






Comprehensive whole-genome analyses of the UK Biobank reveal significant sex differences in both genotype missingness and allele frequency on the X chromosome

Desmond Zeya Chen ¹, Delnaz Roshandel ¹, Zhong Wang ², Lei Sun ^{3,4}, Andrew D. Paterson ^{1,4,5,*}

¹Program in Genetics and Genome Biology, The Hospital for Sick Children, 686 Bay Street, Toronto, ON M5G 1X8, Canada

²Department of Statistics and Data Science, Faculty of Science, National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore

³Department of Statistical Science, Faculty of Arts and Science, University of Toronto, 700 University Ave., Toronto, ON M5G 1Z5, Canada

⁴Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, ON M5T 3M7, Canada

⁵Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, ON M5T 3M7, Canada

*Corresponding author. Program in Genetics and Genome Biology, The Hospital for Sick Children, 686 Bay Street, Toronto, ON M5G 1X8, Canada.

E-mail: andrew.paterson@sickkids.ca

Abstract

The UK Biobank is the most used dataset for genome-wide association studies (GWAS). GWAS of sex, essentially sex differences in minor allele frequencies (sdMAF), has identified autosomal SNPs with significant sdMAF, including in the UK Biobank, but the X chromosome was excluded. Our recent report identified multiple regions on the X chromosome with significant sdMAF, using short-read sequencing of other datasets. We performed a whole genome sdMAF analysis, with ~410 k white British individuals from the UK Biobank, using array genotyped, imputed or exome sequencing data. We observed marked sdMAF on the X chromosome, particularly at the boundaries between the pseudo-autosomal regions (PAR) and the non-PAR (NPR), as well as throughout the NPR, consistent with our earlier report. A small fraction of autosomal SNPs also showed significant sdMAF. Using the centrally imputed data, which relied mostly on low-coverage whole genome sequence, resulted in 2.1% of NPR SNPs with significant sdMAF. The whole exome sequencing also displays sdMAF on the X chromosome, including some NPR SNPs with heterozygous genotype calls in males. Genotyping, sequencing and imputation of X chromosomal SNPs requires further attention to ensure the integrity for downstream association analysis.

Keywords: GWAS; Sex; association; genotyping; X chromosome

Introduction

Quality control (QC) of data for large-scale genome-wide association studies (GWAS) is crucial to control false positive and negative results, with numerous recommendations being provided [1–4]. However, the specific steps and thresholds are rarely justified and may vary by technology. As part of QC procedures, some studies have examined whether there are SNPs with significant sex differences in minor allele frequency (sdMAF) on the autosomes [5–9], effectively a GWAS of sex, with the goal of identifying loci subject to sex selection. Boraska [6] used GWAS data from ~115 k individuals of European ancestry from 51 studies genotyped with a range of different arrays, with imputation to HapMap phase II, for autosomes and the X chromosome, including the pseudo-autosomal regions (PAR). However, they did not identify any variants with genome-wide significant sdMAF. In addition, only a few of the studies had PAR SNPs, which would result in low power.

A second study examined individuals of European and African ancestry from various GWAS arrays and identified loci on chromosomes 6 and 14 with genome-wide significant sdMAF [8].

Thirdly, Ryu [5] used Affymetrix Genome-Wide Human SNP Array 5.0 genotype data with imputation to HapMap 3 data from ~9 k Koreans for autosomal SNPs, and identified 9 loci with genome-wide significant sdMAF. Two of these loci were replicated in an independent dataset [5], however this work has been criticized since some of these variants map to multiple locations in the genome [7].

Autosomal GWAS of sex were performed separately in BioVU (n = 93 864) and the UK Biobank using array-genotyped variants [9]. The authors did not analyse the centrally imputed data, which is the most analysed data of the UK Biobank, and they also excluded the X chromosome. They performed BLAT analysis using the probe sequences against the genome and identified some autosomal probes which cross-hybridized with the sex chromosomes.

More recently, 23andMe performed an autosomal GWAS of sex in 2.4 M subjects of European ancestry genotyped on five different versions of Illumina arrays and imputed using a combined reference panel of the 1000 Genomes Project phase 3 and UK10K data

Received: July 24, 2023. Revised: November 13, 2023. Accepted: November 14, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

[7]. They identified 158 independent autosomal loci with sdMAF. Even after attempting to exclude sex chromosome interference using bioinformatic approaches as well as other quality checks, 49 out of 97 autosomal loci remained. They suggested that selection bias may be contributing to this phenomenon. However, again, the X chromosome was not examined.

The X chromosome has been excluded from most GWAS [10, 11]. Further, it has been suggested that different QC processes are necessary for X-chromosomal data compared to the autosomes [12], and evolutionary forces may result in sex differences in allele frequencies on the X chromosome [13]. We [14] previously showed that an unexpectedly large proportion of biallelic SNPs on the X chromosome have significant sdMAF in both the 1000 Genomes project data (both phase 3 [15] and the high coverage aligned to GRCh38 [16]), as well as in the gnomAD v 3.1.2 genomes [14, 17]. Since these datasets [15–17] are mostly based on short-read next-generation sequencing data, we wanted to examine the same phenomenon in GWAS array data from a large population-based cohort to determine if there are similar patterns of sdMAF.

The UK Biobank is one of the largest publicly available genetic datasets [18], and therefore has been the focus of GWAS of thousands of traits and diseases [19, 20] (<http://www.nealelab.is-uk-biobank>). The UK Biobank centrally performed imputation using the Haplotype Reference Consortium [21], UK10K [22], and 1000 Genomes project [18] reference panels. This centrally imputed data has been commonly used for GWAS by the community, contributing to many recent meta-GWAS [23], as well as for polygenic risk score assessment [24]. Additionally, the array genotype data has been used as the basis for imputation using exome sequencing from a subset of participants [25], as well as high coverage whole genome sequencing from both the TOPMed program [26] and the 100 000 Genomes project from Genomics England. Since the array data has been the basis of many analyses, it is important to further analyse the quality. In addition, whole exome sequencing (WES) has been generated on participants, [27] from which the majority of variants were also found in whole genome sequencing in a subset of participants [28]. Sex differences in large-scale exome sequence have not been reported previously, thus we analysed this data as well.

