# Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource

## Information for researchers

## v1.2 Oct 2015

## Interim Data Release 2015

# 1 Introduction

## 1.1 UK Biobank

UK Biobank is a prospective cohort study of over 500,000 individuals from across the United Kingdom. Participants, aged between 40 and 69, were invited to one of 22 centres across the UK between 2006 and 2010. Blood, urine and saliva samples were collected, physical measurements were taken, and each individual answered an extensive questionnaire focused on questions of health and lifestyle.

The resource will provide a picture of how the health of the UK population develops over many years and it will enable researchers to improve the diagnosis and treatment of common diseases [1].

A key goal of UK Biobank is to collect genetic data on every participant. This data, combined with the extensive information about medical history and lifestyle choices, will present an unparalleled opportunity to investigate how genetics and other factors impact the onset and development of disease.

The UK Biobank resource is open to the research community and it will grow and develop over time. Findings that use UK Biobank data must be fed back to UK Biobank and made available to other researchers.

## 1.2 Purpose of this document

Here we describe the quality control (QC) procedures applied to the genotype data in the interim UK Biobank data release, which contains ~150,000 samples genotyped at ~800,000 SNPs. We also describe characteristics of the released genotype data, both in terms of content and quality. This document is relevant to researchers accessing and using the genotype data available in the interim release. However, largely the same procedures will be applied in future releases. We also briefly describe the UK Biobank resource, the genotyping array, the sample storage and genotyping procedures, although these are described in more detail in the references.

## 1.3 Data releases

The interim release of genotype data for UK Biobank comprises ~150,000 samples. Work is ongoing on aspects of genotype calling that can utilise the scale of the project to further improve the comprehensiveness of the genetic data. This means that some small number of genotype calls in the interim release may change in subsequent releases. If this occurs, information will be made available about which genotype calls have changed, as a complement to the new genotype data.

Information about the likely timing and extent of future data releases is available from the UK Biobank website, http://biobank.ctsu.ox.ac.uk.

## 1.4 The UK Biobank Axiom genotyping array

The UK Biobank Axiom array from Affymetrix was specifically designed by an expert group, for the purpose of genotyping the UK Biobank participants. Many researchers contributed markers and data during the array design process. There are ~800,000 markers on the array (see [2] for more details).

Briefly, the array design philosophy was to:

- Add markers that are of particular interest because of known associations or possible roles in phenotypic variation.
- Add coding variants across a range of minor allele frequencies (MAFs), principally missense and protein truncating variants.
- Choose the remaining content to provide good genome-wide imputation coverage in European populations in the common (>5%) and low frequency (1-5%) MAF ranges.

The UK Biobank Axiom array is being used to genotype ~450,000 of the ~500,000 UK Biobank participants. The other ~50,000 samples were genotyped on the closely related UK BiLEVE array. The UK BiLEVE project, for which the UK BiLEVE array was designed, aims to study the genetics of lung health and disease, and so those ~50,000 individuals were selected based on lung function and smoking behaviour from participants with self-declared European ancestry. Otherwise, the UK BiLEVE cohort and the rest of UK Biobank differ only in small details of the DNA processing stage (e.g., UK BiLEVE samples were manually transferred from storage to plates for DNA extraction).

The two SNP arrays are very similar with over 95% common marker content. The UK Biobank Axiom array is an updated version of the UK BiLEVE Axiom array, and it includes additional novel markers (such as cancer-related markers), which replaced a small fraction of the markers used for genome-wide coverage. The marker lists for both the UK BiLEVE and the UK Biobank Axiom arrays are available as part of the UK Biobank resource, and further details of the array design are available in the UK Biobank Axiom Array content summary [2].

The ~50,000 samples genotyped on the UK BiLEVE Axiom array are included in the interim release. Since the UK BiLEVE sampling scheme and array design are reported in detail elsewhere [3], in the following sections we describe the DNA extraction and genotyping of the other ~450,000 samples processed on the UK Biobank Axiom array.

A small number of variants (7,104) assayed on the array were known, or suspected to have more than two segregating alleles. Multi-allelic markers require special treatment in array design and genotype calling. A number of these variants (3,690) are particularly complicated and are not currently supported by the Affymetrix analysis pipeline; they have been set to missing in all batches. The remaining (3,414) multi-allelic variants are supported by Affymetrix but care must be taken in the interpretation of the calls provided, as a pair of calls (for the same individual) must be considered together to

reconstruct the actual genotype at the marker. The list of all multi-allelic markers, both supported and unsupported by Affymetrix, is available to download. Furthermore, researchers interested in multi-allelic markers can download either the array intensity files (.cel files) or the processed intensity values, and undertake their own calling, QC and analyses.

The custom-designed UK Biobank Axiom array attempts to assay a large number of SNPs that have not been previously genotyped. As expected, a small number of markers (~38,000, i.e., less than 5% of all markers present on the UK Biobank Axiom array) exhibited sub-optimal and/or complex clustering patterns and hence were excluded from all subsequent QC metrics and statistics, and their corresponding calls were set to missing in the interim data release.

## 1.5 Overview of DNA extraction and genotyping

### 1.5.1 Sample storage and DNA extraction

The samples collected from participants are held at the UK Biobank facility in Stockport, UK. Storage protocols for all samples require 850µl stored in racks of 96 x 1.2ml microtubes, at either -80°C or -196°C (depending on sample type). Generally the racks are populated with samples grouped by sample type, collection centre and collection time. DNA is extracted from buffy coat samples, which (generally) make up 24 of every 96 tubes on the racks in storage. Samples are picked by robot to a 96-position destination rack (a plate) ready for DNA extraction (94 samples per plate leaving two spaces for the addition of controls).

Given the unprecedented sample size of the cohort, special attention was given to ensure that sources of sample collection or extraction variability and other measurement errors do not systematically differ between cases and controls in any future case-control studies. Attempts were made to avoid samples submitted for analysis being grouped or submitted in a sequence which itself exhibits an underlying trend. This was achieved via a sample selection algorithm that ensures a mixture of collection centres on each destination rack [4]. During DNA extraction, the DNA concentration and purity are assessed. Samples failing to meet defined thresholds are not submitted for genotyping; where possible these samples are re-processed at a later date. Further details of the UK Biobank sampling and DNA extraction procedures can be found in [4,5].

### 1.5.2 Genotyping

Samples were genotyped at the Affymetrix Research Services Laboratory in Santa Clara, California, USA. Upon receipt of a 96-well plate containing 94 UK Biobank samples, Affymetrix added two control individuals (from 1000 Genomes) to the same well positions on each plate: HG00097 to well A12 and HG00264 to well E12. See Affymetrix laboratory process documentation for further details [6].

Axiom Array plates were processed on the Affymetrix GeneTitan® Multi-Channel (MC) Instrument. Genotypes were then called from the resulting intensities in batches of ~4,700 samples (~4,800 including the controls) using the Affymetrix Power Tools software and the Affymetrix Best Practices Workflow [7]. Supplementary Table S1 shows the number of samples and plates per batch in the interim release (which includes the 11 UK BiLEVE batches and 22 UK Biobank batches, i.e. 11 batches genotyped on the UK BiLEVE Axiom array and 22 batches genotyped on the UK Biobank Axiom array).

Individuals with the same genotype at any given SNP will cluster together in a two-dimensional intensity space (one dimension for each targeted allele). Briefly, genotype calling involved inferring properties of these clusters within each batch and assigning each sample a genotype (or leaving the call missing) based on its position in intensity space. For the interim data release, Affymetrix performed further rounds of genotype calling using algorithms customised for the UK Biobank project. These algorithms targeted very rare SNPs with 6 or fewer minor alleles in a batch, and a subset of SNPs for which the generic calling algorithm did not perform optimally [8]. After genotype calling, Affymetrix performed quality control in each batch separately, to exclude SNPs with poor cluster properties. If a SNP did not meet the Affymetrix prescribed QC thresholds in a given batch, it was set to missing in all individuals from that batch. Affymetrix also checked sample quality (such as DNA concentration) and genotype calls were provided only for samples with sufficient DNA metrics. More information about the Affymetrix calling algorithms and quality control protocols are available in [6,7,8].

# 2 Additional quality control

## 2.1 Our approach

We undertook QC in several stages. First we used several SNP-based metrics to flag SNPs with less reliable genotyping results, to be set to missing in the batches where they failed our filters. Then we identified poor quality samples using only high quality SNPs (defined as SNPs that passed QC filters in all 33 batches in this interim release). We also performed other sample-based inference such as principal component analysis and relatedness inference. Properties of UK Biobank (such as its large cohort size) mean that some quality control metrics commonly used in genome-wide association studies (GWAS) are not sufficient in this context. We used a variety of approaches in our QC procedures to account for the effects of population structure and batch-based genotyping, which we discuss below.

