Deviations from Hardy Weinberg Equilibrium at $CCR5$-$\Delta32$ in Large Sequencing Data Sets

Xinzhu Wei[1] and Rasmus Nielsen*[1,2]

[1] *Department of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, CA, 94720, USA*

[2] *GeoGenetics Centre, University of Copenhagen, 1350 Copenhagen, Denmark*

Abstract

Previous analyses of the UK Biobank (UKB) genotyping array data in the $CCR5$-$\Delta32$ locus show evidence for deviations from Hardy-Weinberg Equilibrium (HWE) and an increased mortality rate of homozygous individuals, consistent with a recessive deleterious effect of the deletion mutation. We here examine if similar deviations from HWE can be observed in the newly released UKB Whole Exome Sequencing (WES) data and in the sequencing data of the Genome Aggregation Database (gnomAD). We also examine the reliability of the genotype calls in the UKB array data. The UKB genotyping array probe targeting $CCR5$-$\Delta32$ (rs62625034) and the WES of $\Delta32$ are strongly correlated ($r^2 = 0.97$). This contrasts to tag SNPs of $CCR5$-$\Delta32$ in the UKB which have high missing data rates and imputation errors rates. We also show that, while different data sets are subject to different biases, both the UKB-WES and the gnomAD data have a deficiency of homozygous $CCR5$-$\Delta32$ individuals compared to the HWE expectation (combined P-value < 0.01), consistent with an increased mortality rate in homozygotes.

*rasmus_nielsen@berkeley.edu

Finally, we perform a survival analysis on data from parents of UKB volunteers, that, while underpowered, is also consistent with the original report of a deleterious effect of $CCR5$-$\Delta 32$ in the homozygous state.

Introduction

The theory of Hardy-Weinberg Equilibrium (HWE) asserts that genotype frequencies can be predicted as simple products of allele frequencies when neutrally segregating in a randomly mating population[1;2]. Therefore, deviations from HWE indicate the presence of other evolutionary forces[3], such as assortative mating, population structure, or natural selection. Many different factors can cause deviations from HWE that will affect the entire genome. However, natural selection will cause deviations from HWE that only affect a local region. Such local deviations from HWE can, therefore, be used to identify specific loci affected by natural selection. However, very few studies have been published that use deviations from HWE to detect selection because it typically requires extremely large sample sizes[3;4]. As HWE is restored in each generation of random mating, the magnitude of deviations from HWE that can be observed will be on the same order as the selection coefficient. However, selection usually works through relatively minor fitness effects. Mutations that have a very strong effect on fitness are not expected to segregate in the population for an extended period and are, therefore, unlikely to be observed at any appreciable frequency. For this reason, in loci with intermediate or high-frequency alleles, we would only expect to observe very weak deviations from HWE due to selection. Furthermore, selection may be working on fertility, or may be working additively, thereby not causing any deviations from HWE. Despite these challenges, it may now be possible to examine hypotheses about selection using deviations from HWE due to the recent availability of large data sets of hundreds of thousands of human genomes made available through projects such as the UK Biobank[5].

A previous study[6] used the genotyping data of the white British individuals in the UK Biobank to investigate deviations from HWE and fitness effects of $CCR5$-$\Delta 32$, a mutation that is known for its protection against HIV in Europeans[7]. Compared to other markers of similar Minor Allele Frequency (MAF), Δ32 showed a significant deviation from the HWE expectation due to a deficiency of individuals who are homozygous for the Δ32 allele at the time of recruitment, suggesting that Δ32/Δ32 causes increased mortality[6]. Consistent with this observation, Δ32/Δ32 individuals had a 3.4-42.6% increase in the all-cause mortality rate compared to individuals who had at least one copy of the wild-type allele, based on survivorship data after recruitment.

The previous study[6] was based entirely on the marker denoted rs62625034. The probe for this marker was initially included in the genotyping array to capture a Denisovan specific Single Nucleotide Polymorphism (SNP)[8]. While this SNP has never been found segregating in the 1000 Genome Project[8] and is unlikely to be present in the UK Biobank, the probe for this SNP fully covers the 32 base pairs of the Δ32 deletion and should, therefore, directly genotypes Δ32. Consistent with this prediction, rs62625034 had a pattern of linkage disequilibrium (LD) with other markers compatible with that observed in previous studies of Δ32[6]. However, a comparison between the genotyping result of rs62625034 and direct sequencing data of Δ32 was not possible at the time of study because direct sequencing data had not been released for the UK Biobank. The recent release of Whole Exome Sequencing (WES) of 49,960 exomes from UKB participants[9], now allows additional validation of rs62625034 as a marker for Δ32. Moreover, although the sample size of the UKB-WES is only approximately 10% of the size of the UKB genotyping data, it is still the largest WES data of a single population available for general research, allowing additional analyses of patterns of deviation from HWE for Δ32. This is particularly important as genotype calling errors could strongly affect analyses of deviations from HWE. Other databases of direct sequencing data with smaller cohorts, such as the gnomAD database[10], provide additional opportunities for examining the patterns of deviations from HWE in Δ32. While the gnomAD database does not provide access to individual-level data for other researchers (`https://gnomad.broadinstitute.org/faq`)[10], it does make the total number of genotype calls for each possible genotype available for each locus.

The objective of this study is threefold. First, we investigate patterns of deviations from HWE in Δ32 in the data produced using Next Generation Sequencing (NGS) in the gnomAD[10] and UKB-WES data[9]. To do so, we first examine general patterns of deviations from HWE in the two data sets for markers with a frequency similar to that of Δ32. Secondly, we take advantage of the new WES data to examine the genotyping accuracy of rs62625034 as a Δ32 marker and compare it to possible tag SNPs that also have been genotyped or imputed in the UKB cohort. Finally, we examine the robustness of the survival analysis of Wei and Nielsen[6] using additional covariates, and we compare it to a survival analysis using the information of the parents of UK Biobank volunteers, similar to the analyses done by Mostafavi and colleagues[11].

Materials and Methods

The UK Biobank data

We use the UKB genotyping data, UKB imputation data, and the UKB-WES data under the UKB application number 33672. We download the following files of the 'v3' version of the UKB imputed data: the bgen, bgi, and mfi files. We first use plink2[12] to convert the files into pgen, psam, pvar format (for format description, see: `https://www.cog-genomics.org/plink/2.0/formats`), and extract the dosage information of the imputed SNPs from these files. The UK Biobank has released two versions of WES variant call data sets (named SPB and FE by the UK Biobank), and the one used in this study (bed, bim, fam files) is the UKB Population-level FE variants data (field 23160) that have the highest variant calling confidence according to the UKB (https://www.ukbiobank.ac.uk/wp-content/uploads/2019/08/UKB-50k-Exome-Sequencing-Data-Release-July-2019-FAQs.pdf). The 'v2' version of bed, bim, fam files of the UKB genotyping data are used for analyses involving directly genotyped data.

The UKB phenotype (annotation) data used include the year of birth, month of birth, age at death, age at recruitment, time attending recruitment center, genetic principal component (PC), sex, father's age at death, mother's age at death, and center. Unless otherwise noted, when analyzing array data, analyses are done on the UKB genotyping array data without imputation. In analyses of deviations from HWE in the UKB-WES data, for reasons described in the Results section, markers with $B_i \leq 0.5$ and $B_i \geq 2$ are removed. The definition of B_i is given in the section entitled 'Deviations from HWE, P-values, and CI'.