Results

UK Biobank Axiom array genotypes

There were 165 737 males and 199 079 females with array genotype data. [Supplementary Table 1](#) provides the total number of genotyped SNPs per chromosome (and within the X chromosome separately by PARs and X-transposed region (XTR)), as well as SNPs that: failed UK Biobank centralized QC, were not biallelic, monomorphic, with minor allele count (MAC) < 10, and those that remained for the analysis. In total, 710 601 variants were analysed across the nuclear genome.

We first examined whether there was sex difference in the genotyping missing rates (sdMISSING). [Supplementary Fig. 1](#) shows the p value of the test, along with the direction and magnitude of the missingness on the X chromosome. In general, females have higher missing rates than males across the NPR. There was no significant sex difference in missingness in PAR1 or PAR2. There are far fewer autosomal ([Supplementary Fig. 2](#)) than X chromosomal SNPs with significant sex difference in missingness. The pattern of sex difference in missingness on the autosomes appears relatively sporadic with little clustering, in comparison to a Manhattan plot from a typical association study. In contrast to the X chromosome, the missingness on the autosomes is smaller in magnitude and does not have strong directional bias.

[Supplementary Fig. 3](#) shows the UK Biobank centrally determined batch passing rate for genotyped SNPs in PAR1 and PAR2. Of note, SNPs close to the boundaries with NPR show a higher rate of failing one or more of the QC metrics at both regions.

[Figure 1](#) shows the X chromosome sdMAF Manhattan plot, along with detailed views of PAR1 and PAR2. Significant sdMAF is present at the PAR1/NPR boundary as well as throughout PAR2, and is also scattered throughout the NPR. The XTR (also called PAR3, indicated in grey, around 90 Mb) behaves similarly to NPR. These findings are consistent with previous sdMAF observations in the 1000 Genomes Project and gnomAD v3.1.2 data, which were based mostly on short-read next-generation sequencing [14]. [Supplementary Fig. 4](#) provides histograms of sdMAF p values stratified by region, indicating an excess of small p values across all regions. [Supplementary Fig. 5](#) shows the direction and magnitude of sdMAF across the X chromosome.

[Figure 2](#) is autosomal sdMAF Manhattan plot, showing occasional isolated SNPs with significant sdMAF. [Supplementary Fig. 6](#) shows histograms of the P values, stratified by chromosome, indicating slight inflation on most chromosomes, but nowhere as marked as the inflation observed on the X chromosome. [Supplementary Fig. 7](#) shows the corresponding QQ plot for all autosomes stratified by MAF and shows some modest inflation (genomic control $\lambda = 1.067$). On the X chromosome, we separately examined whether the QQ plot differed by MAF, and observed no striking difference ([Supplementary Fig. 8](#)). However, the PARs account for only a small proportion of X, so we separately examined the relationship MAF and sdMAF p value by region ([Supplementary Fig. 9](#)) for SNPs with sdMAF $P < 5 \times 10^{-8}$. This shows that SNPs with sdMAF in PAR1 and PAR2 have a wide range of MAF, while those in NPR tend to have low MAF.

We examined whether these autosomal variants with significant sdMAF overlap with those previously reported [7, 14]. In general they did not, with only one, rs11710967 (chr 3: 16 640 075), overlapping with one region previously identified by the 23andMe study [7], and one (rs76712070, chr1: 243 050 0615) overlapping with a region previously identified in the 1000 Genomes Project phase 3 data [14].

For genotyped mitochondrial variants there were none that met genome-wide significance criteria for sdMAF ([Supplementary Data 1](#)).

Centrally imputed data

There were 165 389 males and 198 709 females in this analysis. [Supplementary Table 2](#) provides the counts of imputed SNPs on autosomes and the X chromosome. Isolated SNPs have sdMAF on chromosomes 1, 5, 11, 17 and 18 with [Supplementary Fig. 10](#). Other autosomes do not contain variants with significant sdMAF.

sdMAF on the X chromosome for the imputed data is shown in [Fig. 3](#) ([Supplementary Fig. 11](#) provides histograms of p values). These results show significant sdMAF throughout the X chromosome with clustering at the PAR1/NPR boundary and throughout PAR2, as well as around the centromere among other regions. [Supplementary Fig. 12](#) has the direction and magnitude of the sdMAF, with higher female MAF throughout, especially marked flanking the centromere.

For imputed HLA haplotypes, none showed genome-wide significance sex difference in haplotype frequency ([Supplementary Data 2](#)).

Whole exome sequencing

There were 180 852 males and 212 611 females. [Supplementary Table 3](#) has counts of variants by chromosome. There is a large proportion of X chromosome variants with significant sex

