legend: MCS young people principal components 1 and 2. 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 visualisation.

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 2B).

The self-described ethnic group of 7,822 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, among 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 =  $3.9 \times 10^{-6}$ ), but there are notable disagreements (Figure 3). 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.

## Accessing the data

The genetic data are available for access for research, both on their own and alongside the rich phenotype data collected in the MCS. Application is via the Data Access Committee of the Centre for Longitudinal Studies, details at: <https://cls.ucl.ac.uk/data-access-training/genetic-data-and-biological-samples/>

## Notes

<sup>1</sup> <http://www.dnagenotek.com/ROW/products/OG500.html>.

<sup>2</sup> <http://www.dnagenotek.com/US/pdf/PD-PR-012.pdf>.

<sup>3</sup> For details of issued and achieved samples at previous MCS waves, see Table 1 of Connelly and Platt (2014).

<sup>4</sup> <https://www.well.ox.ac.uk/~gav/qctool/>.

<sup>5</sup> <https://imputationserver.sph.umich.edu>.

<sup>6</sup> <https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.2.11.zip>.

<sup>7</sup> <https://www.well.ox.ac.uk/~wrayner/tools/>.

<sup>8</sup> <http://www.haplotype-reference-consortium.org>.

<sup>9</sup> <https://www.internationalgenome.org>.

## 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.

## Conflict of interest

The authors declare that there is no conflict of interest.

## References

- Abraham, J.E., Maranian, M.J., Spiteri, I., Russell, R., Ingle, S., Luccarini, C., Earl, H.M., Pharoah, P.P.D., Dunning, A.M. and 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(1): 1–6, <https://doi.org/10.1186/1755-8794-5-19>. doi: 10.1186/1755-8794-5-19
- Birnboim, H. (2004) *DNA Stability with Oragene<sup>TM</sup>*, Ottawa: DNA Genotek, Inc.
- Bruinsma, F.J., Joo, J.E., Wong, E.M., Giles, G.G. and Southey, M.C. (2018) The utility of DNA extracted from saliva for genome-wide molecular research platforms, *BMC Research Notes*, 11(1): 1–6, <https://doi.org/10.1186/s13104-017-3110-y>. doi: 10.1186/s13104-017-3110-y
- Calderwood, L., Rose, N., Ring, S. and 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*, <http://surveyinsights.org/?p=3723>.
- Clemens, S., Given, L. and 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*, [http://www.aapor.org/AAPOR\\_Main/media/AnnualMeetingProceedings/2012/01\\_-SAM-CLEMENS\\_B2\\_slides.pdf](http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2012/01_-SAM-CLEMENS_B2_slides.pdf).