The gnomAD data

We download the following data files from the gnomAD website (https://gnomad.broadinstitute.org) 'gnomad.exomes.r2.1.1.sites.vcf' and 'gnomad.genomes.r2.1.1.exome_calling_intervals.sites.vcf'. The second file contains exonic variants called from Whole Genome Sequencing (WGS). There are five European populations [Finnish (FIN) and non-Finnish European (NFE)] in the gnomAD where the MAF of Δ32 is at least 0.1 – Finnish (FIN), Swedish (SWE), Estonian (EST), North-western European (NWE), and Other non-Finnish European (ONF). A MAF of at least 0.1 would translate to having at least 1% of Δ32/Δ32 individuals under HWE. For each SNP, we extract the following fields from the gnomAD data: AC_fin, AN_fin, AF_fin, nhomalt_fin, AC_nfe_swe, AN_nfe_swe, AF_nfe_swe, nhomalt_nfe_swe, AC_nfe_est, AN_nfe_est, AF_nfe_est, nhomalt_nfe_est, AC_nfe_nwe, AN_nfe_nwe, AF_nfe_nwe, nhomalt_nfe_nwe, AC_nfe_onf, AN_nfe_onf, AF_nfe_onf, nhomalt_nfe_onf. In the gnomAD files, 'AC' stands for alternate allele count for samples, 'AN' stands for total number of alleles in samples, 'AF' stands for alternate allele frequency in samples, 'nhomalt' stands for count of homozygous individuals of the alternative allele in samples, and 'nfe' stands for non-Finnish European. For marker i in population j, we estimate its MAF using the count numbers AC_{ij}/AN_{ij}, and confirm it is the same as AF_{ij}. We then merge the counts for SNPs that are called in both vcf files (WES and WGS). The majority of individuals are in the WES data, but four (FIN, EST, NWE, ONF) out of five populations also have additional WGS data. The EST population has the smallest sample size (2408 at Δ32) even after merging the WES with WGS data set. The number of individuals with Δ32 genotype calls in each cohort is given in **Table 1**. In both the gnomAD and the UKB-WES data, Δ32 has been directly called and we rely on the genotype calls provided. As we do not have access to the raw sequencing reads, we do not attempt to recall genotypes from the reads directly.

Identification of markers for Δ32

Δ32 is a 32 bp deletion in the $CCR5$ gene. Two different annotations are used for this variant: 'GTCAGTATCAATTCTGGAAGAATTTCCAG**ACA**' for rs333 is the most common annotation, and it is used in dbSNP[13] and SNPedia[14]. The alternative annotation '**ACA**GTCAGTATCAATTCTGGAAGAATTTCCAG' is used in gnomAD and UKB, which does not have an RS number. Because the UKB and gnomAD data use the alternative annotation, we look up the imputed Δ32 (3:46414943_TACAGTCAGTATCAATTCT GGAAGAATTTCCAG_T) from the files of the genotype calls by this 32 bp sequence. Both annotations identify the same deletion because the 3 bp 'ACA' motif, which appears both at the 5' and 3'end of the 'GTCAGTATCAATTCTGGAAGAATTTCCAG' motif could be either the 'ACA' at the beginning of the deletion or the 'ACA' at the end of the 32 bp deletion. In both cases, the Affymetrix probe for rs62625034 'CCAT**ACA**GTCAGTAT CAATTCTGGAAGAATTTCCA[G/T]**ACA**TTAAAGATAGTCATCTTGGGGCTGGT CCTGCC' fully covers the 32 bp deletion.

Deviations from HWE, P-values, and CI

We will use the following notation: f_i is the MAF in locus i. In the UKB, it is estimated from the data as $\hat{f}_i = (x_{i01} + 2x_{i00})/(2n_i)$, where x_{i00} and x_{i01} are the counts of the minor allele homozygous and heterozygous individuals, respectively, and n_i is the total number of individuals with genotype calls in locus i. In the gnomAD data, it is, as previously mentioned, inferred from the 'AC' and 'AN' counts. Similarly, p_i is the minor allele homozygous gentoype frequency, which is estimated as $\hat{p}_i = x_{i00}/n_i$ (given by $nhomalt_{ij}/(AN_{ij}/2)$ in the gnomAD sample). By convention, we polarize alleles such that $\hat{f}_i < 0.5$ for each data set analyzed. This does not cause any ambiguity for Δ32, which has a frequency within 10.45-13.51% in all data sets analyzed.

For each data set, we select a set of control SNPs with similar allele frequencies and sample sizes as that of Δ32 in order to obtain empirical distributions of the measure of deviations from HWE used in this study (B_i defined below). For the gnomAD data, we choose markers whose MAFs are within ±0.01 of that of Δ32 from the same study population. We also require the number of genotyped individuals at the SNPs to be at least 90% of the number genotyped at Δ32, to partially avoid increased noise from SNPs with very small sample sizes. This procedure results in between 2635 and 7850 control SNPs depending on the cohort (**Table 1**). Because Δ32 and other SNPs can have different MAF in different populations, using population specific control allows us to strictly compare Δ32 to SNPs with similar MAFs in the same study population. For the UKB-WES data, we also choose SNPs with MAFs within ± 0.01 of the MAF of Δ32, while requiring the number of white British ancestry individuals that are sequenced at those SNPs to be at least 90% of the number genotypes at Δ32. Using this procedure we obtain 6609 control SNPs (**Table 1**). We similarly select SNPs from the UKB-Array data, requiring that the number of white British ancestry individuals that are sequenced at those SNPs to be at least 90% of the number of individuals with genotype calls at Δ32. However, we require their MAFs to be within ±0.0025 of that of Δ32. This gives us 5897 SNPs from the UKB array. We use a narrower interval for the UKB array data because there are more SNPs available in the relevant frequency range for the array data than for the sequencing data from UKB-WES or the gnomAD database.

Under HWE, $p_i = f_i^2$ so we use the following statistic to measure deviations from HWE: $B_i = \hat{p}_i/(\hat{f}_i^{\,2})$. Notice that this statistic differs from the traditional F statistic[15] as it is defined for the minor allele homozygotes and, therefore, scales differently. We define the vector of all values of B_i for the S control loci as $B = \{B_1, B_2, , , B_S\}$. We quantify deviations from HWE using $B_{\Delta 32}$ and obtain 95% Confidence Intervals (CIs) for $B_{\Delta 32}$ using the bootstrap percentile interval method (see Efron and Tibshirani[16] pp. 170) with 10,000 replicates for each data set. This is done by sampling individuals with replacement to obtain bootstrap samples with the same sample size as the original sample. $f_{\Delta 32}$ and, subsequently, $B_{\Delta 32}$ are then recalculated for each of these bootstrap data sets to provide a new set of values: $B_{\Delta 32}^{b1}, B_{\Delta 32}^{b2}, ..., B_{\Delta 32}^{b10,000}$. We then sort these values and use the 2.5th and 97.5th percentiles as the limits of the 95% CI.