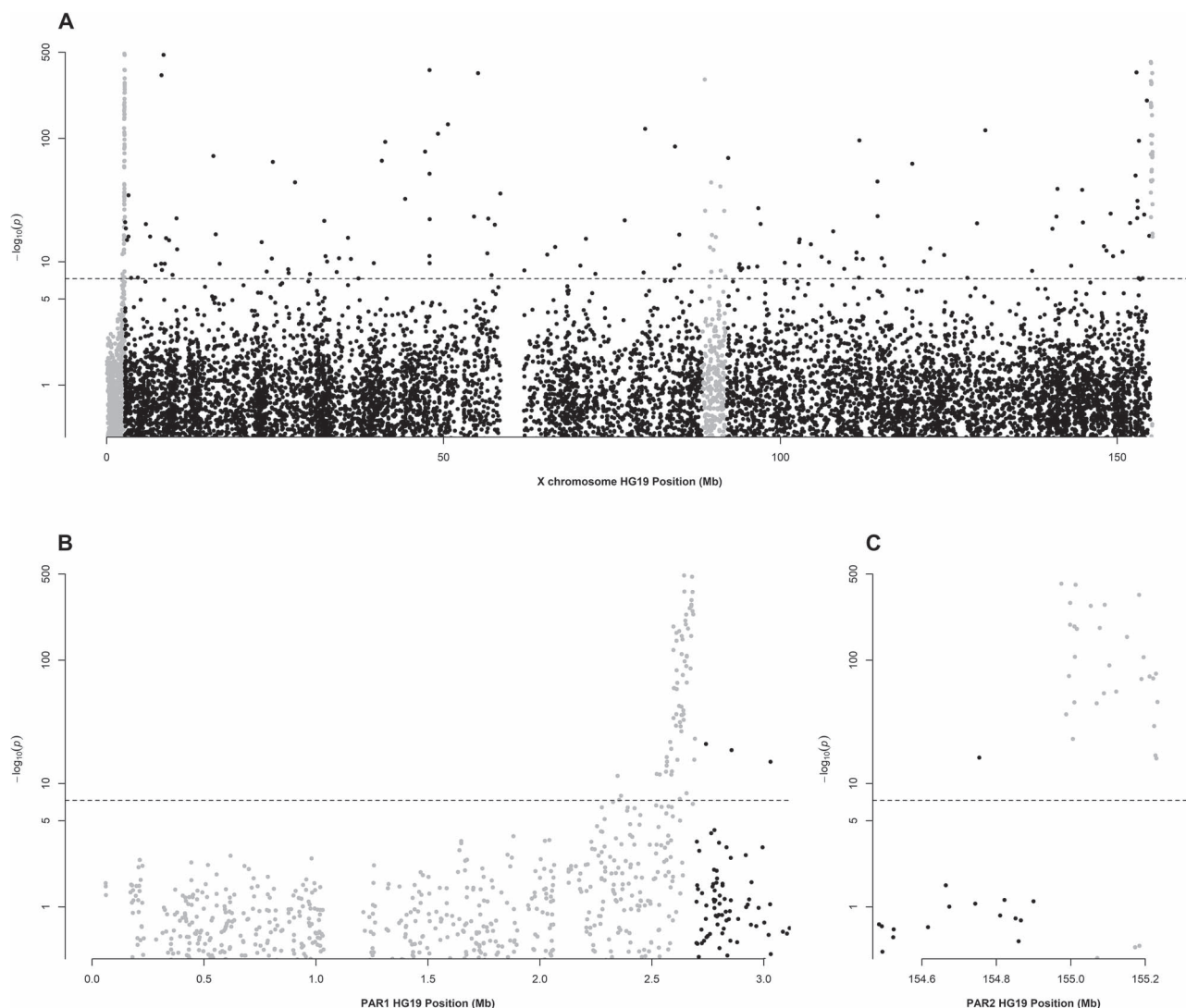


Figure 1. X chromosome sdMAF Manhattan plot from UK biobank axion array. (A) Manhattan plot of X chromosome sdMAF results in UK biobank; Y-axis is the P-value; X-axis is the physical position in build 37. The dashed line is the genome-wide significant threshold $5E-8$. Regions marked in grey from left to right are respectively PAR1, PAR3, and PAR2, while region in black is NPR. (B) Regional Manhattan plot of PAR1 with X-axis limited from 0 to 3 Mb; grey region is PAR1, and black region is NPR. (C) Regional Manhattan plot of PAR2 with X-axis limited from 154.6 to 155.2 Mb; grey region is PAR2, and black region is NPR.

difference in missingness (Supplementary Fig. 13) across the NPR. There are few variants in PAR2 in the exome sequence data, and for both them and those in PAR1 there is no significant difference in the missingness by sex. Compared to the X chromosome, much fewer autosomal variants have significant sex difference in missingness, even though four SNPs on chromosomes 11 and 16 have $P < 10^{-100}$ (Supplementary Fig. 14). These SNPs have higher missing rate in females than males. Additionally, there is a notable region around the centromere of chromosome 2 with significant sex differences in missingness, with different SNPs having higher and lower missingness by sex.

Next, we examined the count of unexpected male heterozygotes per SNP on the X chromosome NPR. This revealed hundreds of SNPs with high male heterozygote counts, especially around 52.8 Mb where nearly all males are heterozygous at multiple SNPs. Other regions include 50 Mb, 136.9 Mb and 142 Mb, where > 50 000 males have unexpected heterozygote calls (Supplementary Fig. 15). These male heterozygote calls in NPR clearly indicate genotyping error.

Examining sdMAF on the chromosome X: (Fig. 4) shows regional features, typically clustered within genes. There is marked sdMAF for variants around the PAR1/NPR boundary, similar to the GWAS array data. In contrast, compared with the GWAS array data, only part of one gene (exon 3 of *SPRY3*) in PAR2 has SNPs that passed QC in WES, of which only one SNP has significant sdMAF. None of the other six PAR2 genes have variants that are captured and/or pass QC in the WES.

Analysis of sdMAF on the autosomes (Fig. 5), indicates that there is a region with sdMAF around the centromere of chromosome 2 that overlaps with the sex difference in missingness (e.g. GRCh38 2:95 935 488:G:A, missing $P < 10^{-323}$, sdMAF $P = 10^{-480}$). There are also isolated SNPs with sdMAF on other chromosomes.

We then examined whether sdMAF on the X chromosome was related to difficulty of sequencing. Regions that have been designated as difficult to sequence by the Genome in a Bottle consortium are indicated by direction and magnitude of sdMAF (Supplementary Fig. 16), many of the variants with larger sdMAF effect sizes are located in such regions.