### 2.1.1 Diverse ancestries

UK Biobank consists of ~500,000 UK individuals. Participants were asked to choose from a set of predefined ethnic categories, or 'Other', and ~470,000 reported their ethnicity as 'White'. Other individuals come from a wide variety of ethnic groups (Table 1).

| Self-reported ethnicity | | Representation (%) |
|---|---|---|
| White | | 94.06 |
| | British | 88.07 |
| | Irish | 2.63 |
| | Any other white background | 3.36 |
| Asian | | 2.28 |
| | Indian | 1.18 |
| | Pakistani | 0.37 |
| | Bangladeshi | 0.05 |
| | Chinese | 0.31 |
| | Any other Asian background | 0.37 |
| Black | | 1.61 |
| | African | 0.68 |
| | Caribbean | 0.90 |
| | Any other Black background | 0.03 |
| Mixed | | 0.59 |
| | White and Asian | 0.17 |
| | White and Black African | 0.08 |
| | White and Black Caribbean | 0.12 |
| | Any other mixed background | 0.22 |
| Other/Unknown | | 1.46 |

**Table 1** Self-reported ethnic groups in the ~500,000 UK Biobank participants. Of these, ~150,000 were genotyped for the interim data release.

The inclusion of samples with diverse ancestry can confound standard QC metrics. For

instance, individuals with unusual heterozygosity are typically excluded from a GWAS, but heterozygosity is correlated with ancestry as allele frequency distributions can vary across populations. Similarly, testing that Hardy-Weinberg Equilibrium (HWE) holds is a common approach for identifying poor quality SNPs, but departures from HWE can be expected in the context of strong population structure, again because of differences in allele frequency distributions.

To account for the effects of population structure, we proceeded in two phases. For SNP-based QC metrics we used only individuals with similar ancestry (so that, for example, HWE is expected). To do this we identified a set of individuals with European ancestry by projecting individuals onto principal components computed from the 1000 Genomes project. We also characterised the population structure unique to UK Biobank by computing principal components using only UK Biobank individuals (after applying SNP QC). We used the UK Biobank-specific principal components analysis (PCA) results to account for population structure in all our sample-based QC metrics.

### 2.1.2 Batch-based genotype calling

In view of UK Biobank's large cohort size, Affymetrix carried out the genotyping and initial SNP QC in batches of around 4,800 samples, effectively treating each batch as an independent experiment. However, the availability of multiple batches, processed under the same strict guidelines, provides new opportunities for SNP QC: we can check the consistency of genotype calling between batches. In rare instances, the Affymetrix calling algorithm might incorrectly call a SNP in one batch but not others.

Affymetrix assays genetic markers using "probesets" which target a particular variant. A probeset is a set of probes whose signal is summarised to make the genotyping call. A small fraction of variants (mostly those that are novel to the UK Biobank Axiom array) are genotyped using multiple probesets, and in this case more than one call is made for the same marker. For these markers Affymetrix recommends a single "best" probeset in each batch separately and the interim release includes only calls from the "best" probesets. We did not use these markers in our sample QC analyses as a different probeset can be recommended for the same SNP across batches.

## 2.2 SNP QC

Due to the size of the UK Biobank cohort, genotyping was performed in a large number of batches (33 batches of ~4800 individuals for the interim data release). This provides additional opportunities to study and ensure data consistency. Affymetrix routinely undertakes SNP QC [7,8], and we adopted the Affymetrix recommendations throughout, for each given batch. In addition, we performed quality checks that are appropriate for a large-scale dataset genotyped in batches. For the reasons described above, we computed all SNP QC metrics using a homogeneous subset of individuals drawn from the largest ancestral group in the cohort (which is European in UK Biobank). To identify these individuals, we projected UK Biobank samples on the two major principal

components computed by analysing the CEU, YRI, CHB and JPT populations from the HapMap3 reference panel (with genotypes provided by 1000 Genomes, phase 1, release v3). Then we selected samples that were projected in the neighbourhood of the CEU cluster, as shown in Figure 1.

The UK BiLEVE batches have a higher proportion of samples with European ancestry by design, as participants were selected in part based on self-declared ethnicity. In those 11 batches we used ~97% samples for SNP QC. In the UK Biobank batches we used 91%-93% samples for SNP QC, as these batches are more ethnically diverse. Appendix A1 describes the analysis we used to choose a homogeneous subset of samples for SNP QC.

In samples drawn from the same population we would not expect differences in genotype frequencies, either between batches or between plates within a batch, at the same marker. Such differences might indicate that the SNP was not genotyped as accurately as other SNPs, in the batch (or plate) which exhibits unusual genotype frequencies. We refer to these cases as batch or plate effects. For example, batch effects can occur when the sample intensities in one batch shift relative to the intensities in other batches. In rare cases, such a shift can cause the Affymetrix calling algorithm to miscall a genotype cluster that is not detected by the routine Affymetrix SNP QC. Similarly, plate effects can occur when the intensities in one plate shift relative to the intensities in other plates, in the same batch.

To look for effects in a particular batch we tested whether we can reject the null hypothesis that the given batch has the same genotype frequencies as all other batches combined. To look for effects in a particular plate we tested whether we can reject the null hypothesis that the given plate has the same genotype frequencies as all other plates, within the same batch, combined. In both cases we used Fisher's exact test on the 2 × 3 table of genotypes. (Since there are several plates in a batch, we performed Fisher's exact test for each plate that is at least half-full, i.e., with 48 samples or more, and then took the smallest p-value.) See Appendix A2 for more details.

We also performed an exact test for Hardy-Weinberg equilibrium for each batch [9]. Again, selecting a homogeneous subset of samples makes the procedure more conservative, as Hardy-Weinberg equilibrium does not necessarily hold in the presence of population structure.

If a SNP did not pass any of these tests (with a p-value of less than $10^{-12}$), this might indicate that the genotypes have not been called correctly in the corresponding batch and the SNP is flagged. For the current interim data release, genotypes at such flagged SNPs were set to missing in batches where the tests suggested issues with the initial calls. With the aim to improve genotype calling in subsequent data releases, SNPs that were filtered out in at least one batch are the subject of ongoing advanced analysis work by Affymetrix. Preliminary data generated by Affymetrix advanced analysis workflow indicates that a substantial number of SNP flagged in the interim release will be released in the final release.

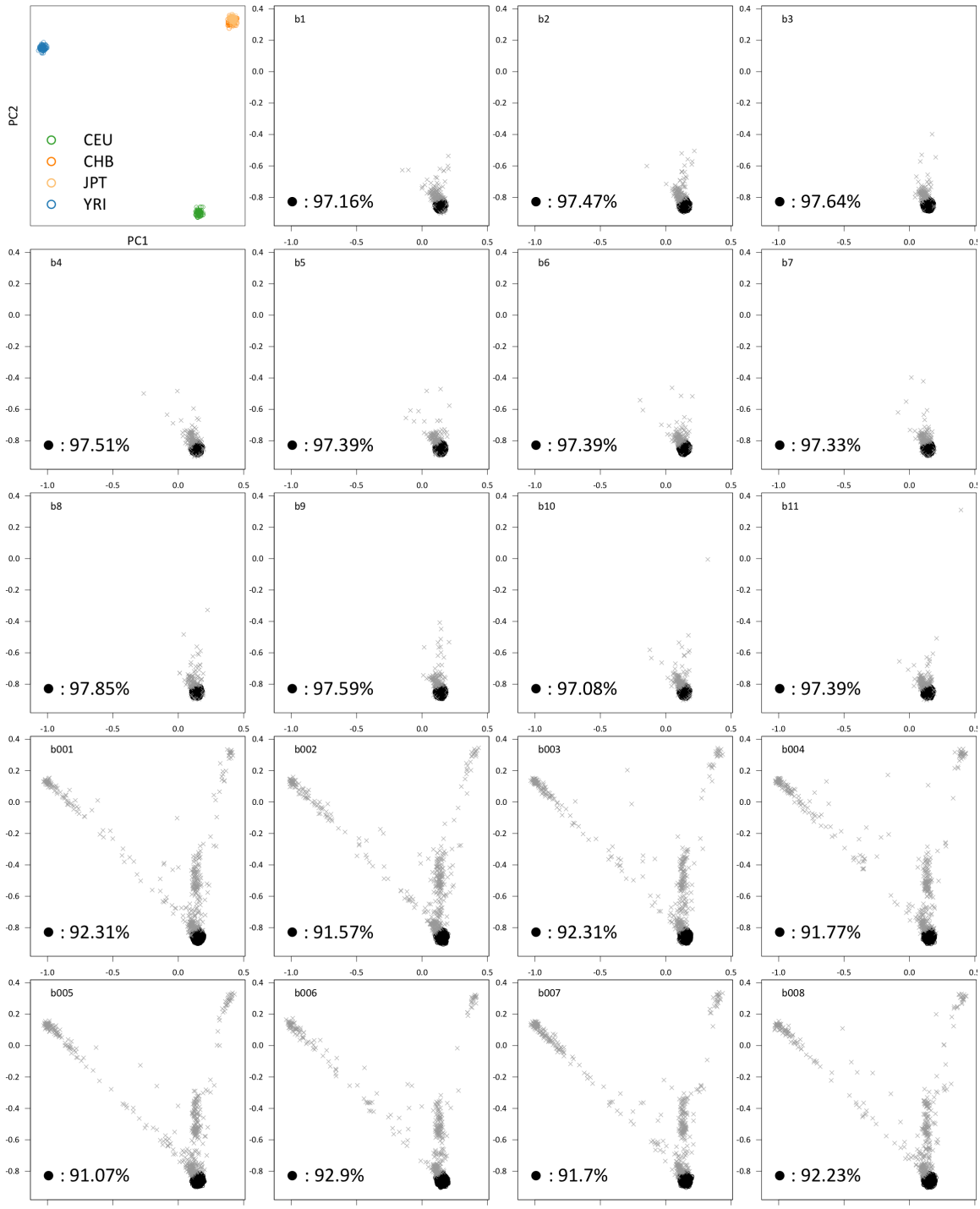


**Figure 1** We used 1000 Genomes data for four HapMap populations (CEU, CHB, JPT, YRI) to compute PCA loadings for ~40,000 SNPs on the UK Biobank Axiom array. In the top left panel, these HapMap samples are projected onto the 1$^{st}$ and 2$^{nd}$ principal components and are coloured by population. In the other panels, all 11 UK BiLEVE batches (labeled b1 to b11) and an arbitrarily chosen subset of 8 UK Biobank batches (labeled b001 to b008) are projected into the same principal component space. The samples are coloured according to whether they were used in SNP QC procedures or not (in black and gray, respectively). For each batch the proportion of samples used for SNP QC is also reported.