We perform three different one-sided tests of deviations from HWE. First, we compute an empirical p-value as the proportion of control markers with $B_i < B_{\Delta 32}$, i.e. for S genomic control loci the p-value is calculated as $P1 = 1/S \sum_{i=1}^{S} I(B_i \leq B_{\Delta 32})$. This test has the advantage that it controls for other genome-wide factors affecting deviations from HWE such as the Wahlund-Effect[17]. However, it has the disadvantage of reduced power if multiple other loci also are affected by selection. The second test addresses this problem by testing for significant deviations from the median value of B. In this test, we calculate the number of bootstrap samples using the aforementioned bootstrap method, for which the bootstrap value of $B_{\Delta 32}$ is bigger or equal to the median and we then define the test statistic as $P2 = 1/10000 \sum_{i=1}^{10000} I(B_{\Delta 32}^{bi} \geq median[B])$. This p-value is a one-sided dual of the procedure used to construct CIs for $B_{\Delta 32}$. We use the median instead of the mean to protect against the effect of outliers that may disproportionately affect the mean.

Thirdly, for the sake of comparison, we conduct a test against the standard null hypothesis of $B_{\Delta 32} = 1$ using the bootstrap procedure by calculating $P3 = 1/10000 \sum_{i=1}^{10000} I(B_{\Delta 32}^{bi} \geq 1)$. If the bootstrap value of $B_{\Delta 32}$ is smaller than the median of B_i or 1, for test 2 and 3, respectively, in all 10,000 bootstrap samples, we define $P < 0.0001$. Notice that this third test does not control for genome-wide effects such as the Wahlund-Effect[17].

We will mainly focus discussion on the results from test 2 ($P2$) but will also show the results of the other tests for the sake of comparison.

Survival analyses while controlling for kinship

To obtain a set of unrelated white British individuals, we first remove every individual who has ten or more relatives in the UKB file ‘ukb3367_rel_s488364.dat’ (which contains pairs of relatives up to 3rd-degree kinship). For each remaining pair, we remove the person carrying the least Δ32 alleles and keep the person with more Δ32 alleles to maximize power. If two people in a pair have an equal number of Δ32 alleles (e.g., both are heterozygous), we randomly remove one of the two. By doing so, at rs62625034, we obtain 329,227 unrelated individuals and 9,494 death events. We apply the same approach to SNP rs113010081 and obtain 307,457 unrelated individuals and 8,609 death events. These two data sets of unrelated individuals are used to investigate the effects of rs62625034 and rs113010081, respectively.

The Cox model we apply to the unrelated white British people is largely identical to the Cox model used in our previous study[6], except that in addition to ancestry (the first 40 PCs), we also control for sex and recruitment center. We use the function 'coxph' in the R package 'survival'[18], and provide the start time, end time, event, and predictors. We use the function 'summary' in R to output the point estimate, standard error (s.e.), and 95% CI of the Cox regression coefficients and the exponentiation of these regression coefficients. We use the individual's age as time in the Cox model (see Kleinbaum and Klein[19] pp. 131-135), so the start time is the estimated age at recruitment (A1). A1 is calculated using the UKB date of recruitment (with year Y1, month M1, and day D1), and the year of birth (Y0) and month of birth (M0) of each individual. We assume all individuals were born on the 15th of each month. We then estimate A1 as A1 = [datenum(Y1,M1,D1) - datenum(Y0,M0,15)]/365.25, where the MATLAB function datenum() returns the number of days from January 0, 0 CE, and 365.25 is the average number of days in a year. Our estimated A1 is more precise than the UKB field age at recruitment, which is a round-down integer in the unit of year.

For volunteers who have passed away, the age at death (A2) in the unit of year to the precision of day and the date at death are provided in the UKB data. The latest date of death among all registered deaths in the downloaded UKB data is 2016-02-16, and we use this date to approximate the time of last death entry, assuming that after 2016-02-16 we have no mortality/viability information of the volunteers. For volunteers who do not have a recorded age at death in the data, the event is coded as 0, and the estimated age on 2016-02-16 (A2'), A2' = [datenum(2016,02,16) - datenum(Y0,M0,15)]/365.25, is used as the end time. For those who have a recorded death, the event is coded as 1, and A2 is the end time. The Cox function[20] assumes that each predictor has a homogeneous effect on the event (event = death outcome) across all times, but the effect of the predictors could, in fact, differ among age groups. We test a recessive deleterious model by coding the $\Delta 32/\Delta 32$ individuals as 1, and the two other genotypes as 0. The first 40 PCs, Sex, and recruitment center (as a factor variable) are also used as predictors in the Cox model. Each predictor, i, has an estimated effect β_i from the Cox model, and $\exp(X_i\beta_i)$ measures the fold change of predictor i relative to the baseline mortality rate, where X_i is the value at predictor i. Since the binary coded genotype is the predictor we care about ($X_g = 1$ for $\Delta32/\Delta32$ homozygotes), we obtain its associated coefficient, β_g, from the Cox-model and report $\exp(\beta_g)$-1, which measures the relative increase in all-cause mortality in $\Delta32/\Delta32$ homozygotes compared to individuals with at least one wild-type allele of $CCR5$ (the relative increase in all-cause mortality, henceforth denoted m).

In another survival analysis (see Results section), we use the ages at death of the unrelated white British parents as events, assuming that the parents of unrelated white British volunteers (329,227 individuals at rs62625034) are also unrelated white British. In this case, we use both the father's age at death and the mother's age at death in the same Cox regression analysis. The ages at death at the parents were recalled and entered by the volunteers, and are thus subject to errors. For parents age at death, there are three records (one at recruitment, and the other two instances from later center visits) and the first one has the most complete entry. We merge the three instances to maximize the power; when a later entry disagrees with an earlier one, we assume the earlier record, which is closer to the occurrence of the event, is more accurate. By doing so, we obtain 442,778 records of death for parents of unrelated volunteers (244,889 female and 197,889 male). The average age at death of the parents is 74.36. Similarly to the analysis for offspring, we use the function 'coxph' in the R package 'survival'[18], and provide the time (age at death), event (death), and predictors (genotype, sex of the parents, center and the first 40 PCs of the children).

Given the frequency of $\Delta32$, $f_{\Delta32}$, and assuming random mating, the probability that a parent has genotype $\Delta32/\Delta32$ is $f_{\Delta32}$, $f_{\Delta32}/2$, and 0, for $\Delta32/\Delta32$, $\Delta32/+$, and $+/+$ individuals, respectively. We apply an additive model, by coding the genotype predictor for $\Delta32/\Delta32$ individuals as 2, $\Delta32/+$ individuals as 1, and $+/+$ individuals as 0. The regression coefficient, β_g', estimates the effect of offspring genotype on parents mortality. Assuming that the effect on mortality of $\Delta32$ is strictly recessive and equal to m, we expect an increase in mortality in parents with $\Delta32/+$ offspring of $mf_{\Delta32}/2$ and an increase in parents with $\Delta32/\Delta32$ of $mf_{\Delta32}$. We notice that the effect might be slightly different in the Cox model depending on whether we measure $\Delta32/\Delta32$ or $\Delta32/+$ individuals as the relation between the regression coefficients in the Cox model and the mortality rates is non-linear. We nonetheless, as an *Ad hoc* approach, translate the joint estimate of β'_g for the Cox model with the offspring genotype coding of 0, 1, and 2, into an estimate of the increase in mortality of $\Delta32/\Delta32$ individuals using $m = exp(2\beta'_g/f_{\Delta32}) - 1$. Below we will present simulations to evaluate the adequacy of this approximation. We also examine the consistency of the Cox model estimates of m between the parent and the offspring analyses using the Z-score $= \frac{\beta_g - 2\beta'_g/f_{\Delta32}}{\sqrt{\sigma^2_{\beta_g} + 4\sigma^2_{\beta'_g}/f^2_{\Delta32}}}$, where σ stands for the s.e., and then convert the Z-score into a two-tailed p-value using a standard normal distribution (notice here that β'_g and β_g are the coefficients estimated from the parents and offspring mortality data, respectively).