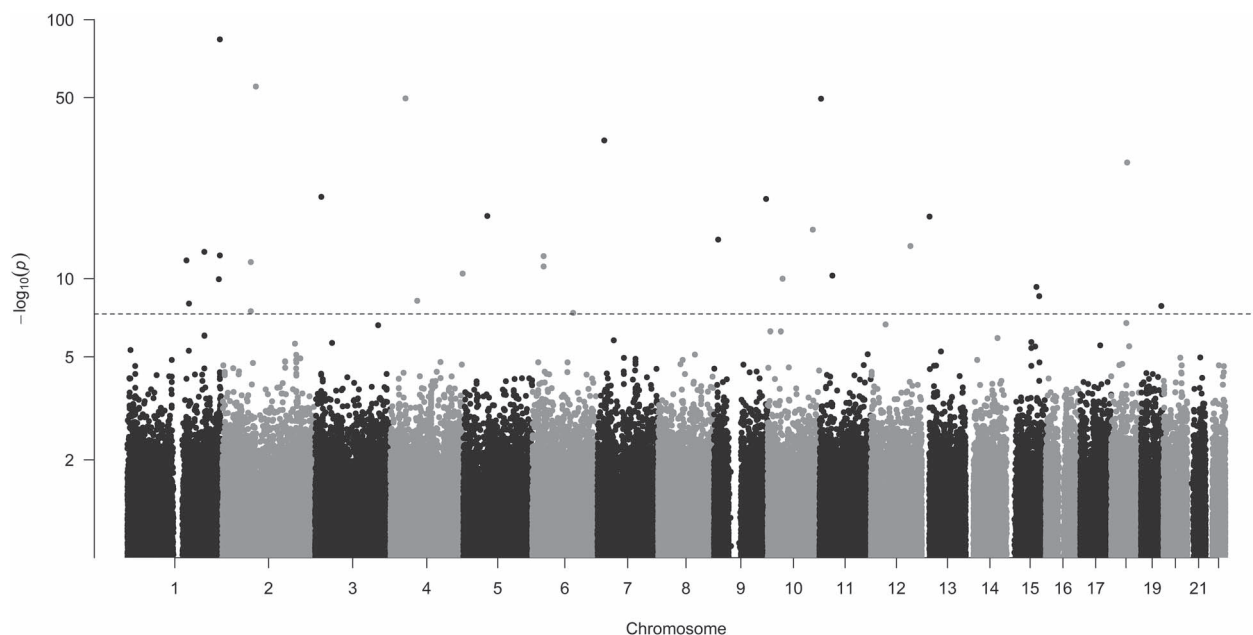


Figure 2. Autosomal sdMAF Manhattan plot from UK biobank axion array. Manhattan plot of sdMAF results in UK biobank axion array genotype data; Y-axis is the P-value; X-axis is the physical position in build 37. The dashed line is the genome-wide significant threshold 5E-8.

In WES, we examined whether there was a relationship between sdMAF and missing rate on the X chromosome NPR. We defined a binary outcome for sdMAF based on sdMAF $P < 5 \times 10^{-8}$ threshold. There were 289 variants with significant sdMAF, compared to 42525 without. We fitted three separate logistic regression models with predictors as missing rate in: 1; male; 2; female; 3; both sexes. Prior to the model fit, we first pruned the variants for LD based on the female data. Male and female missing rates are significantly positively associated with sdMAF, respectively in Models 1 and 2, but when both sexes are included in Model 3, only the male missing rate is significantly positively associated with the presence of sdMAF (Table 1).

Discussion

We examined the UK Biobank from the perspective of sdMAF, essentially a GWAS of sex, using three data types: array genotypes, centrally imputed data, as well as whole exome sequencing. First, we examined whether there was sex difference in genotyping call rate, and we identified major differences on the X chromosome from both the array and WES (Supplementary Figs 1 and 13) with typically lower call rate in males than females. As for the autosomes, there were also sex differences in missingness across the autosomes in the array data (Supplementary Fig. 2), while in the WES only one autosomal region (chromosome 2) had sex differences in missingness (Supplementary Fig. 14).

Next we examined sdMAF, separately for the 3 different data types. Among them, the imputed data had the most marked sdMAF on the X chromosome (Fig. 3), even when only analysing those variants with high imputation quality. This is important since imputed data is the most widely used data type of the UK Biobank. The imputation resources used by the UK Biobank are mostly based on low-coverage whole genome sequence, with either ensemble variant calling (1000 Genomes Project phase 3 [15]) or methods that are not tailored for NPR X chromosomal variants (UK10K and HRC [21]). We previously showed that some of the sdMAF observed on the X chromosome in phase 3 of the 1000

Genomes Project was resolved by use of the high coverage whole genome sequence data [14]. These observations indicate that the UK Biobank centrally imputed X chromosome data should be re-evaluated using high-coverage whole genome sequence data as the reference panel [26].

Our sdMAF analysis of the UK Biobank data is considerably different from that previously performed [9]. Although both used data from the UK Biobank, they ignored the X chromosome, whereas we observed the most striking findings on the X chromosome. They analysed only the array genotype data, but we additionally examined the imputed data (which is most used by researchers), as well as the whole exome sequence. Finally, they included data from both arrays used by the UK Biobank, while we restricted our analysis to the array that was used to genotype most of the individuals (UK Biobank Axiom array), since we observed that the missing patterns differ between the two arrays (iBiLEVE had 50.1% males, compared to 45.4% on the Axiom array).

Earlier reports have suggested that sdMAF could have either biological (true sdMAF) or bioinformatic (spurious sdMAF) causes [7, 14]. SNPs with true sdMAF, especially at the PAR1/NPR and PAR2/NPR boundaries, may have failed QC in the UK Biobank Axiom array data, since SNPs were tested for sex difference in genotype/allele frequency in batches, and SNPs meeting such threshold in any batch were then excluded from the cleaned data [18]. Such variants would have been excluded from a subset of individuals, which then could impact the imputation. Unfortunately, since the UK Biobank only provided a binary indicator as to whether a SNP failed any one of multiple QC tests, we could not investigate this further.