## 2.3 Sample QC

To carry out QC on samples, we first applied SNP QC (as described above) and selected a set of high quality autosomal SNPs. The analyses described below are based on ~600,000 autosomal SNPs which are on both the UK Biobank and UK BiLEVE arrays, and passed SNP QC in all 33 batches.

### 2.3.1 Population structure

To capture population structure specific to the UK Biobank cohort, we performed principal component analysis of ~150,000 UK Biobank samples using ~100,000 SNPs. These PCs can be used to identify samples with similar ancestry or to control for population structure in association studies. Metrics for sample quality control can be sensitive to population structure as well, so we used the principal components in the process of identifying poor quality samples. The four major PCs are shown in Figure 2. The next sixteen PCs (from PC5 to PC20) are shown in Figure S1 and details of the analysis are presented in Appendix A3.

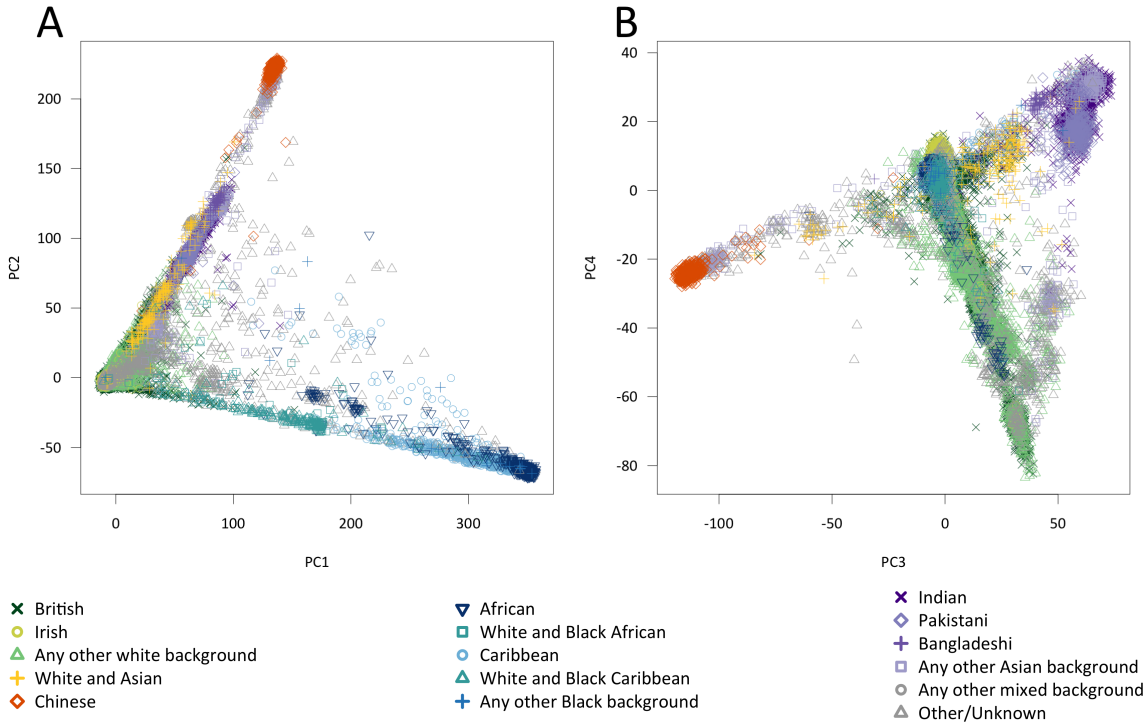


**Figure 2** Genetic principal components in UK Biobank, computed from 141,0670 samples and 101,284 SNPs using flashPCA [10]. **(A)** The 1st principal component (PC1) on the *x*-axis and the 2nd principal component (PC2) on the *y*-axis. **(B)** The 3rd principal component (PC3) on the *x*-axis and the 4th principal component (PC4) on the *y*-axis. In both panels, samples are coloured according to self-reported ethnicity. The legend indicates the coloured symbol used for each predefined ethnicity throughout this document.

### 2.3.2 Heterozygosity and missing rates

Extreme heterozygosity and/or low call rate can be indicators of poor sample quality [11]. However, heterozygosity is sensitive to population structure because allele

frequency distributions (and thus heterozygosity) can differ between populations. Figure 3A shows the effect of SNP ascertainment on heterozygosity: since the UK Biobank array was designed to provide good imputation coverage in European populations, samples with non-European ethnicity tend to have lower heterozygosity. We control for this by fitting a linear regression model with heterozygosity as the outcome and the four major PCs as the predictors (see Appendix A4 for details). The corrected heterozygosity is plotted in Figure 3B.

Some samples can have naturally extreme heterozygosity, even after accounting for population structure. Specifically, individuals with mixed ethnicity tend to have higher heterozygosity (which is not captured by the principal components), and individuals whose parents are closely related tend to have lower heterozygosity. Therefore, we attempted to flag as outliers samples whose extreme heterozygosity is not explained by mixed ancestry or increased levels of marriage between close relatives.

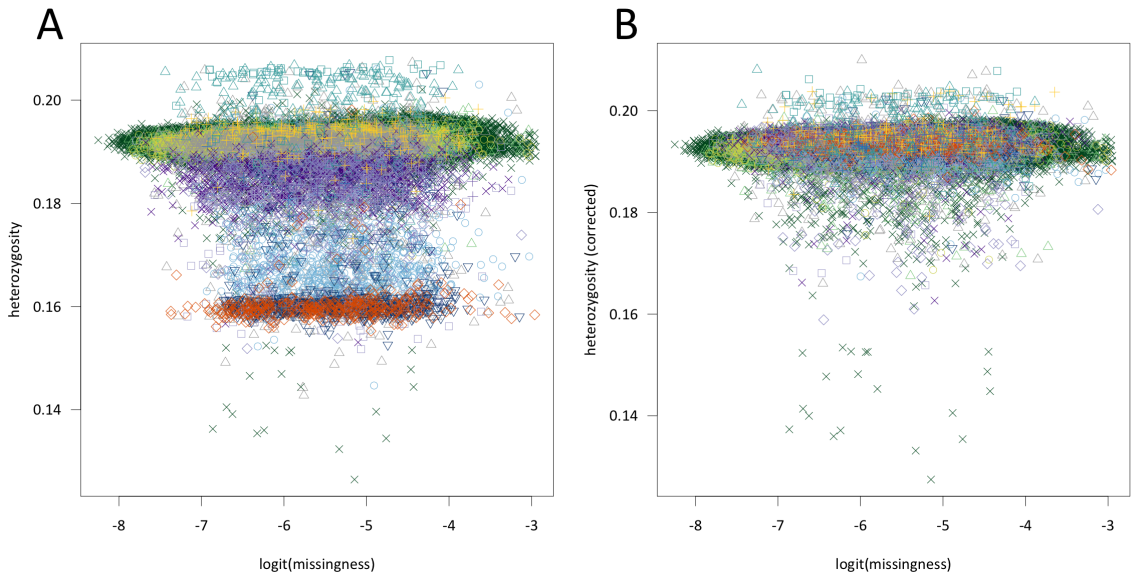


**Figure 3** Heterozygosity and missingness for 152,256 samples in the interim UK Biobank data release, after removing 480 outliers. (Section 2.3.2 details the procedure to flag outliers.) Points are coloured by self-reported ethnicity, using the coloured symbols in the legend of Figure 2. **(A)** Heterozygosity (proportion of autosomal heterozygous calls) on the *y*-axis against logit-transformed missingness (proportion of genotypes not called) on the *x*-axis. The logit transformation, defined as logit($x$) = log($x$/(1-$x$)), is applied to normalise the missingness values. **(B)** Ancestry-corrected heterozygosity on the *y*-axis against logit-transformed missingness on the *x*-axis. The heterozygosity values are corrected for systematic differences due to population structure using four genetic principal components, as described in Appendix A4.