Using simulations we investigate three different possible transformations between m and β'_g: $m_a = exp(2\beta'_g/f_{\Delta32}) - 1$, $m_b = 2(exp(\beta'_g) - 1)/f_{\Delta32}$, and $m_c = (exp(2\beta'_g) - 1)/f_{\Delta32}$. In each simulation, we simulate $N_f = 225000$ and $N_m = 225000$ paternal and maternal genotypes based on the observed frequency of $f_{\Delta32}$ of 0.1159 in the UKB genotyping array. We then randomly sample one allele from each parent to form $N_c = 225000$ offspring genotypes. The male and female mortality rates per year, $h_m(t)$ and $h_f(t)$, are imported from the national life tables in the UK ("nltuk1517reg.xls") from the Office of National Statistics (`https://www.ons.gov.uk`) which contain the mortality per year from age 0 to age 100 for the entire UK population each year from 1980 to 2017, we use the mortality data from the period 1980-1982 as baseline mortality rates in the simulations. For age interval t to $t+1$, deaths of parents that are not $\Delta32$ homozygotes are sampled at random from these empirical tables of values of $h_m(t)$ and $h_f(t)$. $\Delta32/\Delta32$ parents are similarly sampled with probabilities $(1+m)h_m(t)$ and $(1+m)h_f(t)$ where $m = 0.2$. We use two different approaches to deal with the missing information regarding people over the age of 100: 1) we assume people who survive to age 100 all die at age 101 and we include all parents. 2) We exclude people who survive to age 100. We then analyze the simulated data using a Cox regression with the parent's age at death as the response variable and with offspring genotype coded as 0,1,2 as the predictor while controlling for sex. We estimate the allele frequency based on the genotypes used in the Cox regression, $f_{1,\Delta32}$ and $f_{2,\Delta32}$, respectively for the two simulation approaches. After estimating $\beta'_{1,g}$ and $\beta'_{2,g}$ in the two different Cox regressions, we transform them into $\hat{m_{a1}}$, $\hat{m_{b1}}$, and $\hat{m_{c1}}$ using $f_{1,\Delta32}$, and $\hat{m_{a2}}$, $\hat{m_{b2}}$, $\hat{m_{c2}}$ using $f_{2,\Delta32}$, respectively.

Based on 1000 simulations, we obtain $\bar{\hat{m}}_{a1} = 0.1941$ and s.e.$(\bar{\hat{m}}_{a1}) = 0.0021$, $\bar{\hat{m}}_{a2} = 0.1768$ and s.e.$(\bar{\hat{m}}_{a2}) = 0.0018$, $\bar{\hat{m}}_{a3} = 0.1778$ and s.e.$(\bar{\hat{m}}_{a3}) = 0.0018$, $\bar{\hat{m}}_{b1} = 0.1912$ and s.e.$(\bar{\hat{m}}_{b1}) = 0.0021$, $\bar{\hat{m}}_{b2} = 0.1743$ and s.e.$(\bar{\hat{m}}_{b2}) = 0.0018$, $\bar{\hat{m}}_{b3} = 0.1753$ and s.e.$(\bar{\hat{m}}_{b3})$ $= 0.0018$. As expected, there is very little difference in the estimates of m depending on how the missing information regarding deaths after age 100 is dealt with, as a very small proportion of the population reaches this age. Given the $f_{\Delta32}$, and $m = 20\%$ in the simulated data, all three m underestimate the mortality, but m_a is more accurate than m_b and m_c. We will, therefore, in the following report results using m_a. We note that the general problem of estimating dominance effects in a Cox regression from parental mortality data using offspring genotypes warrants further statistical work.

Results

Deviations from HWE in genotyping and sequencing data sets

As described in the Materials and Methods section, to control for various genome-wide cohort-specific factors, we define a set of control markers with similar allele frequencies to Δ32 for each data set. We quantify deviations from HWE in locus i in terms of the statistic B_i which takes on the value 1 if there are no deviations from HWE, < 1 if there is a deficiency of homozygous minor allele genotypes, and > 1 if there is an excess of homozygous minor allele genotypes. The majority of the SNPs in the UKB genotyping data have $B_i < 1$ (**Fig.1a**). This observation was a primary justification for the use of an empirical p-value in[6] to correct for the apparent systematic under-calling of homozygous minor allele genotypes in the UK Biobank array data. In the UKB-WES, the