On the other hand, SNPs with spurious sdMAF may have been missed by the central UK Biobank analysis, as their sdMAF test was performed (a): separately in each of the 106 batches of ~4700 samples/batch, and (b) using the conservative Fisher's exact test, and without adjusting for Hardy-Weinberg disequilibrium. Thus, a variant that did not meet the threshold for exclusion in any individual batch could still have significant sdMAF in the whole

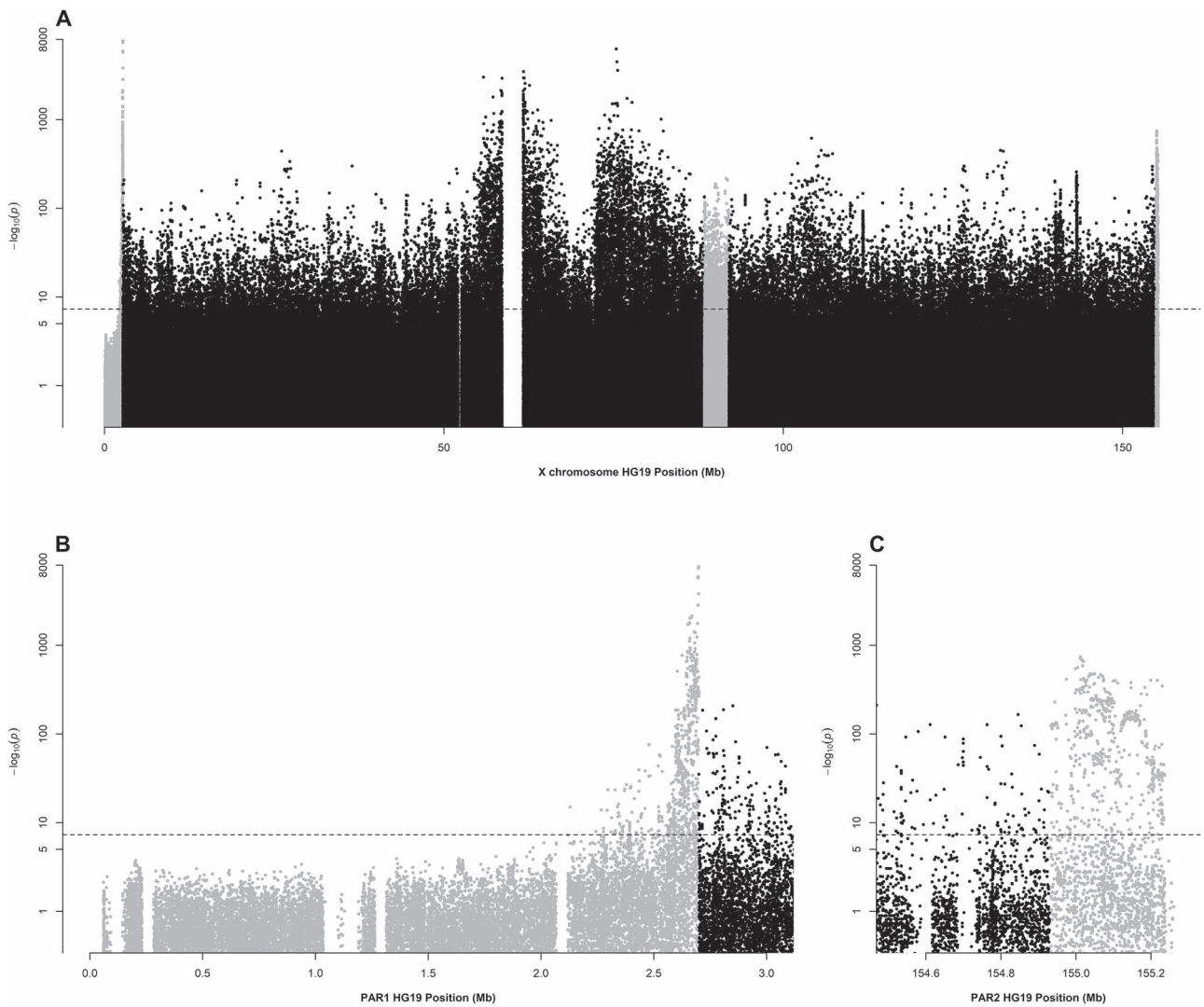


Figure 3. X chromosome sdMAF Manhattan plot from imputed UK biobank. (A) Manhattan plot of X chromosome sdMAF results in array imputed data; Y-axis is the P-value; X-axis is the physical position in build 37. The dashed line is the genome-wide significant threshold 5×10^{-8} . Regions marked in grey from left to right are respectively PAR1, PAR3, and PAR2, while region in black is NPR. (B) Regional Manhattan plot of PAR1 with X-axis limited from 0 to 3 Mb; grey region is PAR1, and black region is NPR. (C) Regional Manhattan plot of PAR2 with X-axis limited from 154.6 to 155.2 Mb; grey region is PAR2, and black region is NPR.

dataset. Indeed, we observed variants across the autosomes and X chromosome with genome-wide significant sdMAF (Figs 1 and 2).

The UK Biobank had a participation rate of 5.5% at baseline [29]: females were more likely to participate than males, as were older and healthier individuals. Although participation bias has been suggested as a cause of sdMAF [7], ironically much of the earlier work omitted the X chromosome [6–9]. Here we show that the majority of sdMAF is located on the X chromosome, although untangling the exact causes (selection bias, genotyping error, sex-specific biology) for each variant remain an open question.

Here we limited our analysis to the white British population in the UK Biobank: the sample sizes for other ancestries are considerably smaller and therefore power could be low. We did not include PCs as covariates in our sdMAF analysis, as unlike GWAS, unadjusted population structure does not lead to false positives [30]. In addition, assessment centre and age at baseline could be additional factors that could be adjusted for.

We also limited analysis to variants with $MAC > 10$ (i.e. $MAF \sim 2 \times 10^{-5}$) to ensure reliability of the sdMAF test. Quality of rare variants genotyped on the UK Biobank array has been examined

previously [31]. For rare variants that change predicted protein sequence and were also present in ClinVar, the authors found a higher rate of likely genotyping error on the X chromosome than the autosomes (45/49 vs 2883/4123, Fisher's exact test $P = 4 \times 10^{-4}$), consistent with our sdMAF findings.

X chromosome results are generally missing from GWAS [10, 11]. Combating this exclusion phenomenon, first and foremost, requires good quality data to ensure the integrity of downstream analysis and interpretation. Thus, in addition to addressing the well-documented challenges in the association analysis of the X chromosome [12, 32, 33], more attention is required to identify and understand the causes of sex differences in genotype missingness and allele frequency that are present on the X chromosome.

Materials and methods

This work was part of ethics approval: HSC #1000073707. We used three different genetic datasets from the UK Biobank to examine sdMAF: UK Biobank Axiom array, centrally imputed data, and whole exome sequencing.