- Connelly, R. and Platt, L. (2014) Cohort profile: UK Millennium Cohort Study (MCS), *International Journal of Epidemiology*, 43(6): 1719–25. doi: [10.1093/ije/dyu001](https://doi.org/10.1093/ije/dyu001)
- Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. et al. and 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools, *Bioinformatics*, 27(15): 2156–8. doi: [10.1093/bioinformatics/btr330](https://doi.org/10.1093/bioinformatics/btr330)
- Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M. et al. (2016) Next-generation genotype imputation service and methods, *Nature Genetics*, 48(10): 1284–7. doi: [10.1038/ng.3656](https://doi.org/10.1038/ng.3656)
- Fitzsimons, E. (2017) Millennium cohort study, sixth survey 2015–2016, User Guide, 1st edn, London: UCL Institute of Education, Centre for Longitudinal Studies.
- Fragile Families (2019) *Fragile Families Genetic Component / DNA Restricted Use Data Appendage*, Princeton, NJ: Bendheim-Thoman Center for Research on Child Wellbeing and Columbia Population Research Center.
- Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. (2015) Minimac2: faster genotype imputation, *Bioinformatics*, 31(5): 782–4. doi: [10.1093/bioinformatics/btu704](https://doi.org/10.1093/bioinformatics/btu704)
- Gassó, P., Pagerols, M., Flamarique, I., Castro-Fornieles, J., Rodriguez, N., Mas, S., Curran, S., Aitchison, K., Santosh, P., Lafuente, A. and The Stop Consortium (2014) The effect of age on DNA concentration from whole saliva: implications for the standard isolation method, *American Journal of Human Biology*, 26(6): 859–62.
- Gudiseva, H.V., Hansen, M., Gutierrez, L., Collins, D.W., He, J., Verkuil, L.D., Danford, I.D., Sagaser, A., Bowman, A.S., Salowe, R. et al. (2016) Saliva DNA quality and genotyping efficiency in a predominantly elderly population, *BMC Medical Genomics*, 9(1): 1–8: art 17, <https://doi.org/10.1186/s12920-016-0172-y>.
- Hansen, T.V., Simonsen, M.K., Nielsen, F.C. and 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 Epidemiology, Biomarkers & Prevention*, 16(10): 2072–6. doi: [10.1158/1055-9965.EPI-07-0611](https://doi.org/10.1158/1055-9965.EPI-07-0611)
- Haplotype Reference Consortium (2016) A reference panel of 64,976 haplotypes for genotype imputation, *Nature Genetics*, 48(10): 1279–83. doi: [10.1038/ng.3643](https://doi.org/10.1038/ng.3643)
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, *Nature Genetics*, 44(8): 955–9. doi: [10.1038/ng.2354](https://doi.org/10.1038/ng.2354)
- Ipsos MORI (2017) *Millennium Cohort Study Sixth Sweep (MCS6), version 2*, Technical Report, London: UCL Institute of Education, Centre for Longitudinal Studies, [https://doc.ukdataservice.ac.uk/doc/8156/mrdoc/pdf/mcs6\\_report\\_on\\_response.pdf](https://doc.ukdataservice.ac.uk/doc/8156/mrdoc/pdf/mcs6_report_on_response.pdf).
- Joshi, H. and 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): 409–30. doi: [10.14301/llcs.v7i4.410](https://doi.org/10.14301/llcs.v7i4.410)
- Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., et al. (2016) Reference-based phasing using the haplotype reference consortium panel, *Nature Genetics*, 48(11): 1443–8. doi: [10.1038/ng.3679](https://doi.org/10.1038/ng.3679)
- McFall, S.L., Conolly, A. and Burton, J. (2014) Collecting biomarkers and biological samples using trained interviewers: lessons from a pilot study, *Survey Research Methods*, 8(1): 57–66.

- Nishita, D.M., Jack, L.M., McElroy, M., McClure, J.B., Richards, J., Swan, G.E. and Bergen, A.W. (2009) Clinical trial participant characteristics and saliva and DNA metrics, *BMC Medical Research Methodology*, 9(1): 1-8: art 71, <https://doi.org/10.1186/1471-2288-9-71>. doi: [10.1186/1471-2288-9-71](https://doi.org/10.1186/1471-2288-9-71)
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. et al. (2007) PLINK: a tool set for whole-genome association and Population-based linkage analyses, *American Journal of Human Genetics*, 81(3): 559–75. doi: [10.1086/519795](https://doi.org/10.1086/519795)
- Rogers, N.L., Cole, S.A., Lan, H.C., Crossa, A. and Demerath, E.W. (2007) New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies, *American Journal of Human Biology*, 19(3): 319–26. doi: [10.1002/ajhb.20586](https://doi.org/10.1002/ajhb.20586)
- Sun, F. and Reichenberger, E.J. (2014) Saliva as a source of genomic DNA for genetic studies: review of current methods and applications, *Oral Health and Dental Management*, 13(2): 217–22.



## Appendix: Fieldwork Interviewer Instructions

### *Saliva collection instructions*

#### *Advance instructions for parents*

##### **How do you give a saliva sample?**

You and your child will be asked to spit your saliva into a small container. It is very easy and can be done in private. About half a teaspoon of saliva is needed. This typically takes about 5 minutes. There is no risk of harm to you or others when giving a saliva sample. You should not eat, drink, smoke or chew gum for 30 minutes before giving a saliva sample.

#### *Advance instructions for young people*

##### **How?**

Giving a saliva sample is very easy – the interviewer will explain how. You will be given a small container and asked to spit into it. You can do it in private. You should not eat, drink, smoke or chew gum for 30 minutes before giving a saliva sample. There is no risk of harm to you or others when giving a saliva sample. Your parent(s) will be asked to do the same thing.