After taking into account mixed ethnicity, we identified 472 outliers (0.3% of total samples) with high missingness or high heterozygosity (plotted in red in Figure 4A), by visually inspecting the scatterplots of heterozygosity and missingness for each self-reported ethnicity (see Figure S2). To distinguish between poor quality samples and samples with naturally low heterozygosity, we looked for long runs of homozygosity (ROH). We computed the total length of long ROH using *plink* [12] (see Appendix A5 for

details), and identified 8 samples with total ROH that is unusually short, compared to other samples with similar heterozygosity (Figure 4B).

In total, we identified 480 samples (0.3% of total samples) with high missingness or for which heterozygosity rates were not explained by ROH analysis nor mixed ethnicity. These samples are not excluded from the data release and instead a list of outlier IDs for these samples is provided to researchers along with the genotype data.

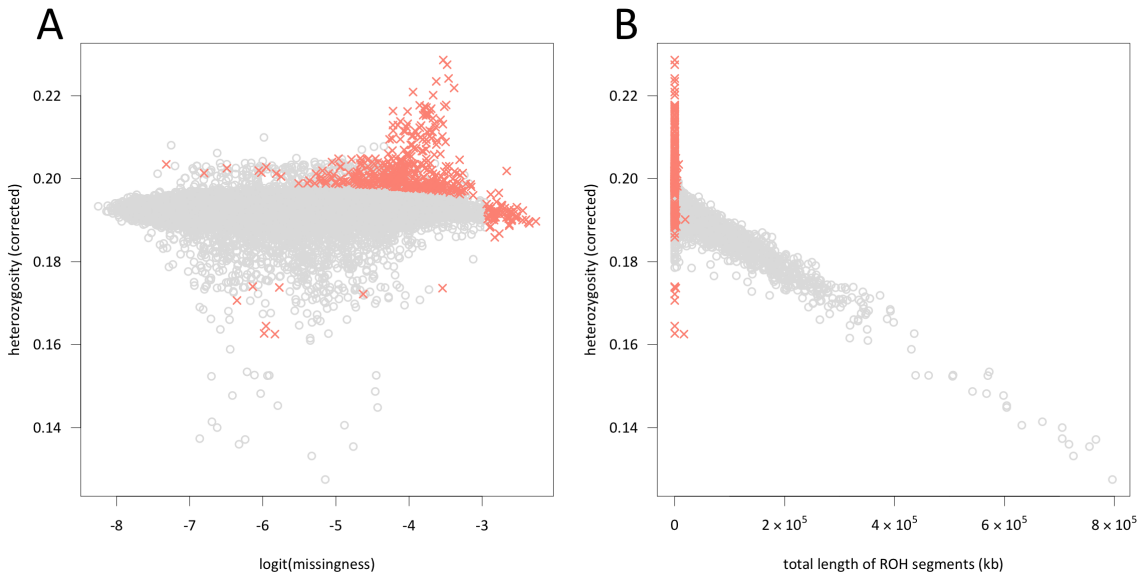


**Figure 4** A total of 152,736 UK Biobank samples were genotyped for the interim data release. (Intended and unintended duplicates are excluded from this count.) Of these, there are 480 outliers, shown in red; the rest of the samples are shown in gray. **(A)** Ancestry-corrected heterozygosity on the *y*-axis and logit-transformed missingness on the *x*-axis. This plot emphasizes that some outliers have high missingness or high heterozygosity. (Samples with mixed ancestry tend to have increased heterozygosity as well, but this is expected and such samples are not flagged as outliers based on heterozygosity alone.) **(B)** Ancestry-corrected heterozygosity on the *y*-axis and total length (in kb) of long runs of homozygosity (ROH) on the *x*-axis. This plot emphasizes that some outliers with low heterozygosity have unusually short total ROH.

## 2.4 Summary

After QC procedures were applied, the interim UK Biobank data release contains genotypes for 152,736 samples that passed sample QC (~99.9% of total samples), and 806,466 SNPs that passed SNP QC in at least one batch (>99% of the array content). As noted above, Affymetrix is pursuing ongoing development work on genotype calling in extremely large multi-batch settings. Therefore, some genotype calls may change between this interim data release and the final data release, and we anticipate that the various metrics will improve further.

# 3 Properties of the UK Biobank genotype data for Interim Release

The interim data release of UK Biobank genetic data consists of 152,736 samples. Of those, 102,754 were genotyped on the UK Biobank array (split into 22 batches) and 49,982 were genotyped on the UK BiLEVE array (split into 11 batches). In addition to computing principal components, we analysed several aspects of the interim release data after quality control had been applied.

## 3.1 Properties of samples

### 3.1.1 Related Individuals

We identified related samples by calculating kinship coefficients for all pairs of samples using KING's robust estimator [13]. We used this estimator as it is robust to population structure and it is implemented in an algorithm efficient enough to consider all $n(n-1)/2$ (~11,250,000,000) pairs in a practicable amount of time. Parent-child and full sibling pairs have the same expected kinship coefficient but can be distinguished by their IBS0 fraction, defined as the proportion of SNPs at which two samples have no alleles in common (see Figure 5). We excluded some samples from the kinship calculation because KING's robust estimator is not reliable for individuals with high heterozygosity or high missingness [13]. See Appendix A6 for details.

We only report relatives to the 3rd, 2nd and 1st degree and monozygotic twins (Table 2).

| *Relationship* | Monozygotic twins | Parent-offspring | Full siblings | 2nd degree | 3rd degree |
|---|---|---|---|---|---|
| *Pairs* | 18 | 619 | 2,183 | 1,061 | 5,811 |

**Table 2** Related pairs (3rd degree or closer) for ~150,000 UK Biobank participants genotyped in the interim UK Biobank data release. (The counts are derived from the kinship information presented in Figure 5.)

We detected 1,856 individuals that are related (to the 1st degree or as monozygotic twins) to more than one person, and thus will occur in more than one pair in Table 2. Seventy-two of these individuals are within a trio (child with two parents) in which checking of the sex and ages of both parents and age of the child was consistent with the inferred relationship. There are 6 instances of two siblings and a parent, and in one of these the siblings are monozygotic twins. The others are individuals within sets of 3 or 4 siblings.

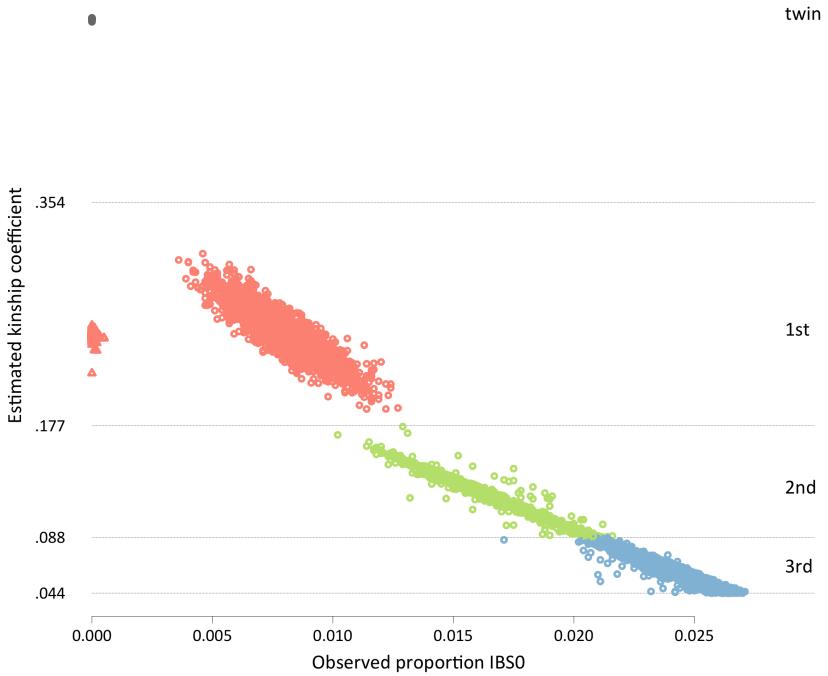


**Figure 5** Close relationships for ~150,000 UK Biobank participants genotyped in the interim release. Each point represents a pair of related individuals and the colours indicate the degree of relatedness: monozygotic twins in black (in the upper left corner), 1st, 2nd and 3rd degree relatives in red, green and blue, respectively. There are two groups of 1st degree relatives: parent-child pairs (red triangles) and full siblings (red circles). For all pairs, the *y*-axis shows the kinship coefficient, defined as the probability that two alleles sampled at random (one from each individual) are identical by descent. The *x*-axis shows the proportion of zero identity-by-state (IBS0), defined as the proportion of SNPs at which one sample carries the minor homozygote and the other sample – the major homozygote, so that they share no alleles.) The degree of relatedness is inferred from the estimated kinship coefficient using KING's criteria [13].

### 3.1.2 Sex mismatches

Affymetrix infers an individual's sex prior to genotype calling (but after measuring allele intensities) so that it can use an appropriate algorithm to call SNPs on the sex-linked chromosomes, X and Y. For this purpose, Affymetrix uses special probes for non-polymorphic sites on the X and Y chromosomes, which produce large differences in intensity between males and females. Self-reported sex (recorded at recruitment) and genetically inferred sex are available for all samples. Out of the ~150,000 samples in the interim release, the self-reported sex does not match the genetically inferred sex in 191 cases (0.1% of total samples).