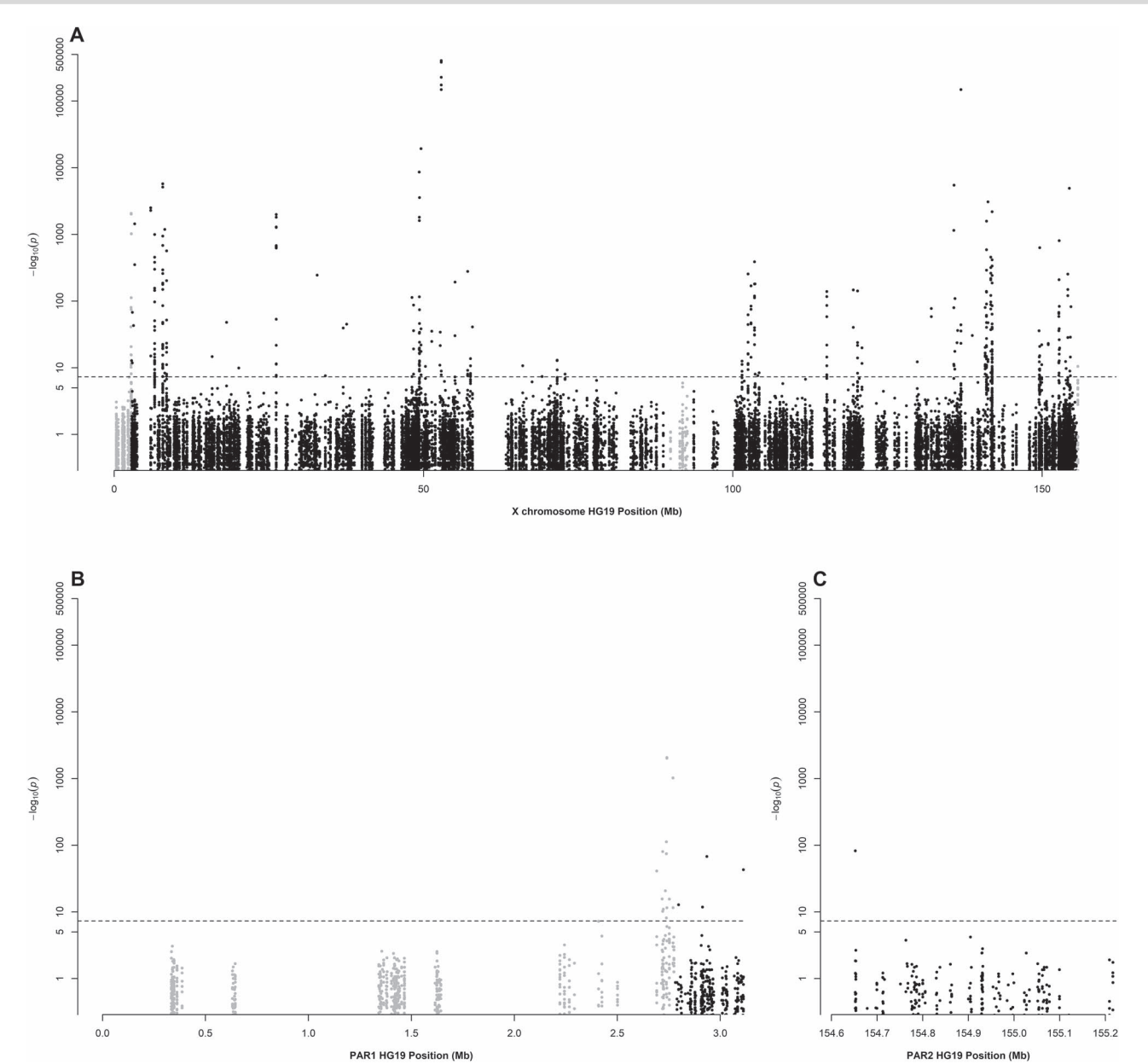


Figure 4. X chromosome sdMAF Manhattan plot from UK biobank whole exome sequence. (A) Manhattan plot of X chromosome sdMAF results; Y-axis is the P-value; X-axis is the physical position in build 38. The dashed line is the genome-wide significant threshold 5E-8. Regions marked in grey from left to right are respectively PAR1, PAR3, and PAR2, while region in black is NPR. (B) Regional Manhattan plot of PAR1 with X-axis limited from 0 to 3 Mb; grey region is PAR1, and black region is NPR. (C) Regional Manhattan plot of PAR2 with X-axis limited from 155 to 155.8 Mb; grey region is PAR2, and black region is NPR.

Table 1. Association between sex-specific missing rate and presence of significant sdMAF on the X chromosome NPR in whole exome sequence.

Model	Missing rate Predictor	Beta (SE)	P value
1	Male	3.9 E-4 (0.065)	< 2E-16
2	Female	3.8 E-4 (2.6 E-5)	< 2E-16
3	Male	3.6 E-4 (4.1 E-5)	< 2E-16
	Female	3.3 E-5 (4.5 E-5)	0.47

Study and subjects

The white British (Field=21 000) ethnic background and “Caucasian” (Field=22 006) genetic grouped individuals from the UK Biobank [18] were used since they are the largest population subgroup, and the most often used sample for GWAS. Individuals with sex chromosome abnormalities (Field 22 019) were excluded from all analyses.

UK biobank axiom array genotypes

Participants were genotyped using either the UK BiLEVE Axiom or the UK Biobank Axiom arrays (Field 22 000). Since the majority (~90%) were genotyped on the UK Biobank Axiom array, only that data was used. The original genotype calls (Category 100 315, Field 22 418) were used, and Batch Fail QC indicator was obtained from Category 263, Resource 1955.