#### *Instructions for interviewers in the household*

##### 4.4.1. Preparing for the saliva sample collection

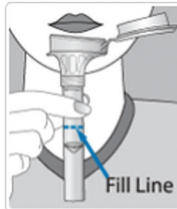



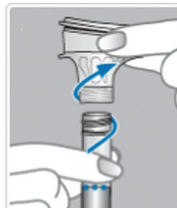
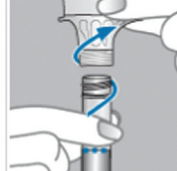




Please ensure that the person giving the sample has not eaten, drunk, smoked, or chewed gum in the 30 minutes prior to providing the sample. Contaminated samples may not be usable.

The front page of the consent booklet has peel-off barcode labels for the main respondent, partner and the young person saliva tubes. You must stick the relevant label on the saliva collection tube immediately before sample collection. The barcode must be stuck along the length of the tube to enable it to be scanned at the lab; it must NOT be wrapped around the tube. Please ensure that you do not cover the fill line with the label.

##### 4.4.2. Collecting the saliva sample

This is a diagram of the Oragene collection tube.



<p>1. Ask the respondent to spit into the container until the amount of liquid saliva (not bubbles) has reached the fill line marked on the side of the tube. This usually takes about 5 minutes.</p>	
<p>2. Please encourage them to try to complete the collection within 30 minutes. If they are having difficulties producing enough saliva, it sometimes helps if people close their mouths and wiggle their tongues or rub their cheeks (or think of their favourite food!)</p>	
<p>3. Please put on your disposable gloves.</p>	
<p>4. When the amount of liquid saliva has reached the fill line on the side of the tube, ask the respondent to pass the tube to you.</p>	
<p>5. Hold the tube upright with one hand. If there are a lot of bubbles in the tube, tap the tube gently against a hard surface. Close the lid with the other hand by firmly pushing the lid until you hear a loud click. The liquid in the lid will be released into the tube to mix with the saliva. <b>MAKE SURE THE LID IS CLOSED TIGHTLY.</b></p>	
<p>6. Hold the tube upright. Unscrew the tube from the funnel.</p>	
<p>7. Pick up the small cap for the tube. Use the small cap to close the tube tightly.</p>	
<p>8. Shake the capped tube for 5 seconds to mix.</p>	
<p>9. Put the tube into the small plastic bag with the absorbent material and seal tightly.</p>	
<p>10. Place the funnel, Oragene packaging and gloves in the disposable bag for waste. Ask the respondent to dispose of it in the household rubbish.</p>	
<p>11. Use the hand gel to clean your hands.</p>	

#### 4.4.3. Packaging the saliva sample

You need to write the barcode number from the consent form onto the despatch form (please also make sure that you have entered your interviewer number in the space provided).

Place the sample and the despatch form in one of the white jiffy bags.

The maximum number of saliva samples you should post in one jiffy bag is 15.

#### 4.4.4. Recording saliva collection information on the consent booklet

You will also be required to record the date each sample was collected on the front of the consent booklet. You will also be required to confirm that the respondent hasn't eaten, drunk, smoked or chewed gum within the 30 minutes prior to collection.

#### 4.4.5. Recording the saliva barcode number in CAPI

You must input the saliva barcode numbers into CAPI in the Final Element module, which you will complete at home. You will need to use the consent form to do this. This is the only way we will know which saliva sample belongs to which respondent. You will also be asked to record any reasons for refusal, any problems with the saliva sample collection and the information you recorded on the front of the consent booklet in the Final Element module.

#### 4.4.6. Storing and despatching the samples

The samples should be stored at room temperature so please do not store them in the fridge. Once a week during fieldwork, please gather all of the samples that you have collected together, check that you have entered the saliva barcode number for each sample on the despatch note(s), and post them back to the Bristol address. You can post them back using a normal post box. You can post samples to Bristol even if you have not completed the Final Element module in CAPI (for example, if you expect to return to the household to complete outstanding survey elements).