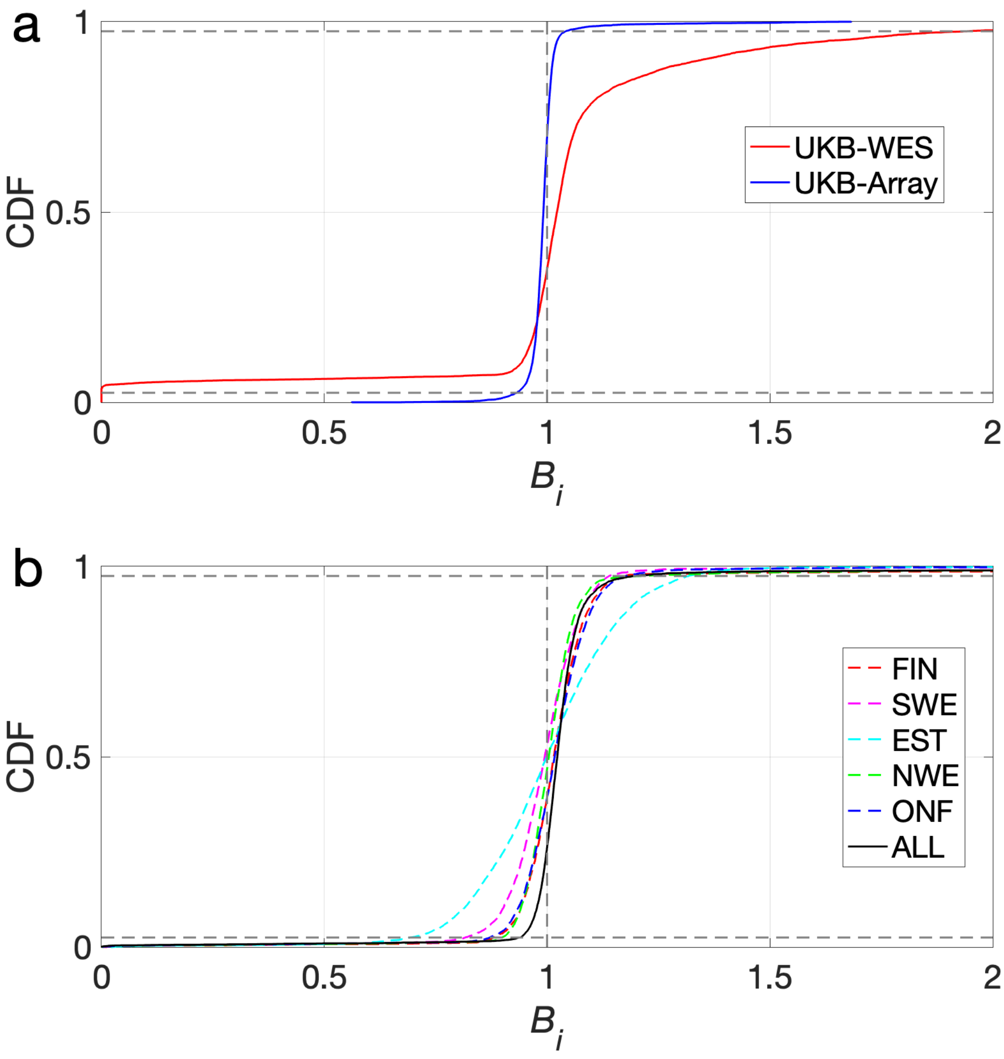

Figure 1: The cumulative distribution function (CDF) of B_i in the UKB and gnomAD data for SNPs with allele frequencies similar to Δ32. The x-axis is $B_i = \hat{p}_i/(\hat{f}_i^{\,2})$ and the y-axis is the proportion of control markers with a value of B_i less than the corresponding value on the x-axis. See Materials and Methods for the choice of control markers. All: stands for the combined sample of five gnomAD populations: FIN, SWE, EST, NWE, and ONF (see Materials and Methods section for an explanation of acronyms). The two horizontal dashed lines mark the 2.5th and 97.5th percentiles of the empirical distribution, respectively, and the vertical dashed line marks $B_i = 1$.

opposite trend is observed. The difference between the UKB array data and WES data could be due to genotyping errors, or biased genotype drop-out, in either the array data (causing a relative deficiency of minor allele homozygous calls) or the sequencing data (causing a relative excess of minor allele homozygous calls). Very few natural processes are likely to cause genome-wide deviations from HWE in the direction of a deficiency of homozygous calls. For example, by the Wahlund-principle[17], population structure will always cause an excess of homozygosity. Processes such as dis-assortative mating, or systematic outbreeding beyond that expected under random mating, can cause systematic deficiency of homozygotes, but are rarely seen in natural populations. For this reason, the most likely explanation for the overall deficiency of minor allele homozygotes in the UKB array data is a systematic under-calling of homozygous minor allele genotypes. However, we also note that the distribution of B_i is very heavy-tailed in the UKB-WES data. In particular, there are 268 among the 7215 control markers in the UKB-WES data that have a value of $B_i = 0.0$. Given the sample size (between 36969 and 41250 individuals per marker), these are almost certainly genotyping errors. Presumably, some bioinformatical artifact has caused a complete drop-out of homozygous minor allele genotypes in some loci in the WES data.

To analyze the gnomAD data, we combine the WES and the coding region of the Whole Genome Sequencing (WGS) for each population (FIN, SWE, EST, NWE, and ONF), and then apply the same method as used in the UKB data to get the empirical distribution of B_i. The majority of the markers show $B_i > 1$ in three (FIN, NWE, ONF) out of five gnomAD populations (**Fig.1b**) as expected due to the Wahlund-effect[17]. The pattern of markers with completely missing minor allele homozygous genotypes observed in the UKB sequencing data, is not seen in these data. We also notice that the UKB-WES data contain a significant fraction of markers with $B_i > 2$. This signature likely also indicates higher error rates in genotype calling in the UKB-WES data, compared to other sequencing data sets. Therefore, in our later analysis (**Table 1** and **Fig. 2**) we use only the UKB-WES SNPs with $0.5 < B_i < 2$, presuming that these outliers likely are caused by genotype calling errors.

Deviations from HWE at Δ32 in sequencing data sets

A deficiency of Δ32/Δ32 individuals, relative to the HWE expectation, at the time of recruitment, provided evidence of a deleterious effect of the Δ32/Δ32 mutation[6]. However, given the possibility of genotype calling errors, we investigate if similar patterns of deviations from HWE are observed in the newly available UKB-WES data[9] and the gnomAD data[10]. Out of 41,250 white British individuals with WES data in the UK Biobank, 41,059 have variant calls at Δ32, among which 537 are Δ32/Δ32, 8537 are Δ32/+, and 31985 are +/+. The Δ32 allele has a MAF of 0.1170. Under HWE, the expected number of Δ32/Δ32 individuals is 562, giving a value of B_i of $B_{\Delta 32} = 0.9555$.

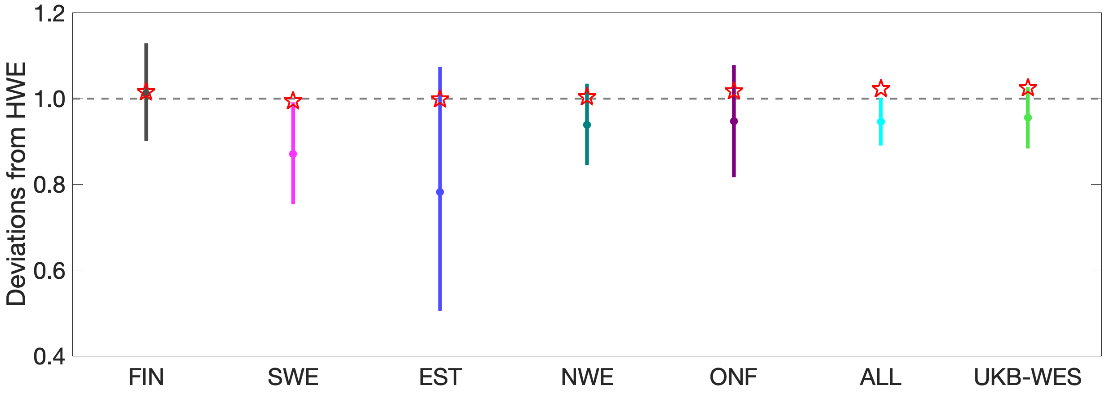

Figure 2: The value of the statistic $B_{\Delta 32}$, which measures deviation from HWE and its 95% CI in the gnomAD and the UKB-WES data. For each data set, the dot and vertical line show the value of $B_{\Delta 32}$ and its 95% bootstrap CI. The star shows the median of B_i among the control SNPs in the data set, and the horizontal dashed line marks $B_i = 1$.

We perform three different tests to determine the significance of this value. First, we compute an empirical p-value as the proportion of control markers with $B_i < 0.9555$ ($P1$). Secondly, we perform a bootstrap test (see Materials and Methods) of the null hypothesis that the true value of $B_{\Delta 32}$ equals the median of the control SNPs ($P2$). Finally, we test against the traditional HWE null hypothesis of $B_{\Delta 32} = 1$ using a similarly constructed bootstrap test ($P3$). This latter test is equivalent to standard tests for deviations from HWE. We consider test 1 and test 2 to be the most appropriate tests because they correct for genome-wide deviations from HWE caused by the Wahlund-effect and other factors and we focus on $P2$ because this test is more robust to the possibility of the presence of other outliers due to selection or other factors.