There are three possible explanations for sex mismatches:

- Clerical error: Either the DNA sample was associated with the wrong individual (mislabelling) or sex was recorded incorrectly at recruitment
- Sex determined by chromosomal make-up does not match gender identity (and thus self-reported sex)

- Sex chromosome aneuploidy (i.e., abnormal number of sex chromosomes, for example – XXY)

Analysis of the X and Y-chromosome average intensities (which are available to download) can be used to identify instances of the third possible explanation. After the interim release, UK Biobank intends to extract DNA (where possible) and reprocess samples with unexplained gender mismatches.

Figure 6 reports two measures that can be used to infer gender. X-chromosome heterozygosity is informative because males carry a single copy of the X chromosome and thus cannot be heterozygous. The ratio of Y-chromosome to X-chromosome average intensity is informative because females carry no copy of the Y chromosome and thus their average Y intensity should be lower (not necessarily zero but at background level). The two measures are not mutually redundant and can be used to identify possible cases of sex chromosome aneuploidy. For example, samples with XXY aneuploidy are expected to have female-like heterozygosity on the X chromosome, but also have male-like intensity values for the Y chromosome. Such samples should not be used in downstream analysis, or used with caution, especially in conjunction with their phenotypic data.

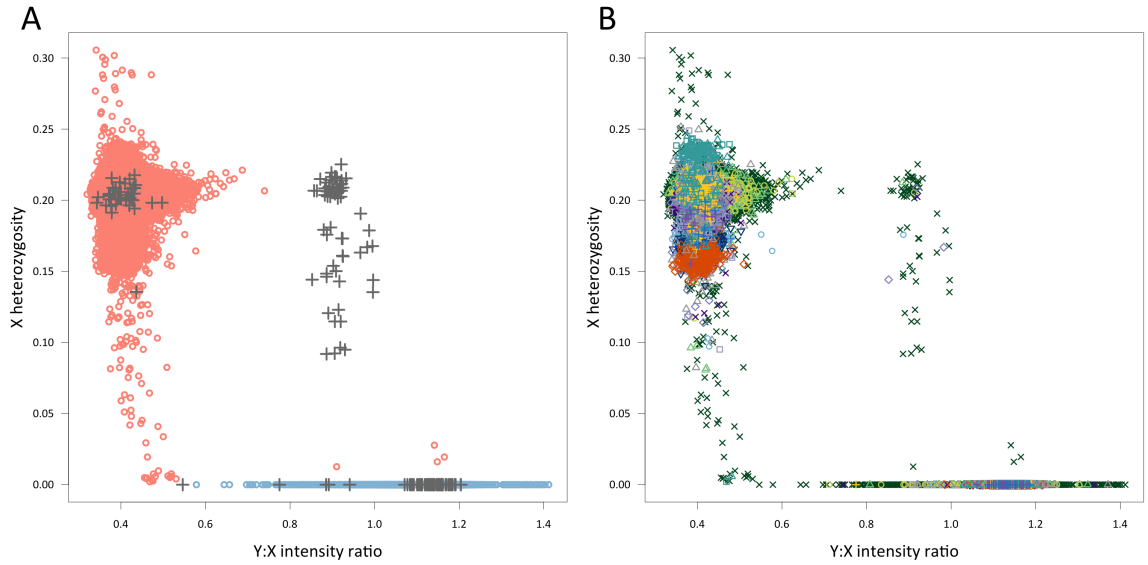


**Figure 6** X-chromosome heterozygosity and ratio of Y-chromosome to X-chromosome average intensity, for 152,736 UK Biobank samples. The X-chromosome heterozygosity is computed from all X-chromosome SNPs outside the PAR regions. The intensity values are measured at the probes used for determining sex prior to genotype calling. **(A)** Samples are coloured by gender: if the self-reported and genetically inferred sex agree, then females are plotted in red and males in blue; otherwise, mismatches are plotted in black. Points in centre of the plot (separated from the blue and red clusters) are possible cases of XXY aneuploidy. **(B)** The same points are coloured by self-reported ethnicity, using the coloured symbols in the legend of Figure 2. X-chromosome heterozygosity exhibits ascertainment bias due to population structure, similarly to autosomal heterozygosity. (Compare the systematic offset in heterozygosity between samples with different ethnic background in this figure and in Figure 3A).

## 3.2 Properties of SNPs

Figures 7, 8 and 9 illustrate various quality metrics and properties of SNPs genotyped on the UK Biobank Axiom array, across multiple batches. Affymetrix processed and genotyped the batches separately and we applied the same filters (the tests for batch or plate effects and Hardy-Weinberg equilibrium described in Section 2.2), independently, multiple times. Therefore, the number of times a SNP passed these filters is an extremely strict measure of its genotype calling quality. This and the call rate are reported in Figure 7.

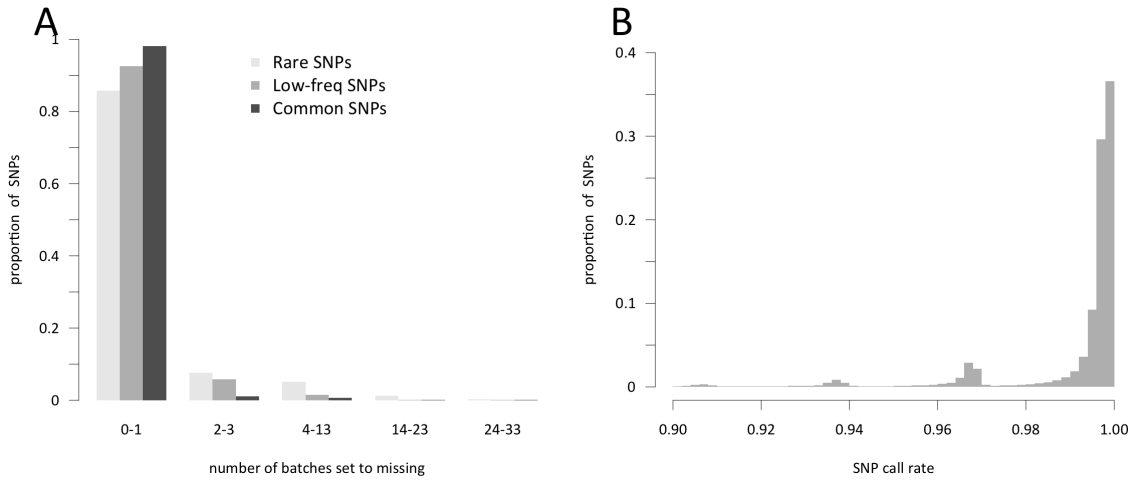


**Figure 7** Overall quality of the genotype data in the interim UK Biobank release, after all SNP QC steps have been applied. **(A)** Number of batches in which a SNP is set to missing (out of 33 batches), for common, low-frequency and rare SNPs genotyped on both the UK BiLEVE and UK Biobank Axiom arrays. The shading indicates one of three minor allele frequency (MAF) categories of SNPs: common (MAF>5%); low frequency (5%>MAF>1%); rare (MAF<1%). MAFs in UK Biobank were estimated from samples with inferred European ancestry. **(B)** SNP call rate for common, low frequency and rare SNPs combined.

The small peaks in the call rate in Figure 7B are due to SNPs set to missing in just a few batches. For example, if a SNP did not pass a QC threshold in exactly one batch in $n$ batches but otherwise has a high call rate in the remaining batches, its call rate is ~($n$-1)/$n$. Since there are 33 batches in the interim release, there is a subset of SNPs with call rate ~32/33 = 0.97 and a smaller subset with SNPs with call rate ~31/33 = 0.94.

Another measure of genotyping quality, reproducibility of calls, was assessed in two controls from 1000 Genomes which were added to every plate (in the same well on each plate) and were genotyped multiple times. Low discordance between calls for the same individual across different plates indicates high quality genotyping. The discordance for a particular SNP is computed as:

$$1 - \frac{\max\{ n_{AA},\ n_{AB},\ n_{BB} \}}{n_{AA} + n_{AB} + n_{BB}}$$

where $n_{AA}$, $n_{AB}$, $n_{BB}$ is the number of times the genotype AA, AB, BB is called, respectively. For concreteness, suppose that max{ $n_{AA}$, $n_{AB}$, $n_{BB}$ } = $n_{AA}$. That is, $n_{AA}$ is the

mode of the set { $n_{AA}$, $n_{AB}$, $n_{BB}$ } and therefore AA is the consensus call. The discordance is the proportion of calls that are not the consensus call; in the example, this is the proportion of AB or BB calls. Figure 8 shows the discordance rates for the two 1000 Genomes controls. In both cases, there is a small number of SNPs with discordance > 0.05 (282 for HG00097 and 143 for HG00264, or 417 (0.05%) in total). These SNPs are included in the interim release but the list can be downloaded. Some might be subject to exclusion in the final release after further analysis has been performed.