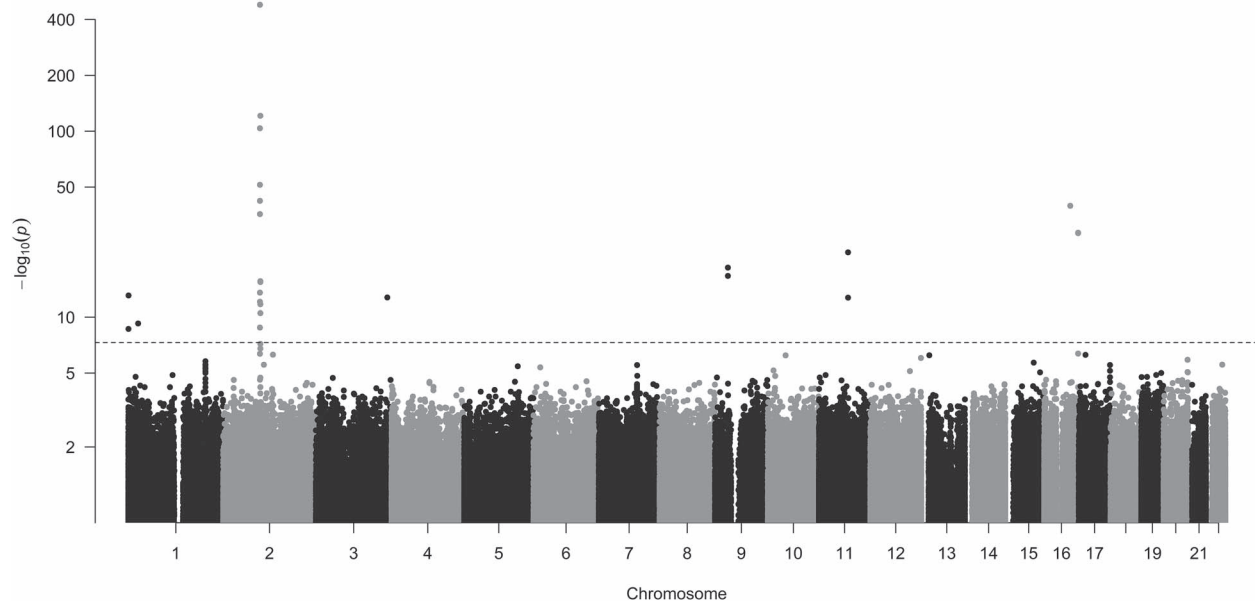


Figure 5. Autosomal sdMAF Manhattan plot from UK biobank whole exome sequence. Y-axis is the P-value; X-axis is the physical position in build 38. The dashed line is the genome-wide significant threshold 5E-8.

A minimum sex-combined allele count (MAC) of 10 was required.

The UK Biobank performed centralized quality control of the array genotyping data [18]. Specifically, samples were assigned to plates to ensure that samples of both sexes (among other factors) were present on each plate. An extensive QC process included genotype calling separately by 95 batches (Field 22 000) of ~4700 samples/batch. For SNPs in the non-pseudo-autosomal region (NPR) of the X chromosome, genotyping was performed separately by sex (see S2.3 in [18]). For each SNP they evaluated genotype frequency by sex (similar to sdMAF), Hardy-Weinberg Equilibrium (HWE), and effects of plate (94 samples) and batch, as well as examining discordance of genotypes across control replicates. Sex, plate and HWE tests were applied within each batch, and if a SNP failed any one of them based on a stringent p value threshold ($< 1 \times 10^{-12}$), then that SNP was set to missing in that batch. However, the raw results for each of these tests were not provided, instead a binary indicator for whether each SNP passed all QC criteria for each batch was available (Resource 1955).

Centrally imputed data

To determine if sdMAF also exists in the centrally imputed data, we used a subset of biallelic SNPs on the X chromosome and autosomes from the imputed data (category 100 319, Field 22 828), selected to have a sex-combined minimum MAC of 10. sdMAF is based on hard calls obtained using PLINK 2.0 with imputation information score > 0.3 . Imputed HLA variants were also examined (Field 22 182). Variant positions for array and imputation were provided on build GRCh37.

Whole exome sequencing

The data came from the UK Biobank Field 23 158 (Population level exome OQFE variants, PLINK format—Final exome release) [27, 34] in the white British and “Caucasian” subset, and were limited to variants with a sex-combined minimum MAC of 10. Exome sequence is on GRCh38. For QC, we applied the suggested

filter that $> 90\%$ of all genotypes at a variant had a read depth ≥ 10 (90pct10dp), as described previously (https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/UKB_WES_AnalysisBestPractices.pdf). Difficult to sequence regions were based on the Genome in a Bottle consortium (GIAB) v3.1 (<https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.1/GRCh38/>) on GRCh38 [35, 36].

Statistical methods

Sex difference in genotype missingness was examined using χ^2 with Yate’s continuity correction applied to the 2×2 contingency table for each variant [37]. Consistency of P value for SNPs with low missing counts was confirmed using Fisher’s exact test. The missingness tables were generated using PLINK 1.9 and 2.0 [38, 39], respectively for the X chromosome and autosomes. The sdMAF test was performed as described previously [14]. Genome-wide significant threshold was used ($P < 5 \times 10^{-8}$) [40]. To examine male heterozygote calls in NPR we changed the chromosome # to an autosome, since by default PLINK uses the location of the NPR to recode males with heterozygous calls to having missing genotypes.

Code availability Paterson-Sun-lab (<https://github.com/Paterson-Sun-Lab/sdMAF-UKB>).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 48839. This work was funded by CIHR #450360. The analyses were conducted on the Research Analysis Platform (<http://ukbiobank.dnanexus.com>).

Author contributions

L.S. and A.D.P. conceptualized the study.

L.S. and A.D.P. supervised the study and drafted the manuscript.

D. Z.C. performed the analyses and summarized the results.

D. Z.C., Z.W., and D.R. reviewed and edited the manuscript.

Supplementary data

Supplementary data is available at HMG Journal online.

Conflict of interest statement: None of the authors have conflicts of interest.

Funding

This work was funded by the Canadian Institutes for Health Research Project Grant #450360.