The p-values for the tests for deviations from HWE are shown in **Table 1** together with other relevant statistics, and the confidence intervals for B_i are illustrated in **Fig. 2**. All sample sizes are considerably smaller than for the original UKB genotyping data, and the CIs are consequently quite wide. Even after combining UKB-WES with gnomAD, the sample size in the UKB array data set is still approximately four times larger. Nonetheless, all data sets tend to show the same trend as in the original analyses of the UK Biobank array data. For example, in the combined gnomAD data, the observed value of $B_{\Delta 32}$ is 0.946, while the mean and the median of the control SNPs are 1.076 and 1.023, respectively. $B_{\Delta 32}$ tends to be smaller than both the median and 1 for all gnomAD cohorts, although, only for one of the cohorts is $P2 < 0.05$ (SWE). However, when considering the pooled sample of all the gnomeAD populations, p-values of $P2 = 0.0034$ and $P3 = 0.0274$ are obtained. Notice that $P3$ is quite conservative because it does not account for the substantial Wahlund effect observed when pooling populations. Test 1 similarly gives a p-value of $P1 = 0.0295$. The $P2$ p-value for the UKB-WES is also significant at the 5% level ($P2 = 0.0272$), while tests 1 and 3 are not ($P1 = 0.0784$, $P3 = 0.1095$). An alternative approach for combining the results from the five gnomAD populations, rather than pooling the data, thereby subjecting it to additional Wahlund effects, would be to combine the p-values using Fisher's combined probability test[21]. Doing so for test 2, we get $P2 = 0.0192$. Using Fisher's method, we can also combine the p-values from UKB-WES with the p-values in five populations from gnomAD, resulting in a combined p-value of 0.0095 for Test 2. While the sample sizes of these data sets are so small that they provide little power to detect selection from deviations from HWE, the results overall confirm the previously reported observation of a deficiency of $\Delta 32$ homozygous individuals.

Imputation quality of the $CCR5$ region in the UKB

Rs62625034 is annotated to be a Denisovan SNP that does not segregate in the modern European population[8]. However, the probe for this SNP included in the UKB genotyping

Table 1. Deviations from HWE at Δ32 in the UKB and gnomAD NGS data.

Population	FIN	SWE	EST	NWE	ONF	All	UKB-WES[10]
Control size [1]	2635	4512	7850	6121	7093	5219	6609
95% range [2]	0.88-1.17	0.82-1.14	0.69-1.32	0.90-1.21	0.87-1.18	0.95-1.20	0.92-1.69
Median of B_i [3]	1.015	0.995	0.999	1.004	1.017	1.023	1.025
Mean of B_i	1.081	0.994	0.999	1.044	1.018	1.076	1.079
$f_{\Delta 32}$	0.1341	0.1212	0.1152	0.1127	0.1045	0.1163	0.1170
Sample size [5]	12524	13064	2408	25309	16520	69825	41059
$B_{\Delta 32}$	1.012	0.871	0.782	0.939	0.947	0.946	0.9555
95% CI $B_{\Delta 32}$ [6]	0.9012-1.1293	0.7555-0.9940	0.5005-1.0712	0.8458-1.0346	0.8171-1.0757	0.8906-1.0013	0.8835-1.0270
$P1$ [7]	0.4822	0.0567	0.0692	0.0934	0.1355	0.0295	0.0784
$P2$ [8]	0.4887	0.0223	0.0670	0.0882	0.1442	0.0034	0.0272
$P3$ [9]	0.5886	0.0186	0.0640	0.1047	0.2133	0.0274	0.1095

[1] Number of control loci whose MAFs are within ±0.01 of that of Δ32 in the sample.

[2] 2.5th and 97.5th percentiles of B_i in the control loci.

[3] The median value of B_i in the control loci.

[4] The mean value of B_i in the control loci.

[5] Number of individuals that have a genotype call at Δ32.

[2] The 95% bootstrap CI for $B_{\Delta 32}$.

[7] $P1$ is the fraction of genomic control loci with $B_i < B_{\Delta 32}$.

[8] $P2$ is the size of the smallest bootstrap confidence set that includes the median of B_i.

[9] $P3$ is the size of the smallest bootstrap confidence set that includes 1.

[10] In the UKB-WES, only SNPs with $0.5 < B_i < 2$ are used as genomic control.

data fully covers the 32 bp deletion region of $CCR5$. The previous study[6] used rs62625034 (coordinate 3:46414975 in GRCh37) in the UKB array genotyping data to identify Δ32 because no other probe for Δ32 was included in the genotyping panel[6]. However, the new availability of Whole Exome Sequencing (WES) data from a subset of UK Biobank individuals provides an opportunity to assess the accuracy of the array genotyping results. There is an overlap of 39,715 white British individuals with Δ32 genotype calls in the UKB-WES data and rs62625034 calls in the UKB genotyping array data. Comparing the genotype calls in these 39,715 individuals between the array data and the WES data, we find a high correlation with an $r^2 = 0.968$ (or $r = 0.984$), confirming that the probe of rs62625034, though not designed for Δ32, in fact measures Δ32 dosage quite accurately.

A recent study used the UKB imputation data to investigate pleiotropic effects of Δ32 and provided supporting evidence for a deleterious effect of Δ32[22]. However, markers in the Δ32 region may suffer from imputation inaccuracies because of mis-annotations in the region. Rs62625034, which in actuality measures Δ32 dosage, is incorrectly annotated and is reported to have a frequency of 0.1159 in the UKB[6] and a frequency of 0 in the 1000 Genome[8], which could potentially distort the imputation accuracy for other markers in the region. To address this question, we investigated correlations between Δ32 in the direct sequencing data on the one hand, and the imputed Δ32 and other imputed tag SNPs in the UK Biobank array data on the other hand. Δ32 in the imputed version of the UKB array data[5], has a MAF of 0.09 (in 'ukb_mfi_chr3_v3.txt'), which is more than 0.02 lower than the reported MAFs for Δ32 in the British population[23], suggesting the imputation quality of Δ32 is poor. In addition, the r^2 between the 35,324 white British individuals for which there are both WES sequencing calls and imputation based genotype calls with high imputation confidence in the array data is relatively low ($r^2 = 0.81$). High confidence genotype calls are here defined using the plink2[12] default for converting dosage into genotypes which uses the following bins for the conversion: 0-0.1,0.9-1.1,1.9-2 into 0,1,2, respectively, and the remaining treated as missing data (14%). Similarly, the imputed MAF of rs62625034 is only 0.0259 (in 'ukb_mfi_chr3_v3.txt'), which is approx. 0.09 lower than the directly genotyped rs62625034 MAF and the MAF expected for the UK population[6], and also much higher than the MAF of 0 for the supposed Denisovan SNP[8]. Because rs62625034 measures Δ32 dosage, but was annotated as a rare variant, the imputation accuracy of rs62625034 is poor. In contrast to the direct genotyping data for rs62625034, neither the imputed Δ32 nor the imputed rs62625034 data can be confidently used.