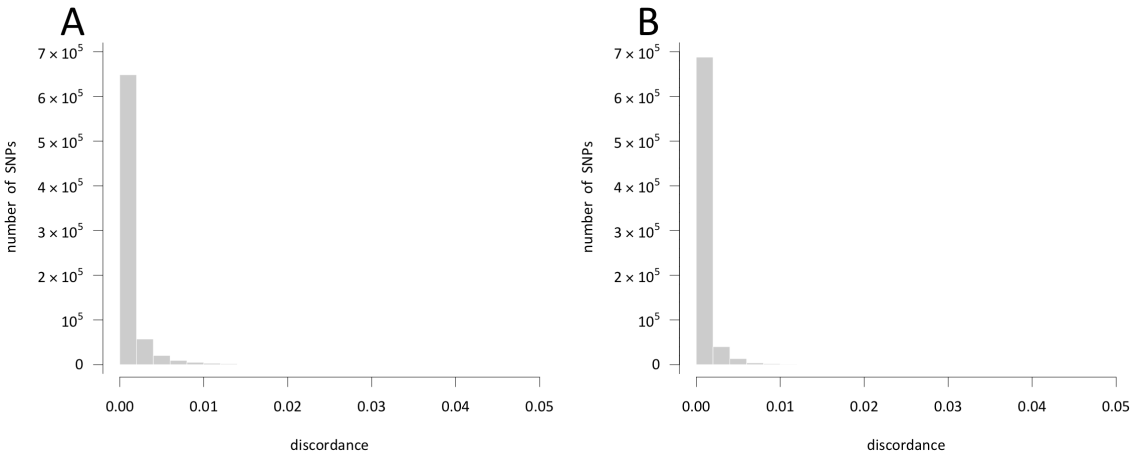


**Figure 8** Rates of discordance from the consensus call, for the two 1000 Genomes controls genotyped multiple times on the UK Biobank array. **(A)** Discordance for HG00097. **(B)** Discordance for HG00264.

Figure 9 shows the distributions of minor allele frequency and missingness, across SNPs that passed all SNP QC filters in all 33 batches in the interim release.

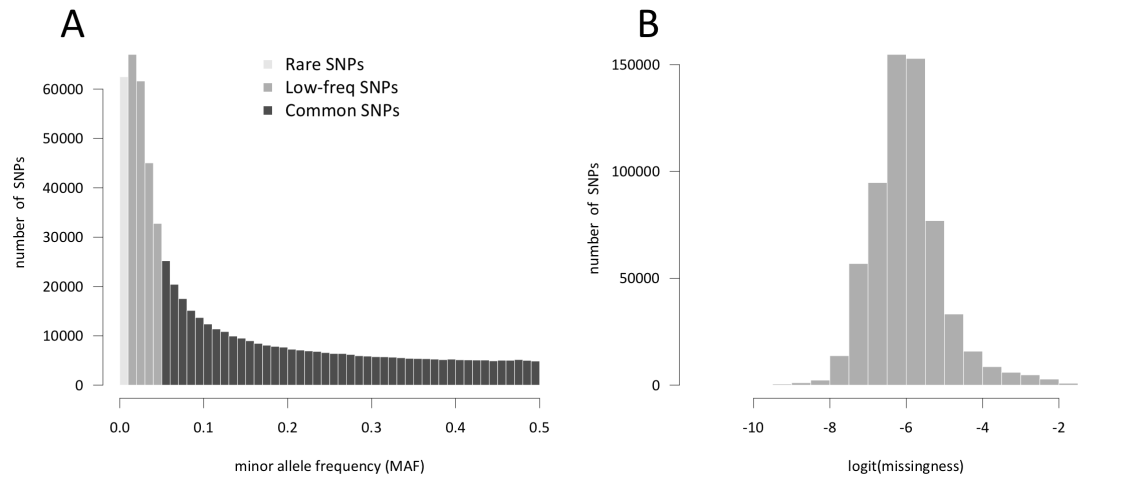


**Figure 9** Distributions of minor allele frequency and missingness across a set of 626,445 SNPs genotyped on both the UK BiLEVE and UK Biobank Axiom arrays, which passed all SNP QC filters in the 33 batches of the interim release. **(A)** Histogram of minor allele frequencies estimated from samples with inferred European ancestry. The shading indicates one of three MAF categories: common SNPs with MAF>5%; low frequency SNPs with 5%>MAF>1%; rare SNPs with MAF<1%. **(B)** Histogram of logit-transformed missingness for common, low frequency and rare SNPs combined. For reference, logit(-8) corresponds to 0.033% missingness; logit(-6) to 0.247% missingness; logit(-4) to 1.799% missingness.

A small number of genotyped autosomal SNPs (65) have been found which show significantly different allele frequencies between the UK BiLEVE array and the UK Biobank array. These SNPs are in the interim data release but should be excluded from analyses. A number (27) of these SNPs were used in phasing and imputation. We strongly recommend conditioning on array in association tests to ameliorate the effect of these SNPs. There could still be a subtle bias in the neighbourhood of these SNPs after conditioning, but this will depend upon the phenotype being tested for association. We recommend looking carefully at any results with imputed SNPs in the regions of the affected SNPs, including confirming any GWAS hits with the genotyped-only data and looking at cluster plots of the genotype data. Additionally, there are a number of SNPs (46) on chromosome X which show a significant allele frequency difference between males and females or show differences between arrays. We recommend that these SNPs be excluded from all analyses. The full list of these markers is available to download. These SNPs were identified as those with a p-value less than $10^{-40}$ in a Fisher exact test on genotype counts.

## References


[1] N. Allen, C. Sudlow, P. Downey, T. Peakman, J. Danesh, P. Elliott, J. Gallacher, J. Green, P. Matthews, J. Pell, T. Sprosen, and R. Collins, "UK Biobank: Current status and what it means for epidemiology," *Health Policy and Technology*, 1(3):123-126, 2012.

[2] The UK Biobank Array Design Group, "UK Biobank Axiom array: content summary", 2014. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf

[3] L.V. Wain *et al.*,"Novel insights into the genetics of smoking behavior, lung function and chronic obstructive pulmonary disease in UK Biobank," *Submitted*, 2015.

[4] UK Biobank, "Genotyping of 500,000 participants: Description of sample processing workflow and preparation of DNA for genotyping", 20 April, 2015.

[5] UK Biobank, "DNA extraction at UK Biobank", 2014. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/DNA-Extraction-at-UK-Biobank-October-2014.pdf

[6] Affymetrix, "UKB_WCSGAX: UK Biobank 500K Samples Processing by the Affymetrix Research Services Laboratory", April, 2015.

[7] Affymetrix, "Axiom® Genotyping Solution Data Analysis Guide", 2014. http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf

[8] Affymetrix, "UKB_WCSGAX: UK Biobank 500K Genotyping Data Generation by the Affymetrix Research Services Laboratory", April, 2015.

[9] J.E. Wigginton, D.J. Cutler and G.R. Abecasis,"A note on exact tests of Hardy-Weinberg equilibrium," *The American Journal of Human Genetics*, 76(5):887-893, 2005.

[10] G. Abraham and M. Inouye, "Fast principal component analysis of large-scale genome-wide data," *PLoS ONE,* 9(4):e93766, 2014.

[11] IMSGC and Wellcome Trust Case Control Consortium, "Genetic risk and the role of cell mediated immune mechanisms in multiple sclerosis," *Nature*, 476(7539):214-219, 2011.

[12] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly MJ, Sham PC (2007) "PLINK: A Tool Set for Whole-Genome and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, 81(3): 559–575, 2007.

[13] A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, and W.-M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, 26(22):2867-2873, 2010.

[14] C. Bellenguez, A. Strange, C. Freeman, Wellcome Trust Case Control Consortium, P. Donnelly, and C.C. Spencer, "A robust clustering algorithm for identifying problematic samples in genome-wide association studies," *Bioinformatics*, 28(1):134-135, 2012.

[15] A.L. Price *et al*. Long-range LD can confound genome scans in admixed populations. *The American Journal of Human Genetics*, 83(1):132-135, 2008.