Online resources

<https://www.ukbiobank.ac.uk/>

Data and code

All data used are publicly available through the UK Biobank (<https://www.ukbiobank.ac.uk/>). All codes used for the analyses are available at: <https://github.com/Paterson-Sun-Lab/sdMAF-UKB/> Datasets with variant-specific results are available: https://drive.google.com/drive/folders/1g8aRXcn_3twNNzUNvpauu2QGipe1FF8D?usp=sharing

The following variable names are defined:

CHROM—Chromosome #

ID—Variant ID

A1—Reference Allele

A2—Alternative Allele

BP—Base-pair Position from genotype calls files

Mmissing—Number of missing genotypes per SNP in males

Fmissing—Number of missing genotypes per SNP in females

F_A1A1—Female Homozygous Reference Count per SNP

F_A1A2—Female Heterozygous Ref-Alt Count per SNP

F_A2A2—Female Homozygous Alternative Count per SNP

M_A1A1.A1—Male Homozygous/Haploid Reference Count per SNP

M_A1A2—Male Heterozygous Ref-Alt Count per SNP

M_A2A2.A2—Male Homozygous/Haploid Alternative Count per SNP

LOG10P—Log10 transformed P-Value of sdMAF

Ffreq—Female Alt Allele Frequency

Mfreq—Male Alt Allele Frequency

DIFmaf—Female—Male Frequency Difference of Sex Combined Minor Allele

sdMISSING—Female—Male Missing Rate Difference

Pmissing—Log10 transformed P-Value of sdMISSING computed using Chi squared test with Yate's correction

References

- Laurie CC, Doheny KF, Mirel DB. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* 2010;**34**:591–602.
- Marees AT, de Kluiver H, Stringer S. et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res* 2018;**27**:e1608.
- Anderson CA, Pettersson FH, Clarke GM. et al. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;**5**:1564–73.
- Lam M, Awasthi S, Watson HJ. et al. RICOPIIL: rapid imputation for Consortias PipeLine. *Bioinformatics* 2020;**36**:930–3.
- Ryu D, Ryu J, Lee C. Genome-wide association study reveals sex-specific selection signals against autosomal nucleotide variants. *J Hum Genet* 2016;**61**:423–6.
- Boraska V, Jerončić A, Colonna V. et al. Genome-wide meta-analysis of common variant differences between men and women. *Hum Mol Genet* 2012;**21**:4805–15.
- Pirastu N, Cordioli M, Nandakumar P. et al. Genetic analyses identify widespread sex-differential participation bias. *Nat Genet* 2021;**53**:663–71.
- Zuo L, Wang T, Lin X. et al. Sex difference of autosomal alleles in populations of European and African descent. *Genes Genomics* 2015;**37**:1007–16.
- Kasimatis KR, Abraham A, Ralph PL. et al. Evaluating human autosomal loci for sexually antagonistic viability selection in two large biobanks. *Genetics* 2021;**217**:1–10.
- Wise AL, Gyi L, Manolio TA. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* 2013;**92**:643–7.
- Sun L, Wang Z, Lu T. et al. eXclusionary: 10 years later, where are the sex chromosomes in GWASs? *Am J Hum Genet* 2023;**110**:903–12.
- König IR, Loley C, Erdmann J. et al. How to include chromosome X in your genome-wide association study. *Genet Epidemiol* 2014;**38**:97–103.
- Monteiro B, Arenas M, Prata MJ. et al. Evolutionary dynamics of the human pseudoautosomal regions. *PLoS Genet* 2021;**17**:e1009532.
- Wang Z, Sun L, Paterson AD. Major sex differences in allele frequencies for X chromosomal variants in both the 1000 genomes project and gnomAD. *PLoS Genet* 2022;**18**:e1010231.
- Auton A, Brooks LD, Durbin RM. et al. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
- Byrska-Bishop M, Evani US, Zhao X. et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* 2022;**185**:3426–3440.e19.
- Karczewski KJ, Francioli LC, Tiao G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43.
- Bycroft C, Freeman C, Petkova D. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9.
- Sinnott-Armstrong N, Tanigawa Y, Amar D. et al. Genetics of 35 blood and urine biomarkers in the UK biobank. *Nat Genet* 2021;**53**:185–94.
- Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK biobank. *Nat Genet* 2018;**50**:1593–9.
- McCarthy S, Das S, Kretschmar W. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**:1279–83.
- Huang J, Howie B, McCarthy S. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015;**6**:8111.
- Sakaue S, Kanai M, Tanigawa Y. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021;**53**:1415–24.
- Khera AV, Chaffin M, Aragam KG. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–24.
- Barton AR, Sherman MA, Mukamel RE. et al. Whole-exome imputation within UK biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* 2021;**53**:1260–9.
- Taliun D, Harris DN, Kessler MD. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* 2021;**590**:290–9.

27. Backman JD, Li AH, Marcketta A. *et al.* Exome sequencing and analysis of 454,787 UK biobank participants. *Nature* 2021;**599**: 628–34.
28. Halldorsson BV, Eggertsson HP, Moore KHS. *et al.* The sequences of 150,119 genomes in the UK biobank. *Nature* 2022;**607**: 732–40.
29. Fry A, Littlejohns TJ, Sudlow C. *et al.* Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am J Epidemiol* 2017;**186**:1026–34.
30. Wang Z, Paterson AD, Sun L. *Annals of Applied Statistics*, in press., pp. <https://arxiv.org/abs/2212.12228>.
31. Wright CF, West B, Tuke M. *et al.* Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am J Hum Genet* 2019;**104**:275–86.
32. Chen B, Craiu RV, Strug LJ. *et al.* The X factor: a robust and powerful approach to X-chromosome-inclusive whole-genome association studies. *Genet Epidemiol* 2021;**45**:694–709.
33. Keur N, Ricaño-Ponce I, Kumar V. *et al.* A systematic review of analytical methods used in genetic association analysis of the X-chromosome. *Brief Bioinform* 2022;**23**:bbac287.
34. Van Hout CV, Tachmazidou I, Backman JD. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK biobank. *Nature* 2020;**586**:749–56.
35. Zook JM, McDaniel J, Olson ND. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019;**37**:561–6.
36. Krusche P, Trigg L, Boutros PC. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* 2019;**37**:555–60.
37. *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL [https://www/R-project.org](https://www.R-project.org) and *argparse* v2.1.6 package.
38. Purcell S, Neale B, Todd-Brown K. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
39. Chang CC, Chow CC, Tellier LC. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7.
40. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;**32**: 227–34.