We further investigate whether the poor imputation quality in the region might have affected other SNPs. SNP rs113010081 is the only directly genotyped SNP that is in strong Linkage Disequilibrium (LD) with Δ32 in the European populations, and it is also in strong LD with rs62625034 in the UKB genotyping data. However, the directly genotyped rs113010081 has 11% missing genotype calls in the white British, which is very high in the UKB data (only 2.2% of SNPs of similar MAF have higher missing rates compared to rs113010081). However, if rs113010081 is correctly imputed, it could potentially be used to study the effect of Δ32. We, therefore, compare the correlation between the imputed data of rs113010081 and the Δ32 WES data (**Table 2**). In the UKB imputation data, high-quality imputed genotypes will be reported with integer dosages (0, 1, or 2), whereas low-quality genotypes will have dosage in the intervals between (0,1) or (1,2). The r^2 between the imputed genotypes of rs113010081 with integer dosages and Δ32 in the UKB-WES is as high as 0.942 (r = 0.971). However, the imputed genotypes of rs113010081 with decimal dosage only poorly correlates with Δ32 in UKB-WES (r^2 = 0.760, r = 0.872). As about 75% of the C/C genotype (minor allele homozygotes) have non-integer imputation dosage (**Table 2**), the imputation errors primarily affect the imputation of Δ32/Δ32 homozygotes. Therefore, despite that rs113010081 has an imputed MAF similar to the true MAF of Δ32 and a high information score of 0.96 (in 'ukb_mfi_chr3_v3.txt'), the imputed rs113010081 in UKB cannot be used to study the effect of Δ32 either.

We further investigate the imputation quality in the $CCR5$ region using another SNP, rs113341849. Rs113341849 is an intergenic SNP in strong LD with Δ32 (R^2 = 0.929 in CEU and GBR) that has not been directly genotyped in the UKB array genotyping data, but has been imputed. Since we cannot compare the rs113341849 imputation result to its WES result (as the SNP is intergenic), we instead compare it to Δ32 in the WES data. Similarly to the results for rs113010081, about 80% of the minor allele homozygotes have non-integer imputation dosage, where the correlation between imputed genotype and Δ32 in the WES data is low (r^2 = 0.786, r = 0.886), although the correlation between the integer dosage genotype and Δ32 is still quite high ($r^2 = 0.973$, $r = 0.986$). The poor imputation quality for minor allele homozygotes could be a general property in the $CCR5$ region. These results suggest that in the UKB array data, only the directly genotyped rs62625034 can be used to study the effect of Δ32 with high confidence.

Table 2. Individuals with each genotype in imputed SNPs and Δ32-WES.

Imputation	Genotype	Δ32/Δ32 [6]	Δ32/+ [6]	+/+ [6]
rs113010081 decimal dosage [2]	C/C	373	36	0
	C/T	30	4024	161
	T/T	0	127	901
rs113010081 integer dosage [3]	C/C	126	5	0
	C/T	7	4178	85
	T/T	0	152	30770
rs113341849 decimal dosage [4]	A/A	404	24	0
	A/G	28	4319	149
	G/G	0	193	1301
rs113341849 integer dosage [5]	A/A	96	2	0
	A/G	8	3927	39
	G/G	0	57	30428

[1] The genotype of the imputed SNPs.

[2] Individuals with imputed dosage (0,0.5] as C/C, (0.5,1.5) as C/T, and [1.5,2) as T/T.

[3] Individuals with imputed dosage 0 as C/C, 1 as C/T, and 2 as T/T.

[4] Individuals with imputed dosage (0,0.5] as A/A, (0.5,1.5) as A/G, and [1.5,2) as G/G.

[5] Individuals with imputed dosage 0 as A/A, 1 as A/G, and 2 as G/G.

[6] Individuals with Δ32/Δ32, Δ32/+, and +/+ genotypes in the UKB-WES data.

Δ32 is associated with increased mortality rates in the homozygous state after controlling for kinship

Wei and Nielsen previously applied a Cox-model and showed that Δ32/Δ32 individuals have a 21% (95%CI: [3.4%,42.6%]) increase in all-cause mortality rate in the white British cohort after controlling for ancestry[6]. Here, we further investigate whether the increased mortality in $\Delta32/\Delta32$ individuals could be affected by statistical artifacts due to kinship, sex, or recruitment center, which were not controlled for in the original study. To this end, we first remove individuals up to 3rd-degree kin (see Materials and Methods), and then apply a Cox model (see e.g., Kleinbaun and Klein[19] on the remaining unrelated individuals while controlling for sex, recruitment center, and ancestry. Despite that the sample size is smaller (329,227 with 9,497 deceased individuals versus 395,699 with 11,532 deceased individuals) after removing related individuals, and that more covariates are being controlled for, we still observe a point estimate of a 19.7% (95%CI: [1.7%,40.9%]) increase in all-cause mortality (Z-score = 2.16, and one-tailed P-value = 0.0154). The estimates and the Z-score are comparable to the ones previously reported (Z-score = 2.37, one-tailed P-value = 0.0091).

Despite the high LD between rs113010081 and $\Delta32$, applying the same analysis to directly genotyped rs113010081 does not result in a significantly increased mortality in the Cox model, though the result is generally in the same direction (9.4% increase in all-cause mortality, Z-score = 1.09 one-tail P-value = 0.138, 95% CI [-6.9%, 28.6%]). This insignificant result could potentially be caused by the high proportion of missing data (11%), compared to other SNPs at similar MAF (only 2.2% SNPs have higher missing rate than rs113010081), and compared to the 3.4% missing rate at rs62625034.

Cox regression on parents

Although $\Delta32/\Delta32$ individuals have a higher all-cause mortality rate in all the Cox models, the p-values are only moderately small, and the CIs for the increase in mortality rate are very wide. The power of survival analyses is limited by many factors, including the sample size and the number of deaths accumulated during the follow-up time. The number of people who are deceased is still fairly low in the UK Biobank, in part also because the all-cause mortality rate in this cohort is lower than in the general population, presumably due to 'healthy volunteer' effect[24] as discussed by Wei and Nielsen[6]. In the case of $\Delta32$, which has a presumed recessive effect, the power is further limited by the frequency of the Δ32/Δ32 individuals (which is only approximately 1%). An alternative strategy for investigating hypotheses regarding mortality in the UKB is to use the age at death of the parents of study participants. The advantage of this strategy is that there are many more death events recorded for the parents of study participants than for the study participants themselves. This strategy was used effectively by Mostafavi and colleagues[11] to identify loci affecting mortality in a genome-wide scan. For the case of Δ32, the presumed effect is a dominance effect in which only individuals homozygous for the minor allele have increased morality, leaving approaches using the age of death of the parents with very little power. Nonetheless, assuming an allele frequency of $f_{\Delta32}$, the probability that a parent is homozygous for the minor allele is $f_{\Delta32}$, $f_{\Delta32}/2$, and 0 for the homozygous minor, heterozygous, and homozygous major genotypes, respectively. In other words, if Δ32 is associated with a 20% increase in mortality rate, the parents of +/Δ32 individuals should on average have an approx. 1% increase in mortality rate relative to the parents of +/+ individuals. This leaves an opportunity, albeit with very limited power, to provide an alternative examination of the effect on mortality of Δ32. We repeat the survival analysis as for the offspring, but using the parents' age at death instead, and treating the two parents as independent of each other while controlling for parents sex instead of the offspring. This analysis otherwise uses the same covariates as for the offspring analysis and also filters for relatedness based on offspring kinship (as information regarding parent kinship is not available). We use an additive model as the expected mortality rate of parents increases proportionally to the Δ32 dosage in the offspring and obtain an estimate of 0.5% increased mortality (95% CI: [-0.15%,1.16%]) and a one-tailed p-value of 0.065 (Z-score = 1.516). The estimate from the parents corresponds to a point estimate and 95% CI for the effect of Δ32/Δ32 of 9.05% and [-2.5%,22.0%]. The CI largely overlaps the one obtained from the offspring, and the two estimates are consistent with each other (Z-score = 0.92, and two-tail $P = 0.367$).