## Appendices

The interim UK Biobank data release consists of 11 UK BiLEVE batches and 22 UK Biobank genotyped (in this order) by Affymetrix using calling algorithms specifically adapted to the UK Biobank project [7,8].

| Batch | Number of genotyping plates | Number of UK Biobank samples | Number of control samples |
|---|---|---|---|
| UK BiLEVE b1 | 52 | 4592 | 195 |
| UK BiLEVE b2 | 53 | 4598 | 186 |
| UK BiLEVE b3 | 58 | 4587 | 210 |
| UK BiLEVE b4 | 52 | 4601 | 196 |
| UK BiLEVE b5 | 59 | 4596 | 197 |
| UK BiLEVE b6 | 61 | 4573 | 216 |
| UK BiLEVE b7 | 63 | 4589 | 199 |
| UK BiLEVE b8 | 53 | 4593 | 202 |
| UK BiLEVE b9 | 54 | 4594 | 198 |
| UK BiLEVE b10 | 59 | 4597 | 184 |
| UK BiLEVE b11 | 72 | 4600 | 199 |
| UK Biobank b001 | 52 | 4710 | 90 |
| UK Biobank b002 | 74 | 4657 | 134 |
| UK Biobank b003 | 85 | 4648 | 141 |
| UK Biobank b004 | 91 | 4652 | 142 |
| UK Biobank b005 | 87 | 4661 | 141 |
| UK Biobank b006 | 64 | 4689 | 113 |
| UK Biobank b007 | 75 | 4678 | 118 |
| UK Biobank b008 | 186 | 4755 | 41 |
| UK Biobank b009 | 73 | 4693 | 104 |
| UK Biobank b010 | 85 | 4713 | 85 |
| UK Biobank b011 | 97 | 4704 | 95 |
| UK Biobank b012 | 83 | 4706 | 89 |
| UK Biobank b013 | 56 | 4692 | 106 |
| UK Biobank b014 | 201 | 4710 | 82 |
| UK Biobank b015 | 469 | 4714 | 71 |
| UK Biobank b016 | 177 | 4605 | 87 |
| UK Biobank b017 | 134 | 4600 | 91 |
| UK Biobank b018 | 111 | 4621 | 93 |
| UK Biobank b019 | 131 | 4627 | 72 |
| UK Biobank b020 | 88 | 4637 | 119 |
| UK Biobank b021 | 182 | 4582 | 175 |
| UK Biobank b022 | 79 | 4719 | 40 |

**Table S1** Number of genotyping plates and processed samples per batch for the interim UK Biobank data release. (These numbers exclude samples with low DNA quality but include intended/unintended duplicates and sample outliers.) The 11 UK BiLEVE batches, labelled b1 to b11, were genotyped on the UK BiLEVE Axiom array; the 22 UK Biobank batches, labelled b001 to b022, were genotyped on the UK Biobank Axiom array.

## A1 Selecting samples with European ancestry for SNP QC

Here we describe the procedure to identify samples with European ancestry and thus construct the homogeneous subset used in computing SNP QC metrics. The procedure includes principal component analysis and two-way clustering.

We first downloaded 1000 Genomes data in Variant Call File (VCF) format and extracted 714,168 SNPs (no INDELs) that are genotyped on the UK Biobank Axiom array as well. We selected 355 unrelated samples from the populations CEU, CHB, JPT, YRI, and then chose SNPs for principal component analysis using the following criteria:

- MAF ≥ 5% and HWE p-value > $10^{-6}$, in each of the populations CEU, CHB, JPT and YRI.
- Pairwise $r^2 \leq 0.1$ to exclude SNPs in high LD. (The $r^2$ coefficient was computed using *plink* [12] and its 'indep-pairwise' function with a moving window of size 1000 bp).
- Removed C/G and A/T SNPs to avoid unresolvable strand mismatches.
- Excluded SNPs in several regions with high PCA loadings (after an initial PCA).

With the remaining 40,538 SNPs we computed PCA loadings from the 355 1,000 Genomes samples, then projected the UK Biobank samples onto the 1st and 2nd principal components. All computations were performed with Shellfish, http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php.

Finally, we applied an outlier detection algorithm (*aberrant* [14], with the lambda parameter set to 20), to isolate the largest cluster of samples from the rest, based on the two leading PCs. In UK Biobank, the largest cluster is composed of individuals with European ancestry.

## A2 Testing for batch effects

The interim UK Biobank data release consists of 33 batches: there are 11 UK BiLEVE batches labeled b1, ..., b11 and 22 UK Biobank batches labeled b001, ..., b022. To perform a batch effect test, we compared the genotype counts in one batch to the genotype counts in other batches combined, using Fisher's exact test. For concreteness and for a specific probeset, we write b1 + b2 to mean ($n_{AA:b1} + n_{AA:b2}$, $n_{AB:b1} + n_{AB:b2}$, $n_{BB:b1} + n_{BB:b2}$) where $n_{AA:b1}$ is the number of called AA genotypes in batch b1. It is straightforward to generalise this notation to aggregate the genotype counts in multiple batches. Furthermore, after the initial, batch-specific QC by Affymetrix, all the calls in a batch might be set to missing, e.g., it might be case that $n_{AA:b1} = n_{AB:b1} = n_{BB:b1} = 0$.

We used a two-test approach to check for calling consistency between the UK BiLEVE and UK Biobank batches. Suppose that we want to check that the genotypes in UK Biobank batch b001, for a specific probeset, are consistent with the genotypes in the other 32 batches, for the same probeset.

- Use Fisher's exact test to compare b001 to b002 + … + b022, i.e., check for batch

effects within the UK Biobank batches.
- Use Fisher's exact test to compare b001 to b1 + … + b11, i.e., check for batch effects across the UK BiLEVE and UK Biobank batches.

We performed the second test (the comparison across the two arrays) only for probesets that uniquely genotype a SNP. (There are SNPs that are genotyped using multiple probesets for which Affymetrix recommended, separately for each batch, the best probeset to genotype the SNP.) If the p-values from the tests performed are smaller than the significance threshold used throughout, $10^{-12}$, then the calls - in batch b001 in the example above - are set to missing.

## A3 Principal components analysis of UK Biobank samples

We characterised population structure unique to UK Biobank using PCA. First we selected a subset of SNPs from those that passed all QC filters in 33 out of 33 batches, using the following criteria:

- Minor allele frequency ≥ 2.5% and missingness ≤ 1.5%. (Checking that HWE holds in a subset of samples with European descent was part of the SNP QC procedures.)
- Pairwise $r^2 \leq 0.1$, to exclude SNPs in high LD.
- Removed C/G and A/T SNPs to avoid unresolvable strand mismatches.
- Excluded SNPs in several regions with long-range LD [15]. (The list includes the MHC and 22 other regions.)

We also removed samples who were related to multiple other samples (to the 1$^{st}$, 2$^{nd}$ or 3$^{rd}$ degree), one sample from each remaining related pair (chosen randomly), as well as removing all twins and gender mismatches and samples with a high missing rate. These filters resulted in 101,284 SNPs for 141,070 samples. We used flashPCA [10] rather than Shellfish to compute loadings and principal components, because flashPCA – which uses an efficient randomised algorithm – is more scalable. Finally, in this computation, it is important to use only SNPs successfully genotyped in all batches; otherwise, differential patterns of missingness across batches mean that the major PCs will distinguish between batches, not between groups with distinct ancestry.