While the survival analyses presented here obviously are underpowered, the results from both parents and offspring are quite compatible with each other and both provide moderate additional evidence in favor of a deleterious effect of Δ32/Δ32.

Discussion

In this study, we estimate deviations from HWE in the gnomAD and UKB. We observe that different data sets have different biases and apply cohort-specific genomic control to partially control for these biases. Compared to the HWE expectation, $\Delta32/\Delta32$ individuals are underrepresented in both the UKB and gnomAD data. These results further validate the claim of a recessive deleterious effect of $\Delta32$ alleles in some study groups and suggest that the effect is not limited to the British population. After a detailed comparison among the UKB genotyping, WES, and imputation data for the $CCR5$ region, we conclude that only the unimputed genotypes from the probe targeting rs62625034 in the UKB genotyping array data can be used with any confidence to identify $\Delta32$, and we recommend future studies use this marker to investigate the effect of $\Delta32$ in the UKB data. Studies based on the released imputed SNPs in the $CCR5$ region may likely lead to erroneous conclusions.

We repeat the survival analysis of Wei and Nielsen[6] using additional covariates and filtering for kinship, and obtain results comparable to those originally reported. An underpowered survival analysis of the parents is also compatible with a deleterious effect of $\Delta32$ in the homozygous state. The survival analysis generally has very low power due to the fact that the frequency of $\Delta32/\Delta32$ individuals is only approx. 1.1% and there are only approx. 10,000 relevant deaths recorded in the UKB. Future studies of the UKB data, as more deaths are recorded, will provide higher accuracy in the estimates of mortality rates and might allow investigations of specific causes of mortality and of differential mortality among age groups.

It is very difficult to detect selection using deviations from HWE, because such deviations are caused by the effect of selection within one generation, and because selection is only one of the many factors affecting the deviations from HWE[3]. Variants under strong selection are expected to segregate at very low frequencies and large sample sizes combined with sample specific genomic control are necessary to obtain sufficient power. Nonetheless, $\Delta32$ shows a consistent pattern of deviations from HWE across cohorts, choice of statistical methods, and data types. In addition, survival analyses, while providing mortality rate estimates with large confidence intervals, also point to a deleterious effect. Clearly, the p-values observed in the different analyses, which are typically on the order of $10{-}2$ or $10{-}3$ are very moderate for genomic data. After all, if these analyses were repeated for all loci in the genome we would expect a very large number of false positives with similarly small p-values due to the many tests that have been carried out. However, Δ32 is not a random mutation in the genome. It is in fact a target for gene-editing experiments, which raises the question as to whether we see any possible adverse effects of the mutation in epidemiological data. When testing for an adverse effect of a new drug, we would not require a correction for multiple tests for all possible compounds that could be used as a drug. Similar standards should be used for investigating potential negative effects of gene-editing targets. The specific hypothesis of a deleterious effect of Δ32 has been raised by the use of gene-editing on Δ32 as a medical instrument.

Inducing mutations in $CCR5$, such as Δ32, has been proposed as a strategy for providing a cure for HIV and has also been the subject of the only reported incidence of embryonic gene editing leading to live births in humans. We emphasize that our results, while clearly being relevant for germ-line editing, cannot be used directly to argue for or against the use of somatic gene-editing in $CCR5$. However, we suggest that careful examination of genetic epidemiological evidence, similarly to the analyses of this paper and that of Wei and Nielsen[6], should be taken into consideration in any human gene-editing experiments aimed at inducing segregating mutations into human populations.

Acknowledgements

The authors would like to thank Ed Hollox and Luke Jostins-Dean for helpful information. The authors would like to thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource. A full list of contributing groups can be found at `https://gnomad.broadinstitute.org/about`. We also thank the volunteers of the UK Biobank.

Data and code availability

The genotype and death registry information are available with the permission of the UK Biobank (`https://www.ukbiobank.ac.uk`). The gnomAD genotype counts are available from their website (https://gnomad.broadinstitute.org).

Analytical results and scripts are available at https://github.com/AprilWei001/Deviations-HWE-CCR5-Delta32.

References

[1] Godfrey Harold Hardy et al. Mendelian proportions in a mixed population. *Classic papers in genetics. Prentice-Hall, Inc.: Englewood Cliffs, NJ*, pages 60–62, 1908.

[2] Wilhelm Weinberg. Über den nachweis der vererbung beim menschen. *Jahres. Wiertt. Ver. Vaterl. Natkd.*, 64:369–382, 1908.

[3] Oliver Mayo. A century of hardy–weinberg equilibrium. *Twin Research and Human Genetics*, 11(3):249–256, 2008.

[4] David Modiano, Gaia Luoni, Bienvenu Sodiomon Sirima, Jacques Simporé, Federica Verra, Amadou Konaté, Elena Rastrelli, Anna Olivieri, Carlo Calissano, Giacomo Maria Paganotti, et al. Haemoglobin c protects against clinical plasmodium falciparum malaria. *Nature*, 414(6861):305, 2001.

[5] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018.

[6] Xinzhu Wei and Rasmus Nielsen. Ccr5-δ 32 is deleterious in the homozygous state in humans. *Nature medicine*, page 1, 2019.

[7] Michel Samson, Frédérick Libert, Benjamin J Doranz, Joseph Rucker, Corinne Liesnard, Claire-Michèle Farber, Sentob Saragosti, Claudine Lapouméroulie, Jacqueline

Cognaux, Christine Forceille, et al. Resistance to hiv-1 infection in caucasian individuals bearing mutant alleles of the ccr-5 chemokine receptor gene. *Nature*, 382(6593):722, 1996.

[8] Kara C Hoover. Intragenus (homo) variation in a chemokine receptor gene (ccr5). *PloS one*, 13(10):e0204989, 2018.

[9] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua X Hoffman, Bin Yi, Ashutosh Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Daren Liu, Nilanjana Banerjee, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the uk biobank. *bioRxiv*, page 572347, 2019.

[10] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

[11] Hakhamanesh Mostafavi, Tomaz Berisa, Felix R Day, John RB Perry, Molly Przeworski, and Joseph K Pickrell. Identifying genetic variants that affect viability in large cohorts. *PLoS biology*, 15(9):e2002458, 2017.

[12] Christopher C Chang, Carson C Chow, Laurent C A M Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2015.

[13] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[14] Michael Cariaso and Greg Lennon. Snpedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic acids research*, 40(D1):D1308–D1312, 2011.

[15] Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

[16] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[17] Sten Wahlund. Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11(1):65–106, 1928.

[18] Terry M Therneau and Thomas Lumley. Package 'survival'. *R Top Doc*, 128, 2015.

[19] David G Kleinbaum and Mitchel Klein. *Survival analysis*, volume 3. Springer, 2010.

[20] David Roxbee Cox. *Analysis of survival data*. Routledge, 2018.

[21] RA Fisher. Statistical methods for research workers. 1930.

[22] Ting Li and Xia Shen. Pleiotropy complicates human gene editing: Ccr532 and beyond. *Frontiers in Genetics*, 10:669, 2019.

[23] John Novembre, Alison P Galvani, and Montgomery Slatkin. The geographic spread of the ccr5 δ32 hiv-resistance allele. *PLoS biology*, 3(11):e339, 2005.

[24] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology*, 186(9):1026–1034, 2017.