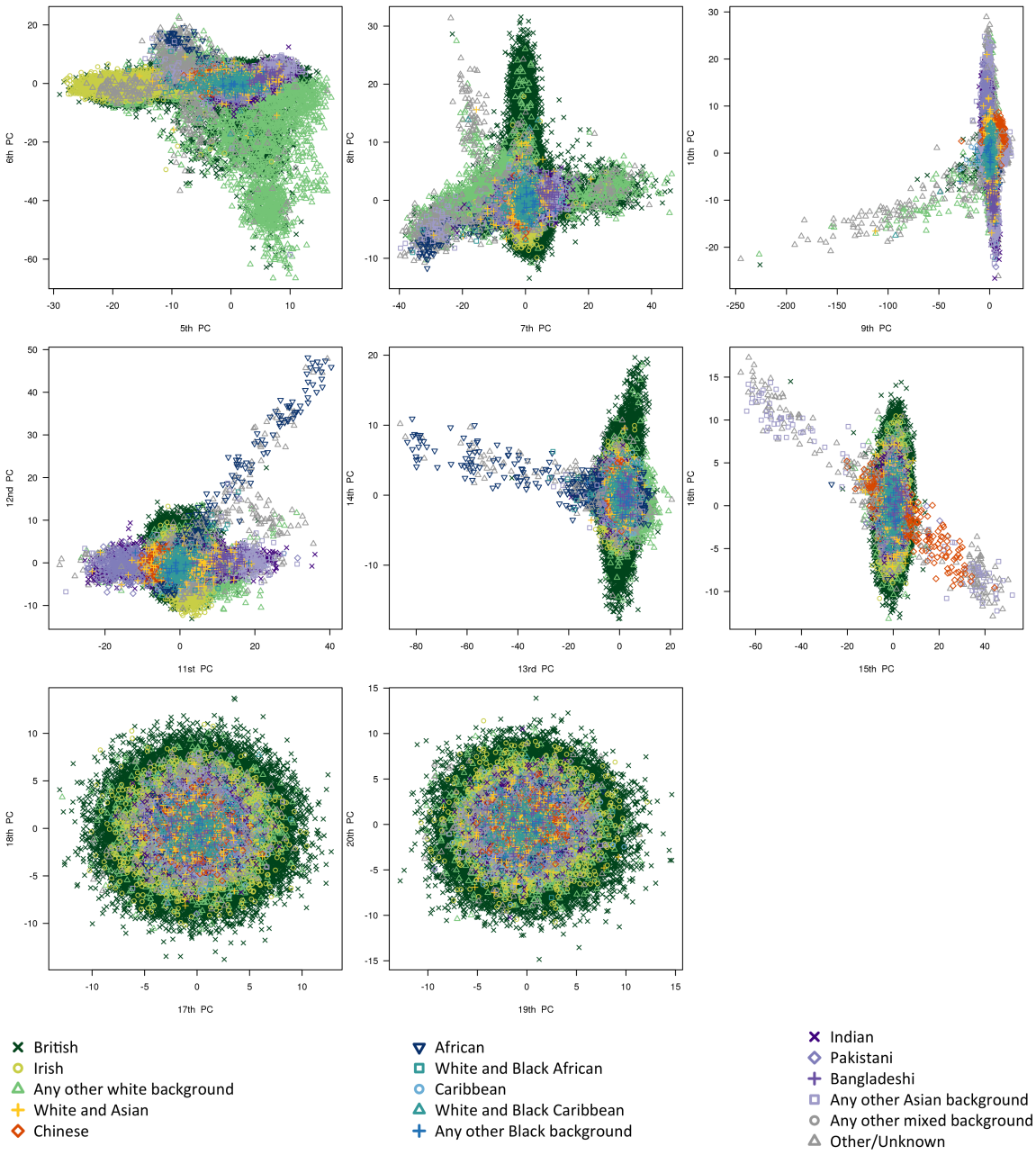


**Figure S1** Genetic principal components in UK Biobank. This figure shows principal components PC5 to PC20, and it complements Figure 2, which shows principal components PC1 to PC4. PCs are plotted in pairs, from PC5 and PC6 in the top left panel, to PC19 and PC20 in the last panel on the 3rd row. In each panel, samples are coloured by self-reported ethnicity, using the same coloured symbols as in Figure 2. The later principal components (PC16 to PC20) do not appear to distinguish any subsets in UK Biobank and only PC1 to PC15 are reported as part of the interim release.

## A4 Accounting for the heterozygosity bias explained by population structure

Heterozygosity (computed from either autosomal or X-chromosome SNPs) is sensitive to population structure because of ascertainment bias: a majority of SNPs on the UK Biobank Axiom array were chosen to satisfy certain properties – imputation coverage,

for example – in European populations. Here we describe the details of a regression model to adjust heterozygosity by accounting for the effects of population structure.

Let $h$ denote the heterozygosity and let $x$ be a set of features correlated with ancestry. We used the projections onto the four major UK Biobank principal components to characterise ancestry, writing $x = (x_1, x_2, x_3, x_4)$ for these four principal component values. Consider the following model for heterozygosity under population structure:

$$h(x) = h_0 + \beta(x)$$

where $h(x)$ is the raw heterozygosity, which depends on the features $x$, $h_0$ is the ancestry-adjusted heterozygosity and $\beta(x)$ is a bias term due to population structure. We chose a quadratic form for $\beta(x)$, which includes all linear and quadratic terms $x_i$ and $x_i^2$ as well as all cross terms $x_i x_j$, and we estimated $h_0$ with ordinary least squares. More specifically, the bias was assumed to have the following functional form:

$$\beta(x) = \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{33}x_3^2 + \beta_{44}x_4^2 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{14}x_1x_4 + \beta_{23}x_2x_3 + \beta_{24}x_2x_4 + \beta_{34}x_3x_4 .$$

The fitted value $\hat{h}_0$ is the ancestry-corrected heterozygosity, plotted on the $y$-axis in Figure 3B (all ethnicities combined) and in Figure S2 (each predefined ethnic group separately).

### A5 Detecting long runs of homozygosity

We used *plink* [12] to detect long ROHs (runs of homozygous genotypes), using the `homozyg-kb` command with a homozygous run required to span at least 1000 kb distance.

### A6 Detecting familial relationships

To detect relatedness among UK Biobank individuals, we used the robust kinship coefficient estimator implemented in KING [13]. This estimator is robust to population structure and computationally practicable even on the scale of the UK Biobank cohort. On the other hand, it is not reliable for samples with high heterozygosity or high missing rate, and a single poorly genotyped individual could lead to a cluster of inflated relationships [13]. Therefore, to minimise false positives in the detection of related samples we excluded individuals using the following filters:

1. Individuals with self-reported 'mixed' ethnicity (which tends to increase heterozygosity) were excluded from the kinship inference. That is, individuals in one of the following categories of self-reported ethnic background (~700 individuals):

- Any other mixed background
- Mixed
- White and Asian
- White and Black African
- White and Black Caribbean

2. After inferring pairs that are related to 3rd degree or closer, we excluded pairs for which at least one of the pair had either of the following properties (~800 individuals):

- Heterozygosity (PC-adjusted) > 0.1951154 (equivalent to 1.28 standard deviations from the mean)
- Missing rate > 0.02

For every individual a flag has been provided which indicates whether they have been excluded from kinship inference.

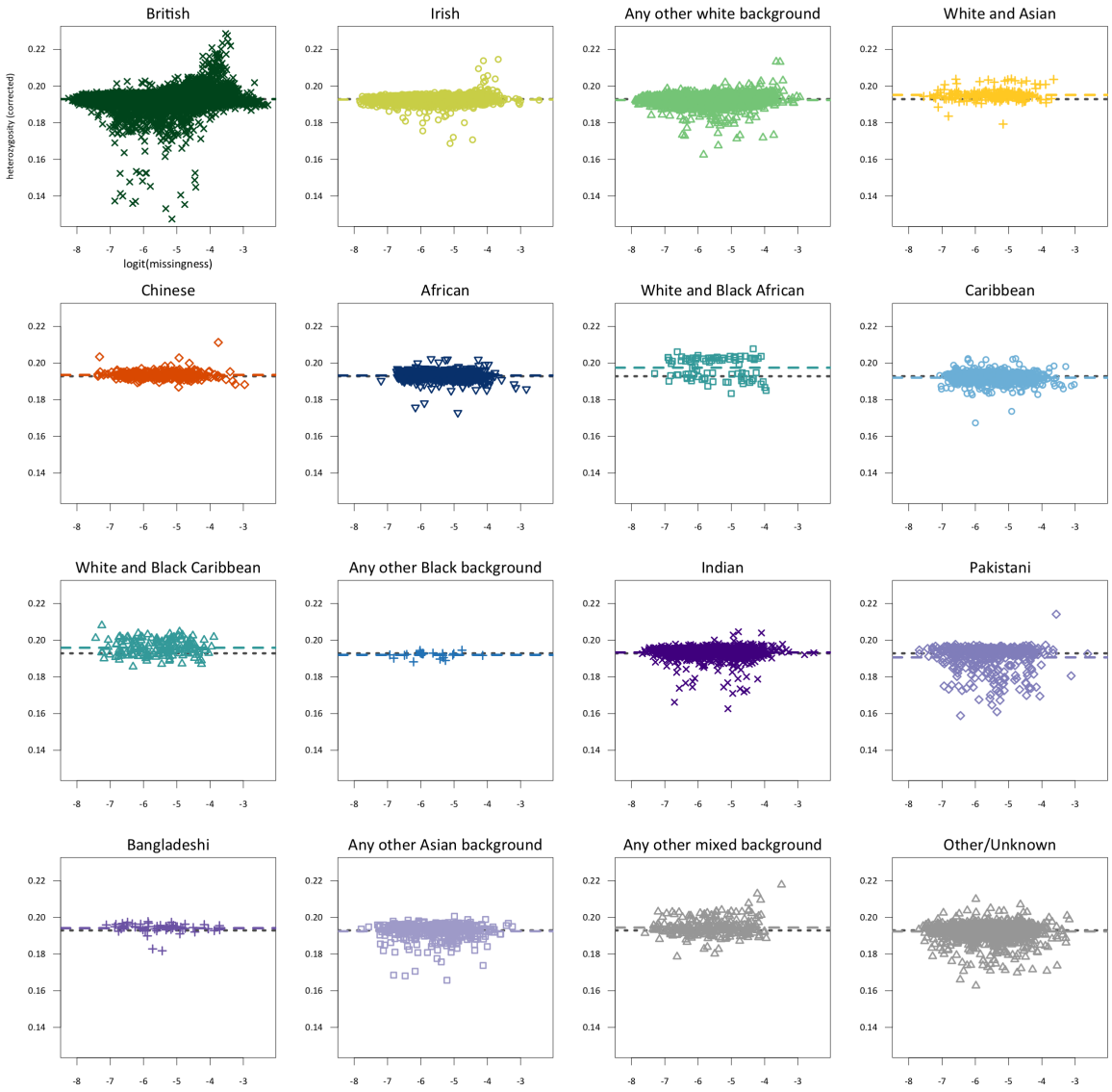


**Figure S2** Ancestry-corrected heterozygosity and missingness, for each predefined ethnic group in UK Biobank. The axes are the same in every panel: heterozygosity after correcting for bias due to population structure on the *y*-axis, and logit-transformed missingness on the *x*-axis. The logit function is defined as logit($x$) = log($x$/(1-$x$)). The coloured symbols for each ethnicity are those used in the legend of Figure 3 (and throughout this document). In all panels, the black dotted line indicates the overall mean heterozygosity; in each panel, the coloured dashed line indicates the mean heterozygosity for the respective ethnicity. The individuals with mixed ancestry (particularly, those who self-identified as "White and Black African" or "White and Black Caribbean") tend to have increased heterozygosity, even after correcting the bias due to population structure.